British Journal of Political Science (2023), **53**, 163–182
doi:10.1017/S0007123421000661

British Journal of Political Science

ARTICLE

The Right Accounting of Wrongs: Examining Temporal Changes to Human Rights Monitoring and Reporting

Daniel Arnon[1]*, Peter Haschke[2] and Baekkwan Park[3]

[1]University of Arizona, [2]University of North Carolina at Ascheville and and [3]East Carolina University
*Corresponding author. Email: danielarnon@email.arizona.edu

(Received 11 December 2020; revised 21 June 2021; accepted 17 November 2021; first published online 7 February 2022)

Abstract

Scholars contend that the reason for stasis in human rights measures is a biased measurement process, rather than stagnating human rights practices. We argue that bias may be introduced as part of the compilation of the human rights reports that serve as the foundation of human rights measures. An additional source of potential bias may be human coders, who translate human rights reports into human rights scores. We first test for biases via a machine-learning approach using natural language processing and find substantial evidence of bias in human rights scores. We then present findings of an experiment on the coders of human rights reports to assess whether potential changes in the coding procedures or interpretation of coding rules affect scores over time. We find no evidence of coder bias and conclude that human rights measures have changed over time and that bias is introduced as part of monitoring and reporting.

Keywords: human rights measurement; bias; machine learning; natural language processing

Introduction

Over the last few decades, the proportion of people living in democracies has increased markedly—as has the number of people escaping poverty (Roser 2018; Roser and Ortiz-Ospina 2018). Yet, despite these positive developments, global human rights conditions have stagnated according to common standards-based human rights measures, such as the Political Terror Scale (PTS) and the Cingranelli–Richards Physical Integrity Rights Index (CIRI) (Cingranelli, Richards, and Clay 2014; Gibney et al. 2019). This inconsistency constitutes a puzzle, as scholarship on human rights practices has long established a strong association between democratic governance, economic development, and human rights conditions (see, for example, Bueno de Mesquita et al. 2005 [2003]; Davenport 2007a; Haschke 2018; Henderson 1982; Poe and Tate 1994). It also prompted a lively debate among scholars questioning whether human rights conditions are really stagnating and whether the absence of a positive trend in human rights practices since the mid- to late 1980s is possible when other indicators of human welfare have improved so significantly over the same period.

While some have insisted that global human rights conditions have improved little, if at all, (for example, Cingranelli and Filippov 2018; Haschke and Gibney 2018; Richards 2016), others try to resolve the discrepancy by arguing that the appearance of stagnating human rights conditions is an artifact of bias affecting standards-based measures of human rights practices (for example, Clark and Sikkink 2013; Fariss 2014; Fariss 2018).[1] Among others, the most commonly

[1]Recent scholarship questioning the validity of human rights indicators continues previous work that identified sources of bias affecting human rights measurement efforts (for example, Poe and Tate 1994; Poe, Carey, and Vazquez 2001; Simmons 2009; Wood and Gibney 2010).

articulated mechanisms of bias can be summarized under the headers of "changing norms" and "information effects" (Clark and Sikkink 2013; Fariss 2014). Although scholars have referred to the bias processes using different terms, they are in agreement that the production process of human rights reports by monitoring organizations has changed over time.

Although, scholars have proposed "fixes" and developed alternative and ostensibly unbiased measures of human rights conditions (for example, Human Rights Dependence Scores [see Fariss and Schnakenberg 2014]), the bias mechanisms themselves remain largely assumed and untested.[2] In this article, we fill this void and examine evidences of the bias processes outlined earlier.[3] To do so, we begin by characterizing human rights measurement as a two-stage process. The first stage, which we call the "monitoring and reporting stage," consists of the monitoring of human rights conditions in states by both state and nonstate human rights organizations, and the subsequent compiling of country reports.[4] The second stage comprises the coding of the country reports into standards-based human rights scores by teams of academics. We call this later stage, the "coding stage," where bias may also enter (Haschke and Arnon 2020).

With this characterization in place, we seek to identify the point in the measurement process at which bias is most likely to be introduced. A changed information environment, combined with changing classification standards, may have introduced bias during the monitoring and reporting stage, as human rights monitoring and reporting agencies today may apply systematically different standards as to what counts as human rights abuse than in the past. Similarly, monitors and reporting agencies today operate in a vastly different information environment. We call bias that appears in the first stage "events-to-reports" bias, to reflect changes and biases that appear in the monitoring and reporting stage. Alternatively, human coders' subjective interpretation of human rights or changed coding standards may point to the coding stage as the point at which bias is introduced into human rights measures. We call bias introduced in the second stage "reports-to-scores" bias to indicate biases emanating from coders, after the compilation and publication of reports, rather than bias emanating from the reporting agencies themselves. Based on the logic of exclusion, we first examine the presence of bias in the measurement process with supervised machine learning and natural language processing (NLP), which may appear as either events-to-reports bias, or reports-to-scores bias. We then examine if bias can be attributed specifically to the human coders during the coding stage and conduct an experiment with the PTS research team coding human rights reports to produce the PTS human rights scores.

While we find that human rights measures are indeed biased, we find no evidence of reports-to-scores bias entering human rights measures as part of the coding stage. By specifying the data-generating processes of human rights measures and by identifying potential sources of bias that may distort human rights scores, our article contributes to the understanding of bias processes in human rights measures and advances an ongoing debate among human rights scholars. In addition, the implications of our theory and empirical analyses are generalizable and help us better understand standards-based measures used in social science research more generally.

The article proceeds as follows. In the next section, we review the literature on human rights measurement and describe the two-stage process of quantifying human rights conditions. We then disentangle several sources of bias that might affect standards-based measures in the process. The next section presents findings from our experiment that allows us to rule out the presence of reports-to-score bias introduced in the second coder stage. A final section concludes with recommendations for future scholarship on human rights and social science research using standards-based measures.

[2]However, see Park, Greene, and Colaresi (2020).

[3]Bias concerns, of course, are not unique to human rights measures, and our discussion extends to other standards-based measures in political science research.

[4]Specifically, the US State Department, Amnesty International, and Human Rights Watch.

https://doi.org/10.1017/S0007123421000661 Published online by Cambridge University Press

Observing Human Rights Conditions

The measurement of state-sponsored human rights violations, specifically and states' human rights records more generally, has been an integral part of scholarship on political violence, repression, and human rights since the early 1980s. Measurement efforts began when nongovernmental organizations (NGOs) such as Amnesty International (AI) and governmental agencies such as the US Department of State began to systematically monitor and record the degree to which internationally recognized human rights are protected or violated, and political scientists began to quantify the reports that were disseminated by these organizations. Measurement projects such as the PTS and CIRI are arguably the most widely used standards-based measures that were developed to track and analyze states' commitment to the protection of basic human rights (for example, the right to physical integrity).

The production of standards-based human rights measures such as the PTS or the CIRI scores can be characterized as a *two-stage process*. We call the first stage the "monitoring and reporting stage." It consists of the monitoring of human rights conditions within states or territories by organizations such as the US State Department (SD), AI, or Human Rights Watch (HRW), and the subsequent compiling of qualitative country reports. In cooperation with local and international human rights organizations, human rights monitoring organizations observe, record, verify, and document human rights events (for example, instances of torture, political imprisonment, and extrajudicial killings) that appear in violation of human rights law, such as the Convention of Torture (CAT) or the Universal Declaration of Human Rights (UDHR). The second stage, which we call the "coding stage," involves the conversion of the qualitative country reports produced by the monitoring organizations into standards-based human rights scores by scholars. This stage includes the identification of key concepts to be measured (for example, ill-treatment and torture, forced disappearances, and so on), the operationalization of indicators for scoring, and finally the application of coding rules to country reports.

Both the PTS and CIRI research projects rely on the annual country reports as their source material on human rights practices (Wood and Gibney 2010). Whereas the PTS produces three separate sets of scores—one for each monitoring and reporting organization's reports[5]—CIRI produces five sets of scores: an overall additive index and four composite scores for extrajudicial killings, forced disappearances, torture, and political imprisonment, respectively.[6] PTS scores range from 1, indicating good human rights conditions, to 5, signifying systemic and pervasive human rights abuses. CIRI's individual indicators range from 0 to 2, where higher scores indicate fewer reported violations of the respective category.[7]

Bias Processes

Validity questions have followed standards-based measures of human rights for decades,[8] and scholarship has raised serious concerns about the ability of standards-based measures to track human rights conditions across time and space. Fariss (2014; Fariss 2017; Fariss 2018) and Potz-Nielsen, Ralston, and Vargas (2018), for example, argue that due to changing monitoring and reporting standards over time, standards-based human rights measures are entirely inappropriate for temporal comparisons and allow for only cross-sectional comparisons. Keck and Sikkink (1998), Clark and Sikkink (2013) and Fariss (2014) argue that temporal comparison is likely fraught due to changing monitoring and reporting capacity, and Eck and Fariss (2018)

[5]For further instruction on how to use each of the three separate scores, see: https://www.politicalterrorscale.org/About/FAQ/

[6]For a more detailed review of common standards-based human rights measures, see Wood and Gibney (2010) and Landman and Carvalho (2010).

[7]CIRI's overall additive index ranges from 0 to 8.

[8]See Poe, Carey, and Vazquez (2001), Poe et al. (1994), Simmons (2009), and Wood and Gibney (2010).

https://doi.org/10.1017/S0007123421000661 Published online by Cambridge University Press

caution against cross-sectional comparison because monitors are confronted with vastly different levels of access to countries, which biases reports. Building on this growing body of work, Haschke and Arnon (2020) propose a typology of bias processes, distinguishing bias processes that affect the monitoring and reporting stage of measurement from those that affect the coding stage. They also distinguish bias processes that vary temporally from those that prohibit spatial or cross-sectional comparison. It should be noted that the source of bias in each stage is different. In the first stage, during the compilation of human rights reports by the reporting agencies, the events-to-reports bias may appear as a result of the interests of the reporting agencies to either highlight or erase certain violations and human rights conditions for geopolitical or publicity reasons (Hill, Moore, and Mukherjee 2013). In the coding stage, the translation of reports to scores is performed by teams of academics (for example, CIRI and PTS), who use seemingly consistent standards but may be biased in their application over time. Figure 1 illustrates the potential sources of bias in these two stages. It is important to note that we refer here to bias in a broad sense, that is, as systematic changes, whether intentional or not, that result in the inconsistent mapping of text from human rights reports to standards-based rights scores.

Events-to-reports bias (monitoring and reporting stage)

The most robust challenge to standards-based human rights measures has been advanced by Clark and Sikkink (2013) and Fariss (2014; Fariss 2017; Fariss 2018). They argue that the very "definition of what constitutes torture or state-sponsored killing has expanded over the years" and behaviors that were not considered human rights concerns in the past are considered violations today (Clark and Sikkink 2013, 546). In early years, SD reports focused almost exclusively on the most heinous violations of human rights, such as extrajudicial killings or forced disappearances. Today, reports provide in-depth detail of violations including excessive use of force, stealth torture, or stress and duress methods—practices that arguably would not have been included in earlier reports. Increasingly stricter standards and changing expectations used by monitoring organizations "mask real improvements to the level of respect for human rights," as these changing standards translate into increasingly detailed and harsher reports (Fariss 2017, 239–40). Moreover, they argue that these additional details in the reports are picked up during the coding stage, and resulting human rights scores will be biased because coders do not consider what Fariss (2014) calls the "changing standards of accountability."

It is important to note that two distinct mechanisms may bias reports in the first stage. "Changing norms" refers to changes of the classification strategy or changes of monitors' "subjective views of what constitutes a 'good' human rights record" (Fariss 2014, 299). These are distinct from "information effects," which refer to changes in monitoring capacity or changes to the level of access to information about human rights conditions, or to other political pressures.[9] Information effects encompasses both what Clark and Sikkink (2013) refer to as "changes in the quality and availability of information" on human rights violations, and what Fariss (2014) calls the ability to "look harder for abuse [and] look in more places."

The application of changing standards by monitoring organizations is likely attributable to a changing and dynamic body of international law, and the development of new human rights norms over time. As international law changes and international "norm entrepreneurs" add new rights to the human rights discourse, reporting agencies respond by incorporating information pertaining to new rights into the reports or by refocusing the reports' emphases. Bagozzi and Berliner (2018) locate topical changes by means of structural topic models that identify the underlying topics of attention and scrutiny in SD reports. They find that new topics or rights

[9]It should be noted that the allegation of politicized reporting to support the strategic interests of the reporting organization is among the oldest claims of bias affecting human rights reports. Reports, of course, are not created in a neutral environment and may be constructed to align with the reporting organizations' interests.

https://doi.org/10.1017/S0007123421000661 Published online by Cambridge University Press

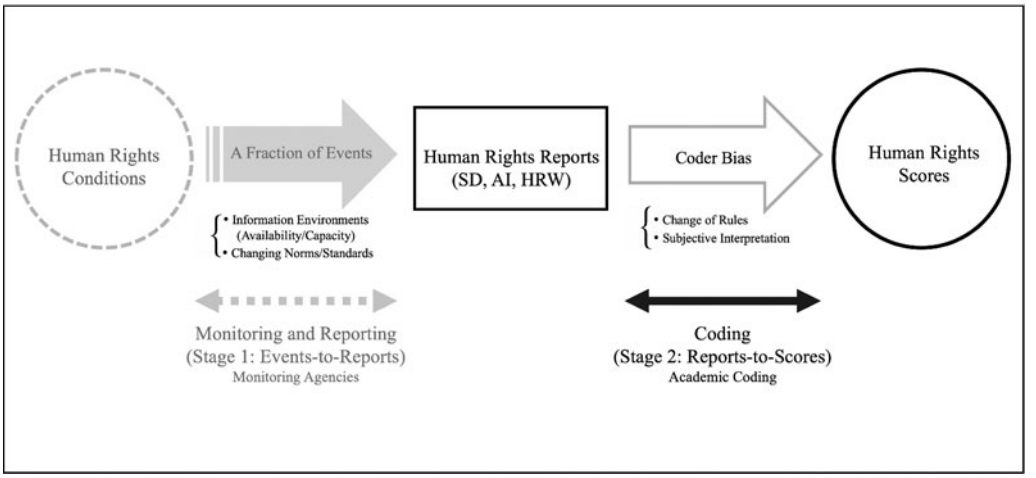

Figure 1. Sources of bias in each stage.

do indeed appear in more recent human rights reports.[10] Park, Murdie, and Davis (2019) similarly find that there is an evolution of topic coverage over time in various human rights organizations' reports.[11] Analyzing the hierarchical structure of human rights reports from the SD, AI, and HRW, Park, Greene, and Colaresi (2020) also find that there has been a significant increase in human rights topics or concepts. Information effects are likely a function of increased budgets allocated to the SD or AI to compile reports, increased collaborations with local and international human rights NGOs, and the development of information and telecommunication technologies (Clark and Sikkink 2013; Fariss 2014).

While scholars have identified broad changes in the text of human rights reports, the untested argument put forward by these scholars pertains to the *mapping* of the text to standards-based human rights measures. This is primarily because of the secretive nature in which human rights violations are carried out. Abuses are difficult to detect and to observe, and even more difficult to verify (Davenport 2007b; Roth 2004; Simmons 2009). Especially as states democratize, they commonly transform their use of repression from overt to covert strategies (Conrad and DeMeritt 2011; Conrad and Moore 2010). Consequently, although many human rights organizations may have a good sense of a state's human rights practices, observing, monitoring, and reporting *all* human rights violations is impossible. Thus, human rights reports are inevitably far from complete. A number of scholars have indicated that human rights reports are merely a collection of selective observations (Hill, Moore, and Mukherjee 2013; Murdie, Davis, and Park 2020; Ron, Ramos, and Rodgers 2005). It is impossible to directly examine or pinpoint exactly which human rights events are reported and accounted for in human rights reports, and which are not. Thus, we argue that the monitoring and reporting stage constitutes a potential source of bias. It should be noted that the biases we are addressing here are not benchmarked by the "universe" of human rights violations; instead, the biases considered are downstream and appear once reports are being compiled in the monitoring and reporting stage, with the changes occurring at the reporting agency level. In the following, we assess whether events-to-reports bias—understood as both increases in the amount of information and changes in the classification standards—is associated with more stringent human rights scores. Thus, we suggest the following implication:

[10]Bagozzi and Berliner (2018) note that emerging human rights topics include human trafficking and labor rights.

[11]The authors identify new topics such as lesbian, gay, bisexual, trans, and queer (LGBTQ) rights and international justice mechanisms handling war crimes, genocide, and crimes against humanity.

https://doi.org/10.1017/S0007123421000661 Published online by Cambridge University Press

Implication 1: If the arguments put forward by scholars about monitoring and reporting bias are correct, we should expect to see systematic changes in the translation of the texts in human rights reports into standards-based human rights scores.

Reports-to-scores bias (coding stage)

Bias can also be introduced at the second stage, that is, as part of the translation of compiled and published human rights reports into human rights scores. This second stage includes both the development of coding rules or standards and the application by human coders of a set of coding rules to the source material (that is, the human rights reports). Bias can then enter the coding stage at two points: (1) bias can be introduced when the coding rules or standards that coders use to convert reports into human rights scores change over time; and (2) bias could be a function of varying interpretations and applications of the coding rules by coders. Bias in the second stage is thus introduced not through changes to the reports, which are already compiled and published, but through changes to the coding rules or procedures and through changes in coders' interpretation of the coding rules.

If coding rules follow changes to international human rights law, or if the development of new human rights norms prompts changes to the standards coders are asked to apply to human rights reports, the resulting human rights scores will be biased. Expanding the scope of what coders must consider as evidence of abuse we call "reports-to-score bias" in the coder stage. Even if coding rules and standards remain fixed, the coders themselves could introduce bias into human rights measures. If coders interpret the coding rules and standards differently today than 25 years ago, or if there is regular turnover among coders such that older cohorts of coders are replaced with new coders, bias could be introduced. While asked to recognize their own biases, human coders are, of course, human; as such, current human rights norms and other contemporaneous information might color a coder's reading of a human rights report.

Certainly, for the PTS, the rules for coding human rights reports have been in place and unchanged since the early 1980s. Thus, we are confident in our ability to rule out that the coding rules or standards coders are asked to apply to human rights reports have changed. However, to assess the possibility that coders themselves or the context in which reports are coded could lead to bias at this second stage, we conducted an experiment as part of the annual coding efforts of the PTS. We consider a difference of human rights scores between two coders, A and B, as evidence of *coder bias* if Coder A at time t assigns a different human rights score to a human rights report produced at time than Coder B coding at time $t+1$. As both the coding rules and the reports themselves remain constant or fixed, this difference can only be attributed to the coders themselves or to the changing context in which the scores were produced. Thus, we propose the following implication:

Implication 2: If our argument about coder bias is correct, we should expect that coder A at time t assigns a different human rights score to a human rights report produced at time t than Coder B coding at time $t+1$.

In sum, we describe that there are largely two stages in which biases can be introduced. Since it is impossible to directly observe and test all the human rights practices in the first stage, we rely on the logic of exclusion. If we cannot find the evidence that the bias is introduced in the coding stage, we can reasonably conclude that biases are most likely being brought into the process in the monitoring and reporting stage by systematic changes in the reports, rather than through the systematic translation of reports to scores. We describe the detail and findings of the experiment later, following our assessment of the possibility of monitoring and reporting bias affecting scores.

https://doi.org/10.1017/S0007123421000661 Published online by Cambridge University Press

Methods and Data

In order to test the presence of bias in either the first or the second stage, we use NLP and supervised machine-learning algorithms to gather information about the texts of the annual human rights reports produced by the SD and AI, and then to "learn" how the words and content of the texts map onto the standards-based scores given to each report. Once the algorithm has learned how the text maps onto scores, we use the trained algorithms to predict contemporary scores.

The large temporal span of reports allows us to first train the classifiers on an early subset of SD reports as a training set and then use the trained classifiers to predict contemporary scores (out-of-window texts). We then retrain the algorithm by pushing the training set one year forward and using the newly trained algorithm to predict *the same* contemporary reports again. This method is iterated for the entire temporal span of reports in ten-year windows.[12]

Scholarly arguments regarding monitoring and reporting bias hold that both human rights norms and the information environment in which human rights monitoring organizations operate and produce reports are dynamic and have changed significantly over time. In our first test, we substitute the manual procedure of assigning scores to reports with algorithms. The main advantage of this method is that we can systematically track how well the algorithm performs over the entire span of the data. If overall bias is present in the processes, then the algorithm trained on 1977–87 should not perform as well as the algorithm trained on 2001–11, for example. This is because the mapping of reports to scores is biased by increased information and changing norms.

Most recently, applying a forecasting method based on machine learning, Greene, Park, and Colaresi (2019) test the assumptions of changing norms in SD reports, which are used to produce PTS scores. Using the entire texts of SD reports from 1977 to 2010 and supervised machine-learning algorithms, Greene, Park, and Colaresi (2019, 229) conclude that "there is some underlying change in coding of human rights measures from texts over time." In other words, they argue that there has been some change, but they do not specify if the change is derived from changes in the reports (monitoring and reporting bias) or from changes in "translation" to scores (coder bias). *If some underlying change has occurred, the question still remains: what exactly is the source of this change?* More to the point, PTS is based only on Section 1 of the SD reports, but by using the entire reports instead of using the relevant section of the reports in their analyses, they misrepresent how PTS is generated. In the following section, we discuss our approach to examining the presence of the changes and the possible source of the changes.

Data and Research Design

Data

To evaluate the arguments in the previous section, we use the annual SD country reports on human rights practices and the AI human rights annual reports from 1976 to 2016. In order to more accurately understand and analyze the reports and the corresponding PTS and CIRI scores, we break down the reports. Table 1 summarizes the different sets of analyses. PTS consists of two sets of scores: PTS-SD and PTS-AI. PTS-SD is based on the SD reports, more specifically, "Section 1. Respect for the Integrity of the Person"; PTS-AI is based on the entire AI annual reports.

First, we use Section 1 of the SD reports with PTS-SD (1) and the AI reports on PTS-AI (2). CIRI is based on Section 1 of the SD report and the entire AI report combined; thus, we use them with the aggregated CIRI score (CIRI/PHYSINT) (3). In addition, we also run similar models on the disaggregated CIRI scores for each respective subsection of the report—torture (CIRI/TORT) (4), political imprisonment (CIRI/POLPRIS) (5), extrajudicial killings (CIRI/KILL) (6), and

[12]A full explication of the method appears in the next section.

https://doi.org/10.1017/S0007123421000661 Published online by Cambridge University Press

Table 1. Data (texts) and labels

	Texts	Labels	Years
(1)	SD (Section 1)	PTS-SD (1–5)	1978–2016
(2)	AI (all)	PTS-AI (1–5)	1976–2016
(3)	SD (Section 1) + AI (all)	CIRI/PHYSINT (0–8)	1981–2011
(4)	SD (Section1/torture)	CIRI/TORT (0–2)	1981–2011
(5)	SD (Section1/imprisonment)	CIRI/POLPRIS (0–2)	1981–2011
(6)	SD (Section1/kill)	CIRI/KILL (0–2)	1981–2011
(7)	SD (Section1/disappearance)	CIRI/DISAP (0–2)	1981–2011

Notes: PTS data are available from 1976 to 2016 and CIRI scores are available from 1981 to 2011. SD reports are available from 1978 and AI reports are available from 1976.

enforced disappearances (CIRI/DISAP) (7)—in order to examine if the changing standards apply uniformly to all physical integrity violations.

Fixed-rolling-window forecasting

We take a supervised machine-learning approach to examine how language or the text of human rights reports maps onto human rights scores (that is, PTS and CIRI scores). This essentially constitutes a classification task, where human rights reports are assigned to human rights scores. Assume that our training dataset $D_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{n}$, with n annotated data (for example, country reports). Here, x_i denotes a string of texts (tokens), $x_i^1 \ldots x_i^{l_i}$, where l is the total number of words; x_i is drawn independently from the total set of words X and $x_i \in X$, according to an unknown distribution on X, called P_X. Also, y_i denotes the annotation of c classes of x_i and $y_i \in \{1, \ldots, c\}$ (for example, PTS or CIRI scores). We also assume an unknown function $f{:}X \rightarrow \{1, \ldots, c\}$ and that $D_{\text{train},y_i} = f(x_i)$ for all $i = 1, \ldots, n$. Thus, the main goal is to estimate f on the X given the training dataset D_{train} and generalizing to the entire X. A classification model $\hat{f}$ is an estimate of unknown f based on the training data D_{train}, and we get a classifier function, $\Phi = \hat{f}(D_{\text{train}}){:}X \rightarrow \{1, \ldots, c\}$, after a training. To evaluate the trained model, Φ, a testing dataset (D_{test}) is set aside, where $D_{\text{test}} = \{(\hat{x}_i, \hat{y}_i)\}_{i=1}^{\hat{n}}$. One of the ways to evaluate Φ is accuracy, $(1/\hat{n}) \sum_{i=1}^{\hat{n}} \delta(\Phi(\hat{x}_i), \hat{y}_i)$, where δ refers to the Kronecker delta and $\hat{x}_i$ and $\hat{y}_i$ are from the testing dataset. Here, $\Phi(\hat{x}_i)$ is the prediction from the trained classification model. Thus, accuracy approximates the probability of having $\Phi(x)$ equal $f(x)$, that is, $P_X(\Phi(x)) = f(x)$.

In the fixed-rolling-window-forecasting approach, we divide the training dataset D_{train} by a series of training windows $\mathcal{W}_t$, for time $t \in (1, \ldots, T)$, $D_{\text{train},\mathcal{W}_t}$. Then, we evaluate a classifier model for each window $\Phi_{\mathcal{W}_t} = \hat{f}(D_{\text{train},\mathcal{W}_t})$ on the set-aside testing $D_{\text{test},\mathcal{W}_{\text{out}}}$.

As illustrated in Figure 2, by dividing the training data by ten years, we create twenty-six ten-year in-window sets from the training sets (1977–2011), $D_{\text{train},\mathcal{W}_t}$, $t \in (1, \ldots, 26)$. For example, $\mathcal{W}_1$: 1977–86, $\mathcal{W}_2$: 1978–87, … $\mathcal{W}_{26}$: 2002–11 for *in-window sets*. For a trained classification model for each in-window set, $\Phi_{\mathcal{W}_t} = \hat{f}(D_{\text{train},\mathcal{W}_t})$, we estimate the extent to which Φ approximates the unknown function f on the *out-of-window testing set* (2012–16), $D_{\text{test},\mathcal{W}_{\text{out}}}$.[13] For example, at $\mathcal{W}_1$, an algorithm based on the texts from 1977 to 1986 (ten-year fixed window) learns the function of $\Phi_{\mathcal{W}_1}$ and is tested on $D_{\text{test},\mathcal{W}_{\text{out}}}$. At $\mathcal{W}_2$, algorithms based on the texts from 1978 to 1987 (ten-year fixed window) learns the function of $\Phi_{\mathcal{W}_2}$ and is tested on the same out-of-sample window. We continue this until $\mathcal{W}$, in which the texts are from 2002–11 (the final window), and compare performance for both in-window sets and out-of-window sets for each window. Therefore, if the changes in the information environment and/or norms had no influence on the SD reports and PTS/CIRI over time, then we would assume that $\Phi_{\mathcal{W}_1} = \Phi_{\mathcal{W}_2} \ldots = \Phi_{\mathcal{W}_{26}}$ and, thus, expect

[13]We simply chose the most recent five years for the testing set. As a robustness check, we also tried different window sizes such as three years (2014–16) and seven years (2010–16). The results are consistent with the five-year testing models.

https://doi.org/10.1017/S0007123421000661 Published online by Cambridge University Press

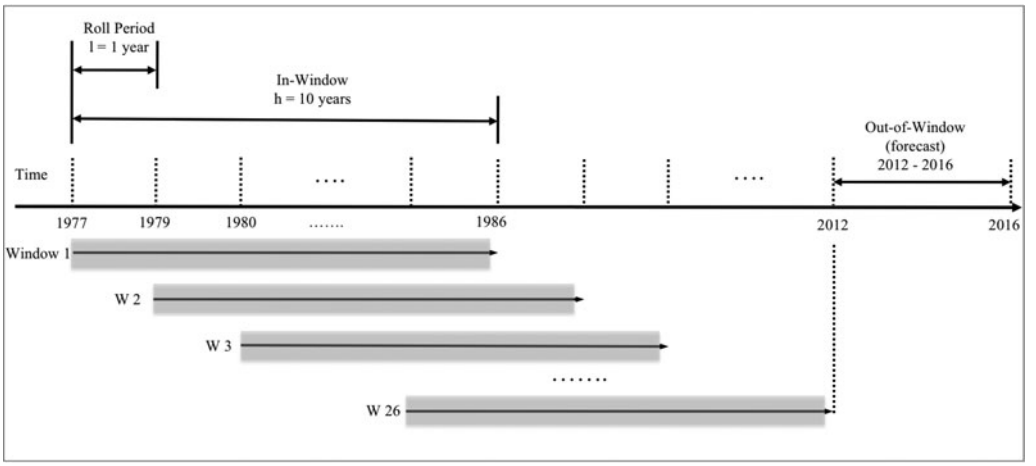

Figure 2. Fixed-rolling-window forecasting.

$P_{X_{out}}(\Phi_{\mathcal{W}_1}(x) = f(x)) = P_{X_{out}}(\Phi_{\mathcal{W}_2}(x) = f(x)) \ldots = P_{X_{out}}(\Phi_{\mathcal{W}_{26}}(x) = f(x))$. On the other hand, if more and better information and norm changes lead to an increased stringency in the issuance of standards-based scores, we would observe changes in out-of-window accuracy over time.[14]

We train a number of traditional supervised machine-learning models and neural network models based on deep learning (Φ). First, we represent each human rights report (document) by a feature count vector. The text of reports is modeled as a "bag of words," that is, a set of content words without any word order or syntactic or relational information, leading each unique word to a separate word. We use all words available with both term frequency (Tf) and term frequency inverse document frequency (Tf-Idf) weighting to locate informative or relevant features. We also explore the role of higher-order *n*-grams (the occurrence of a word based on the occurrence of its previous words) as features for discerning the subtleties reflecting human rights ratings. It is possible that higher-order *n*-grams contain greater relevant information than simple unigrams. As suggested by Pang, Lee, and Vaithyanathan (2002), employing higher-order *n*-grams and combining them (unigram, bigram, and trigram together) could give us better performance than using them separately. We train a number of linear and nonlinear machine-learning algorithms, such as naive Bayes (NB), logistic regression (LR), support vector machines (SVM), and random forests (RF).[15] In addition, we also train a number of convolutional neural network (CNN) models with word embeddings. Due to the capability of capturing local correlations of spatial and temporal structures, along with recurrent neural networks (RNN) models, CNNs have achieved remarkable results in NLP tasks, such as semantic parsing (Yih et al. 2011), search query retrieval (Kalchbrenner, Grefenstette, and Blunsom 2014), and other traditional NLP works (Collobert et al. 2011; Zhou et al. 2015). Similar to ordinal neural networks, CNNs are made up of neurons that have learnable weights and biases, with several layers of convolutions with nonlinear activation functions. In particular, these convolutions are used over the input layer to compute the output to result in local connections, where each region of the input is connected to a neuron in the output. A key aspect of CNNs is the use of pooling layers, typically applied after the convolutional layers. One property of pooling is that it provides a fixed-size output matrix, which typically is required for classification. This allows the use of variable-size

[14]In this context, a more stringent score refers to a worse human rights score than would have been observed; had there been no bias in the monitoring and reporting stage.

[15]For each of these algorithms, see Lewis (1998), McCullagh and Nelder (1989), Cortes and Vapnik (1995), and Breiman (2001).

https://doi.org/10.1017/S0007123421000661 Published online by Cambridge University Press

sentences and variable-size filters but by obtaining the same output dimensions to feed into a classifier. Moreover, pooling also reduces the output dimensionality while keeping the most salient information. Then, during the training phase, a CNN automatically learns the values of its filters based on the task to be performed. In this article, we train CNNs with one layer of convolution on top of word vectors obtained from an unsupervised neural language model. Initializing word vectors from an unsupervised neural language model, we use the publicly available "global vectors for word representation" (GloVe) (Pennington, Socher, and Manning 2014) that were trained on 6 billion words from Google News. The vectors have dimensionality of 200. Words not present in the set of pretrained words are initialized randomly. We ran four different models of CNNs: (1) CNN-Random Model, a baseline model, in which all words are randomly initialized and then modified during training; (2) CNN-Static Model, a model with pretrained vectors from GloVe, in which all words—including the unknown ones that are randomly initialized—are kept static and only the other parameters of the model are learned; (3) CNN-Dynamic Model, the same model as the static model but the pretrained vectors are finetuned for each task; and (4) CNN-Multi Channel Model, a model with two sets of word vectors, where each set of vectors is treated as a "channel" and each filter is applied to both channels, but gradients are backpropagated only through one of the channels. Thus, the CNN-Multi Channel Model is able to finetune one set of vectors while keeping the other static; both channels are initialized with GloVe. In total, we run 29 models for each report (data).

Results and Discussion

Figure 3 shows the results of the first sets of analyses. The top panel in Figure 3 displays the out-of-window prediction accuracy for each of our algorithms trained on ten-year rolling windows. As discussed in the previous section, if there were no changes or biases affecting monitoring/reporting and the production of human rights reports over time, and human coders translated reports consistently to standardized human rights scores, we would expect all the models trained in each rolling window to predict the data (texts) in the out-of-window sets equally and the accuracy measure ($P_{X_{\text{out}}}(\Phi_{\mathcal{W}_t}(x) = f(x))$) should be the same across years. In other words, there would be no meaningful changes in out-of-window accuracy over time.

However, as illustrated in Figure 3 (top panel in each plot), trained models perform better as they get closer to the out-of-window test set, as indicated by the consistent upward slope of the out-of-window accuracy over the years. In general, across Models 1, 2, and 3, we see a 20 to 30 per cent increase in model performance. Model 3 (Section 1 with CIRI scores) seems to show a little lower performance throughout the years compared to Models 1 and 2, but given that CIRI/PHYSINT has nine classes, the baseline accuracy is much lower (0.11) and shows consistent changes over the years. In order to determine that the changes are statistically significant, we perform McNemar's test (Raschka 2018) to compare the first trained model ($\Phi_{\mathcal{W}_1}$) and the last trained model ($\Phi_{\mathcal{W}_t}$) on the out-of-window test set for each model.[16] From Models 1 to 3, we reject the null hypothesis that the accuracies from these two models are equal.[17] Substantively, it means that the latest training models can predict about 190 to 280 human rights scores more accurately than the earliest models. The mapping functions trained in later years are more likely to predict the human rights scores accurately in the most recent years (2012–16). There were substantial changes in translating the reports to standards-based human rights measures. The middle panel in each plot in Figure 3 shows the in-window accuracies (performance) at each training window (tested in-window test set). Although there were some fluctuations, they do not appear to be increasing or decreasing across time. In other words, there is no meaningful change in the ability of the algorithms to estimate the unknown mapping function $f(\cdot)$ of texts to

[16]We choose the majority voting classification model for the test.

[17]Model 1: $p = 3.735 \times 10^{-50}$, Model 2: $p = 9.313 \times 10^{-16}$, and Model 3: $p = 7.764 \times 10^{-9}$.

https://doi.org/10.1017/S0007123421000661 Published online by Cambridge University Press

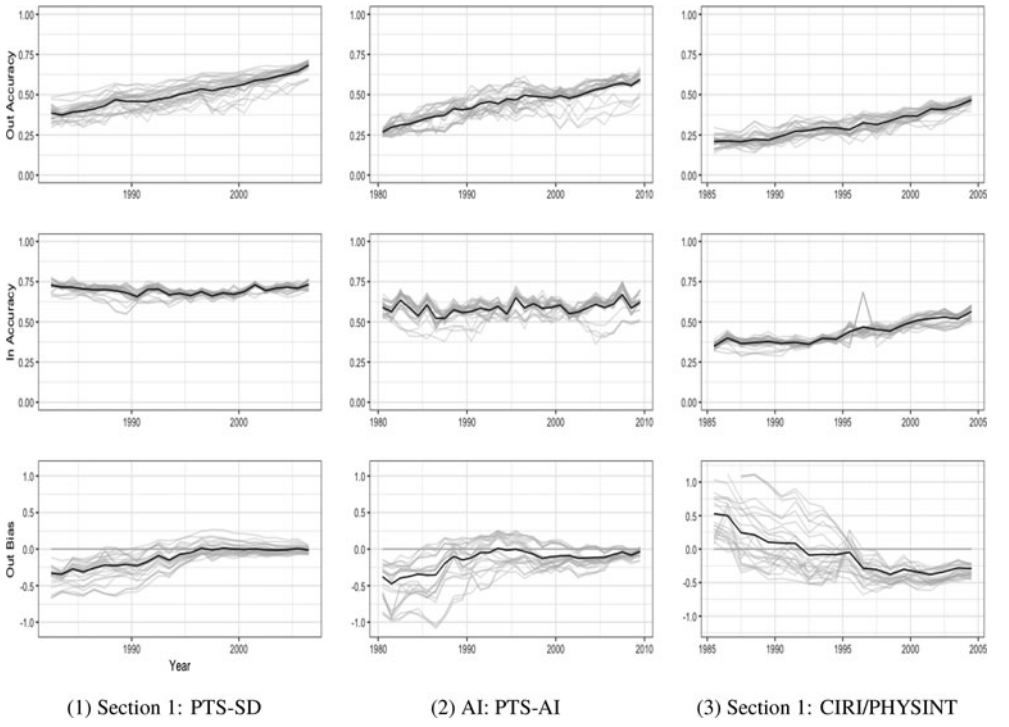

Figure 3. Accuracy and bias across algorithms over time.
Notes: Shown are the measured in- and out-of-window accuracy and bias for 29 machine-learning models. The gray-colored lines represent 29 models and the solid black line is an average value across all the models. Left panels use Section 1 of SD report on PTS-SD scores. Middle panels use AI reports on PTS-AI scores. Right panels use only Section 1 of the SD reports on CIRI aggregate scores. The top panels for out-of-window accuracy show an increasing slope, indicating that for all of these measures, standards have changed. It should be noted that each year in the plots represents the midpoint of the ten-year training window.

scores over time. The changes in these in-window model performances indicate how trained models in each window fit properly, but they do not directly evaluate the out-of-sample validity of the model predictability.[18] The bottom panel in each plot in Figure 3 indicates the average bias from the out-of-window test set, $(1/n)\sum_{i=1}^{n}(\Phi_{\mathcal{W}_t}(x_i) - y_i)$, where $\Phi_{\mathcal{W}_t}(x_i)$ is a model prediction and y_i is the true label for each data point x_i in the test set. In other words, it is the average difference between the actual labels and the model predictions. For Models 1 and 2, the average bias across all classification algorithms gradually increase as they get closer to the end. In early years, there is *negative bias* (underprediction) on the testing set. As CIRI's scores are a reversed scale, Model 3 shows *positive bias* (overprediction). This means that, on average, the trained algorithms get more stringent over the years. The trained model in the first window ($\Phi_{\mathcal{W}_1}$) evaluates states' human rights conditions more "leniently" than the actual evaluation in later years. In other words, evaluating standards have become more "stringent."[19]

Next, Figure 4 shows the results of out-of-window accuracy for each of the disaggregated CIRI physical integrity indicators (for example, torture, political imprisonment and so on). In order to emulate the data-generating process more accurately, we used only the relevant subsections from

[18]We note that the in-window accuracy for the CIRI measure appears to be increasing, as opposed to the PTS measures. We believe that this is due to the CIRI's limited data availability (shorter time span) and more classification categories.

[19]As a robustness check, we also did rolling-backward forecasting, that is, the model training starts from the later years and predicts the first year set (1978–82), for example. The findings are consistent with the forward forecasting. As the models roll backward, the model performance decreases. For more detail, see the Online Appendix.

https://doi.org/10.1017/S0007123421000661 Published online by Cambridge University Press

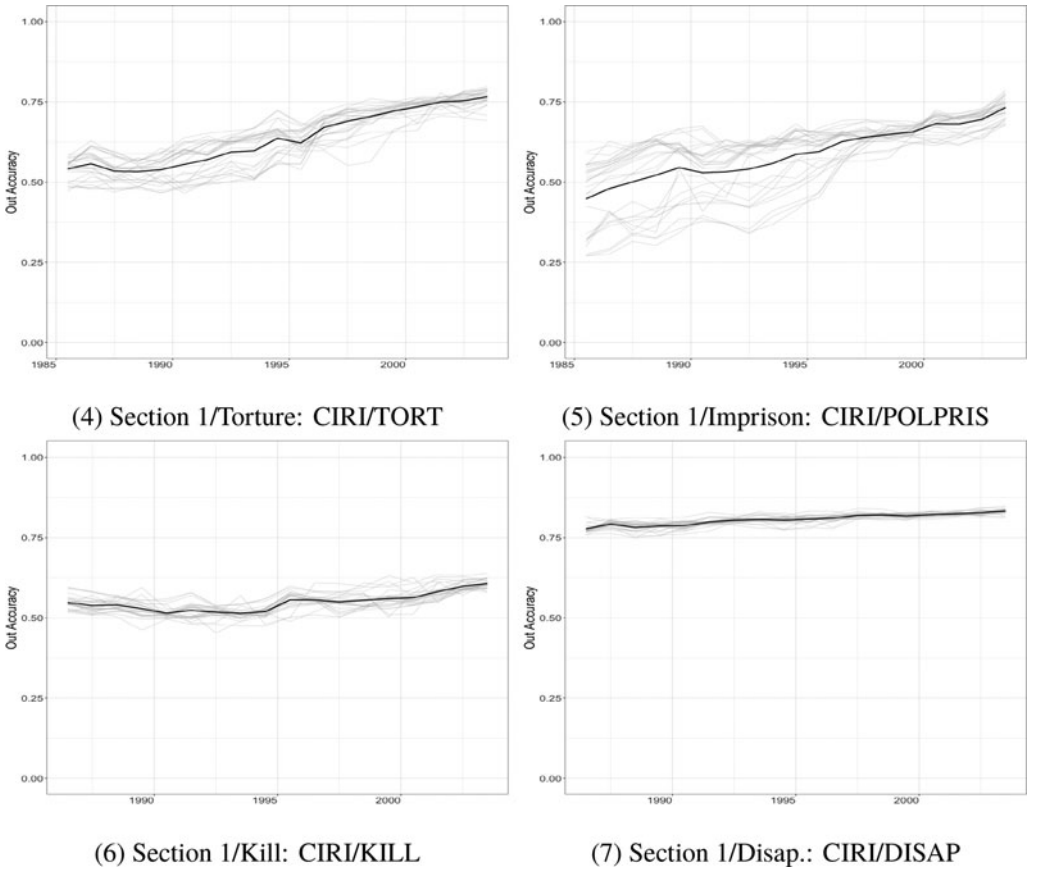

Figure 4. Accuracy and bias across algorithms over time.
Notes: Shown are the measured in- and out-of-window accuracy and bias for 29 machine-learning models. The gray-colored lines represent 29 models and the solid black line is an average value across all the models: (4) torture (top-left), (5) political (top-right), (6) imprisonment (bottom-left), and (7) political disappearances (bottom-right). We use only the relevant sections of the SD reports, based on the measures' coding rules.

the SD reports pertinent to each physical integrity score coded by CIRI, in accordance with the CIRI codebook. It is noticeable that in Figure 4, torture (Model 4) and political imprisonment (Model 5) display increasing accuracy over time (by about 25 to 40 per cent), whereas accuracy for extrajudicial killings (Model 6) and enforced disappearances (Model 7) remains relatively constant, and we observe little, if any, change.[20] This suggests that monitoring and reporting bias primarily affects CIRI's torture and political imprisonment indicators, rather than those for extrajudicial killings and enforced disappearances. In other words, changes in monitoring/reporting are driven primarily through torture and political imprisonment. We run a similar exercise for robustness, using AI reports on the disaggregated CIRI score, and results remain largely the same.[21] Unlike previous findings, according to which all CIRI components except the political imprisonment indicator are affected by monitoring and reporting bias (Fariss 2014), we show here that despite this historic emphasis on political imprisonment, its norms have changed,

[20]We do McNemar's test (Raschka 2018) and find that Models 4 and 5 also reject the null hypothesis, $p = 6.054 \times 10^{-20}$ and $p = 1.7904 \times 10^{-18}$.

[21]CIRI only uses the AI reports as a secondary source material for their scores, and our robustness test thus does emulate CIRI's coding scheme.

https://doi.org/10.1017/S0007123421000661 Published online by Cambridge University Press

whereas the texts in the reports regarding extrajudicial killings and disappearances changed little, if at all. Again, as a robustness check, we also tried five-year training window sizes and different out-of-window testing sizes (three years and seven years), and the results are largely consistent with our findings. Also, unlike informal texts, such as social media or online contents, the SD and AI documents are official annual reports that have organized structures (for example, writing styles) that rarely change over time. It is difficult to think that the results are simply derived from the changes in document styles.

So far, we have shown evidence of changes within the reports that directly bias the mapping of reports to scores. *However, what is unclear is whether this bias may be introduced either in the first stage, through reporting and monitoring bias, or in the second stage, through coder bias.* In the next section, we directly test whether coder bias is present, using an experiment on PTS coders that examines whether *internal* standards of coding are changing over time. To preview our results, we find no evidence of coder bias.

Experiment: Identifying Coder Bias

In his analysis of human rights measures, Fariss (2014) assumes that the human coders apply coding rules or standards consistently over time, and both the PTS and CIRI project report high intercoder reliability within a given reporting year. Haschke and Arnon (2020) raise coder bias as an additional potential source of temporal bias, but up to this point, the absence (or presence) of reports-to-scores bias in the coder stage has never been demonstrated. Here, we discuss our efforts to evaluate whether coders or the context in which coders produce human rights scores introduce bias into human rights measures.

To test whether reports-to-scores bias is introduced during the coding stage, we run an experiment with the PTS coders who translate the annual SD reports into human rights scores. The experiment is intended to discover whether coders today code differently than coders in the past, or whether the context in which coders operate today is different than that of past coders. Furthermore, the analysis described in "Methods and Data" cannot determine if bias is introduced in the first or second stage, and the experiment described in the following is designed to pinpoint the location of bias. If we can rule out coder bias, any bias must have been introduced in the first stage—as part of monitoring and reporting.

For the coding of the SD reports for 2015, we assigned to each coder a random sample of SD reports from 2005, such that half of the assigned reports a coder received were from 2005 and half covered events that took place in 2015.[22] All reports were detemporized, meaning that any temporal information in the report was redacted for both the 2005 and 2015 reports, with the hope that coders would not know whether they were coding a contemporary report (covering human rights conditions in 2015) or a historic one (covering human rights condition in 2005).[23] We alerted all coders to the fact that we were conducting an experiment; however, they were not informed of the purpose of the experiment.[24]

If current coders assign different scores to SD reports from 2005 than coders who originally scored the 2005 reports in 2006, we interpret this difference as evidence for coder bias. As such, it is reasonable to conclude that current coders are applying different standards to human rights reports or operate in a different information environment or context than coders in 2006. However, if there is no difference between the scores assigned to the 2005 SD reports in 2006 and the scores produced in 2016, we are able to rule out coder bias, at least for the more recent

[22]This took place in the Fall of 2016.

[23]In some cases, the contextual information in the report may have tipped off the coder. A report describing stable political conditions in Syria, for example, is easily identified as an old one, whereas a report covering abuses in South Sudan could not have been published in 2005. Additional robustness tests are conducted to exclude the possibility of biased inference.

[24]Asking coders to read partially redacted reports, of course, necessitated informing coders of some of the details of the experiment being conducted.

https://doi.org/10.1017/S0007123421000661 Published online by Cambridge University Press

years.[25] As such, any bias present in human rights scores would most likely have been introduced during the monitoring and reporting stage, consistent with the evidence presented in the previous section.

PTS accounting

To translate human rights reports into human rights scores, the PTS research group randomly assigns each human rights report to three coders. Random assignment ensures that no one coder codes only reports covering human rights conditions from countries of a specific region, size, or level of human rights abuse. Reports are scored on an ordinal scale ranging from 1 to 5—with higher scores indicating worse human rights conditions (see the Online Appendix). After all coders have submitted scores for their assigned reports, reports on which coders disagreed are identified and coded in a second (or third) round of coding until disagreements are resolved and scores are published.[26]

The experimental design

As part of the annual coding efforts in 2016, we conducted an *audit experiment*. An audit study is "a specific type of field experiment primarily used to test for discriminatory behavior when survey and interview questions induce social desirability bias" (Gaddis 2018, 3). The researcher typically randomizes one or more characteristics of an auditor (for example, their race or age) to evaluate the effect those characteristics have on some outcome.

In our experiment, coders were randomly assigned 50 to 60 detemporized country reports from 2005, in addition to 50 to 60 redacted reports covering human rights conditions in 2015, as part of the PTS project's routine annual coding. We then compared human rights scores for human rights reports from 2005 coded in 2016 to human rights scores for human rights reports from 2005 coded in 2006. As such, the altered characteristic in our experiment is the "identity" of the coder—is the coder a coder in 2006 or 2016. We refer to this experiment as audit-like because we do not directly randomize characteristics of an individual, but rather use randomization to ensure that coders do not know whether they are coding a current or historic human rights report. This produces an appropriate comparison of two coders (with varying characteristics) evaluating the same report.

All reports (reports covering human rights conditions in 2005 and reports pertaining to 2015) were redacted to exclude all obvious temporal identifiers so that coders could not determine whether they were coding a current report (that is, a report covering events of 2015) or an old one (that is, a report covering human rights conditions in 2005). We chose the year 2005 with the hope that human rights conditions and consequently the content of the reports themselves had changed sufficiently over a ten-year period. We initially considered choosing reports from the 1990s or 1980s but were wary that reports from that period would have easily been identified and distinguishable from current reports given stylistic differences, as well as references to geopolitical contexts (for example, the Cold War). Other possible strategies were also considered, for example, choosing multiple treatment years. This possibility was discarded, as the number of additional reports coders would be required to read and code to ensure sufficient power would have overwhelmed coders. As described earlier, all reports—in this case, roughly 200 reports from 2005 and 200 reports from 2015—were randomly assigned to three coders each, ensuring that no coder received two reports covering the same country or territory.[27] Once all

[25]This also requires the assumption that 2005 reports are representative of reports around this period.

[26]Disagreement is typically rare and fewer than 5 per cent of reports have to be recoded. In exceptionally rare cases where disagreement cannot be resolved, an additional coder serves as a tiebreaker (Wood and Gibney 2010). Coders almost never diverge by more than one level in their assigned scores.

[27]For example, a coder would not be asked to code both the Afghanistan report for 2005 and the Afghanistan report for 2015.

https://doi.org/10.1017/S0007123421000661 Published online by Cambridge University Press

coders had submitted their scores for every assigned report, all disagreements over scores were adjudicated and scores were tabulated.

The procedure outlined earlier and summarized in Figure 5 provides us with two sets of scores for human rights conditions covered in the 2005 human rights reports. For every country report from 2005, we have one *original* score produced in 2006 and one *new* score produced in 2016. Comparison of the two sets of scores allows us to assess whether coders (or the context in which scores are assigned to reports) have changed over time and are introducing bias into human rights scores.

We compute a simple two-sample Welch's unequal variances *t*-test to test the hypothesis that the mean of PTS scores for the 2005 SD reports originally coded in 2006 is equal to the mean of the 2005 SD reports coded in 2016.[28] A nonzero difference of means would indicate the presence of coder bias. A positive difference would suggest that coders in 2016 assign, on average, lower scores (indicating better human rights conditions) than coders coding in 2006. In other words, current coders code less harshly. A negative difference would suggest that scores are biased in the opposite direction, with coders in 2016 coding the 2005 human rights reports more stringently than coders coding in 2006. Estimates and standard errors are presented in Table 2.

Findings of audit experiment

Figure 6 shows the distribution of scores produced in 2016 and the distribution of scores originally coded in 2006. Over two thirds of reports were coded identically (in 2006 and 2016), with 21 reports receiving a higher score and 38 reports receiving a lower score from coders in 2016 than from the original coders coding in 2006. This evidence suggests that coding standards may indeed have changed slightly over time. However, the average effect is small and positive, suggesting that more recent coders actually code *less* rather than more stringently. Moreover, and as shown in Table 2, the difference of means is not distinguishable from zero.

Discussion and Conclusion

Other conditions that are closely related to human rights, such as democracy, the rule of law, and economic development, have made a remarkable improvement over the past few decades. It is difficult to deny that there exist biases in standards-based human rights scores that do not properly reflect actual human rights conditions. This article pays close attention to, and explores the potential sources of bias in, standards-based human rights measures. We first discuss biases that originate at the monitoring and reporting stage, which comprise both "changing norms" and "information effects." According to the changing norms argument, bias is a function of changing human rights norms. Information effects are biases attributable to the increased capacity of reporting agencies to gather and report information on human rights violations. Importantly, these events-to-reports biases affect the actual human rights reports and their content over time.[29] We also explore the possibility that bias is introduced during a second stage after the reports are compiled and published—that is, as part of the translating of reports into human rights scores. We call this "reports-to-score bias" in the coder stage, where the coders' interpretations and application of coding rules may change over time. This mechanism is distinct from events-to-reports bias, which occurs as part of the process of compiling the reports themselves.

To evaluate whether bias, from the first stage and/or the second stage, is introduced, we employ a supervised machine-learning approach to first test for the presence of bias. This allows us to uncover whether the mapping of the report text to the standards-based human rights scores

[28]Our estimator, the difference of means (that is, $\overline{\text{Score}}_{\text{original}} - \overline{\text{Score}}_{\text{new}}$) is equivalent to the average treatment effect (ATE).

[29]While we distinguish between these two conceptual mechanisms of bias, we do not disentangle the two in our empirical tests.

https://doi.org/10.1017/S0007123421000661 Published online by Cambridge University Press

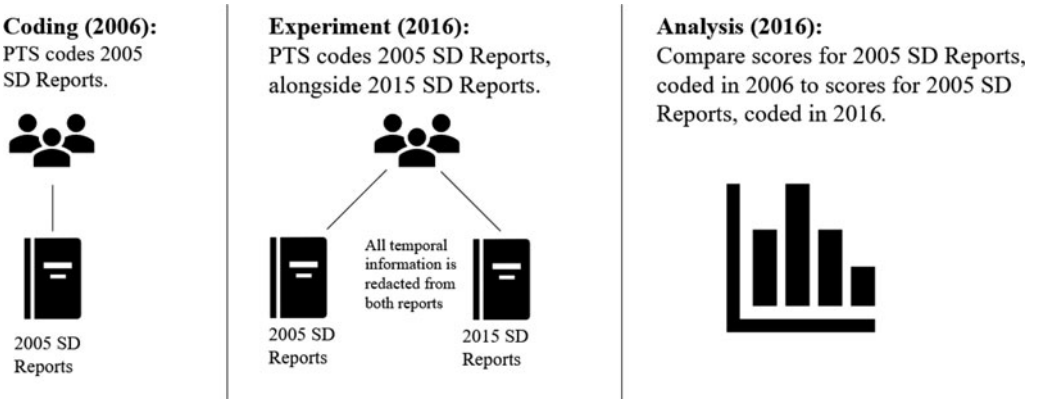

Figure 5. Experimental design.
Notes: In 2006, reports were first coded contemporaneously. Ten years later, during the coding of 2015 reports, coders were given reports from both 2015 and 2005, with all temporal information redacted. We then compared the originally coded 2005 reports with the scores of the same reports assigned in 2016. Results of the comparison are reported in the following.

Table 2. Difference of means between original and new scores

Difference in means $\overline{\text{Score}}_{\text{original}} - \overline{\text{Score}}_{\text{new}}$	SE Type	SE	p-value	95% CI Lower bound	95% CI Upper bound
0.12	Welch *t*-test	0.12	0.32	−0.35	0.11
0.12	Bootstrap	0.12	0.32	−0.32	0.16

Notes: Shown are the difference in means, standard errors (SE) (estimated using bootstrapping and Welch's nonequal variance modification to the degrees of freedom), p-values, and 95 per cent confidence intervals (CIs).

have changed over time, which includes both events-to-reports bias *and* reports-to-scores bias. We find substantially and statistically significant evidence that the mapping of reports to scores has changed. Specifically, we find a 20–40 per cent increase in model performance, indicating that meaningful changes have occurred in the translation of reports to scores. We also find that there has been negative bias across the years, meaning that evaluating standards have become harsher. When re-employing the same method on the disaggregated CIRI scores for torture, political imprisonment, extrajudicial killings, and enforced disappearances, we find that the biases are concentrated primarily in torture and political imprisonment, where we again find a 25–40 per cent increase in model performance. On the other hand, the models for extrajudicial killings and enforced disappearances did not show meaningful increases in model performance, indicating that the source of bias in CIRI scores is more likely to emanate from the former rather than the latter. Importantly, this departs from previous findings that all but political imprisonment are affected by these mechanisms of bias.

While the supervised machine-learning approach indicates that biases are introduced temporally to standards-based human rights scores, it is unclear whether their source is in the monitoring and reporting stage or emerge from coder bias. As such, we employ an experiment on the human coders of PTS. If the interpretations and application of rules have changed over time, even unknowingly, then we would expect reports coded in a later period to change significantly from reports coded in an earlier period. As part of the experiment, while coding the reports for 2015, PTS coders were randomly assigned to additional reports from 2005. All temporal information from reports were redacted to ensure that coders were using the same set of standards to code the reports. We then compared the scores given to the original 2005 reports with the same 2005

https://doi.org/10.1017/S0007123421000661 Published online by Cambridge University Press

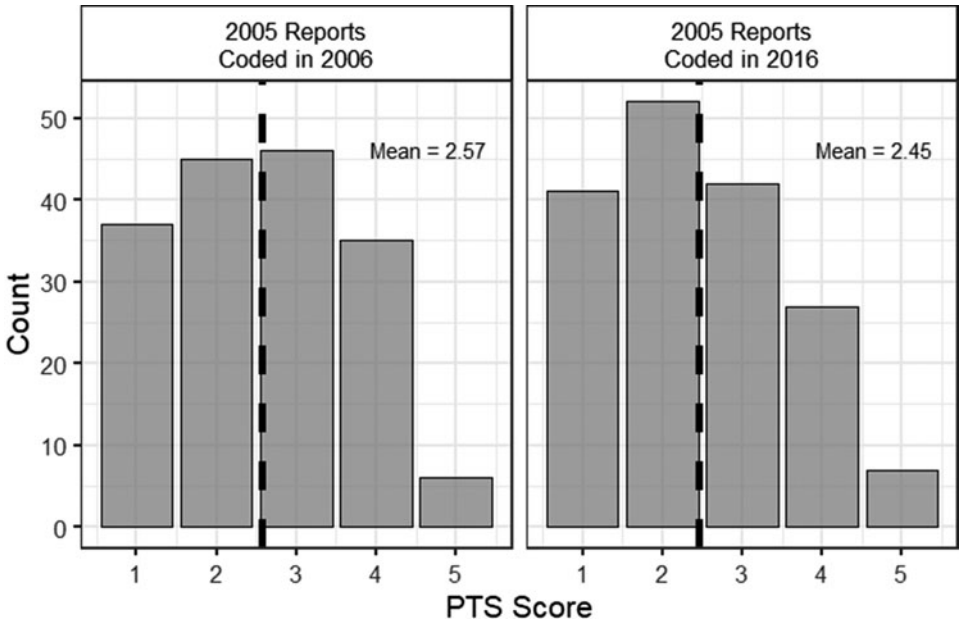

Figure 6. Distribution of scores: 2006 v. 2016.
Notes: Shown is the distribution of PTS scores: scores assigned in 2016 (left panel); and original scores coded in 2006 (right panel).

reports coded in 2016. We find no significant changes in the reports, indicating that reports-to-scores bias is less likely to be a source of bias. Therefore, bias is most likely to be introduced in the monitoring and reporting stage.

These findings lead us to examine the first stage more closely to understand the presence of bias in standards-based human rights scores. Analyzing the hierarchical structure of human rights reports from SD, AI, and HRW, Park, Greene, and Colaresi (2020) find that for the past few decades, there has been a significant information increase in the reports. In this context, although existing literature tends to separate the information effects and changing norms, as we discussed earlier, they inevitably interact with each other. For example, in our analyses of the disaggregated CIRI physical integrity rights scores, we find that bias processes were more likely in torture and political imprisonment (see Figure 4). The reports not only contain more information over time, but also have more hierarchical structures in later years (see Table 6 in the Online Appendix). Especially, the increase in subcategories is noticeable for torture and political imprisonment (arbitrary arrest or detention), whereas extrajudicial killing and disappearance show little to no change. This is consistent with our findings that the source of bias in PTS is more likely to be derived from torture and political imprisonment.

This article interrogates a widely held assumption in the human rights literature regarding changing standards or norms and standards-based measures of human rights. We contribute to this discussion by using novel methods to investigate how and which biases are present in our widely used measures. Scholars have increasingly become aware of the biases that result from the underlying data-generating process, which may systematically skew our understanding of the world through data. For example, Kerner and Crabtree (2018) explore how political interests shape macroeconomic data, and Harmon, Arnon, and Park (2022) explore the political motivations in trafficking-in-persons data. We join this line of inquiry to expand the exploration of bias processes beyond purely "political" motivations to include changing norms and information effects. Arriving at an unbiased human rights measurement is critical for our ability to correctly assess the state of human rights globally. *The right accounting of these wrongs is a critical step in*

https://doi.org/10.1017/S0007123421000661 Published online by Cambridge University Press

this assessment. Methodologically, the article provides useful analytical frameworks that can also be used to explore broader trends and possibilities of bias in similar measures by theorizing and understanding the underlying data-generating process, and simulating the process using machine-learning techniques. Future researchers could use these methods to examine the underlying biases of other standards-based measures, such as performance on religious freedom, labor rights, or women's empowerment. Similarly, the audit experiment can be replicated for similar measures using human coders, or across a longer time span than the current experiment, to ensure that these results generalize out to earlier periods.

More substantively, future research should continue to disentangle potential sources of bias in standards-based human rights measures. Scholars must pay particular attention to: (1) how the reports are created by the reporting agencies; and (2) how the PTS and CIRI research teams produce the scores. Additionally, future research would benefit from attention to how the reports have grown in their conceptualizations of human rights and how this impacts the reports. Lastly, this article has focused only on biases from temporal changes. Future research should also focus on biases in standards-based data that vary with geographic coverage.

Supplementary Material. Online appendices are available at: https://doi.org/10.1017/S0007123421000661

Data Availability Statement. Replication data for this article can be found in Harvard Dataverse at: https://doi.org/10.7910/DVN/KHBT9D

Acknowledgements. The authors would like to thank David Davis, Mark Gibney, Amanda Murdie, Chad Clay, discussants and audience at APSA 2017, GAHRNET 2018 and 2019, and the anonymous reviewers for their helpful comments and suggestions.

Financial Support. None.

Competing Interests. None

References

Arnon D, Haschke P and Park B (2022) Replication data for: The Right Accounting of Wrongs: Examining Temporal Changes to Human Rights Monitoring and Reporting, https://doi.org/10.7910/DVN/KHBT9D, Harvard Dataverse, V1

Bagozzi BE and Berliner D (2018) The politics of scrutiny in human rights monitoring: evidence from structural topic models of US State Department human rights reports. *Political Science Research and Methods* **6**(4), 661–677.

Breiman L (2001) Random forests. *Machine Learning* **45**(1), 5–32.

Bueno de Mesquita B et al. (2005 [2003]) *The Logic of Political Survival* (pbk edn). Cambridge, MA: MIT Press.

Cingranelli D and Filippov M (2018) Problems of model specification and improper data extrapolation. *British Journal of Political Science* **48**(1), 273–274.

Cingranelli DL, Richards DL and Clay KC (2014) The CIRI human rights dataset. Available from http://www.humanrightsdata.com

Clark AM and Sikkink K (2013) Information effects and human rights data: is the good news about increased human rights information bad news for human rights measures? *Human Rights Quarterly* **35**(3), 539–568.

Collobert R et al. (2011) Natural language processing (almost) from scratch. *Journal of Machine Learning Research* **12** (August), 2493–2537.

Conrad CR and DeMeritt JHR (2011) Human rights advocacy and state repression substitutability. Paper presented at the 2011 Annual Meeting of the American Political Science Association, Seattle, USA.

Conrad CR and Moore WH (2010) What stops the torture? *American Journal of Political Science* **54**(2), 459–476.

Cortes C and Vapnik V (1995) Support-vector networks. *Machine Learning* **20**(3), 273–297.

Davenport C (2007a) *State Repression and the Democratic Peace*. Cambridge, MA: Cambridge University Press.

Davenport C (2007b) State repression and the tyrannical peace. *Journal of Peace Research* **44**(4), 485–504.

Eck K and Fariss C (2018) Ill treatment and torture in Sweden: a critique of cross-case comparison. *Human Rights Quarterly* **40**, 591–604.

Fariss C (2014) Respect for human rights has improved over time: modeling the changing standard of accountability. *American Political Science Review* **108**(2), 297–318.

Fariss CJ (2018) The changing standard of accountability and the positive relationship between human rights treaty ratification and compliance. *British Journal of Political Science* **48**(1), 239–271.

Fariss C (2019) Yes, human rights practices are improving over time. *American Political Science Review* **113**(3), 868–881.

Fariss C and Schnakenberg KE (2014) Measuring mutual dependence between state repressive actions. *Journal of Conflict Resolution* **58**(6), 1003–1032.

Gaddis SM (2018) An introduction to audit studies in the social sciences. In Gaddis SM (ed.), *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*. Cham, Switzerland: Springer International Publishing, 3–44.

Gibney M et al. (2019) The Political Terror Scale 1976–2019. Available from http://www.politicalterrorscale.org

Greene KT, Park B and Colaresi M (2019) Machine learning human rights and wrongs: how the successes and failures of supervised learning algorithms can inform the debate about information effects. *Political Analysis* **27**(2), 223–230.

Harmon R, Arnon D and Park B (2022) TIP for Tat: Political Bias in Human Trafficking Reporting. *British Journal of Political Science* **52**(1), 445–455.

Haschke P (2018) *Human Rights in Democracies*. New York, NY: Routledge.

Haschke P and Arnon D (2020) What bias? Changing standards, information effects, and human rights measurement. *Journal of Human Rights* **19**(1), 33–45.

Haschke P and Gibney M (2018) Are global human rights conditions static or dynamic? In Backer DA, Bhavnani R, and Huth PK (eds), *Peace and Conflict 2017*. New York, NY: Routledge, 88–100.

Henderson C (1982) Military regimes and rights in developing countries: a comparative perspective. *Human Rights Quarterly* **4**, 110–123.

Hill DW Jr, Moore WH and Mukherjee B (2013) Information politics versus organizational incentives: when are Amnesty International's "naming and shaming" reports biased? *International Studies Quarterly* **57**(2), 219–232.

Kalchbrenner N, Grefenstette E and Blunsom P, (2014) A convolutional neural network for modelling sentences. In Kristina, T and Wu H (eds.), *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore, Maryland: Association for Computational Lin (Volume 1), 58–68. arXiv preprint arXiv:1404.2188.

Keck ME and Sikkink K (1998) *Activists beyond Borders: Advocacy Networks in International Politics*. Ithaca, NY: Cornell University Press.

Kerner A and Crabtree C (2018) The Political Economy of Data Production. Unpublished Manuscript.

Landman T and Carvalho E (2009) *Measuring human rights*. London, UK: Routledge.

Lewis D (1998) Naive (Bayes) at forty: The independence assumption in information retrieval. In Nédellec C., Rouveirol C. (eds.), *Machine Learning: ECML-98. ECML 1998.* Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence). Springer: Berlin, Heidelberg, vol. 1398, 4–15. https://doi.org/10.1007/BFb0026666

McCullagh P and Nelder J (1989) *Generalized Linear Models*. 2nd Edition, London, UK: Chapman and Hall. http://dx.doi.org/10.1007/978-1-4899-3242-6

Murdie A, Davis DR and Park B (2020) Advocacy output: automated coding documents from human rights organizations. *Journal of Human Rights* **19**(1), 83–98.

Pang B, Lee L and Vaithyanathan S (2002) Thumbs up? Sentiment classification using machine learning techniques. In Jan H and Matsumoto Y (eds.), *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*. Prague, Czech Republic: Association for Computational Linguistics, 79–86. arXiv preprint cs/0205070. https://aclanthology.org/W02-1011

Park B, Greene K and Colaresi M (2020) Human rights are (increasingly) plural: Learning the changing taxonomy of human rights from large-scale text reveals information effects. *American Political Science Review* **114**(3), 888–910.

Park B, Murdie A and Davis DR (2019) The (co)evolution of human rights advocacy: understanding human rights issue emergence over time. *Cooperation and Conflict* **54**(3), 313–334.

Pennington J, Socher R and Manning CD (2014) GloVe: global vectors for word representation. In Alessandro A, Pang B and Daelemans W (eds.) *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 1532–1543. http://www.aclweb.org/anthology/D14-1162

Poe S et al. (1994) Human rights and US foreign aid revisited: the Latin American region. *Human Rights Quarterly* **16**(3), 539.

Poe SC and Tate NC (1994) Repression of human rights to personal integrity in the 1980s: a global analysis. *American Political Science Review* **88**(4), 853–872.

Poe SC, Carey SC and Vazquez TC (2001) How are these pictures different? A quantitative comparison of the US State Department and Amnesty International human rights reports, 1976–1995. *Human Rights Quarterly* **23**, 650–677.

Potz-Nielsen C, Ralston R and Vargas TR (2018) Recording abuse: how the editing process shapes our understanding of human rights abuses. Working paper.

Raschka S (2018) Model evaluation, model selection, and algorithm selection in machine learning. Available from http://arxiv.org/abs/1811.12808

Richards DL (2016) The myth of information effects in human rights data: response to Ann Marie Clark and Kathryn Sikkink. *Human Rights Quarterly* **38**(2), 477–492.

Ron J, Ramos H and Rodgers K (2005) Transnational information politics: NGO human rights reporting, 1986–2000. *International Studies Quarterly* **49**(3), 557–588.

Roser M (2018) Democracy. Available from https://ourworldindata.org/democracy

Roser M and Ortiz-Ospina E (2018) Global extreme poverty. Available from https://ourworldindata.org/extreme-poverty

Roth K (2004) Defending economic, social and cultural rights: practical issues faced by an international human rights organization. *Human Rights Quarterly* **26**, 63–73.

Simmons BA (2009) *Mobilizing for Human Rights: International Law in Domestic Politics*. Cambridge, MA: Cambridge University Press.

Wood RM and Gibney M (2010) The Political Terror Scale (PTS): a re-introduction and a comparison to CIRI. *Human Rights Quarterly* **32**(2), 367–400.

Yih W-T, Toutanova K, Platt JC and Meek C (2011) Learning discriminative projections for text similarity measures. In Pradhan S (ed.), *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*. Oregon, USA: Association for Computational Linguistics, 247–256. https://aclanthology.org/W11-0329

Zhou C, Sun C, Liu Z and Lau F (2015) A C-LSTM neural network for text classification. arXiv preprint arXiv:1511.08630.

Cite this article: Arnon D, Haschke P, Park B (2023). The Right Accounting of Wrongs: Examining Temporal Changes to Human Rights Monitoring and Reporting. *British Journal of Political Science* **53**, 163–182. https://doi.org/10.1017/S0007123421000661