Draft Submission to B.J.Pol.S. **XX,** X–XX *Cambridge University Press, XXXX*
doi:doi:10.1017/XXXX

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

Daniel Arnon, Peter Haschke, and Baekkwan Park

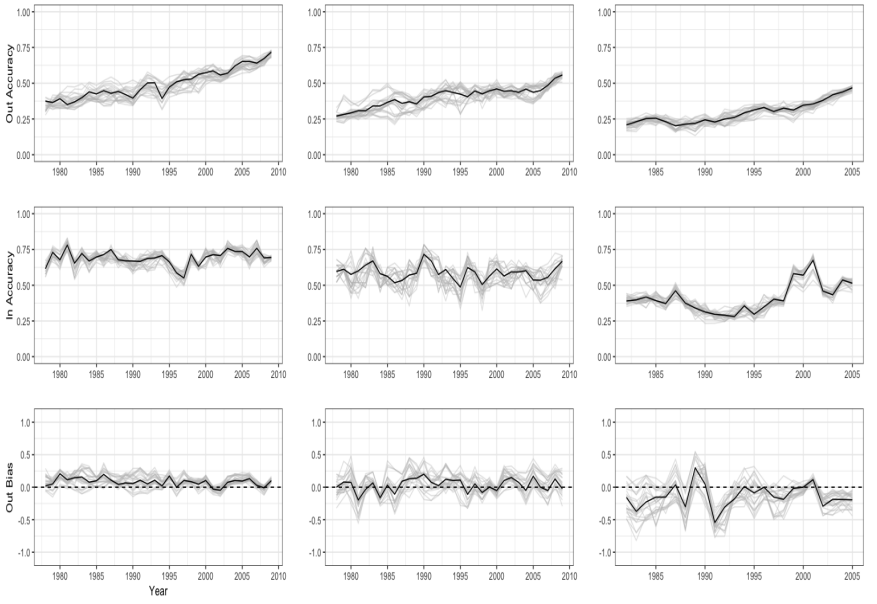
Forward forecasting

We also run different rolling-window sizes: 3 year, 5 year, and 7 year. The overall settings are the same as the 10 year models. Regardless of the window size, the overall results are consistent with the results in the main paper.

	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

TABLE 1: Data (Texts) and Labels for 3 year, 5 year, and 7 year window

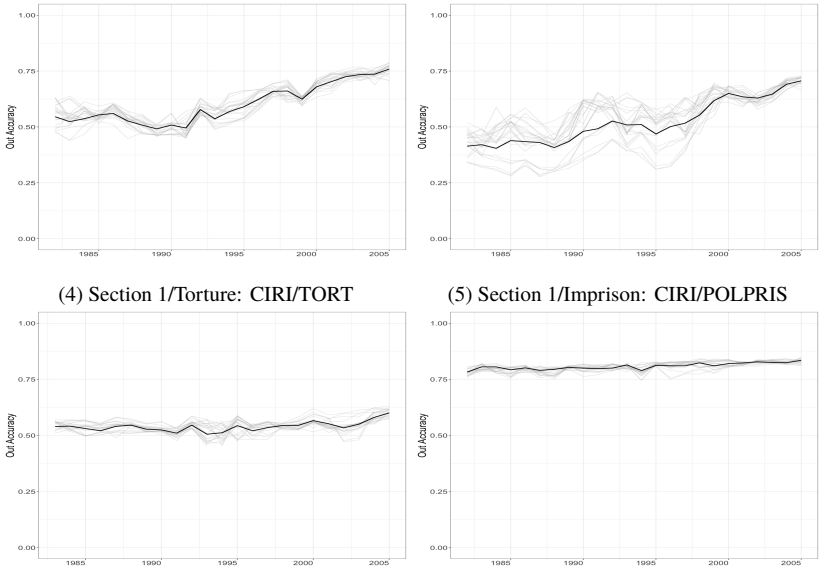
Note: SD (State Department Reports), AI (Amnesty International Annual Reports). PTS are available from 1976-2016 and CIRI Scores are available from 1981 to 2011. SD reports are available from 1978 and AI reports are available from 1976.



(1) Section 1: PTS-SD (2) AI: PTS-AI (3) Section 1: CIRI/PHYSINT

Figure 1: Accuracy and Bias across Algorithms over Time (3 year)

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

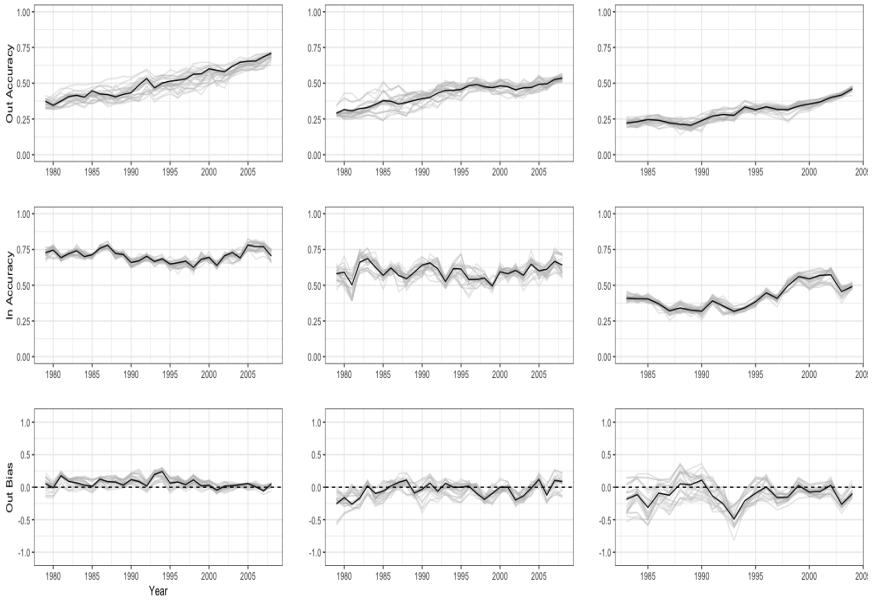


(4) Section 1/Torture: CIRI/TORT (5) Section 1/Imprison: CIRI/POLPRIS

(6) Section 1/Kill: CIRI/KILL (7) Section 1/Disap.: CIRI/DISAP

Figure 2: Accuracy and Bias Across Algorithms over time (3 year)

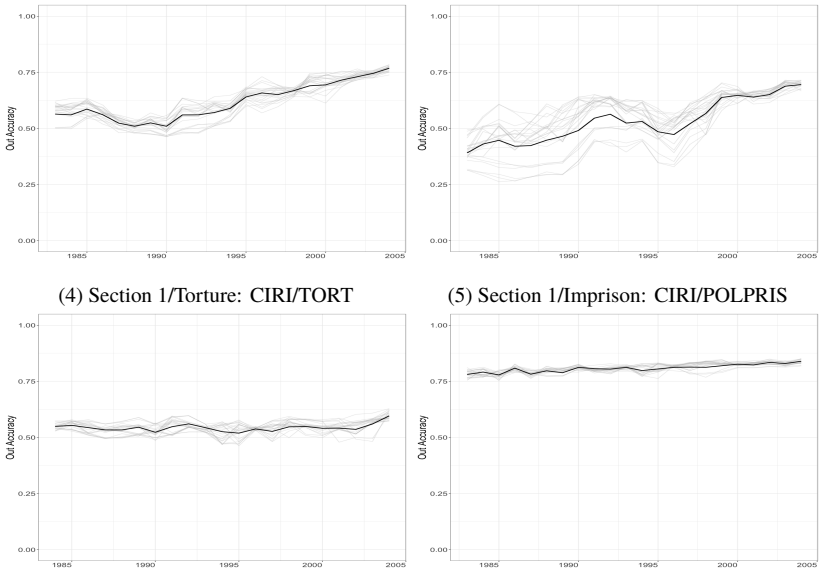
Note: Shown are the measured in- and out-window accuracy and bias for 29 machine learning models. The gray colored lines represents 29 models and the solid black line is an average value for across all the models: (4) torture (top-left), (5) political (top-right), (6) imprisonment (bottom-left), (7)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



(1) Section 1: PTS-SD (2) AI: PTS-AI (3) Section 1: CIRI/PHYSINT

Figure 3: Accuracy and Bias across Algorithms over Time (5 year)

Note: Shown are the measured in- and out-window accuracy and bias for 29 machine learning models. The gray colored lines represents 29 models and the solid black line is an average value for across all the models. Left panel uses Section 1 of SD report on PTS-SD scores. Middle panel uses AI reports on PTS-AI scores. Right panel uses only Section 1 of the SD report on CIRI aggregate scores. The top panels, out-window-accuracy, show an increasing slope indicating that for all of these measures, standards have changed. Note that each year in the plots represents the midpoint of the 5-year training window.



(4) Section 1/Torture: CIRI/TORT

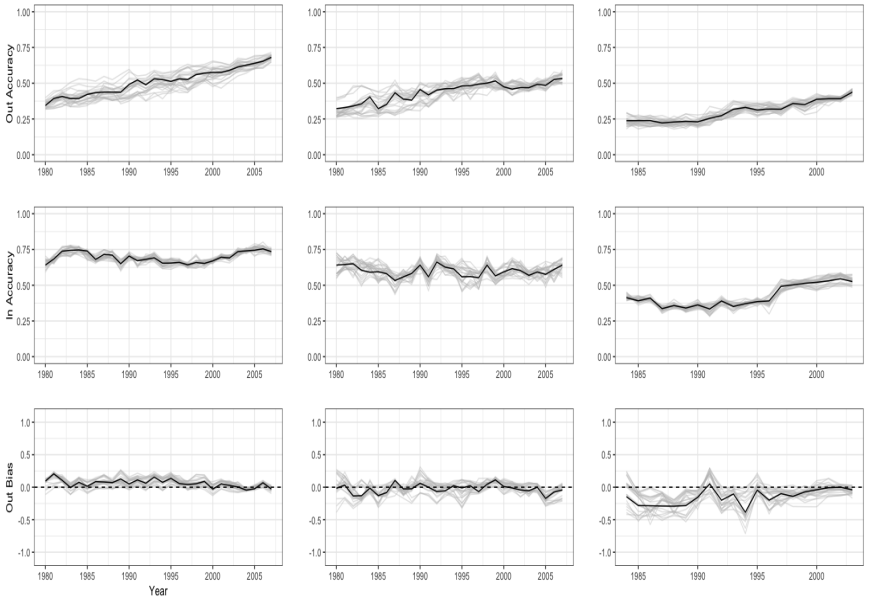
(5) Section 1/Imprison: CIRI/POLPRIS

(6) Section 1/Kill: CIRI/KILL

(7) Section 1/Disap.: CIRI/DISAP

Figure 4: Accuracy and Bias Across Algorithms over time (5 year)

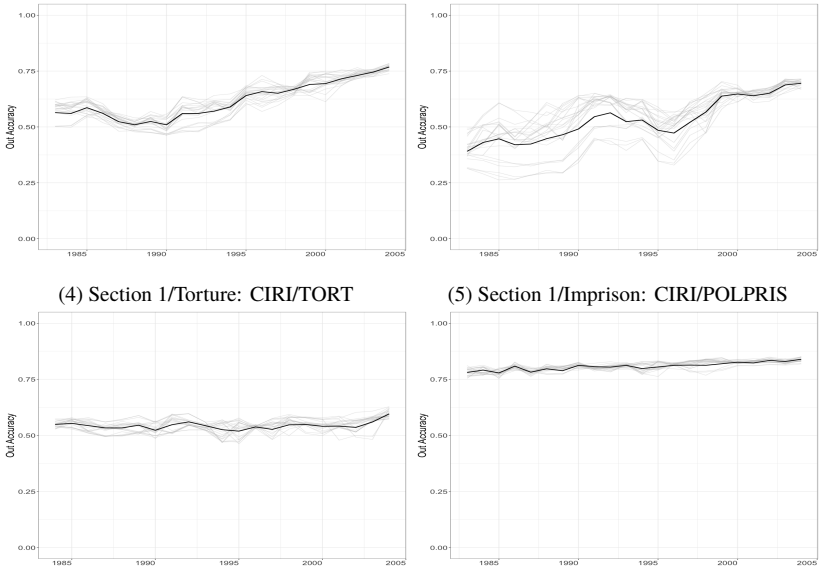
Note: Shown are the measured in- and out-window accuracy and bias for 29 machine learning models. The gray colored lines represents 29 models and the solid black line is an average value for across all the models: (4) torture (top-left), (5) political (top-right), (6) imprisonment (bottom-left), (7)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



(1) Section 1: PTS-SD (2) AI: PTS-AI (3) Section 1: CIRI/PHYSINT

Figure 5: Accuracy and Bias across Algorithms over Time (7 year)

Note: Shown are the measured in- and out-window accuracy and bias for 29 machine learning models. The gray colored lines represents 29 models and the solid black line is an average value for across all the models. Left panel uses Section 1 of SD report on PTS-SD scores. Middle panel uses AI reports on PTS-AI scores. Right panel uses only Section 1 of the SD report on CIRI aggregate scores. The top panels, out-window-accuracy, show an increasing slope indicating that for all of these measures, standards have changed. Note that each year in the plots represents the midpoint of the 7-year training window.



(4) Section 1/Torture: CIRI/TORT (5) Section 1/Imprison: CIRI/POLPRIS

(6) Section 1/Kill: CIRI/KILL (7) Section 1/Disap.: CIRI/DISAP

Figure 6: Accuracy and Bias Across Algorithms over time (7 year)

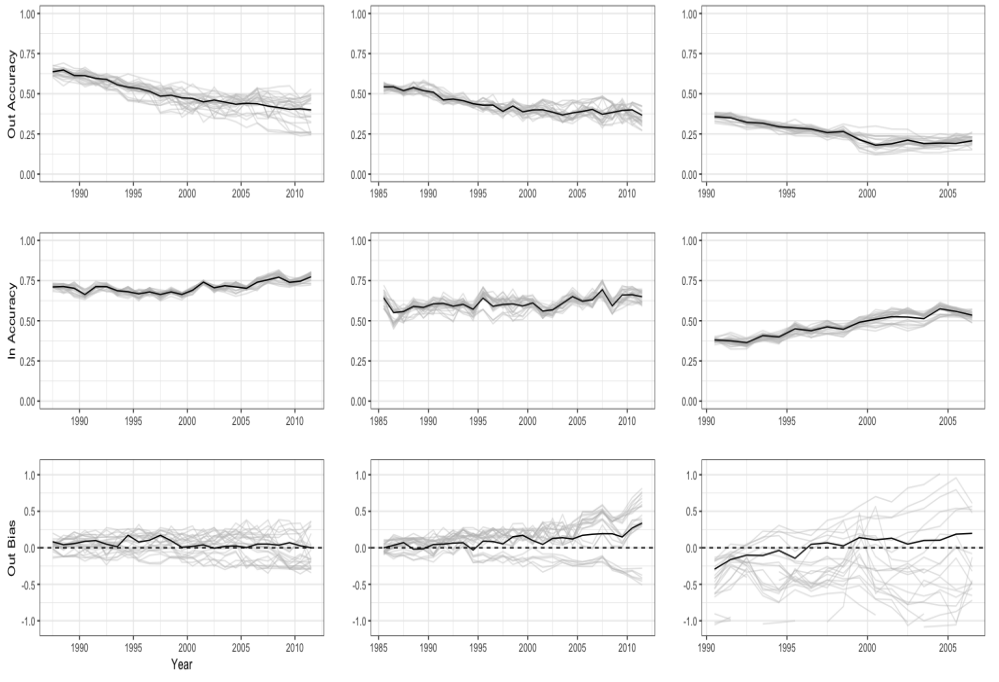
Note: Shown are the measured in- and out-window accuracy and bias for 29 machine learning models. The gray colored lines represents 29 models and the solid black line is an average value for across all the models: (4) torture (top-left), (5) political (top-right), (6) imprisonment (bottom-left), (7)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

Fixed Backward Rolling Window Forecasting

In order to forecast backward, we reverse the training and testing processes. By dividing the training data by 10 years, we create twenty six 10-year-in-window sets from the training sets (2015-1982), $D_{\text{train},\mathcal{W}_t}$, $t \in (1, ..., 26)$. For example, $\mathcal{W}_1$: 2015-2006, $\mathcal{W}_2$: 2014-2005, ... $\mathcal{W}_{26}$: 1991-1982 for *in-windows*. For a trained classification model for each in-window, $\Phi_{\mathcal{W}_t} = \hat{f}(D_{\text{train},\mathcal{W}_t})$, we estimate the extent to which Φ approximates the unknown function f on the *out-of-window testing set* (1977-1981), $D_{\text{test},\mathcal{W}_{\text{out}}}$. For example, at $\mathcal{W}_1$, an algorithm based on the texts from 2015-2006 (10 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 2014-2005 (10 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 2002-2011 (final window) and compare performance for both *(*in-window) and *out-of-window* for each window. Therefore, if the changes in the information environment or/and 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 $\mathrm{P}_{X_{\text{out}}}(\Phi_{\mathcal{W}_1}(x) = f(x)) = \mathrm{P}_{X_{\text{out}}}(\Phi_{\mathcal{W}_2}(x) = f(x)) \ldots = \mathrm{P}_{X_{\text{out}}}(\Phi_{\mathcal{W}_{26}}(x) = f(x))$.

Figure 7 below shows the results of the first sets of analyses. The top panel in Figure 7 displays the out-of-window prediction accuracy for each of our algorithms trained on 10 year rolling windows. As discussed in the main paper, 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 equally and the accuracy measure ($\mathrm{P}_{X_{\text{out}}}(\Phi_{\mathcal{W}_t}(x) = f(x))$) should be the same across years. That is, there would be no meaningful changes in out-of-window accuracy over time.

As illustrated in Figure 7 (top panel in each plot), trained models perform better as they get closer to the out-of-window test set, indicated by the consistent downward slope of the out-window accuracy over the years. In general, across models (1), (2), and (3), we see a 10 to 30% deccrease in model performance. Model (3) (Section 1 with CIRI-Physical Integrity Scores) seems to show a little lower performance throughout the years compared to Model (1) and (2), but given that CIRI-PHYSINT has 9 classes, the baseline accuracy is much lower (0.11), and shows consistent



(1) Section 1: PTS-SD (2) AI: PTS-AI (3) Section 1: CIRI/PHYSINT

Figure 7: Accuracy and Bias across Algorithms over Time-Backward

Note: Shown are the measured in- and out-window accuracy and bias for 29 machine learning models. The gray colored lines represents 29 models and the solid black line is an average value for across all the models. Left panel uses Section 1 of SD report on PTS-SD scores. Middle panel uses AI reports on PTS-AI scores. Right panel uses only Section 1 of the SD report on CIRI aggregate scores. The top panels, out-window-accuracy, show an increasing slope indicating that for all of these measures, standards have changed. Note that each year in the plots represents the midpoint of the 10-year training window.

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.[1] From Model (1) to (3), we reject the null hypothesis that the accuracies from these two models are equal.[2] Substantively, it means that the earliest training models can predict about 180 to 280 human rights scores more accurately than the latest models. The mapping functions trained in earlier years are more likely to predict the human rights scores accurately in the testing years (1977-1981).

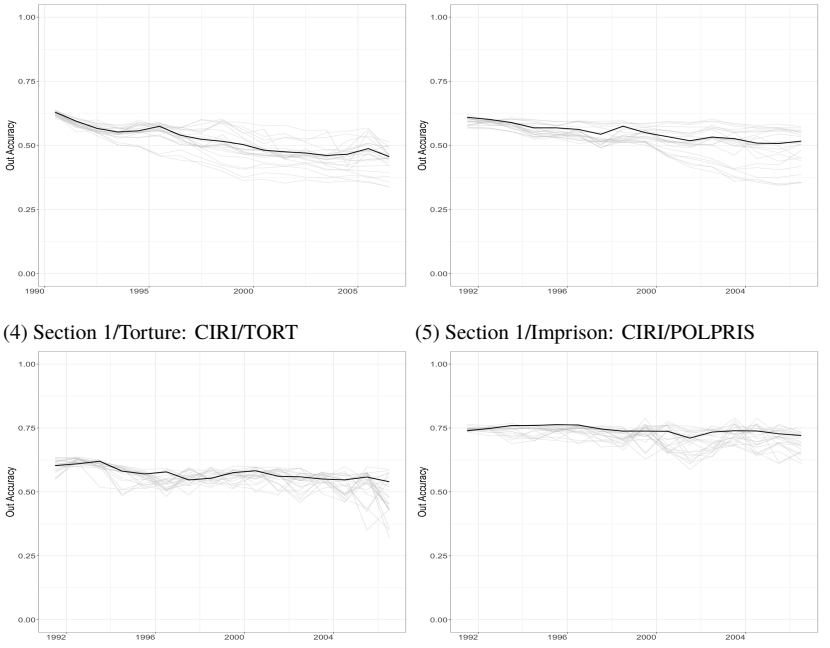
[1]We choose the majority voting classification model for the test.
[2]Model (1): $p = 3.104 \times 10^{-50}$, Model (2): $p = 9.227 \times 10^{-16}$ Model (3): $p = 7.041 \times 10^{-9}$

There were substantial changes in translating the reports to standards-based human rights measures. The middle panel in each plot in Figure 7 shows the in-window accuracies (performance) at each training window (tested in-window test set). Although there were some fluctuation, they do not appear to be increasing or decreasing across time. That is, there is no meaningful change in the ability of the algorithms to estimate the unknown mapping function $f(\cdot)$ of texts to scores over time. The bottom panel in each plot in Figure 7 indicates the average bias from the out-window test set, $\frac{1}{n}\sum_{i=1}^{n}(\Phi_{W_t}(x_i) - y_i)$, where $\Phi_{W_t}(x_i)$ is a model prediction and y_i is the true label for each data point x_i in the test set. That is, it is the average difference between the actual labels and the model predictions. For Model (1) and Model (2), the average bias across all classification algorithms gradually decreases as they get closer to the end. In early years, there is ***positive bias*** (over-prediction) on the testing set. Because CIRI's Physical Integrity Rights are reversed scale, Model (3) shows ***negative bias*** (under-prediction).

Next, Figure 8 shows the results of out-window accuracy for each of the dis-aggregated CIRI physical integrity indicators (e.g., torture, political imprisonment). In order to emulate the data generating process more accurately, we used only the relevant subsections from the SD reports pertinent to each physical integrity score coded by CIRI, in accordance with the CIRI code-book. Notice that in Figure 8, torture (4) and political imprisonment (5) display decreasing accuracy over time (by about 25 to 30%), whereas accuracy for extrajudicial killings (6) and enforced disappearances (7) remains relatively constant and we observe little if any change.[3] 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. That is, changes in monitoring/reporting are driven primarily through torture and political imprisonment.

Overall, the findings in the backward rolling forecasting is consistent with the findings of the forward rolling forecasting (just the opposite direction). These findings are expected because there is no meaningful differece between the forward and the backward forecasting and testing the same dynamics in reversed orders.

[3]We do McNemar's Test (Raschka 2018) and find that model (4) and model (5) also reject the null hypothesis $p = 4.190 \times 10^{-19}$ and $p = 1.382 \times 10^{-15}$



(4) Section 1/Torture: CIRI/TORT

(5) Section 1/Imprison: CIRI/POLPRIS

(6) Section 1/Kill: CIRI/KILL

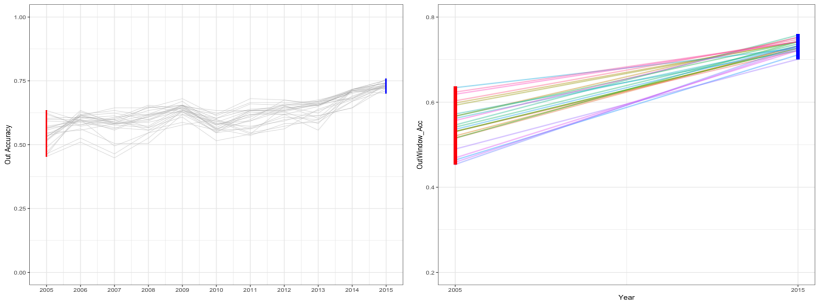
(7) Section 1/Disap.: CIRI/DISAP

Figure 8: Accuracy and Bias Across Algorithms over Time-Backward

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

Model Training 2015 and Testing on 2005

In order to directly compare the results from the experiments, we also run additional ML models.



(1) Comparison Over Time (2) Parallel Comparison 2015 to 2005

Figure 9: Accuracy Across Algorithms over Time-Backward (2015)

Note: Shown are the measured in- and out-window accuracy and bias for 29 machine learning models. The gray colored lines represents 29 models and the solid black line is an average value for across all the models. (1) illustrates the models trained with the 2015 reports and testing on each years report from 2005 to 2014 for comparison over tome. (2) shows direct comparison between 2005 and 2015.

The results are basically consistent with the backward rolling forecasting. Models trained with the 2015 reports do not predict the PTS scores for the 2005 reports very well. As discussed in the main paper and the section above, this is due to the changes in the reports from 2005 to 2015. Figure 9(1) shows these gradual changes over time. and Figure 9(2) shows the direct comparison between 2015 and 2005. In the experiment sections, we show that there is no empirical evidence that these changes are derived from the coding stage. We suggest that these changes are more likely to be from the monitoring and reporting stage.

Experimental Protocol and Additional Tests

The Experimental Design

For the coding of the State Department reports on 2015, we ran a randomization experiment in which each coder received roughly 60 of the countries from 2015, and 60 from 2005. The reason for choosing 2005 as the treatment is to allow for a long enough period to have passed in order for the possibility of substantial differences in government policies towards its citizens to have taken place. Other possible randomization strategies could have been to choose multiple treatment years, but this possibility was discarded due to the number of observations required to ensure sufficient power in the experiment (exact power calculations will follow).

We edited the reports such that they would not have any clear indicator of the year in which they were published. Using the same assignment procedure used in previous years, each coder was assigned 60 countries to code which are (unknowingly) from 2015 and 60 reports from 2005. Countries for which a report is released are coded by two or three coders. Using a similar assignment mechanism, while ensuring that no coder receives the same country twice, each coder received an additional 60 countries to code, not knowing which countries' reports are contemporary and which are from the past. After coders submit their scores for each country, the same process of adjudication of scores occurs for both treatment and control observations, and final scores are reported. Though we coded all 194 reports released in 2015, in 2005 PTS did not yet code small Island nations with small populations (under 1 million), and these 14 observations were discarded, since they had no appropriate comparison.

In short, this experimental design allows us to compare two sets of coding for 2005. The first, is the *original* 2005 scores, which were coded in 2006 as part of the annual PTS coding procedures. The second set of scores are the same reports from 2005, *recoded* in 2016. The randomization ensures that the coders are unaware of which year they are actually coding, and therefore cannot adapt the standards according to a supposedly different standard to the 2005 reports and to the 2015 reports. This allows for an apt comparison of whether the coders have

changed the standards of accounting for PTS scores over time. This will allow for a valid test of coder bias within human rights measures.

Hypothesis and Estimators

Experiment Hypothesis: Current standards of coding PTS scores have changed between 2005 and 2015.

The estimator we use is the Average Treatment Effect (ATE). The ATE is formally defined as $E[Y_i(1) - Y_i(0)]$, and is simply calculated by taking the average of scores from the *recoded* 2005 scores (treated units) and subtracting the average scores from the *recoded* 2005 scores (control units). If the result is positive and statistically significant, it would indicate that there is a temporal bias and present coders give the same report a higher score (more physical integrity violations) than past reports on average and the temporal bias is in the direction of 2015 coders holding a *more stringent* standard than in the past. If the result is negative and statistically significant, it would indicate that there is also temporal bias, but in the opposite direction. Present coders give the same report a lower score (less physical integrity violations), and the temporal bias is in the direction of 2015 coders holding a *less stringent* standard than in the past. If the result is not statistically significantly different from zero, we would fail to reject the null that the standards of accounting have not changed over time. Standard errors for the point estimate will be estimated using both bootstrapping and with Welch's t-test, which does not assume constant variance across samples. [4].

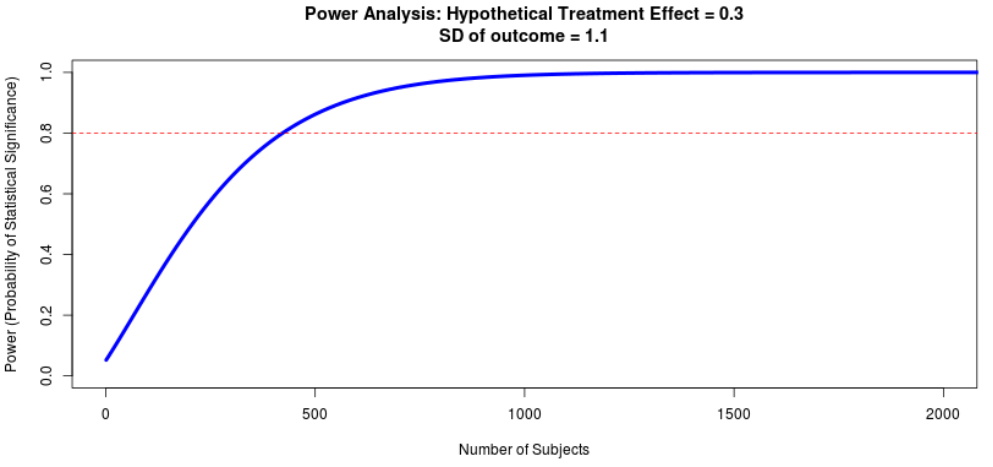
Power Calculation

In order to ensure that the experiment has sufficient power, we run a pre-test power analysis. The outcome variable, PTS scores given to all countries each year, has a mean of 2.49 since 1990. The average standard deviation from scores each year is 1.1. For a two-tailed test, with significance level at 0.05, we hypothesize an effect size of 0.3 and find that the minimum number of observations required for a power level of 0.8 is 423 (see Figure 10). Since we have eight

[4] $\hat{SE} = \sqrt{\frac{\hat{Var}(Y_i(0))}{N-m} + \frac{\hat{Var}(Y_i(1))}{m}}$

coders coding 60 countries under control and treatment each, we conclude that our sample size is sufficiently large. While in the experimental design we have sufficient observations to ensure sufficient power, after the accounting procedure is complete we remain with only 396 observations. As a result the actual power of the analysis will be .78, assuming the observed effect will be 0.3. It is possible that the anticipated effect size is relatively large, and that the estimated ATE from the experiment will be smaller.

Figure 10: Power Calculation



Results

As can be seen in the bar plots in Figure 6 and in Table **??** in the main manuscript the overall results of the experiment indicate no significant changes in the results. The majority of countries (67 %) were coded identically between the times. While 38 countries received a higher score in the original 2005 coding, 21 countries received a lower score in the original 2005. This would indicate that while coding standards may have changed slightly over time, they are not overwhelmingly in one direction or the other, but with almost twice as many observations applying a less stringent standard than a more stringent standard in 2015 compared to 2005.

TABLE 2: Scores from Original 2005 and Recoded 2005 (in 2015)

PTS Score	# of Obs. in Original 2005	# of Obs. in Recoded 2005	Difference in Scores (Recoded - Original)	# of Obs. with Difference
1	38	42	-2	2
2	49	55	-1	36
3	49	45	0	118
4	35	28	1	21
5	6	7		

In order to test the hypothesis of whether standards of coding have changed systematically over time we use several forms of difference in means tests to estimate the ATE. The point estimate for the simple difference in means:

$$ATE = E[Y_i(1) - Y_i(0)] = E[Y(Recoded) - Y(Original)] =$$

$$E[Y(Recoded)] - E[Y(Original)] = 2.452 - 2.56 = -0.11$$

The negative result of the ATE indicates that the direction of the bias is in the direction of holding more stringent standards over time. But in order to test the hypothesis, we estimate the standard error using both a bootstrap and the conventional standard errors with different variances. We find that the result is not statistically significant using either methods.

Another way to test for bias amongst coders is to test whether a single senior coder has a dominant effect that biases the entire sample. To estimate this we removed scores coded by senior coders, one at a time, and evaluated whether this may have any effect on the results, but the results remain similar. Full results of the standard errors and confidence intervals are displayed below in Table **??**.

Threats to Validity

The main threat to validity of this experiment is its inability to be fully double-blind. Since the coders are all experts in the field of political science, it is reasonable to assume that reading a

TABLE 3: ATE Estimates and Confidence Intervals

ATE (Difference in Means)	SE (Type)	SE (Result)	p-value	CI (Lower Bound)	CI (Upper Bound)
-0.11	Welch t-test	0.12	0.37	-0.34	0.13
-0.11	Bootstrap	0.12	0.37	-0.18	0.13

report from 2005 would have noticeable difference for some countries. For example, if a coder were assigned to code Syria, which is currently experiencing a civil war resulting in high levels of repression, any informed coder would expect the State Department to report egregious human rights violations. If the coder receives the report from Syria under treatment condition, they will observe the lack of information about the civil war, since it had not yet begun in 2005. While this example shows that in certain cases there is a noticeable difference in reports, this example is particularly exceptional. In most cases, coders reported that they were not be able to distinguish between reports from 2005 and 2015. Due to this possibility that coders observed this difference, they were informed that an experiment is being conducted and that some of the reports they receive may be from past years. They were instructed to code the reports with the same internal standards of coding, even if they noticed a report from past years. In order to overcome this potential bias, coders were instructed to treat all reports given to them as true, and to code using the same standards they apply regularly to any report they receive. Because the most noticeable cases in which reports were not from the current year, were expectedly the ones that also had the highest level of human rights violations additional robustness checks were run. We excluded all category 5 countries (those with the most egregious violations) and results remained similar.

1977	1997	2016
Section 1. Respect for the Integrity of the Person, Including Freedom from:	**Section 1. Respect for the Integrity of the Person, Including Freedom from:**	**Section 1. Respect for the Integrity of the Person, Including Freedom from:**
a. Torture	a. Political and Other Extrajudicial Killing	a. Arbitrary Deprivation of Life and Other Unlawful Politically Motivated Killings
b. Cruel, Inhuman or Degrading Treatment or Punishment	b. Disappearance	b. Disappearance
c. Arbitrary Arrest or Imprisonment	c. Torture and Other Cruel, Inhuman, or Degrading Treatment or Punishment	c. Torture and Other Cruel, Inhuman, or Degrading Treatment or Punishment - Prison and Detention Center Conditions - Physical Conditions - Administration - Independent Monitoring - Improvements
d. Denial of Fair Pubilc Trial	d. Arbitrary Arrest, Detention, or Exile	d. Arbitrary Arrest or Detention - Role of the Police and Security Apparatus - Arrest Procedures and Treatment of Detainees - Arbitrary Arrest - Detainee's Ability to Challenge Lawfulness of Detention before a Court - Pretrial Detention - Protracted Detention of Rejected Asylum Seekers or Stateless Persons - Amnesty
e. Invasion of Home	e. Denial of Fair Public Trial	e. Denial of Fair Public Trial - Trial Procedures - Political Prisoners and Detainees - Civil Judicial Procedures and Remedies - Property Restitution
	f. Arbitrary Interference With Privacy, Family, Home, or Correspondence	f. Arbitrary or Unlawful Interference with Privacy, Family, Home, or Correspondence
	g. Use of Excessive Force and Violations of Humanitarian Law in Internal Conflicts	g. Abuses in Internal Conflict - Killings - Abductions - Physical Abuse, Punishment, and Torture - Child Soldiers - Other Conflict-related Abuse:

TABLE 4: Changes of Contents of Section 1 Over Time in the State Department Reports

Note: The table shows how the content of Section 1 in SD reports have changed over time. In 1977, it had 5 subsections, but it increased dramatically over time, and in 2016, it had 28 subsections (sub-sub sections, sub-sub-sub sections). What has increased/changed is not only the word counts, but the composition and (hierarchical) structure of the reports.

REFERENCES

Raschka, Sebastian. 2018. "Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning." *CoRR* abs/1811.12808. arXiv: 1811.12808. http://arxiv.org/abs/1811.12808.