

Ad Fontes Media’s First Multi-Analyst Content Analysis Ratings Project

White Paper

August 2019

Author: Vanessa Otero

- I. Overview**
- II. Analysts**
- III. Training**
- IV. Structure**
- V. Data Analysis Results**
 - a. Standard Deviations and Outliers**
 - b. Inter-rater reliability**
 - c. Interesting things to study**
- VI. Overall Source Score Weighting**
- VII. Next Steps**

I. Overview

From March to June 2019, Ad Fontes Media designed and conducted a multi-analyst content analysis ratings project of articles, videos and TV shows from just over 100 news sources. The purpose of this project was multi-fold. Its first purpose was to provide ranking data to populate a new version of Ad Fontes Media’s famous Media Bias Chart in such a way that each article, video, or TV show was ranked by at least three analysts of different political affiliations (left, right, and center). The second purpose was to test Ad Fontes Media’s content analysis methodology itself for replicability. The third purpose was to refine the methodology of the project by analyzing the resulting data.

The reason we refer to this endeavor as a “project” rather than a “study” is because we are creating ranking data in the form of scores upon a brand new, invented scale for rating news. Typically, the term “study” implies that numerical data collected are measurements of things that are completely quantifiable. Though some aspects of the news are quantifiable, like the number of *total* words in an article, or the number of *certain* words in an article, those were largely not the things we were judging.

This document is called a “white paper” because that term is broad enough to encompass nearly any sort of detailed technical description of a product, service, or project. It is not published in any academic journals or publications, and it has not been peer-reviewed. Therefore, we are avoiding the terms “study” and “article” because we do not want to imply that this white paper is something that it is not.

All the scores we collected during this project rely on an assumption that the taxonomy of the Media Bias Chart is a valid way of ranking news sources on the dimensions of reliability and political bias. This taxonomy was created by Vanessa Otero, the founder of Ad Fontes Media, and details on how and why she constructed it may be found in [these blog posts](#). Aspects of the systems and methods described herein are patent pending.

Over the past three years, millions of observers of the Media Bias Chart have found the system of classification itself to be useful, regardless of whether they personally agree with where news sources are placed upon the Chart. Some observers have taken issue with aspects of the taxonomy itself; for example, some object that a left-right spectrum doesn't capture the full extent of political positions, or that it is impossible to define anything as truly "neutral" or "center." Such points may be (and likely will be) debated forever but based on the proven utility of this taxonomy for so many, we pressed forward with it.

Additionally, the number of categories and numerical values of the taxonomy are somewhat arbitrary; as a result, the numerical values collected for rankings upon it are also somewhat arbitrary.

The horizontal axis (political bias, left to right) is divided into seven categories, three of which represent the spectrum on the left, three of which represent the spectrum on the right, and one in the middle. Each category spans 12 units of rating, so the total numerical scale goes from -42 on the left to +42 on the right. As previously stated, these values are somewhat arbitrary, though there are some good reasons for them, including that they 1) allow for at least seven categories of bias, 2) allow for more nuanced distinction between degrees of bias within a category (allowing analysts to categorize something as *just a bit* more biased than something else), and 3) they correspond well to visual displays on a computer screen or printed out 18x 24 poster.

The vertical axis (overall reliability, top to bottom), is divided into eight categories, each spanning eight rating units, for a total numerical scale of 0 to 64. Again, these are somewhat arbitrary, but eight categories provided sufficient levels of classification of the types of news sources we are rating and sufficient distinction within the categories.

Observers may argue that numerical scales for bias and reliability should be more or less granular, but this project is based on this scale. Therefore, an underlying premise of the numbers collected is that this scale is useful for conveying understanding of its meaning to its observers.

Those of us involved in this project at Ad Fontes Media, which include me, three formal Advisors of Ad Fontes, our database architect, our data visualization developer, four study coordinators, and twenty analysts, consider this project a good start. It is a good start in the sense that it shows, fairly well, that Ad Fontes' overarching taxonomy is valid and that its methodology for rating news sources is replicable among a group of trained analysts having different political views. It is a good start in the sense that it is not perfect, but that it shows where improvements can be made with future iterations and what kind of data would be most fascinating to pursue.

It is also a good start given the limited resources with which it was conducted. This project was funded via an Indiegogo crowdfunding campaign which raised \$32,000. These funds went towards developing the technological infrastructure for collecting the data and the labor itself, which means that everyone who made this happen spent a lot of time and effort for not very much money. For everyone's efforts, we are immensely grateful, because as you will see, we have laid the groundwork for an extraordinary system of news content analysis.

Social scientists, data scientists, statisticians, and news organizations will likely find areas of this project that are ripe for improvement, and we welcome those criticisms and suggestions. As we increase our resources for funding this ongoing work, we will make those improvements. It is, after all, just a good start.

II. Analysts

For this project, we selected 20 analysts, each of whom applied by submitting the following:

- 1) A resume, CV, or written description of qualifications (this was required)
- 2) A self-reported classification of their political leanings. Each analyst submitted a spreadsheet about their political views overall and per listed political topic. The spreadsheet is shown below (this was required)
- 3) Basic demographic information, also on the attached spreadsheet (this was optional)

Education and Qualifications

We did not have set minimum educational or experience requirements, but the primary qualifications we looked for were:

- 1) High reading comprehension skills
- 2) High analytical skills
- 3) Extensive political knowledge

All of our selected analysts completed at least some college. Two completed some college, eight had bachelor's degrees, seven had at master's degrees, and three had doctoral degrees. Two had journalism backgrounds, seven had professional education backgrounds, and the rest had various professions across industries.

Political Leanings

We wanted to balance our team of raters politically, roughly along the Republican-Independent-Democrat proportions reflected in Gallup surveys found here: <https://news.gallup.com/poll/15370/party-affiliation.aspx>, which is approximately 30-40-30, respectively.

We had our analysts self-rate their political leanings at the beginning. We also calculated their political leanings based on their ratings at the end of the project, which we will discuss in the results section.

Political positions and ideologies are, of course, more complex than merely Republican/Independent/Democrat, and so were the articles and shows we rated. An important part of bias rating is evaluating the “leftness” or “rightness” on a number of different policy positions, and how analysts view those policy positions naturally affects how they view the leftness or rightness thereof. Therefore, we had them rate themselves on 20 different political positions on the Media Bias Chart horizontal axis, (from most extreme left to most extreme right)

Definitions of the categories:

The horizontal categories are defined by the policy positions of current US elected officials. For more on why, see [here](#). There are three important definitions we use for defining areas on the chart, which are as follows:

- a) The line between “Most Extreme Left/Right and Hyper-Partisan Left/Right” is defined by the policy positions of the most extreme *elected* officials. Any position in a news source that is more extreme than what they call for falls in the “Most Extreme” category.
- b) Most of the categories of “Hyper-partisan Left/Right and Skews Left/Right” are defined by the current [Democratic](#) and [Republican](#) party official platforms
- c) Not every political issue has an associated “neutral” or “centrist” position.

Instructions to the analysts were as follows:

- 1) In the attached Excel workbook, on the first sheet, please score yourself by putting an “x” in the row for each political issue in the column corresponding to one of the categories (Most Extreme Left to Most Extreme Right)
- 2) If you wish to elaborate on what your score for a particular political issue means to you, you may do so in the row immediately below your marked box. This is not required, but some people will want to explain what they mean or why they marked themselves there or may be undecided on which category to place themselves between two options. You may elaborate on one or more positions, all of them, or none at all. While we have internally established definitions for the policy positions corresponding to these categories, your input here may also be used to inform those definitions in the future.
- 3) Provide an overall rating for yourself in the last row. This does not have to be a mathematical average of your individual ratings.

These 20 political positions were selected by going to the contact websites of the two Senators from Colorado (the state in which Ad Fontes Media is based). One is a Democrat, Michael Bennet, and one is a Republican, Cory Gardner. Each had a drop-down list of topics about which you could contact them. Though the titles of the political topics varied slightly, each Senator had some version of these 20 topics on their site. Though social scientists and political scientists may prefer more detailed and/or quality-controlled ways to determine an individual’s political bias, this method was granular enough for the purpose of this project.

Based on the analysts’ self-ratings, we classified them into five categories: solid left, lean left, center, lean right, and solid right. Our chosen analysts comprised three solid left, three lean left, eight center, three lean right, and three solid right, corresponding to our desired 30-40-30 distribution.

III. Training

To train analysts on Ad Fontes' Methodology, each analyst read each of Ad Fontes Media's most recent methodology articles, which was four articles in total. Then, we held three videoconference training sessions of approximately an hour to an hour and a half each over a two-week period. In these sessions, we covered how to rate articles and shows for overall reliability and bias. We covered each of the many possible considerations but focused on Ad Fontes' main criteria of Veracity, Expression, Fairness, and Headlines/Graphics for the Reliability metric, and Political Position, Characterization, and Comparison for the Bias metric.

We went through several article examples during these initial training sessions. As we proceeded through the first few weeks of ratings, analysts communicated questions and discussion topics as desired through email and a group Slack channel. We held two additional videoconference training sessions over the next twelve weeks, during which we went through additional article examples. During the videoconferences, analysts provided feedback about their questions, observations, challenges, and requests for clarification. At the end of the ratings project, we held another videoconference to discuss suggestions for improvement for various aspects of the project, including training.

A unique aspect of the project is that each analyst was assigned articles from each source on the chart. Because one of the criteria for rating is comparison between news sources, the experience of reading each source was a training exercise itself; most news observers do not read over one hundred different news sources in a short period of time, so our analysts developed unique experience in that regard.

IV. Structure

a. Overall

This ratings project was conducted over approximately 12 weeks, with 20 analysts and four coordinators. In total, 1818 online articles and 98 cable news shows were fully rated by a minimum of three analysts, though many of those were rated by four analysts (a left, right, center, and "wildcard").

The coordinators' primary role was to select articles for rating and place the article URLs into Ad Fontes' ratings interface for distribution to the analysts. Coordinators also assisted with logistical issues related to training, scheduling, and accessing articles and TV shows.

Each article was accessed by its analysts in its native form, meaning that it was viewed directly from its URL in the same context that any news consumer would be able to read it. There were discussions about potential merits of having the article text rated "blind," meaning independently and stripped out of its digital environment, such that the source itself would be unknown to the reader.

However, three factors led us to decide not to rate articles in this manner: 1) the logistical difficulty of pulling just the text of an article out (which would require web parsing, new document creation, and document storage), 2) the varying decisions that would have to be made on what to include and what to omit (i.e., author name, graphics, captions?), and 3) the fact that news consumers never encounter content in this format, and that reliability and bias clues can be inferred from non-textual elements.

We concluded that it is preferable to rate articles in their full native context primarily because of the reliability and bias clues that can be inferred from non-textual elements. In fact, we instructed our analysts to consider all elements in their ratings, including banners, polls, ads, and sponsored content, as indicators of reliability and bias.

Given that this project was intended to provide enough data to populate the first version of the Interactive Media Bias Chart, the number of sources, articles, and shows were selected to provide a sample large enough to meaningfully represent a cross-section of the news landscape with which many Americans are familiar. Therefore, we focused on rating online articles from over 100 popular news and political information sites, as well as several shows from the three major cable news networks (MSNBC, CNN, and FOX).

b. Source Selection:

The 100 online news sources were selected to include many of the news organizations with the greatest reach of print and online viewership in the United States, but they are not a list of the top 100 by that metric. As we update the Interactive Media Bias Chart through our ongoing ratings, we will fill in additional sources in order of reach (according to Alexa website traffic rankings) and by top requests from our subscribers.

Other sources included in the ratings project included ones with large social media followings, and some smaller ones were selected simply to show representation in the areas of the chart in which they appear. Others were selected because they were among our top requested ones from previous versions of the chart.

Performing content analysis of the news in a systematic fashion is challenging because content formats differ greatly. There is an enormous difference in format between a written online article and a video news program. Though Ad Fontes has developed content analysis methodologies for both written content and TV/video content, conducting analysis across multiple formats presents many logistical challenges. However, we believed that any attempt at mapping the news landscape needed to include a reasonable sample of shows from the three major cable news networks because of their outsized influence and reach.

Our analysts rated a total of 98 cable news shows from CNN, MSNBC, and FOX. We also rated online written articles from both CNN.com and FoxNews.com. We did not rate additional content from MSNBC.com, though because nearly all content on that site is in the format of either clips from MSNBC's television content or written articles from

nbc.com. Therefore, there are two logos for CNN (one for online and one for TV), two for FOX (one for online and one for TV) and one for MSNBC (for TV only).

We rated online content that contained video format only if the video content was relatively short (i.e., five to eight minutes or less). These videos could essentially be rated similarly to articles because of their length and resulted in the inclusion of video content from The Weather Channel and Newsy.

Because of time constraints, we omitted longer-format video content from certain sources, such as The Young Turks. This has been rated on previous versions of the chart, and we plan to include them again in our ongoing ratings, but we did not rate enough samples to include them in this round.

Certain sources that were included in previous versions are no longer included for various reasons; Patriotics because it does not publish content frequently enough to collect meaningful samples (which is a good hint one should not rely on it as a news source); David Wolfe because it now almost exclusively focuses on (bad) health information rather than (bad) political information; and Weekly Standard and Forward Progressives because they no longer exist. Note that articles under the Conservative Tribune logo have Western Journal URLs because it was bought by Western Journal. The news landscape changes quickly.

Drudge Report is not included because its content almost exclusively comprises direct links to other sites. Though it has discernable bias, all of it is contained in the headlines and graphics, which do not appear when simply clicking on the URL that links to the original source. Therefore, this did not fit into our current process for assigning articles via URL. We will include it using different methods in future phases.

c. Article Selection

We set out to select the most articles from the biggest U.S. print publishers, the highest amount being 90 each for the Washington Post and the New York Times. We aimed to select 30 to 45 for an additional tier of large publishers, and 15 each for every news source we rated. In nearly all cases, we fully rated (i.e., with at least three ratings) fewer than the total articles selected for each source. This was because some ratings were not completed for logistical reasons, and some ratings were thrown out as outliers.

Our team of four coordinators conducted the task of selecting articles to be rated from the news sources. They pulled the URLs from the websites of each of the news sources they were assigned and input them into a coordinator interface of our custom-built ratings software. We selected articles during three different time periods over the course of the project. Each time, we selected approximately 600 articles, with each coordinator selecting 150. We did each batch of article selection on the same day so we would be retrieving news stories from the same news cycle from all the publications. We selected them on three different days, spread three to four weeks apart to get variety in

the topics covered. We selected the articles on a weekday each time, because most publications publish more new stories during the week than on the weekend.

We determined that the articles to select would be what we reasonably believed to be the “most prominent” ones for each source. Each coordinator was instructed to select articles listed by the publisher as “most read,” “most shared,” “most popular,” or some other indication that the story had significant reach, if that information was available from the publisher. Otherwise, coordinators selected the stories that were in the biggest print, highest up on the site, and/or most easily noticed and accessible to a site visitor.

We used this “prominence” feature as a proxy for actual reach of the story. In the future, we would like to include actual reach numbers for each article, but obtaining such data is beyond our current resources. Our methodology factors in prominence because many publishers feature highly opinionated or biased articles to drive engagement, even if they mostly publish content that is more fact-based and neutral. Public perceptions of bias of large publishers are often driven by the extensive reach of lower-reliability, highly biased content.

d. Article Assignment

Once articles were pulled into a batch, they were assigned to analysts by our ratings software program such that each article was rated by at least one left (either solid or lean left), right (either solid or lean right), one center, and one “wildcard” analyst, who was either left, right, or center. We aimed to collect four ratings for each article in the event that we had to throw out outliers or if analysts were unable to complete ratings for some reason.

Articles were assigned to analysts at random, and they were not able to see the articles before they were assigned. Each analyst completed between 120-130 articles over a three to four-week period.

e. TV Show assignment

We ran two cycles of cable news show ratings, rating most weekday, prime time, and weekend shows on each of FOX News, MSNBC, and CNN. Analysts were assigned 11-12 hour-long cable news shows to watch over a period of three weeks. Each news show was pulled from the same news cycle when possible. For example, weekday shows were all pulled on the same weekday, and weekend shows were pulled on the same weekend.

TV shows were assigned to analysts similarly to how we assigned articles; a left, center, and right analyst for each.

In total, we collected full ratings of 98 cable TV news shows; 25 for CNN, 41 for FOX, and 32 for MSNBC. For some of these shows, we rated two separate episodes, but for others, we only rated one. Due to some logistical issues, we did not collect as complete a sample of TV show ratings as we initially intended. Each show is an hour, and

therefore a much bigger time commitment for individual analysts than an online article. As a result, they contain much more content to rate than an article. Given the constraints of the project, logistical issues arose that prevented collecting ratings for certain programs. Therefore, not all the shows on each network are rated.

However, we believe we did collect enough ratings to provide a decent representation of the shows on each network, which we used to calculate an overall score for each cable news source.

V. Data Analysis Results

The easiest way to see the resulting ratings for each article and show is, of course, on the new Interactive Media Bias Chart. By clicking on a button for a particular source, you can see a scatter plot of each article rated for that source. The overall score of an article or show was the average of the three (or in some cases, four) individual scores. In future versions of the chart, we may break down the individual ratings further into the left, right, and center ratings.

To see an individual article or TV show and its score, you can search the table function just below the chart. Searching for a source name will pull up all the individual articles for that source along with their scores, so if you like, you can click on the URL to read the story and compare it to the score. For TV shows, only the names and networks of the programs are available.

In total, we fully rated 1818 articles and 98 TV shows.

a. Standard Deviations and Outliers

We had 7116 individual article ratings, and for each, we calculated the standard deviation for each article rating compared to the average rating (the average of three or four ratings for the article).

The standard deviation for bias scores was 7.97—just under 8 units. The width of one bias category (i.e., one column on the chart) is 12 units, which means that the standard deviation from the average score for an article was within one column width.

That means that to the extent analysts disagreed about bias, it was typically at most one column to the left or right of the others (e.g., someone would rate something “skews right” and another would rate it “hyper-partisan right.”) We found this to be heartening, because we believe it shows that analysts of different political views can agree on whether and how something is biased to quite a similar degree.

To determine what we would consider outliers for bias—outliers that were egregious enough to justify throwing out the score, we looked at the ratings with the highest standard deviations and checked them against their scores, moving from the most variant ones down toward the less variant ones.

It became clear that scores outside of three standard deviations (approximately more than 24 points, or two full bias categories) were incorrect due to one or more reasons. These tended to be outright rating mistakes (i.e., the physical entry of a wrong score) or a case in which the analyst was unfamiliar with the politics of the topic and just missed the correct bias classification. In these cases, the deviation was high because the other analysts (at least two of them) scored the article similarly to each other.

At less than three STD (i.e., scores that varied less than 24 units from the average), the variant ratings tended to have more valid reasons why a particular analyst would disagree with the other analysts' ratings, so we kept each of those.

For bias ratings, we had 100 individual ratings out of the 7116, or 1.41% that fell outside of three standard deviations and were subsequently thrown out.

Of the 7116 ratings, 5.4% fell between two and three STD (between 16 and 24 bias units away from the average), and 23.2% that fell between one and two STD (between 8 and 16 bias units away from the average). The remainder, approximately 70%, fell within one STD.

We found these numbers—70% within less than one bias category difference, and 93% within one and one-half bias category difference—to be promising indicators of the workability of our framework for evaluating bias

We analyzed quality (reliability) scores similarly. The scale for quality is smaller: it is 0-64, with each reliability category spanning 8 units. Our standard deviation for reliability was greater than that for bias; it was 8.6—slightly greater than one category. This means that as a percentage of the total scales, the standard deviation for reliability was quite a bit higher than for bias; 13.4% for reliability (8.6/64) and 9.4% (7.97/84) for bias.

Our own observations of analysts' experience and feedback of rating reliability is that it was generally more difficult to rate reliability than bias. We believe that much of this difficulty, and the resulting larger standard deviation, can be resolved by additional training and refinement of our methodology.

In particular, the most difficult distinctions to make were typically between what should be classified as “complex analysis,” “analysis,” and “opinion.” According to our methodology, the distinction between these categories lies in how many facts are used to support a conclusion, and how closely the facts are tied to conclusions. Differentiating these points requires more standardization and examples in our training.

For reliability ratings, we had 66 individual ratings out of the 7116, or 0.93% that fell outside of three standard deviations and were subsequently thrown out.

Of the 7116 ratings, 5.9% fell between two and three STD (between approximately 17 and 26 reliability units away from the average), and 27.5% that fell between one and two STD (between 8.6 and 17 bias units away from the average). The remainder, approximately 65%, fell within one STD.

Again, we found these numbers—65% within approximately one reliability category difference, and 93% within two reliability category difference—to be promising

indicators of the workability of our framework for evaluating reliability, though it shows room for improvement.

After throwing out the outliers, we had 6956 total ratings, and the standard deviations for both bias and reliability fell to approximately 6.97.

b. Inter-rater reliability and bias adjustment

One way we measured inter-rater reliability (i.e., the ability of different raters to rate the same thing in similar manners to each other) was to calculate the average of their individual standard deviations across all of their ratings. Of course, some analysts had lower averages (and therefore, greater inter-rater reliability) than others, but none had average standard deviations that were high enough to justify discarding all of a particular analyst's ratings. Rather, it was sufficient to throw out individually outlying ratings as described above.

We will use our inter-rater reliability measurements for future training and refinement of our methodology; ideally, the lower we can get the differences, the more reliable our ratings will become.

What we did to account for these differences (for example, when an individual analyst's average STD difference across all ratings was -1.5, or +0.8), is that we weighted those analyst's scores by that margin. Essentially, these averages would show that an analyst tended to rate things a little more left or right compared to the other analysts rating the same article. A -1.5 would mean that the analyst tended to rate things just a bit more to the left and +0.8 meant they tended to rate things just a bit more to the right. This was bias we could measure, separate from self-reporting bias. We adjusted for this bias in our overall scores.

For example, if the "-1.5" analyst gave an article a score of -12, we would adjust it to -10.5, and if the "+0.8" analyst gave the same article a score of -6, we would adjust it to -6.8. Therefore, the overall ratings for the same article, though still different, were closer together. We believe this accurately accounts for analysts' individual bias while respecting the actual scores given by the analysts. These adjustments were typically in the range of less than two points (units).

c. Average article ratings

The average article rating for each article shown on the chart and in our searchable set of articles is the per-analyst-adjusted, straight average rating of the three or four individual scores for the particular article. Because the fourth rating was always a "wildcard" rating, this means the articles with four ratings were either (Left, Center, Right, Left), or (Left, Center, Right, Right), or (Left, Center, Right, Center). As described above, because we threw out large outliers and adjusted for individual analyst bias, having two of one type of analyst for a particular article did not significantly affect a four-analyst article's average score compared to a three-analyst article's average score.

d. Interesting things to study

There are several interesting things we can look at, specifically with regard to the analysts, at not only in this data set, but in additional data sets we collect in the future. For example, it would be interesting to see how an individual analysts' ratings changed over time as they became more exposed to articles from across the spectrum. Did they get better at judging reliability? We didn't look at it, but we hypothesize that they would.

Did they rate bias more or less harshly over time as compared to their first ratings? Again, we didn't look at it, but we hypothesize that an analysts' ratings of bias for the first 50 articles might differ as compared to those rated after seeing that many.

Another interesting thing (that we did look at but will write up in a separate post) is how an analyst's self-reported bias affected their reliability ratings. Did analysts tend to rate things they agreed with politically as having higher reliability than things they disagreed with? We found that they did, but not necessarily in the way one would expect. We will write more about this later.

VI. Overall Source Weighting

Close observers of the Interactive Media Bias Chart will notice that, particularly for low-scoring sources, the overall source scores appear to be lower than what would be expected from a straight average. This is because in our overall source-ranking methodology, we weight extremely-low-reliability and extremely-high-bias article scores very heavily.

The reason is this: the lowest rows of the chart indicate the presence of content that is very unreliable, including selective or incomplete stories, unfair persuasion, propaganda, misleading information, inaccurate, and even fabricated information (these are listed in order of egregiousness). Therefore, it is unacceptable for reputable news sources to include this type of content, even if it is infrequent or not the majority of the content. A source that has even 5% inaccurate or fabricated information is highly unreliable. A source that "only" publishes misleading or inaccurate content 33% of the time is terrible. In our system, they do not get credit for the 67% of stories that are merely opinion, but factually accurate.

A straight average, in such cases, would result in a higher overall source score—one that is inconsistent with the judgment of most savvy news consumers. Therefore, article scores of less than 24 for quality and more than +/-36 were weighted very heavily.

All other article scores for sources were straight-averaged. For example, if a news source had a mix of "fact reporting," "complex analysis," "analysis," and "opinion" articles, and a mix of left, right, and center bias scores, those would be straight averaged. As shown, our taxonomy rewards high percentages of fact reporting and complex analysis in sources and slightly downranks them for high percentages of opinion content (via straight averages). It does not punish a source for opinion content, because opinion content does have a useful place in our information ecosystem. However, our system does punish unfair opinion and worse content—that which we view as most polarizing "junk news."

Note that in our methodology, if a source publishes a story and then later finds that some major aspect was incorrect, and they timely issue a correction or retraction, we do not rate that story as a 0-8 (“contains inaccurate/fabricated info). No instances of this occurred in our sample, though.

Those types of low scores are given to sources that publish stories that can reasonably be ascertained by a knowledgeable reader to be false, and the source does not issue a correction or retraction. Typically, sources that are ranked very low on the chart do not have a practice of issuing retractions; that is a distinguishing factor between reputable and disreputable publishers.

We did not weight outlets, articles, or TV shows by reach, although we plan on doing so in future iterations. Forthcoming updates within the next few weeks of the release of the 5.0 version of the chart will have visualizations of reach data.

VII. Next Steps

We learned a great deal of information from this “good start” project. Moving forward, we plan on rating more articles and shows in an ongoing fashion, every week and month, rather than having a finite ratings period. We have built the software and database infrastructure to do so, and we are excited to continue this work

Rating more articles and shows does require financial resources to train and pay analysts. Although many people have a passion for this, and some could do it for a bit of time for free, it is time consuming, and people deserve to be paid for the hard and diligent work they do to rate sources. Therefore, we will continue to roll out products and services based on our Media Bias Chart data to fund our ongoing efforts. We hope you will find them valuable and purchase them if they meet your needs as a news consumer/citizen, an educator, or an organization.

We will refine our training and methodology with what we have learned so far and will continue to iterate and improve. We will implement technology tools, including AI, in order to increase our rating capability.

Our priorities for content we plan to rate are (generally in this order) 1) all major network and cable TV news shows, 2) more national online and print publications, and 3) local news publications. We do plan on expanding geographically (to other countries) and by content type (including radio, podcasts, and online video channels such as Facebook and YouTube channels).

Essentially, anything that presents itself as a source of news or news-like information—including sources that solely exist on social media—we want to rate it. It’s a big task, but that is what is required to help people navigate the news landscape, clean up our information ecosystem, and transform our democracy.

More to come in the future.