COGITATIO

Media and Communication (ISSN: 2183–2439)
2020, Volume 8, Issue 3, Pages 39–49
DOI: 10.17645/mac.v8i3.3022

Article

Automated Journalism as a Source of and a Diagnostic Device for Bias in Reporting

Leo Leppänen [1,*], Hanna Tuulonen [2] and Stefanie Sirén-Heikel [3]

[1] Department of Computer Science, University of Helsinki, 00014 Helsinki, Finland; E-Mail: leo.leppanen@helsinki.fi
[2] Swedish School of Social Science, University of Helsinki, 00014 Helsinki, Finland; E-Mail: hanna.tuulonen@helsinki.fi
[3] Media and Communication Studies, University of Helsinki, 00014 Helsinki, Finland; E-Mail: stefanie.siren-heikel@helsinki.fi

* Corresponding author

Submitted: 15 March 2020 | Accepted: 10 June 2020 | Published: 10 July 2020

Abstract
In this article we consider automated journalism from the perspective of bias in news text. We describe how systems for automated journalism could be biased in terms of both the information content and the lexical choices in the text, and what mechanisms allow human biases to affect automated journalism even if the data the system operates on is considered neutral. Hence, we sketch out three distinct scenarios differentiated by the technical transparency of the systems and the level of cooperation of the system operator, affecting the choice of methods for investigating bias. We identify methods for diagnostics in each of the scenarios and note that one of the scenarios is largely identical to investigating bias in non-automatically produced texts. As a solution to this last scenario, we suggest the construction of a simple news generation system, which could enable a type of analysis-by-proxy. Instead of analyzing the system, to which the access is limited, one would generate an approximation of the system which can be accessed and analyzed freely. If successful, this method could also be applied to analysis of human-written texts. This would make automated journalism not only a target of bias diagnostics, but also a diagnostic device for identifying bias in human-written news.

Keywords
algorithmic journalism; automated journalism; bias; diagnosis; journalism; news automation

Issue
This article is part of the issue "Algorithms and Journalism: Exploring (Re)Configurations" edited by Rodrigo Zamith (University of Massachusetts–Amherst, USA) and Mario Haim (University of Leipzig, Germany).

1. Introduction

In the current news media landscape, examining and acknowledging underlying bias is an important step in strengthening newswork and rectifying trust in journalism. As media is becoming reliant on metrics and personalization, striving for balance in issues such as gender, race, age, socioeconomic status, and story topics become increasingly poignant. Particularly when considering the expectations of the public of news as a representation of 'reality' (Reese & Shoemaker, 2016, p. 393). While working towards this goal, it is somewhat common to view automated journalism as a savior: an 'unbiased,' 'fair' and 'objective' decision-making system in comparison to the seemingly biased decision-making of humans. From this point of view, increased automation in the newsroom sounds like a match made in heaven, as newsrooms strive to be bastions of objectivity (Mindich, 2000, p. 1). As such, it comes as no surprise that many newsrooms are either already employing automated journalism or are interested in doing so (Sirén-Heikel, Leppänen, Lindén, & Bäck, 2019).

While the literature on automated journalism has presented various partially conflicting definitions (cf. Graefe, 2016), a very useful one is provided by Dörr (2016) and Caswell and Dörr (2018), who approach au-

tomated journalism through the technology employed. In their view automated journalism is about the employment of Natural Language Generation methods for producing news text. Natural language generation is a "subfield of artificial intelligence and computational linguistics that is concerned with the construction of computer systems that can produce understandable text in English or other human languages from some underlying nonlinguistic representation of information" (Reiter & Dale, 1997, p. 57). As such, Caswell and Dörr's (2018) definition explicitly excludes, for example, systems that produce summaries of news content written by other humans.

In this article, we use the term automated journalism along the lines of Caswell and Dörr (2018). In our view, automated journalism is the act of automatically producing a complete or near-complete news text from some underlying data. We include the qualifier 'near-complete' as a conscious acknowledgement of the view that a human can—and perhaps should—be included in the journalistic process of publishing. In practice, this means that our definition includes systems that produce story 'blanks,' raw textual material which already contain the main beats of the story but need further human editing before they are ready for audiences.

Irrespective of the precise definition of automated journalism, we believe it to be important to inspect the technology critically. As pointedly demonstrated by the now (in)famous analysis of automated prediction of recidivism (Angwin, Larson, Mattu, & Kirchner, 2016), algorithmic biases can have substantial effects. If the algorithms are viewed with an assumption of fairness, they present a danger of entrenching and hiding pre-existing biases. In the context of journalism, a profession and product defined largely by ideals such as objectivity, neutrality and factuality, it is crucial that unwanted biases are not allowed to entrench themselves unnoticed in the language and the content of news.

Other authors have previously researched both how algorithms can be investigated for journalistic purposes (Diakopoulos, 2015), described how algorithms involved in newswork could be made transparent (Diakopoulos & Koliska, 2017) and provided descriptions of how automation can help reduce bias in reporting (Fischer-Hwang, Grosz, Hu, Karthik, & Yang, 2020). Similarly, some technical works have investigated methods for identifying bias in non-journalistic contexts (e.g., Caliskan, Bryson, & Narayanan, 2017; Knoche, Popović, Lemmerich, & Strohmaier, 2019). In this article, we synthesize how these methods and ideas apply to diagnosing automated journalism itself for bias.

Such diagnoses can serve multiple purposes. First, they would quite naturally be of interest to researchers, as they would increase our understanding of the news media. Second, they would be of interest to third-party interest groups as a method for highlighting potential biases against any one of multiple demographics. Third, they present an opportunity to the newsrooms themselves to highlight the results of the audits as benchmarks or as societal commentary. Statistics on the gender distribution in news stories is already used by some news organizations for benchmarking (Helsingin Sanomat, 2018).

In relation to bias in news journalism, bias has conventionally been studied from the perspective of an autonym to objectivity, having adverse effects on the journalistic ethos to report reality truthfully, and as a symptom of partisanship (Hackett, 1984). As journalism is conceptualized as the fourth estate in democratic societies, bias has largely been tied to politics and ideology, editorial policy and individual journalists. The complexity of journalistic bias has gained a new dimension with digitalization. The shift towards mobile and the changes in audience behavior has increased the role of the audience, affecting news values and journalistic work (Harcup & O'Neill, 2016; Kunert & Thurman, 2019). Personalization, in effect a form of bias, has become a strategy for media organizations and platforms for creating customer value. Catering for audience tastes based on implicit or explicit user information can also increase the value for automated news, for example based on location, as suggested by Plattner and Orel (2019). However, as Kunert and Thurman (2019) found in their longitudinal study, most news organizations remain committed to exposing their audience to a diversity in news stories, reaffirming the prevailing framing of quality journalism.

Distinguishing between 'acceptable' bias, such as exhibited in personalized sports news, and 'unacceptable' bias, e.g., favoring certain ethnicities, is a value ridden process. Both are examples of 'selectivity,' as suggested by Hofstetter and Buss (1978, p. 518), or more generally framing (see Entman, 1993; Scheufele, 1999). Only shared values decide that one is acceptable and the other is not. Encoding such values exhaustively into any automated procedure is extremely difficult. It is unlikely that automated methods will be able to make this distinction outside of the most blatant cases. As such, when we refer to 'detecting bias,' 'causing bias,' etc., we are in fact talking about biases of 'undetermined polarity,' meaning that additional human analysis is required to determine whether the potential biases detected are acceptable or not. Nonetheless, due to the effects of media on audience perceptions, consciousness of bias and embedded values in automated journalism is of paramount importance.

2. Bias in Automated Journalism

Despite increased media attention, the term 'algorithmic (un)fairness' is still unfamiliar for many (Woodruff, Fox, Rousso-Schindler, & Warshaw, 2018, pp. 5–7). This is understandable as the 'unbiasedness' and 'fairness' of algorithms is often expressed as a selling point of automation: The prospect of a perfectly fair and objective computer replacing the biased human as the maker of hiring decisions, arbitrator of loan applications, and judge of those accused of crimes is very enticing.

Automated journalism has mostly been employed in settings where the objectivity standard can be considered the highest, such as weather reports (Goldberg, Driedger, & Kittredge, 1994) and financial news coverage (Yu, 2014). While automation has since been applied to domains where news media often produce more subjective commentary, such as elections and sports (Diakopoulos, 2019, p. 107), to the best of our knowledge even in these domains the systems tend to be applied to what we consider the objective side of the topic, reporting results rather than analysis.

While this positioning of automated journalism in the larger journalistic field is clearly driven by technology to some degree (i.e., the technology being unsuitable for other, more subjective, story types; see e.g., Stray, 2019), it seems that the view that objectivity is the best aspect of automation is also an influencing actor. The views of the media seem to be exemplified by the words of an editor of a regional media company, who stated that automatically produced stories represent "facts…and figures, not someone's manipulated interpretation" (Sirén-Heikel et al., 2019, p. 56). To us, such beliefs indicate two crucial assumptions: that removing the individual—or the first level of hierarchy of influences (Reese & Shoemaker, 2016)—is sufficient to remove bias, and that using automation indeed removes the effect of the individual. We will return to these assumptions in the conclusions of this article.

As increasingly acknowledged both within and without computer science, the use of algorithms is not a panacea to removing bias from society, if such a thing is feasible at all. In fact, automated systems are increasingly recognized as reflecting existing societal biases (Selbst, boyd, Friedler, Venkatasubramanian, & Vertesi, 2019) and due to the 'objective' imagery associated with them they might further systematize these biases. At the same time, it is hard to define what, exactly, it would mean for an algorithm to be unbiased or 'fair' (Woodruff et al., 2018, p. 1), with some notions of algorithmic fairness even being fundamentally incompatible with each other (Friedler, Scheidegger, & Venkatasubramanian, 2016, p. 14). As an example of the complexities of the topic, consider whether a system that simply reflects some underlying societal bias—and would automatically stop doing so if the societal bias was removed—is by itself biased? Due to these difficulties in defining what, precisely, is fair and unbiased, we do not focus our efforts on identifying what is unbiased or proscribing how the world should be. Instead we will next consider a few examples of cases where a system for automated journalism is either clearly biased, or at least raises the question of whether the system or the society it is employed in is biased.

We base our analysis on the observation that, in very broad conceptual terms, natural language generation can be thought of as consisting of three major subprocesses: deciding what to say, deciding how to say it, and actually saying it (Gatt & Krahmer, 2018, p. 84; Reiter & Dale, 2000, p. 59). The distinction between the last two steps is that whereas the second step decides e.g., what words to use, and in which grammatical forms, the actual inflection is done at the third step. It seems clear to us that if a system for automated journalism results in biased output when starting from data considered objective, the bias must have been introduced in either the first or the second step.

At the same time, whether based on human-written rules or machine learning, a system for automated journalism can also produce biased output text if the system inputs are biased. For example, an ice hockey reporting system will only produce news about the male leagues if it is never provided the results for the female leagues. Bias resulting from biased input is, however, distinctly different from biases built into the automated systems, with the operative difference being which part of the process must be modified to address the bias. Any system will misfunction when presented with incorrect inputs, or as the saying goes: 'garbage in, garbage out.' While a system receiving incorrect information indubitably reflects badly on the journalists and editors responsible for the system, it does not necessarily indicate that the system itself is malfunctioning. For this reason, in order to understand the weaknesses of the system, we must first focus on whether it malfunctions in the case of correct, i.e., unbiased, inputs. As such, going forward with our analysis, we will assume that the system is receiving correct, unbiased inputs.

As noted above, biases introduced by the system must be related to either content selection or the language used in the text. We will now consider the kinds of biases that could be introduced in both steps separately.

2.1. Bias in News Content Selection

With bias in news content selection, we refer to any phenomenon where the inclusion and exclusion of pieces of information from a news article reflects a potential bias. A real-life example of such a bias is described by Hooghe, Jacobs, and Claes (2015), who observe that female members of parliament received less speaking time than their male colleagues in Belgian media. Similar phenomena have been observed, for example, in sports reporting, where the coverage of male sports significantly eclipse the coverage of female sports (Eastman & Billings, 2000) and in reporting about same-sex marriages, where male sources were more likely to be quoted than female (Schwartz, 2011).

Phrased in terms of automated journalism, we can imagine biased automated systems that e.g., prioritize reporting election results of male candidates before those of female candidates. However, it is important to note that simply quoting more male politicians than female politicians does not necessitate that the automated system has a gender bias. Instead, it might be simply reflecting underlying societal factors and biases: If there are 99 male politicians to one female politician, a system ran-

domly picking a candidate to quote would mostly quote males. A more nuanced analysis is needed in such cases.

These content selection biases can, however, be more subtle and less obvious. It might be, for example, that a news text categorically only includes the racial background of a suspect if the suspect is part of an ethnic minority. Or similarly, reporting of a car crash might only mention the gender of the driver if they are female. In both cases, such reporting could entrench prior reader biases, either affirming their biased beliefs (those who are part of an ethnic minority commit more crimes, women are worse drivers) or not presenting contradicting evidence (a suspect of unspecified ethnicity committed a crime, a driver of unspecified gender crashed).

These examples show that bias can result not from just exclusion of information (i.e., protected classes being ignored or underrepresented in reporting), but also from highlighting the membership in a protected class.

2.2. Bias in News Language

It is also possible for the language of the news to be biased even in cases where the information content itself is not necessarily so. For example, Eastman and Billings (2000, p. 208) observe a tonal difference in sports reports, where male athletes were discussed in an enthusiastic tone, while female athletes were discussed in a derogatory tone.

Such linguistic bias can manifest in the minor difference in the nuance of the words that are used in the news text. For example, there is significant tonal difference in whether a car accident is described using language where the actor of the event is the pedestrian ('a pedestrian ended up being hit by a car'), the car ('a car ran over a pedestrian') or the driver of the car ('a driver ran over a pedestrian'). Minor changes in the lexical choice presents the driver of the car as having a passive role in the event, almost making them an observer, even if the facts of the event place most of the blame on the driver. Seemingly minor choices such as these can be seen as biased against those of lower socioeconomical status, who are less likely to own a car and more likely to be pedestrians.

These kinds of linguistic biases are very rarely as obvious as the content selection biases defined above but are nevertheless relevant. Minor changes in lexical choice can have significant effect. The same increase in unemployment can be described as an 'increase' or as 'rocketing' with significantly different tone. Similarly, consider the difference between describing a 17-year-old perpetrator of a crime as either 'boy' or 'young man': While neither is significantly more accurate than the other, they carry significantly different tone and can have significant effect on how the reader perceives the perpetrator.

There is nothing inherent to automated journalism that would prevent such biased language from being produced by an automated system, just like there is nothing inherent to the automation that would prevent systems from having biases in content selection. Next, we consider the mechanisms that would allow such biases to appear in the text produced by automated journalism.

3. The Mechanisms for Biased Automated Journalism

The previous sections highlighted ways in which the output of a system for automated journalism could be biased. It did not, however, address the mechanism by which such biases end up in the system. We now turn to this question.

Automated journalism, as in the automated production of news texts, can fundamentally be achieved by two technical methods (Diakopoulos, 2019). The first of these is via algorithms consisting of human-written rules that directly govern the actions of the system. The second is via algorithms that learn the rules from examples provided by the system creators, also known as machine learning. We will next discuss both approach in turn, with special focus on how biases might end up being encoded in such systems.

3.1. Bias in Rule-Based Systems

Rule-based systems for automated journalism are based on explicit rules programmed by human programmers, such as 'start an article on election results by mentioning who is now the largest party, unless some party lost more than 25% of their seats, in which case discuss them first.' Such rules, however, can be implemented using various technical methods and are best defined by the common factor that they are not automatically learned from examples. As these systems are, fundamentally, driven by rules and heuristics produced directly by humans, the principal reason for these systems to produce biased content is by the human-produced rules being biased.

Commercial actors providing systems for content creation or distribution, particularly those involving automation or machine learning, tend to keep their systems' details largely hidden from the research community. Naturally, this also holds true for systems used for automated journalism. However, interviews with media industry representatives indicate that most of the systems employed in the real-world newsrooms are indeed rule-based, rather than based on complex machine learning (Sirén-Heikel et al., 2019). Based on the limited evidence available, such as the few open source systems (e.g., Yleisradio, 2018), these systems are often based on what can be described as 'story templates.' These templates are, in broad terms, the algorithmic equivalent of the combination of a Choose Your Own Adventure book and a Mad Libs word game. The software inspects the input data, and based on human-written rules, selects which spans of text to include in the story and in which order. These 'skeleton' text spans contain empty slots, where values from the input are then embedded to produce the textual output of the story. While significantly more complex rule-based methods exist, espe-

cially in academia (see, e.g., Gatt & Krahmer, 2018, for an overview), the decree to which they have entered use in the industry is not clear to us.

Irrespective of the technical details of the system, the important factor in these types of rule-based systems is that on a fundamental level they work based on explicit instructions that have been manually entered by humans. In simpler systems these rules can then be trivially investigated for potential bias: if some part of the system makes a decision based on a protected attribute, such as gender, it could be considered immediately suspect. This kind of surface-level inspection would reveal trivial cases of bias, such as where a human programmer has encoded in the system that election results pertaining to male candidates are more interesting than similar results pertaining to female candidates.

However, such clear-cut examples are, we hope, rare. We believe it is much more likely that the system incorporates some heuristic that reflect unconscious underlying biases, with unintended results. This becomes increasingly probable as the system complexity and the amount of automated data analysis conducted by the system increase. For example, a system producing news about the local housing market might use the average housing prices of an area as part of its decision making about which areas to discuss in the produced news text, assuming a higher price equates to higher newsworthiness. These housing prices, however, are likely to be well correlated with socioeconomical factors of the area population, resulting in coverage that is biased against populations of lower socioeconomical status as a result of not discussing aspects of the housing market relevant to them.

An even more nuanced example of the same phenomena could be observed if the decisions on what areas to report on were based on the absolute change in the housing prices; if the housing prices changed everywhere by the same percentage, the more well-off areas would see significantly higher absolute changes, which in the case of our hypothetical system would result in the same effect as above. As such, the investigation cannot be limited to only protected attributes, but rather all attributes that correlate with protected attributes must also be inspected.

3.2. Bias in Machine Learning Systems

The other major archetype of systems for automated journalism is presented by systems that employ machine learning. These systems differ from the rule-based systems by the fact that their decision-making is not based on human-written rules and heuristics, but rather on rules identified from training data. Most commonly, in supervised machine learning, this training data takes the form of pairs of 'given this input, the system should produce this output,' such as news texts previously written by human journalists paired with the data that underlies each text. While some works have been published on unsupervised text generation methods where the data is not aligned in this way (e.g., Freitag & Roy, 2018), to our knowledge such systems are still rare and suffer from severe limitations in terms of their applicability to automated journalism. A detailed description of unsupervised automated journalism is thus skipped.

In machine learning systems (see e.g., Flach, 2012, for an introduction to machine learning), the human programmers do not explicitly provide the actual rules of processing, but rather provide a framework and a set of assumptions. For example, in the case of a system for producing automatic textual reports of election results, a programmer might make the assumption that the journalistic process being replicated is, effectively, a 'translation' from the numerical results released by the election organizers to the natural language news report. As such, they might elect to implement a neural machine learning model similar in architecture to those used in machine translation, and train it by using examples of 'given this result data, the system should output this textual description.'

The machine learning process then identifies a specific model (analogous to the ruleset developed by-hand above) that minimizes the average difference between what the model outputs for an input in the training dataset and what the expected output was. In other words, the training attempts to identify a process that mimics the process that generated the training samples as closely as it is able. The degree to which this process succeeds is still limited by factors such as the amount of training data (it is hard to learn things of which there are no examples) and the model architecture (the learned model is restricted by the architecture selected by the human developer, and a badly selected architecture might be fundamentally unable to mimic the process that generated the training data).

Another issue is presented by overfitting, where the learned model might incorporate assumptions that hold for the training data but do not generalize to other cases. Even state-of-the-art machine learning systems for natural language generation suffer from this type of behavior in what is referred to as 'hallucination' in the technical literature. That is, they produce output not grounded in the input data, but based solely on strong correlations found in the training data. Such behavior has been identified in state-of-the-art systems in various domains, ranging from very constrained restaurant description tasks (Dušek, Novikova, & Rieser, 2020) to sports news generation (Puduppully, Dong, & Lapata, 2019).

When discussing bias, the model definition, its architecture, is significantly less important than the examples from which the system is trained. An important aspect of supervised machine learning is that the system truly does its best to mimic a process that could have generated the training data it observes. This means that even if the programmer allowed the system to consider some protected attributes, such as gender, the system only does so if the behavior in the training data seems to be influenced by

said attributes. This, however, also means that if there are any biases in the training data, these are also learnt. This applies whether the biases are intentional or not.

At the same time, however, simply removing a potentially biasing variable from the input is insufficient to ensure that the system does not act in a biased manner and many 'debiasing' techniques can simply hide the issue without solving it (e.g., Gonen & Goldberg, 2019; Kleinberg, Ludwig, Mullainathan, & Rambachan, 2018). As long as the underlying bias exists in the training data, even if the identified variable causing the bias is removed, the system will locate so called proxy variables to encode said bias into the model (Kilbertus et al., 2017). For example, if the training data for a system making loan decision was provided by humans that were discriminatory by providing smaller loans to non-white applicants, a sufficiently complex model might learn to observe whether the name of the applicant or their postal code is indicative of a high likelihood of being non-white as a proxy for the race of the applicant, even if the race was not explicitly provided as input to the system. In the context of automated journalism, a machine learning system would thus learn any biases present in the news stories that were used to train it. As such, the 'unbiased' algorithm would simply be faithfully replicating and entrenching any pre-existing biases in the news text used to train it.

4. Detecting Bias in and with Automated Journalism

As for detecting bias in systems for automated journalism, we see three primary scenarios where such an investigation could be undertaken: a scenario of a clear box system, a scenario of a cooperative operator with a black box system, and a scenario where only system outputs are available. We next discuss each in turn, considering how the system might be diagnosed for bias given the constraints of the scenario.

4.1. Full Transparency

Clear box investigations depend on the ability to inspect the internal workings of the automated journalism system. As such, they are only possible in cases where the operator of the system is cooperative, allowing access to the source code of the system. Furthermore, they are in practice limited to rule-based systems: even if a modern machine learning model was made available to experts, the systems tend to be so immensely complex that they are, in practice, black boxes.

Given access to a rule-based system, it should be in principle possible to investigate the logic and the rules employed by the system and determine whether any of them are blatantly biased. For example, as noted previously, any rules where the system directly considers a variable related to, for example, gender, is immediately suspect of introducing gender bias into the report and can be investigated further. Such an investigation, however, becomes increasingly difficult when one attempts to identify nuanced effects such as those described in the housing price report example shown before.

To identify more nuanced (potential) biases and to investigate systems that are too complex for manual inspection of the system's internal workings, a method based on system input variation might be more practical. Notably, this method still requires some level of cooperation from the system operator but does not require access to the system internals, and as such is also applicable to black box systems. In this process, samples of slightly varied system inputs are prepared, and ran through the system in sequence and the results inspected for differences.

4.2. Cooperative Operator with a Black Box System

An example of such a cooperative black box case would be a machine learning system producing reports of election results. In such a case, one can take the election results that act as the system's input and produce a variation of those results where potentially bias-inducing variables are modified. For example, the researcher could produce a copy of the system input where all the genders of the candidates have been changed but the input is otherwise left as-is. Producing output from both the unmodified and modified inputs would then allow for a comparison of the output texts, so that any differences can be inspected for potential bias. Continuing the example, observing changes between the two datasets in, for example, the order the results are discussed in would give rise to suspicion of potentially biased treatment of the different genders. In fact, any significant changes in lexical and content selection should be investigated in detail.

4.3. Output Only

In cases where the system operator is not cooperating the investigation must be conducted solely based on the available system outputs. From the point of view of the applicable methods, this case is indistinguishable from the case of a researcher conducting an analysis of human-written news, with the potential exception of significantly higher amount of texts available for analysis. We hypothesize, that in this case the role of automated journalism can be reversed, so that automated journalism can help highlight bias in news texts, whether produced by humans or computers.

A relatively simple method for natural language generation is provided by language models. In general terms, a language model is a machine learning model that describes how likely a sentence is based on training data the model was trained on. Consequently, many language models can be used to generate language by querying the model for 'what is the most likely next word, if the preceding words are…' Due to their simplicity, they are currently not very useful for generating real news, even if they do have other applications in the field of natural

language processing. At the same time, if trained on a large collection of news articles, they in effect learn what an 'archetypical' news article looks like and can mimic that style.

Previous technical works, such as those by Sheng, Chang, Natarajan, and Peng (2019), have demonstrated how language models can be interrogated for bias. In their experiment, they construct pairs of sentence starts, such as 'the woman worked as/the man worked as,' and completed the sentences using a language model. Their analysis of the sentence completions revealed the language model had internalized deep societal biases and reflected them in its output.

While standard language models are not suitable for automated generation of real news, we hypothesize that a language model trained on a sufficient amount of training data produced by a news automation system would learn and retain all the biases of the original system, in effect functioning as a proxy. The language model could then be interrogated for bias, for example using the method of Sheng et al. (2019), and any evidence of bias in the language model would be indicative of a potential bias in the underlying system.

While the wide variety of methods for language modelling are too numerous to enumerate here, it is notable that the most recent advances in language modelling take advantage of word embeddings (e.g., Bengio, Ducharme, Vincent, & Jauvin, 2003; more recently, Devlin, Chang, Lee, & Toutanova, 2019; Peters et al., 2018). In word embeddings (e.g., Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), words (or sometimes subword-units) are represented as points in a multidimensional vector space. Due to the way the word embedding model is trained, these spaces have several intriguing properties, a principal one being that words that are used in similar contexts in the observed texts are located close to each other in the vector space. Therefore, the nearness of two words in this vector space approximates the semantic relatedness of the words. This same mechanism, however, means that word embeddings trained on a text corpus internalize biases from said corpus (e.g., Gonen & Goldberg, 2019). This has two important consequences.

First, when training a language model as suggested above, care must be taken to ensure that bias is not introduced into the language model via use of word embeddings pretrained on another corpus. Consider, for example, a situation where a language model trained on news texts shows potential bias. If the language model is based on word embeddings pretrained on a highly biased corpus, it would not be clear to what degree the observed biases were incorporated into the model from the news text and to what degree from the biased word embeddings. This problem can be avoided by either using a language model that is not based on word embeddings, or preferably by training both the language model and the word embeddings from scratch. While this procedure prohibits taking advantage of the state-of-the-art pretrained language models such as BERT (Devlin et al., 2019) and ELMo (Peters et al., 2018), it should ensure that any biases observed in the final model come from the texts being inspected.

Second, the tendency of word embeddings to internalize biases also present an opportunity. Previous works (e.g., Caliskan et al., 2017; Knoche et al., 2019) have trained word embeddings from various textual corpora in order to detect biases in said texts. For example, given a word embedding model trained on a newspaper corpus, it is possible to inspect whether keywords indicating either a positive or negative affect are, on average, close to the word 'white' than to the word 'black.' A situation where positive keywords are on average closer to the word 'white' than to 'black' indicates that the corpus contains potential racial biases.

Notably, neither of these last two methods (training and inspecting either language models or word embeddings) is in any way dependent on the data underlying the model being derived from a news generation system. Rather, they could be applied to all kinds of news texts, including those produced by human journalists. Similarly, these latter methods might be useful even in scenarios where the system operator is cooperating. As noted by Diakopoulos (2015), reverse engineering can "tease out consequences that might not be apparent even if you spoke directly to the designers of the algorithm" (p. 404). Indeed, it seems unlikely that a rule-based system for automated journalism would be biased on purpose, and more likely that any potential biases are subtle and introduced unintentionally.

5. Conclusions

In this work, we have briefly described what automated journalism is, including a description of the two archetypical technical methods to conduct news automation: rule-based and based on machine learning. We have identified two major categories of bias that can appear in the output of such systems: content bias and language bias. We then provided a description of the mechanisms that might result in biased output from systems for automated journalism, as well as mechanisms through which these biases could be identified. An important observation is that while the mechanisms require an underlying human source for the bias, the biases can emerge in the system without human intention and in very subtle manners.

Our investigation of bias in automated journalism highlights that automatically produced text needs to be inspected for bias just as human-written texts do. The applicable methods, however, depend on the level of cooperation from the system operator as well as the technical details of the system. In cooperative cases more rigorous inspections of automated systems are possible, yet in some cases the investigation is not meaningfully distinguishable from an investigation of human-written texts. As a result, we note that methods such as the one proposed above could also be applied to investigating the biases of human-written news.

We observed that the belief of unbiased automated journalism seems to predicated on two assumptions: that removing the individual—or the first level of hierarchy of influences (Reese & Shoemaker, 2016)—is sufficient to remove bias, and that using automation indeed removes the individual's effect.

Starting with the second assumption, our investigation above indicates that while automation can obscure the influence of the individual, which would naturally lead to assumptions such as above, automation does not remove the influence of the individual. In case of a rule-based system, the individuals who influence the output are those who build the system and decide what rules it should follow. In case of machine learning, the individual is further removed but still has immense effects on the system's actions through their role as a producer and selector of the training data. In either case, the individual remains, albeit obscured by the system itself.

As for the other assumption, that removing the individual removes bias, we point to the fact that this assumption ignores the possibility of influences imposed by the higher levels of Reese and Shoemaker's (2016) hierarchy. In other words, the belief that the removal of the individual removes bias is predicated on the assumption that bias is created by the individual. Such beliefs overlook societal and organizational biases and the nature of the organization and the society as a collective of individuals.

It warrants repeating that automated journalism fundamentally requires an individual or a collective of individuals to define (whether explicitly through programming rules or implicitly by producing and selecting the training data that tells the system what to do) a set of frames through which the data underlying the story is portrayed (e.g., Entman, 1993; Scheufele, 1999). Any claim of the resulting system being 'unbiased' implicitly insinuates that the frames employed are also unbiased, or alternatively overlooks their existence in the first place. Unless these frames are highlighted and scrutinized—both in academia and outside of it—they risk being entrenched and becoming axiomatic. It is for this reason that investigating automated journalism for bias is so important: By obscuring the individual, automation risks obscuring the framing, hiding both the underlying individual and structural biases. This also has consequences for researchers investigating automated journalism for bias: significant care must be taken to identify origins, originators and contexts of any identified biases. For example, the use of machine learning does not preclude a bias originating from a specific individual.

We believe future work needs to be undertaken on at least two fronts. First, computational methods for identifying bias should be extensively trialed in terms of applicability to the analysis of journalistic texts, with the aim of producing a clear description of usefulness and usability, especially to those without extensive technical knowledge. Optimally, the work should lead to easy-to-use tools for both technical and non-technical researchers. Second, the methods for user-cooperative scenarios need to be tested in detail on real-world systems for automated journalism to determine best practices for conducting such audits, and for determining the origins of the discovered biases.

Automated journalism raises a multitude of ethical questions without obvious answers. For example, attributing authorship of computer-generated texts is a difficult task (Henrickson, 2018; Montal & Reich, 2017), which in turn raises the question of credit, and responsibility, for the end product. It is our opinion that automated journalism cannot be allowed to become a smoke screen for eluding responsibility. In terms of practical recommendations, we point the reader towards the succinct but well thought out guidelines published by the Council for Mass Media in Finland (2019). In short, we concur with the view that automated journalism is a journalistic product, hence the control and responsibility must always reside with the newsroom, ultimately in the hands of the editor in chief. In order to ensure that editors can take this responsibility, developers of automated journalism are liable for creating systems that are transparent and understandable, with auditing providing one way of achieving this goal.

Acknowledgments

This article is supported by the European Union's Horizon 2020 research and innovation program under grant agreement No 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The second author's work was enabled by a personal grant from The Media Industry Research Foundation of Finland and C. V. Åkerlund Media Foundation.

Conflict of Interests

The corresponding author is employed in a joint research project with various European media companies.

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias. *ProPublica*. Retrieved from https://www.propublica.org

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3, 1137–1155.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186.

Caswell, D., & Dörr, K. (2018). Automated journalism 2.0: Event-driven narratives: From simple descriptions to real stories. *Journalism Practice*, *12*(4), 477–496.

Council for Mass Media in Finland. (2019). Statement on marking news automation and personaliza-

tion. *Council for Mass Media in Finland*. Retrieved from http://www.jsn.fi/en/lausumat/statement-on-marking-news-automation-and-personalization

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human Language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Stroudsburg, PA: Association for Computational Linguistics.

Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism*, *3*(3), 398–415.

Diakopoulos, N. (2019). *Automating the news: How algorithms are rewriting the media*. Cambridge, MA: Harvard University Press.

Diakopoulos, N., & Koliska, M. (2017). Algorithmic transparency in the news media. *Digital Journalism*, *5*(7), 809–828.

Dörr, K. N. (2016). Mapping the field of algorithmic journalism. *Digital Journalism*, *4*(6), 700–722.

Dušek, O., Novikova, J., & Rieser, V. (2020). Evaluating the state-of-the-art of end-to-end natural language generation: The E2E NLG challenge. *Computer Speech & Language*, *59*, 123–156.

Eastman, S. T., & Billings, A. C. (2000). Sportscasting and sports reporting: The power of gender bias. *Journal of Sport and Social Issues*, *24*(2), 192–213.

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, *43*(4), 51–58.

Fischer-Hwang, I., Grosz, D., Hu, X. E., Karthik, A., & Yang, V. (2020). *Disarming loaded words: Addressing gender bias in political reporting*. Paper presented at Computation + Journalism '20 Conference, Boston, MA.

Flach, P. (2012). *Machine learning: The art and science of algorithms that make sense of data*. Cambridge: Cambridge University Press.

Freitag, M., & Roy, S. (2018). Unsupervised natural language generation with denoising autoencoders. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 3922–3929). Stroudsburg, PA: Association for Computational Linguistics.

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im)possibility of fairness. *Cornell University*. Retrieved from https://arxiv.org/abs/1609.07236

Gatt, A., & Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, *61*, 65–170.

Goldberg, E., Driedger, N., & Kittredge, R. I. (1994). Using natural-language processing to produce weather forecasts. *IEEE Expert*, *9*(2), 45–53.

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 609–614). Stroudsburg, PA: Association for Computational Linguistics.

Graefe, A. (2016). *Guide to automated journalism*. New York, NY: Tow Center for Digital Journalism.

Hackett, R. A. (1984). Decline of a paradigm? Bias and objectivity in news media studies. *Critical Studies in Media Communication*, *1*(3), 229–259.

Harcup, T., & O'Neill, D. (2016). What is news? *Journalism Studies*, *18*, 1470–1488.

Helsingin Sanomat. (2018, March 3). *Helsingin Sanomat lisää naisten osuutta artikkeleissaan* [*Helsingin Sanomat* increases the share of women in its articles] [Press release]. Retrieved from https://sanoma.fi/tiedote/helsingin-sanomat-lisaa-naisten-osuutta-artikkeleissaan

Henrickson, L. (2018). Tool vs. agent: Attributing agency to natural language generation systems. *Digital Creativity*, 29(2/3), 182–190.

Hofstetter, C. R., & Buss, T. F. (1978). Bias in television news coverage of political events: A methodological analysis. *Journal of Broadcasting & Electronic Media*, *22*(4), 517–530.

Hooghe, M., Jacobs, L., & Claes, E. (2015). Enduring gender bias in reporting on political elite positions: Media coverage of female MPs in Belgian news broadcasts (2003–2011). *The International Journal of Press/Politics*, *20*(4), 395–414.

Kilbertus, N., Carulla, M. R., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 656–666). Red Hook, NY: Curran Associates.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018). Algorithmic fairness. In W. R. Johnson & K. Markel (Eds.), *AEA papers and proceedings* (Vol. 108, pp. 22–27). Nashville, TN: American Economic Association.

Knoche, M., Popović, R., Lemmerich, F., & Strohmaier, M. (2019). Identifying biases in politically biased wikis through word embeddings. In C. Atzenbeck, J. Rubart, D. E. Millard, & Y. Yesilada (Eds.), *Proceedings of the 30th ACM conference on hypertext and social media* (pp. 253–257). New York, NY: Association for Computing Machinery.

Kunert, J., & Thurman, N. (2019). The form of content personalisation at mainstream, transatlantic news outlets: 2010–2016. *Journalism Practice*, *13*(7), 759–780.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 26* (pp. 3111–3119). Red Hook, NY: Curran Associates.

Mindich, D. T. (2000). *Just the facts: How "objectivity" came to define American journalism*. New York, NY: New York University Press.

Montal, T., & Reich, Z. (2017). I, robot. You, journalist. Who is the author? Authorship, bylines and full disclosure in automated journalism. *Digital journalism*, *5*(7), 829–849.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 2227–2237). Stroudsburg, PA: Association for Computational Linguistics.

Plattner, T., & Orel, D. (2019). *Addressing micro-audiences at scale*. Paper presented at the Computation+Journalism Symposium 2019, Miami, FL.

Puduppully, R., Dong, L., & Lapata, M. (2019). Data-to-text generation with content selection and planning. In P. Stone, P. Van Hentenryck, & Z. Zhou (Eds.), *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, pp. 6908–6915). Palo Alto, CA: AAAI Press.

Reese, S. D., & Shoemaker, P. J. (2016). A media sociology for the networked public sphere: The hierarchy of influences model. *Mass Communication and Society*, *19*(4), 389–410.

Reiter, E., & Dale, R. (1997). Building applied natural language generation systems. *Natural Language Engineering*, *3*(1), 57–87.

Reiter, E., & Dale, R. (2000). *Building natural language generation systems*. Cambridge: Cambridge University Press.

Scheufele, D. A. (1999). Framing as a theory of media effects. *Journal of communication*, *49*(1), 103–122.

Schwartz, J. (2011). Whose voices are heard? Gender, sexual orientation, and newspaper sources. *Sex Roles*, *64*(3/4), 265–275.

Selbst, A. D., boyd, d., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In d. boyd, J. Morgenstern, A. Chouldechova, & F. Diaz (Eds.), *Proceedings of the conference on fairness, accountability, and transparency* (pp. 59–68). New York, NY: Association for Computing Machinery.

Sheng, E., Chang, K. W., Natarajan, P., & Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing* (pp. 3407–3412). Stroudsburg, PA: Association for Computational Linguistics.

Sirén-Heikel, S., Leppänen, L., Lindén, C. G., & Bäck, A. (2019). Unboxing news automation: Exploring imagined affordances of automation in news journalism. *Nordic Journal of Media Studies*, *1*(1), 47–66.

Stray, J. (2019). Making artificial intelligence work for investigative journalism. *Digital Journalism*, *7*(8), 1076–1097.

Woodruff, A., Fox, S. E., Rousso-Schindler, S., & Warshaw, J. (2018). A qualitative exploration of perceptions of algorithmic fairness. In R. Mandryk, M. Hancock, M. Perry, & A. Cox (Eds.), *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1–14). New York, NY: Association for Computing Machinery.

Yleisradio. (2018). Avoin-voitto source code. *Github*. Retrieved from https://github.com/Yleisradio/avoin-voitto

Yu, R. (2014, June 30). How robots will write earnings stories for the AP. *USA Today*. Retrieved from https://eu.usatoday.com/story/money/business/2014/06/30/ap-automated-stories/11799077

About the Authors

Leo Leppänen is a Computer Science PhD Student at the University of Helsinki. He has a MSc in Computer Science and a BA in Language Technology. His research focus is on automated generation of natural language, especially on the generation of factual content such as news and other reports from structured data. He is currently exploring news automation for less-resources European languages.

Hanna Tuulonen is a PhD Student at the Swedish School of Social Science, University of Helsinki. In her PhD dissertation, Tuulonen researches news automation in China, and how Chinese news automation and data-driven media content affect European media practices and content. In 2017, Tuulonen did her MA thesis in Finnish and Swedish news automation practices, and in 2018 she also participated in the Immersive Automation research project (http://immersiveautomation.com).

Stefanie Sirén-Heikel is a PhD Student in Media and Communication Studies at the University of Helsinki. She has a B.Soc.Sc. in journalism studies and a M.Soc.Sc. in media and global communication from the University of Helsinki. Her research is focused on how algorithmic decision-making and automation affects journalistic values, how journalism is defined, and performed. She has particularly interest in the sociotechnical aspects of newsroom innovation and management. She has a background in broadcast, print and digital journalism.