Automated News Recommendation in front of Adversarial Examples & the Technical Limits of Transparency in Algorithmic Accountability

Antonin Descampe[a,b], Clément Massart[a], Simon Poelman[a],
François-Xavier Standaert[a,*], Olivier Standaert[b]

[a] Institute of Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM), Université catholique de Louvain, Louvain-la-Neuve, Belgium.
*[b] Louvain School of Journalism, Université catholique de Louvain, Louvain-la-Neuve, Belgium; * E-mails : antonin.descampe@uclouvain.be, massart.clement@gmail.com, sim.poelman@gmail.com, fstandae@uclouvain.be, olivier.standaert@uclouvain.be*

Abstract. Algorithmic decision making is used in an increasing number of fields. Letting automated processes take decisions raises the question of their accountability. In the field of computational journalism, the algorithmic accountability framework proposed by Diakopoulos formalizes this challenge by considering algorithms as objects of human creation, with the goal of revealing the intent embedded into their implementation. A consequence of this definition is that ensuring accountability essentially boils down to a transparency question: given the appropriate reverse-engineering tools, it should be feasible to extract design criteria and to identify intentional biases. General limitations of this transparency ideal have been discussed by Ananny and Crawford (2018). We further focus on its technical limitations. We show that even if reverse-engineering concludes that the criteria embedded into an algorithm correspond to its publicized intent, it may be that adversarial behaviors make the algorithm deviate from its expected operation. We illustrate this issue with an automated news recommendation system, and show how the classification algorithms used in such systems can be fooled with hard-to-notice modifications of the articles to classify. We therefore suggest that robustness against adversarial behaviors should be taken into account in the definition of algorithmic accountability, in order to better capture the risks inherent to algorithmic decision making. We finally discuss the various challenges that this new technical limitation raises for journalism practice.

Keywords: computational journalism, algorithms, news recommendation systems, accountability, transparency, adversarial machine learning, fair machine learning.

1. Introduction: the need for algorithmic accountability

The availability of large amounts and various types of data combined with recent advances in machine learning are increasingly used as a basis for automated decision making in order to address complex problems in numerous contexts. While automation can reduce some of the subjective aspects inherent to human decisions, it also comes with a potential for opaque and discriminatory impact against certain groups. Fair and accountable machine learning has therefore emerged as an important research topic, at the intersection between computer science and social sciences (Lepri et al., 2018).

Fair machine learning generally aims at avoiding that automated decisions reproduce patterns of discrimination or widespread biases that may persist in society (Crawford and Schultz, 2014). As recently overviewed by Barocas, Hardt and Narayanan (2019), algorithmic fairness is difficult to formalize and therefore to ensure: there is no single definition that can capture all societal concerns raised by machine learning algorithms, and some relevant features for fairness are even impossible to reach concurrently.

In order to simplify the following exposition, we consider fairness as a general goal that applies both to the collected data and the algorithms manipulating this data, and we consider accountability as the more specific challenge (on which we focus) of ensuring that algorithms optimize their publicized criteria. In other words, we view algorithmic accountability as a necessary (but not sufficient) condition for fair machine learning.

In the field of computational journalism, the study of algorithmic accountability, on which we focus in this work, has been initiated by Diakopoulos (2015). “At its core, algorithmic accountability is about providing descriptions, explanations and even sometimes justifications for the behavior of decision making algorithms, particularly in cases where there was a fault or error” (Diakopoulos 2019, 206). Taking into account the important judging decisions that algorithms can make in the news production process, namely prioritizing, classifying, associating and filtering news content, and since “computational processes can produce the news content itself” (Coddington 2015, 336), algorithmic accountability is therefore an important challenge for the increasing number of news organizations that exploit machine learning in their production processes.

Besides, the call for a growing transparency (Karlsson, 2010) and, ultimately, for accountability, is not only an issue for computational journalism: it has to be considered in the wider context of mistrust against news media and the confidence crisis that many studies and trust barometers report (Nielsen, 2016). Nonetheless, computational journalism, as a particular form of thinking built around automation (Wing 2008, Coddington, 2015), embodied by specific interactions between “social actors” and “technological actants” (Lewis and Westlund, 2016), raises issues that the primary social-science rooted practices and epistemologies of journalists are only beginning to address (Dagiral and Parasie, 2017). For example, algorithmic accountability reporting (i.e., the journalistic investigation of algorithmic decisions, their potential biases and the way they shape audiences’ visions of society) was only recently introduced as a new discipline for journalists, and as an important contributor to public accountability in general, for news production or for other applications in various domains (Diakopoulos, 2019).

In this context, transparency (i.e., the understandability of the designers’ intent that guides algorithmic decisions) was pinpointed as a necessary ingredient to reach accountability, leading to technical and non-technical challenges. Non-technically, it is frequent that parts of a machine learning algorithm must remain secret (for privacy reasons, or to protect proprietary information). Technically, the black box audit of algorithms is a challenging task (Sandvig et al., 2014; Datta, Tschantz and Datta, 2015). Furthermore, the possibility for decision making algorithms to update their statistical models and to be constantly fed with new data increases the complexity of any strategy to ensure transparency, as it implies a need to understand the evolving models on-the-fly.

In this paper, we argue that despite ensuring transparency is already difficult, it is not yet a sufficient condition for algorithmic accountability. Precisely, we put forward that even if transparency could be perfectly ensured (i.e., if all the criteria embedded into an algorithm correspond to its publicized intent), it may be that adversarial behaviors make machine learning algorithms deviate from their expected operation. This is for example what happens with so-called *adversarial examples* (Goodfellow, McDaniel and Papernot, 2018). As its name indicates, an adversarial example is an input, unknown to the target machine learning algorithm, that makes it deviate from its public specifications and is purposely generated by an adversary having an interest in such a deviated behavior.

Concretely, we analyze the case of an automated news recommendation system including a classification of more than 5,000 newspaper articles collected from a quality outlet over several months into generic classes (namely: World News, National News, Economy, Culture, Sports, Media, Others). We then show that minor and (importantly) hard-to-notice modifications of the articles may allow an adversary to force their misclassification towards a class of her choice, and we quantify how the perturbation needed to reach this adversarial goal evolves in function of the classes' statistical similarity. One worrying feature of such adversarial examples is that they target the articles to be classified by a recommendation system and do not require any modification of the algorithms' code itself nor access to the articles that are used to train the statistical models of the system. These results indicate that even people who did not design a recommendation system can use adversarial examples in order to bias its outcomes. As discussed in Section 6, this may for example have impact on content personalization systems and amplify the usual concerns of filter bubbles, echo chambers and polarization associated with them.

For discussion purposes, we also mention that other types of attacks against the security of machine learning algorithms (that we do not evaluate experimentally) exist. For example, so-called *data poisoning* aims at infecting the modelling data in order to bias the decisions (Biggio, Nelson and Laskov, 2012; Steinhardt, Koh and Liang, 2017). These investigations indicate that robustness against adversarial behaviors should be taken into account in the definition of algorithmic accountability, in order to better capture the risks inherent to the deployment of automated decision making. In other words, we suggest that adversarial machine learning should be viewed as part of the technical toolbox for algorithmic accountability reporting. We then browse the various challenges that this additional requirement raises and the difficulty to deal with adversarial machine learning in a theoretically sound manner (and therefore, the practical difficulty to design robust and accountable algorithms). We also initiate a discussion of the consequences of these limitations for journalistic practice and propose directions for further research.

2. Related works and literature review

Overall, research on fair and accountable machine learning can be viewed as a technical counterpart to various pieces of non-technical research showing how critical are the changes driven by automation, big data and algorithms, both for news workers and audiences. For example, among the different approaches to a sociology of computational journalism, Anderson argues for a "technologically-oriented study of computational journalism", in order to discuss technology on its own terms (Anderson 2012, 1016). Lewis and Usher also point out the importance of a technology-focused approach to journalism and the need for "understanding how the ideas, practices, and ethos long held by communities of technologists could be applied to rethinking the tools, culture, and normative framework of journalism itself" (2013, 603). In this respect, Lewis, Guzman and Schmidt recently "approached the broader ontological questions of automated journalism" through the prism of Human-Machine Communication, which enabled them to "reposition technology within the social processes of journalism and to develop new research questions better attuned to such a technology" (2020). Our contribution precisely falls within this kind of approaches through the concrete issue of adversarial examples.

As far as our case study is concerned, our results are naturally connected with the line of research analysing content personalisation in the context of digital journalism. In their essay, Kunert and Thurman discuss how "implicit personalisation that infers preferences from data collected" is gaining traction with mobile devices having their way into people's daily routine (2019). Their paper reveals a tension between the possible threat to diversity that personalisation may cause and the positive feedback of consumers regarding algorithmic news selection based on past behaviors (Thurman et al., 2018). Kunert and Thurman also observe that implicit personalisation is almost always implemented without a possibility to opt out for users. The latter is in line with former observations by Kormelink and Costera Meijer that users are unwilling to maintain complex personalization preferences (2014). But the combination of implicit recommendation with the risks of adversarial examples that are not easily comprehended goes against the desire of users' control that their investigations put forward. More generally, our results directly connect with the ethical challenges raised by recommender systems, and in particular with the opacity issue discussed by Milano et al. (2020).

Our research is also deeply grounded in the broad literature analysing the impact of artificial intelligence in different fields, and in particular journalism. Starting with high-level observations, our results can be seen as an attempt to clarify what artificial intelligence can & cannot achieve, which is one of the questions discussed by Broussart's (2018). More connected with the journalism studies, a recent invited forum of Broussart et al. surveys various issues that intersect with our findings (2019). For example, our conclusions are well in line with and even amplify Broussart et al. (2019, Diakopoulos' section) when arguing that "the future of AI in journalism has a lot of people around". We show that even if designers can try to embed values into algorithms, adversaries can sometimes misuse technologies to enforce contradictory values. This extends the role of journalists as an interface between data-intensive technologies and the end users of news applications. The approach we put forward in this paper is also aligned with Broussart et al.'s call for crossing disciplinary boundaries (2019, Guzman's section), which we do by adapting a popular methodology from the field of information security to the context of digital journalism. Namely, we try to nail down a definition iteratively by reasoning based on counter-examples. Besides, our investigations can be connected with recent research on the perception of automated decision making, which highlights that users have mixed opinions about the fairness and usefulness of such systems and are primarily concerned about risk (Araujo et al., 2019). Adversarial examples precisely fall in the category of hard-to-contain technological risks that machine learning algorithms carry.

Eventually, the limitations of the transparency ideal and its application to algorithmic accountability is thoroughly discussed by Anany and Carwford (2018). They list 10 shortcomings of the transparency ideal, among which the discussion of technical caveats is closely related to the issue of machine learning security. Again, their primary focus is on "systems that are inscrutable even to their creators because of the scale and speed of their design" (Burrel, 2016; Crain, 2018). As a result, even engineers of these systems, despite having access to all implementation details, are sometimes unable to explain some problems they may lead to. We amplify this view by showing that adversarial behaviors can make the situation worse, and currently lack a theoretical treatment enabling strong robustness guarantees. In the case of adversarial examples, transparency can even be detrimental to accountability (i.e., simplify the adversarial task of finding minor modifications of an article leading to misclassification). Anany and Crawford also conclude by suggesting an alternative typology of algorithmic governance (recognizing and ameliorating the limits of transparency). We concur by suggesting machine learning algorithms as necessary "decision helpers" when data-intensive analyses have to be conducted rather than (for now hardly accountable) autonomous "decision makers".

We note that in order to delineate our investigations, our focus in this paper is on a quite restricted and design-centric definition of algorithmic accountability. In fact, the notion of accountability as transparency we study would best correspond to the notion of loyalty defined by the French Conseil d'Etat in its 2014 study on "The Digital and Fundamental Rights", which for example requires informing users about ranking and referencing criteria.[1] More user-centric definitions considering the different actors of an automated decision and their responsibilities have been developed, in particular in legal and political sciences (McGregor et al., 2019). We conclude in Section 8 that such user-centric definitions complement the design-centric approach and compensate its limitations.

3. Research question and methodology

The starting point of our investigations is driven by the standard approach of modern cryptography, which recognizes that "security definitions are essential for the proper design, study, evaluation and usage of cryptographic primitives" (Katz and Lindell, 2014, Chapter 1, Section 4). Admittedly, the concept of algorithmic accountability, as for example discussed by Diakopoulos, was not introduced as an operational definition. Yet, we posit that specifying an algorithm's goals in such precise terms can be useful both for designers (so that they know what to target) and for users (so that they know what can be expected from algorithms). Based on these premises, the main research question we investigate in this paper is whether a definition of algorithmic accountability which only requires algorithms and implementations to be transparent is sufficient.[2]

[1] https://www.vie-publique.fr/sites/default/files/rapport/pdf/144000541.pdf

[2] As discussed by Katz and Lindell, "a common mistake is to think that definitions are not needed or are trivial to come with, because everyone has an intuitive idea of what" (for example) "security

We answer this question negatively by following the standard methodology for reasoning about definitions, as for example discussed by Katz and Lindell in the cryptographic context (2014). Namely, we look for a counter-example of machine learning algorithm that satisfies the transparency requirement and is nevertheless not accountable from the user's point-of-view. For this purpose, we consider the exemplary case of a prototype news recommendation system and show that the aforementioned adversarial examples can be used to bias the algorithm results, concretely disproving its accountability.

We note that we make no claim about the optimality of the news recommendation system that we investigate. Our only claim is that the algorithms it uses are based on state-of-the-art Natural Language Processing (NLP) and machine learning tools. In this respect, we insist that, formally, invalidating a definition does not require showing that it fails in many cases, nor that it fails in practically-relevant cases. Quite the opposite, the ambition of a definition is to be general and independent of the case studies. So even a single counter-example is enough to show that the definition does not re fulfil its intended goals. Nevertheless, and in order to put forward that the risks inherent to adversarial machine learning are not only of theoretical nature, we illustrate them by first comparing a few algorithms for news classification and next targeting the one with the best prediction rate. We then use this example as a starting point for a more general discussion, leading to technical challenges (regarding the exploitation of adversarial machine learning against other tasks considered for automation in computational journalism) and non-technical challenges (regarding how journalistic practice can deal with the exhibited risks).

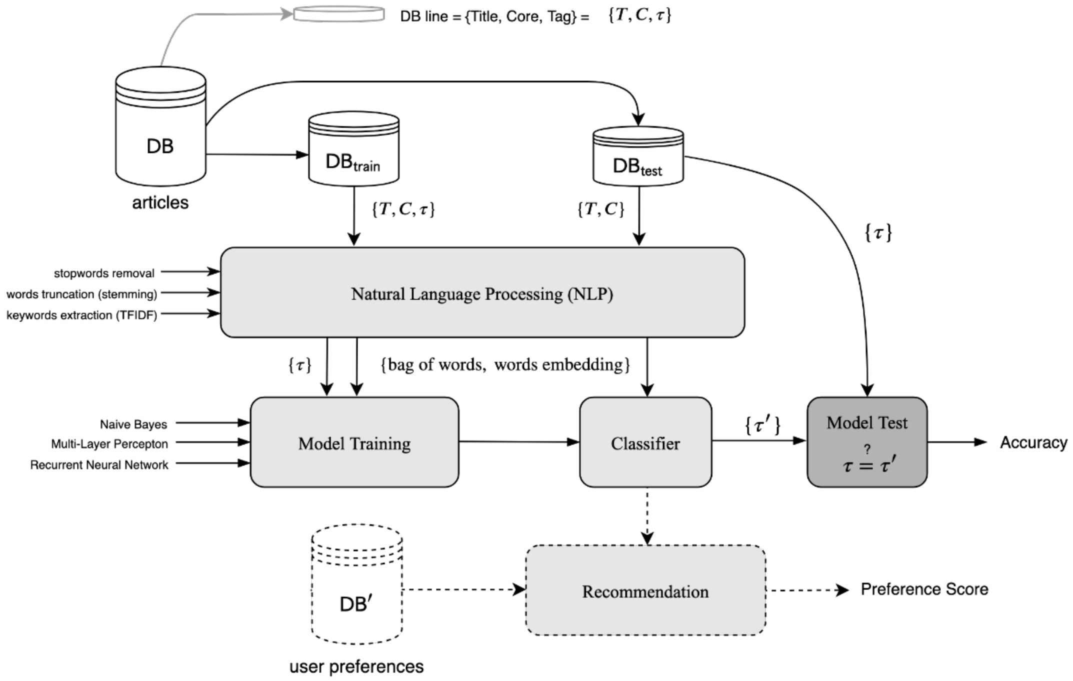
Based on these results, we conclude that robustness against adversarial machine learning should be considered as another necessary requirement for algorithmic accountability. As discussed in Section 7, this extended definition implies new trade-offs since ensuring robustness may be even more challenging in the case of transparent algorithms. We insist that we still do not claim that these two conditions are sufficient: whether other conditions are necessary to ensure algorithmic accountability is an interesting open problem.

4. Case study: a news recommendation system

We consider the case study of a fully automated recommendation system for news articles, illustrated in Figure 1. In such a system, a set of tags (or classes) is pre-defined by journalists and the articles are ingested and automatically classified into one of them. Next, based on the user preferences (including, for instance, the reading history) and the automatically assigned tags, a preference score is computed for each article. Such a score can then be used to prioritize the articles presented to the user in its newsfeed.

means". It turns out specifying what (for example) security means has been an iterative process where definitions were introduced, invalidated thanks to counter-examples and refined. As our following discussions will show, a similar situation holds with algorithmic accountability.

Concretely, we used a database (DB) of more than 5,000 articles collected from https://www.lesoir.be/, which is considered as a quality sheet. Each article (i.e., line of the DB) includes a title, the core of the article and the tag (which is a category manually assigned by the journalist who wrote the article). Tags can be World News, National News, Economy, Culture, Sports, Medias and Others. Articles were collected over a period of six months, between August 2018 and January 2019, with a web crawler. The number of articles collected by category varies between 500 and 800 (except for the Media category that only contains 104 articles). The resulting DB is split in two parts: the $\mathrm{DB}_{\mathrm{train}}$ part is used for *training* a machine learning classifier, the $\mathrm{DB}_{\mathrm{test}}$ part is used for *testing* the quality of the machine learning classifier. The goal of a classifier is to take a fresh article and to assign a tag to it. The manually-assigned tags are used both for the training and testing phases and are considered as the *ground truth* that the classifiers have to reach. The test is then simply done by challenging the classifiers with fresh articles and by comparing the tags output by the classifier and the correct ones. We used *k*-fold cross validation with *k*=5 for estimating the classifiers' accuracy (so we trained each model 5 times with 4/5th of the data in $\mathrm{DB}_{\mathrm{train}}$ and tested it 5 times with the remaining 5th).



DB line = {Title, Core, Tag} = $\{T, C, \tau\}$
DB
articles
$\mathrm{DB}_{\mathrm{train}}$
$\mathrm{DB}_{\mathrm{test}}$
$\{T, C, \tau\}$
$\{T, C\}$
$\{\tau\}$
stopwords removal
words truncation (stemming)
keywords extraction (TFIDF)
Natural Language Processing (NLP)
$\{\tau\}$
{bag of words, words embedding}
Naive Bayes
Multi-Layer Percepton
Recurrent Neural Network
Model Training
Classifier
$\{\tau'\}$
Model Test
$\tau \overset{?}{=} \tau'$
Accuracy
DB′
user preferences
Recommendation
Preference Score

Figure 1. Prototype recommendation system including an article classifier.

In order to avoid privacy issues that the manipulation of personal data inevitably raises, we focus in this paper on the quality and robustness of the article classification task. In this context, the goal of an adversary would be to take an article from $\mathrm{DB}_{\mathrm{test}}$ that is correctly classified and to modify it in the most unnoticeable manner so that the classification is changed towards a tag of his choice. In our case study of a fully automated recommendation system, it implies that the recommendations could be biased arbitrarily without modifications of the classifier nor changes of the code implementing it.

The recommendation systems that we consider start with some NLP. Precisely, we first removed the punctuation and used word stemming (i.e., cut off the suffix of words to make variations of the same stem equal to only one symbol). We then compared two different methods in order to either reduce the total number of words in each article or represent them in a more friendly format for automatic treatment, with the general rationale of comparing simple methods and more advanced ones. The simpler solution is to keep a dictionary of 20,000 most frequent words from our database and to represent each article as a "bag of words" that can be directly fed to machine learning algorithms.[3] The more advanced one is to exploit words embedding, and more specifically the Word2Vec model (Mikolov et al., 2013), which allows the representation of words in a continuous vector space that preserve certain semantic relations between words.

We then selected three exemplary machine learning algorithms: the Naïve Bayes (NB) classifier, the Multi-Layer Perceptrons (MLP) classifier and a Recurrent Neural Network (RNN) classifier (Rumelhart et al., 1986; Bishop, 2007). NB is a simple probabilistic classifier that assigns tags based on their estimated probability distribution. Practically, occurrences of words in articles for which tags are known allow building a probability of occurrence for each word in each class. Then, given an unknown article, it is possible to compute the probability that this article belongs to each class based on these pre-computed word probabilities and on the overall class probability (i.e., the number of articles of a given class related to the total amount of articles). MLP is a slightly more complex classifier, based on several layers of "perceptrons". Each perceptron is a simple mathematical function that outputs '0' or '1' depending on whether a weighted sum of its inputs reaches a given threshold. By combining many such functions, it is possible to obtain a classifier that takes articles as inputs and that outputs the most likely tag for each article. In a nutshell, it works in two steps. During the training phase, the input weights and thresholds of each perceptron are tweaked to ensure the best match between the MLP outputs and the tags of the training articles (which are known). In the testing phase, the MLP can be used to output tags for unknown articles. Such a classifier is a simple form of artificial neural network, a popular type of machine learning algorithm. Finally, RNN is a type of "deep" neural network which is also made of neuron-like nodes organized into successive layers. It enables predicting variable-size structures and is quite popular for speech recognition and text classification (Lai et al, 2015). This model analyses a text word by word and stores the semantics of all the previous text in a fixed-sized hidden layer, which is generally expected to effectively capture contextual information.

Concretely, we trained the NB and MLP classifiers with the fixed-length sequences that are provided by the bag of words NLP, while the RNN was combined with the variable-length vectors provided by the words embedding. We then estimated the accuracy of these three different combinations of tools. We reached 85% of correct classifications (without

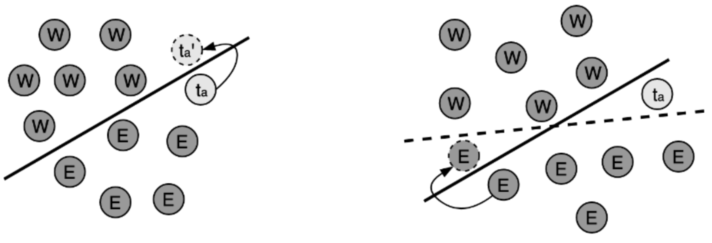
[3] Alternatively, we selected this dictionary based on the "Term Frequency - Inverse Document Frequency" (TFIDF) in order to detect salient words for articles with different tags (Ramos, 2003). Both options gave similar results. Our results are for the most frequent words method.

adversarial examples) for the NB and MLP classifiers, and only 80% with the RNN. These results can be explained by the fact that the goal of these classifiers is to identify the topic of the articles, which is a simple feature reasonably well captured by simple tools. By contrast, the RNN should gain interest to identify more subtle text features such as the meaning of sentences, and would require more data for this purpose. We note that in all cases, we reached an accuracy that is significantly higher than the 16% that corresponds to the best a priori strategy of guessing the most likely class given the distribution of our dataset. So we conclude from these preliminary investigations that the proposed tools satisfy an accountability definition limited to the transparency requirement: when evaluated in a controlled environment, with manually assigned tags considered as ground truth for training and testing, the NB, MLP and RNN (to a slightly lower extent) classifiers reach the expected behavior that they can classify articles nearly as well as journalists would do manually. Technical details about these algorithms and their parameters are not necessary for the understanding of our results: we defer them to Appendix A for the interested reader. The appendix also includes training curves confirming that the 5,000 articles of our database are enough for the NB and MLP classifiers to converge. We note that the use of French articles is not expected to affect our conclusions and all the tools we use next can be directly adapted to any language.

Following our methodology, we now investigate the possibility to manipulate the results obtained with the best (NB and MLP) classifiers thanks to small and hard-to-notice modifications of the articles. We first provide an intuitive description of what adversarial examples are, then give concrete examples and finally evaluate them quantitatively.

5. Adversarial examples against articles classification

5.1. Adversarial examples (intuition). Before describing experimental results of adversarial examples against a tag classifier and concrete illustrations with minimum perturbations, we provide an intuitive description of how adversarial examples proceed in order to lead to misclassifications. We use the left part of Figure 2 for this purpose.



(a) Adversarial example.

(b) Data poisoning.

Figure 2. Adversarial examples and data poisoning: illustration.

It represents an example of classifier built from 5 newspaper articles labeled “W” (for World news) and 5 articles labeled “E” (for Economy). These 10 articles correspond to DB_{train} (i.e., the part of the database used for model training) and are represented in dark grey. The classifier corresponds to the border between the World news and Economy articles. There is also a test article (t_a), represented in lighter grey, which is initially (and correctly) classified as an Economy article. The goal of an adversarial example is to modify the test article minimally into a similar article (denoted t_a' on the figure) so that t_a' is wrongly classified as a World news article (i.e., crosses the classifier's border). The difficulty to craft adversarial examples depends on the statistical distance between the initial and target classes, which is expected to reflect both thematic and semantic differences. For example, intuitively it should be easier to create confusion between the Culture & Media tags than between the Culture and Economy ones. In the following, we will therefore quantify how easy it is to craft adversarial examples by measuring how much one should modify the articles in function of the distance to the target class.

We insist that the goal of adversarial examples is not just to modify articles arbitrarily until they are classified as desired by the adversary (both for the machine learning algorithm and the human perception). The goal of adversarial examples is to modify the articles in a way that modifies their class for the machine learning algorithm while not modifying it for the human perception. A typical example is the case of traffic signals (e.g., speed limits) modified in a way that human drivers will still interpret correctly and that an automated car will interpret as a lower or higher speed (Eykholt et al., 2018).[4]

Note also that the right part of the figure illustrates the complementary issue of data poisoning mentioned in introduction. In this case, the adversary's goal is to modify (i.e., poison) some of the articles in $\mathrm{DB}_{\mathrm{train}}$ in order to move the border so that t_a becomes misclassified. This attack requires stronger capabilities than adversarial examples. Namely, it requires the control of the training data (which may be scrutinized offline) rather than the test data (which is usually fed on-the-fly to machine learning algorithms when deployed in automated applications). In the following, we therefore use adversarial examples as a counter-example to the definition of algorithmic accountability.

5.2. Concrete adversarial example with minimum perturbation. The following text contains parts of an article that was correctly classified as a National news tag by both the NB and MLP classifiers.[5] The bold modification in curly brackets is sufficient to have this article misclassified as an Economy one by the NB classifier:

> Preventive measures against wolf attacks. The Country's Minister for Nature and Agriculture […] has established special measures for breeders whose animals could be attacked by wolves. […] The prevention of damage caused by wolves has received a boost from the authorities who will subsidize a series of preventive measures. Past investments can be subsidized if they meet the conditions. *This* {**Normally, this**} new regulation must be ready for early April. The breeders who equip their land with a pen against the wolves will be able to recover 80% of their investments. […]

This example was found by manually looking for articles with limited perturbations, such that the impacting words were perceptually neutral. It typically corresponds to a sweet spot for adversarial examples, since (as detailed next) the Economy and World news classes actually correspond to the closest classes for this article and our two classifiers. Yet, it illustrates the concrete relevance of the threat and the possibility to bias a

[4] https://www.mcafee.com/blogs/other-blogs/mcafee-labs/model-hacking-adas-to-pave-safer-roads-for-autonomous-vehicles/

[5] Since our study is based on a French-speaking newspaper, the article is translated.

recommendation system in a quite hard-to-notice manner. We complete this example with a second bold modification in curly brackets, which corresponds to a perturbation leading to a misclassification towards the most distant (Sport) class for the MLP classifier:

> Preventive measures against wolf attacks. The Country's Minister for Nature and Agriculture […] has established special measures for breeders whose animals could be attacked by wolves. […] The prevention of damage caused by wolves has received a boost from the authorities […]. Past investments can be subsidized if they meet the conditions. This new regulation must be ready for early April. The breeders *who equip their land with a pen* {**who equip themselves with a pen around their field**} against the wolves will be able to recover 80% of their investments. […]

While the word "field" is less neutral (since quite usual in sport descriptions), it remains that the modification is quite minimal for such a strong change of classification.

5.3. Quantitative Experimental results. As a complement to the previous examples, we describe the systematic experiment we performed, in which we evaluated the feasibility of adversarial examples for the part of the 5,000 articles that we collected and for which the classifier is correct (i.e., 85% of 5,000 articles). For each such article, it proceeds in two main steps: first evaluating the distance to the other classes; then computing the minimum perturbation needed to reach these other classes in function of their distance.

For the first step, we start by computing a score for each word inside each class (i.e., the number of apparitions of the word among the articles of the class, divided by the total number of words in the class). For each article, we next compute a score per class (i.e., the sum over all the article's words of the words' scores in the class). This score is then used as our distance metric to determine close and distant classes. For the second step, and for each article and target class, we first look for the most impacting words by computing the number of times they must be added or removed to reach the target class. We then use this number of added/removed words as our perturbation metric, that we normalize in order to get an idea of the fraction of text that must be modified.

The result of this experiment for the NB classifier is given as a whisker plot in Figure 3 (which represents the quantiles of the minimum perturbations). A similar figure is given in Appendix B, Figure 5, for the MPL classifier, leading to similar outcomes.

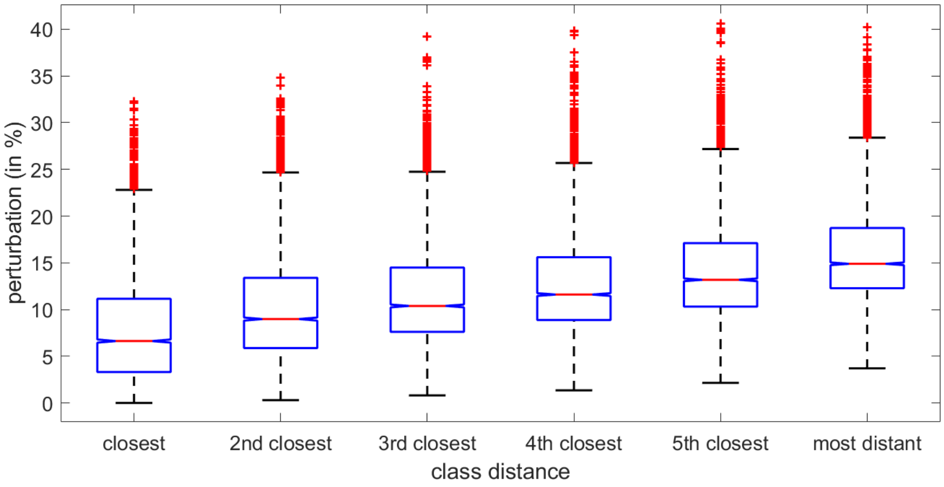


Figure 3. NB classifier: minimum article perturbation in function of the class distance.

The main conclusion of this experiment is that minimum perturbations ranging from ≈6% of the articles' content (for the closest target) to ≈15% of the articles' content (for the most distant target) are needed to misclassify 50% of the articles. However, for some articles, much smaller perturbations are sufficient. For example, the minimum perturbation can be below 1% for the closest target and 3% for the most distant one. Admittedly, this quantitative view does not reflect the fact that whether an adversarial example is critical or not is also a perceptual question. For example, it may be that considering more modifications of seemingly neutral words (i.e., words that look class-independent) makes the adversarial examples harder to detect than considering less modifications of salient words (e.g., strongly connotated adjectives or nouns). Yet, the examples of Section 5.2 confirm that hard-to-notice adversarial examples can be based on this approach, by using perceptually neutral words leading to misclassification.

We note that the closer and more distant target classes are specific to each article. In order to gain some intuition about whether some classes are systematically close, we additionally computed the average perturbation needed to misclassify an article of a given class (represented by the lines in Table 1) into an article of another class (represented by the columns in Table 1). It confirms some expected intuitions such as (1) the "Others" class is close to most classes as it may contain various types of information, and (2) the "Sports" class is more separated (suggesting a specific vocabulary). For the rest, differences are less significant, though the "World", 'Belgium" and "Economy" classes look more clustered and the same holds for the "Medias" and "Culture" classes.

	World	**Medias**	**Others**	**Economy**	**Culture**	**Belgium**	**Sports**
World	0	12,2	8,9	11,8	12,8	9,74	15
Medias	14,8	0	13,8	13,4	14,4	13,7	17,6
Others	6,1	7,8	0	6,9	8,4	4,9	12,2
Economy	11,1	13,3	10,9	0	13,8	8,4	16,2
Culture	12,8	10,6	9,6	12,8	0	11,4	13,5
Belgium	8,9	10,8	7,7	7,1	13,1	0	14
Sports	16,9	17,3	17,3	17,6	19,5	17,6	0

Table 1. NB classifier: average perturbation to misclassify one class into another.

6. Impact: towards robust algorithmic accountability

The adversarial examples put forward in the previous section show that a definition of algorithmic accountability that is solely based on transparency does not capture some of the risks that machine learning based automation carries. In this section, we discuss the impact of this observation in broader terms. For this purpose, we first illustrate how adversarial examples (and machine learning in general) could apply to other uses of algorithmic automation in the news production process, as recently surveyed by Thurman, Lewis and Kunert (2019). While we do not claim that the envisioned applications correspond to an immediate risk, they illustrate a gap in their treatment which, to the best of our knowledge, did not consider adversarial behaviors so far. We then propose an extended definition of algorithmic accountability which formalizes the need to consider such behaviors in the evaluation of automated computational journalism applications.

A first example is news content personalization, which involves selecting, highlighting, filtering and compiling news. As discussed by Bodò (2019), the goals and implementation of personalization may vary, but the extent to which it is generalized is a solid trend over the last decade. In a recent work, Helberger (2019) discussed the democratic role of news recommenders and the concerns raised by machine learning and artificial intelligence in this context. One salient feature of this discussion is that the values to optimize highly depend on the democracy models considered, and so does the level of impact of adversarial examples. For example, from a liberal perspective, the independence of the media from advertisers, political parties, … is paramount. In this respect, the increased control over algorithmic recommendation that adversarial examples enable is problematic. More critically, in a more participatory understanding of democracy, “the role of the media shifts from merely informing to actively educating active citizens” (Helberger, 2019). Hence, democratic recommenders must make a fair and representative selection of news. As a result, avoiding filter bubbles, echo chambers and polarization, all of which could be amplified with adversarial examples, becomes more important. See Perra and Rocha (2019) for an account on how opinion dynamics can be influenced on social media platforms, and how minimum nudging can be sufficient for this purpose. See Kunert and Thurman (2019) for a general discussion of content personalization.

Moreover, creating biases in the treatment of news does not need personalization to be effective. Automation in general primarily depends on the available data which, as mentioned in introduction, may not be diverse enough. First, it is expected that some types of news have a better potential for automation. For example, sports and economy news tend to produce a lot of data on a regular basis, which typically corresponds to “friendly” inputs for machine learning algorithms, possibly leading to a better coverage than less systematic news. Next, the generation of data could also be considered in an adversarial manner, in the sense of a data poisoning attack: for example, by automatically producing large amounts of contents on a topic that one wants to prioritize. To give a simple analogy, it is well known that it is possible to bias the citation count of a scholar with false documents (López-Cózar et al., 2014). A similar risk exists for news, for example with the automated generation of fake news and its propagation by bots.

Other contexts where adversarial examples are problematic include the automated moderation of online forums where the amount of comments prevents manual filtering (Arnt and Zilberstein, 2003) and automated fact checking (Hassan et al., 2017). In both cases, fooling a classifier so that it wrongly classifies offending comments as moderated ones, or fake news as genuine information, are well motivated adversarial goals.

So overall, the application of adversarial examples to a news recommendation system and the more general issue of machine learning security (e.g., including adversarial examples, data poisoning, …) suggest a necessary extension of the transparency-based definition of algorithmic accountability outlined in Section 3. It is not enough that algorithms are accountable in the sense that their execution in a controlled environment corresponds to the designers' publicized intent. The informed and ethical manipulation of any data (in the field of computational journalism and in general) requires robust accountability. That is, the algorithms' behavior has to be guaranteed even in adversarial contexts, where the data to classify, filter, prioritize, …, can also be under adversarial control. We insist that this extended definition is not in contradiction with the transparency requirement, which remains an important ingredient to enable a good understanding of the algorithm's goals.[6] It is rather a strengthening of the technical requirements that a design-centric definition of algorithmic accountability should cover. Adversarial examples in particular, and machine learning security in general, therefore appear as legitimate tools for algorithmic accountability reporting. They also relate to the goal of design orientation for automated news production recently put forward by Diakopoulos (2019b), and in particular with the need of metrics to "measure the alignment of technical implementations with organizational goals and values". Clearly, robustness against various types of adversarial data manipulations should be part of the criteria to be measured for this purpose.

7. Technical challenges and countermeasures

From a technical viewpoint, the results in this paper call for further investigations in different directions. For example, the automated generation of hard-to-notice adversarial examples is an interesting problem. It requires capturing the idea of neutral word that guided the manual generation of the examples in Section 5.2. Understanding such an automated generation and its potential limits would help better understanding the extent to which adversarial examples can be generalized and deployed at a large scale. Note that, as mentioned in Footnote 5, such an automated generation would benefit from transparency: knowing the details of a learning algorithm eases the task of adversaries willing to craft adversarial examples based on fine-tuned model parameters.

[6] Yet, as will be mentioned next, transparency may facilitate the generation and exploitation of adversarial examples, and therefore make the robustness requirement harder to reach.

Similarly, evaluating how the difficulty to craft adversarial examples with neutral words evolves with the size of the training database would be interesting. For example, the reason why the word "normally" caused an adversarial example in Section 5.2 is that its probability of occurrence was different for the classes of our automated recommendation system. As the size of the training set increases, it is expected that this probability of occurrence will become more similar for each class. Yet, since the range of features that machine learning can exploit is quite unlimited, especially for advanced (e.g., deep learning) tools that could ingest articles without NLP simplifications, it remains that many other (seemingly neutral) writing gimmicks could be captured and lead to adversarial examples. For example, the writing style of sport journalists may be based on shorter sentences than that of culture journalists, offering a path to guide misclassifications.

Again, we insist that our claim is not that adversarial examples directly apply in a critical manner to any machine learning application. Our claim is only that this threat has to be part of the goals for accountable algorithms. In other words, the risks we put forward are qualitative, and it is a designers' goal to assess them in a quantitative manner. These risks therefore question the possibility to mitigate adversarial examples with countermeasures. In this respect, the short summary is that various heuristics can be considered, but none of them provides a theoretically satisfying solution so far. One intuitive option is to use multiple models (e.g., NB, MPL and RNN in our case), hoping that adversarial examples are unique to each model so that a majority vote would prevent misclassifications. However, experiments show that adversarial examples tend to transfer across models, which is an expected consequence of the good generalization of these models (Tramèr et al., 2017). A more compelling approach seems to be adversarial training (Tramèr et al., 2018): its main principle is to inject adversarial examples in the training data to increase robustness against them. The study of such countermeasures in the digital journalism context is an interesting open problem. Yet, their understanding is in an early stage, and we are far from a situation where data scientists could offer strong robustness guarantees so that risks of adversarial examples can be ignored by machine learning users.

8. Further research and discussion

As stated by Coddington, "the concept of computational thinking, of abstracting data when approaching complex tasks or objects of news […], does not appear to have a precedent or analog in pre-computer-age journalism" (2015, 344). As a consequence, and despite the growing tendency to bridge newsroom culture and practices to computer science and engineering knowledge, some of the technical problems specifically belonging to computer science still remain unknown, or largely hidden. The case of the adversarial examples presented in this article can be considered as an illustration of the extent to which algorithmic processes raise issues that journalists, and scholars in social sciences in general, are simply not used to consider for themselves. This happens not only because the challenges surrounding the implementation of algorithms are still new and emergent. It also happens because people involved in journalism do not necessarily think of their work in computational terms. In general, they look at how technologies threaten or foster their established missions and practices, which means that it remains difficult to mobilize another background when discussing technology-based issues. In this respect, technical questions raised by adversarial examples and machine learning security, and the risk to easily misclassify data that they imply, can be seen as a form of problematization of the interaction between journalistic performance (grasped here through the notion of algorithmic accountability) and one particular topic in computer science.

This work therefore confirms how better problematizing the power and the functioning of algorithms, and, eventually, to explore how they challenge the fast-moving boundaries of (computational) journalism (Carlson and Lewis, 2015), can take advantage of a deeper integration of particular technical problems studied in other fields of research. As suggested by Coddington (2015), it may not be enough to confine this problematization to material and technical dimensions, as they are both framed by the epistemological values and orientations that each discipline develops. The methodology of challenging and refining definitions with counter-examples, which is used in this paper and standard in cryptography, is an example of positive outcome that such interactions can provide. Journalism, that is frequently presented as a field with blurred and flexible perimeter, sensitive to others' influences, has long demonstrated a capacity to integrate elements of diverse professional cultures (see Lewis and Usher, 2013, for what concerns automation). Here, the technical complexity and the political stakes inherent in the algorithmic process make this capacity for openness particularly desirable when considering the various epistemologies (Ward 2015, 2018) and the important normative roles that journalists assign to themselves (i.e., finding, checking and spreading news items).

More generally, adversarial examples, and the so far limited technical countermeasures that can mitigate them, naturally match Beck's risk society framework (1992), i.e., "a society where we increasingly live on a high technological frontier which absolutely no one completely understands", and where "manufactured risk is created by the very progression of human development, especially by the progression of science and technology" (Giddens, 1992, 3-4). In such a society, the goal of any organization is to manage risks, which is only possible in reference to some values. In the context of computational journalism, it implies that the level of risk that one may accept depends on the values that may be threatened by the biases potentially induced by adversaries, taking into account that contrary to physical risks that leave obvious damage, adversarial machine learning may be exploited surreptitiously, without leaving obvious traces.

Interestingly, this context offers different levels of responsibilities that may lead to different uses of automated technologies by different actors. For example, the direct interaction between a technology and its potentially uninformed end users generally implies hard-to-contain risks, which leads some to conclude that adopting a precautionary principle is the most effective way to cope with manufactured risks in this case. But the situation of a newsroom significantly differs in the sense that journalists may act as an interface between automated news production (and compilation, …) and their readers. By extending their traditional gatekeeper role to an "algorithmic gatekeeping" (Karppinen 2018, 504), able to assess the fairness of the collected data and the robustness of the algorithms manipulating it (i.e., by performing robust algorithmic accountability reporting), journalists can enable a more ambitious and controlled (yet, still not unbounded) use of new technologies, in function of well-specified values, while ultimately remaining the only accountable and responsible news authors.

Such an extended role for newsroom managers and journalists is quite in line with the "human still in the loop" perspective recently put forward by Milosavljević & Vobič (2019), which emphasizes a pressing need to embed innovations into established human-machine relations in the news production process. It also follows Bucher's analysis of how algorithms are discussed and integrated in newsrooms, concluding that "algorithms contest and transform how journalism is performed" and that they "do not eliminate the need for human judgment and know-how in news work; they displace, redistribute, and shape new ways of being a news worker" (Bucher 2018, 145). Importantly, this is especially true as long as robust algorithmic accountability cannot be ensured by technical means only: adversarial machine learning is a young research field and it is not (yet) clear whether strong general guarantees against adversarial examples (or other similar threats) can be obtained. In practice, the latter implies that the extended algorithmic accountability definition that we propose, despite being desirable as a design goal, may be difficult to reach in general and impose restrictions on the automation of the news. A more user-centric (responsibility-based) view of algorithmic accountability therefore appears as a necessary complement to the design-centric approach, in order to compensate its limitations (McGregor et al., 2019). In other words, one could use the design-centric approach to try making algorithms as accountable (i.e., transparent and robust) as possible while also being aware of their limitations, ultimately relying on a user-centric view of accountability defined within a legal framework able to assign responsibilities.

Acknowledgments. François-Xavier Standaert is a senior associate researcher of the Belgium Fund for Scientific Research. This work has been funded in parts by the European Union though the ERC Consolidator Grant SWORD (724725).

References.

Mike Ananny, Kate Crawford, 2018. *Seeing without knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability*, New Media & Society, Volume 20, Number 3, pp 973-989.

Christopher Anderson, 2012. *Towards a Sociology of Computational and Algorithmic Journalism*, New Media & Society, Volume 15, Number 7, pp 1005–1021.

Theo Araujo, Natali Helberger, Sanne Kruikemeier and Claes H. de Vreese, 2019. *In AI we trust? Perceptions about Automated Decision-Making by Artificial Intelligence*, AI & Society, volume 35, pp 611–623.

Andrew Arnt and Shlomo Zilberstein, 2003. *Learning to Perform Moderation in Online Forums*, Web Intelligence 2003, pp 637-641.

Solon Barocas, Moritz Hardt and Arvind Narayanan, 2019. *Fairness in Machine Learning*, https://fairmlbook.org/.

Ulrich Beck, 1992. *Risk Society: Towards a New Modernity*, London, Sage.

Battista Biggio, Blaine Nelson and Pavel Laskov, 2012. *Poisoning Attacks against Support Vector Machines*, Proceedings of ICML 2012, pp 1467-1474.

Christopher M. Bishop, 2007. *Pattern Recognition and Machine Learning, 5th Edition*, Information Science and Statistics, Springer.

Balázs Bodó, 2019. *Selling News to Audiences – A Qualitative Inquiry into the Emerging Logics of Algorithmic News Personalization in European Quality News Media*, Digital Journalism, Volume 7, Number 8, pp 1054-1075.

Meredith Broussard, 2018. *Artificial Unintelligence: How Computers Misunderstand the World*, MIT Press.

Meredith Broussard, Nicholas Diakopoulos, Andrea L. Guzman, Rediet Abebe, Michel Dupagne and Ching-Hua Chuan, 2019. *Artificial Intelligence and Journalism*, Journalism & Mass Communications Quaterly, Volume 96, Number 3, pp 673-695.

Taina Bucher, 2018. *If...Then: Algorithmic Power and Politics*, Oxford University Press.

Jenna Burrel, 2016. *How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms*, Big Data & Society, Volume 3, Number 1.

Matt Carlson and Seth C. Lewis, 2015. *Boundaries of Journalism: Professionalism, Practices, and Participation,* Routledge.

Mark Coddington, 2015. *Clarifying Journalism's Quantitative Turn*, Digital Journalism, Volume 3, Number 3, pp 331-348.

Matthew Crain, 2018. *The Limits of Transparency: Data Brokers and Commodification*, New Media & Society, Volume 20, Number 1, pp 88-104.

Kate Crawford and Jason Schultz, 2014. *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, Boston College Law Review, Volume 55, p. 93.

Éric Dagiral and Sylvain Parasie, 2016. *La « Science des Données » à la Conquête des Mondes Sociaux. Ce que le « Big Data » doit aux Épistémologies Locales*, In Big Data et Traçabilité Numérique. Les Sciences Sociales Face à la Quantification Massive des Individus, pp 85-104, Collège de France, Paris.

Amit Datta, Michael Carl Tschantz and Anupam Datta, 2015. *Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination*, PoPETs, Volume 2015, Number 1, pp 92-112.

Emilio Delgado López-Cózar, Nicolás Robinson-García and Daniel Torres-Salinas, 2014. *The Google Scholar Experiment: How to Index False Papers and Manipulate Bibliometric Indicators*, JASIST Volume 65, Number 3, pp 446-454.

Nicholas Diakopoulos, 2015. *Algorithmic Accountability: Journalistic Investigation of Computational Power Structures*, Digital Journalism, Volume 3, Number 3, pp 398–415.

Nicholas Diakopoulos, 2019. *Automating the News: How Algorithms Are Rewriting the Media*, Harvard University Press.

Nicholas Diakopoulos, 2019b. *Towards a Design Orientation on Algorithms and Automation in News Production*, Digital Journalism, Volume 3, Number 3, pp 1180-1184.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno and Dawn Song, 2018. *Robust Physical-World Attacks on Deep Learning Visual Classification*, Proceedings of CVPR 2018, pp 1625-1634.

Anthony Giddens, 1992. *Risk and responsibility*, The Modern Law Review, Volume 62, Number 1, pp 1-10.

Ian J. Goodfellow, Patrick D. McDaniel and Nicolas Papernot, 2018. *Making Machine Learning Robust against Adversarial Inputs*, Communications of the ACM, Volume 61, Number 7, pp 56-66.

Tim Groot Kormelink and Irene Costera Meijer, 2014. *Tailor-Made News: Meeting the Demands of News Users on Mobile and Social Media*, Journalism Studies, Volume 15, Number 5, pp 632-641.

Naeemul Hassan, Fatma Arslan, Chengkai Li and Mark Tremayne, 2017. *Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster*, Proceedings of KDD 2017, pp 1803-1812.

Michael Karlsson, 2010. *Rituals of Transparency: Evaluating Online News Outlets' Uses of Transparency Rituals in the United States, United Kingdom and Sweden*. Journalism Studies, volume 11, pp 535–545.

Kari E. Karpinnen, 2018. *Journalism, Plusralism and Diversity,* Journalism, De Gruyter.

Jonathan Katz and Yehuda Lindell, 2015. *Introduction to Modern Cryptography, Second Edition*. CRC Press 2014.

Jessica Kunert and Neil Thurman, 2019. *The Form of Content Personalisation at Mainstream, Transatlantic News Outlets*, Journalism Practice, Volume 13, Number 7, pp 759-780.

Siwei Lai, Liheng Xu, Kang Liue and Jun Zhao, 2015. *Recurrent Convolutional Neural Networks for Text Classification*, Proceedings of IAAA 2015, pp 2267–2273.

Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland and Patrick Vinck, 2018. *Fair, Transparent, and Accountable Algorithmic Decision-making Processes*, Philosophy & Technology, Volume 31, Issue 4, pp 611–627.

Seth C. Lewis and Nikki Usher, 2013. *Open Source and Journalism: Toward New Frameworks for Imagining News Innovation*, Media, Culture & Society, Volume 35, Number 5, pp 602–619.

Seth C. Lewis and Oscar Westlund, 2016. *Mapping the Human-Machine Divide in Journalism*, Sage Handbook of Digital Journalism, pp 341-353, Sage.

Seth C. Lewis, Andrea L. Guzman and Thomas R. Schmidt, 2019. *Automation, Journalism, and Human–Machine Communication: Rethinking Roles and Relationships of Humans and Machines in News*, Digital Journalism, pp 409-427, Sage.

Lorna McGregor, Daragh Murray and Vivian NG, 2019. *International Human Rights Law as a Framework For Algorithmic Accountability*, International & Comparative Law Quarterly, Volume 68, Number 2, pp 309-343.

Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean, 2013. *Efficient Estimation of Word Representations in Vector Space*, https://arxiv.org/abs/1301.3781.

Silvia Milano, Mariarosaria Taddeo and Luciano Floridi, 2020. *Recommender systems and their ethical challenges*, AI and Society, Volume 35, pp 957-967.

Marko Milosavljević and Igor Vobič, 2019. *Human Still in the Loop: Editors Reconsider the Ideals of Professional Journalism Through Automation*, Digital Journalism, Volume 7, Number 8, pp 1098-1116.

Brent Mittelstadt, 2016. *Automation, Algorithms, and Politics: Auditing for Transparency in Content Personalization Systems*, International Journal of Communication, volume 10, pp 12.

Rasmus Nielsen, 2016. *The many crises of Western Journalism. A Comparative Analysis of Economic Crises, Professional Crises, and Crises of Confidence*", The Crisis of Journalism Reconsidered, pp 77-97, Cambridge University Press.

Nicolas Perra and Luis E.C. Rocha, 2019. *Modelling Opinion Dynamics in the Age of Algorithmic Personalisation*, Nature Scientific reports, Volume 9, Number 1, pp 7261.

Juan Ramos, 2003. *Using TF-IDF to Determine Word Relevance in Document Queries*, Proceedings of the First Instructional Conference on Machine Learning, pp 133-142.

David E. Rumelhart, Geoffrey E. Hinton & Ronald J. Williams, 1986. *Learning Representations by Back-Propagating Errors*, Nature, Volume 323, Number 6088, pp 533–536.

Christian Sandvig, Kevin Hamilton, Karrie Karahalios and Cedric Langbort, 2014. *Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms*, Data and discrimination: Converting Critical Concerns into Productive Inquiry, 22.

Jacob Steinhardt, Pang Wei W. Koh and Percy S. Liang, 2017. *Certified Defenses for Data Poisoning Attacks*, Proceedings of NIPS 2017, pp 3517-3529.

Neil Thurman, Judith Moeller, Natali Helberger and Damian Trilling, 2018. *My Friends, Editors, Algorithms, and I: Examining audience attitudes to news selection*, Digital Journalism, Volume 7, Number 4, pp 447-469.

Neil Thurman, Seth C. Lewis and Jessica Kunert, 2019. *Algorithms, Automation, and News*, Digital Journalism, Volume 7, Number 8, pp 980-992.

Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh and Patrick McDaniel, 2017. *The Space of Transferable Adversarial Examples*, https://arxiv.org/abs/1704.03453.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh and Patrick D. McDaniel, 2018. *Ensemble Adversarial Training: Attacks and Defenses*, ICLR, 20p.

Stephen Ward, 2015. *Radical Media Ethics: A Global Perspective*, Wiley.

Stephen Ward, 2018. *Epistemologies of Journalism*, Journalism, Volume 19, pp 63-82.

Jeanette Wing, 2008. *Computational Thinking and Thinking about Computing*, *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences,* Volume 366, pp 3717–3725.

Appendix A. Details on the classifiers

The multinomial NB classifier was used directly on the histograms made of 20,000 words provided by the bag of words NLP. The only technical tweak was the use of a smoothing factor of 0.09 in order to deal with words having estimated probability zero.

For the MLP classifier, we selected the best parameters thanks to a grid search set to optimize the classifier's accuracy, with one to four layers and number of neurons per layer ranging from 10 to 1,000). Several solutions led to similar results and we eventually selected a classifier with three layers: the first one uses 141 neurons (corresponding to the square root of the 20,000 words output by the bag of words NLP) with a *relu* activation function, the last one last one uses 7 neurons (i.e., our number of classes) with a *tanh* activation layer, and the hidden layer uses 31 neurons (which corresponds to the geometric mean between 141 and 7, i.e., $\sqrt{141 \times 7}$) and a *relu* activation layer. Using more layers did not lead to significant concrete improvements in our case study.

Finally, we used different number and types of layers for the RNN: bidirectional, convolutional (combined with pooling), dense, LSTM (Long Short-Term Memory) with dropout. The one that provided the best results in our context used the next parameters:

Layer (type)	Output Shape	# of param.
embedding_2	(none,100,200)	4,000,200
bidirectional_2	(none,100,200)	180,600
conv1D_2	(none,98,5000)	3,005,000
global_max_pooling_1d_2	(none,5000)	0

```
dense_3                    (none,100)                 500,100
dropout_2                  (none,100)                 0
dense_4                    (none,7)                   707
----------------------------------------------------------------------------------------------------
Total params: 7,686,607
Trainable params: 3,686,407
Non-trainable params: 4,000,200
----------------------------------------------------------------------------------------------------
```

The implementations of all the machine learning tools that we used are based on the scikit-learn library available at the address: https://scikit-learn.org/stable/. As for the NLP part of the tool, we used the *SnowballStemmer* library for the words stemming (https://kite.com/python/docs/nltk.SnowballStemmer), in which the stopwords removal can be activated as an option, and we used the Word2Vec models made available by Jean-Philippe Fauconnier for the words embedding (http://fauconnier.github.io/). We report the different learning curves of the three combinations of NLP and ML tools in Figure 4. As can be observed, the amount of profiling data (i.e., 4000 given that we estimate the accuracy with 5-fold cross-validation) is sufficient for the NB and MLP classifiers to approach convergence, which confirms the amount of collected data is sufficient for those classifiers to provide meaningful outcomes. The RNN shows slightly worse results in our case, which may be due both to the simple (topic) feature that we aim to capture and to a lack of data for such a more data-demanding machine learning algorithm.

Appendix B. Additional figure

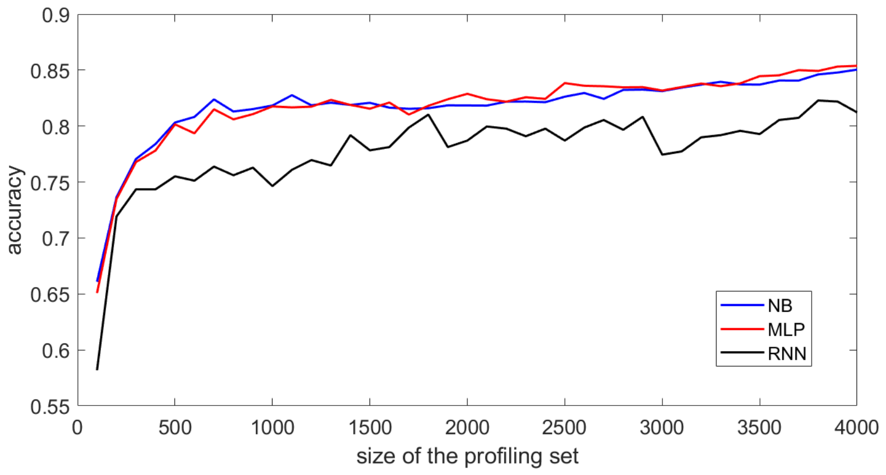


Figure 4. Learning curves of the experimented NLP+ML tools.

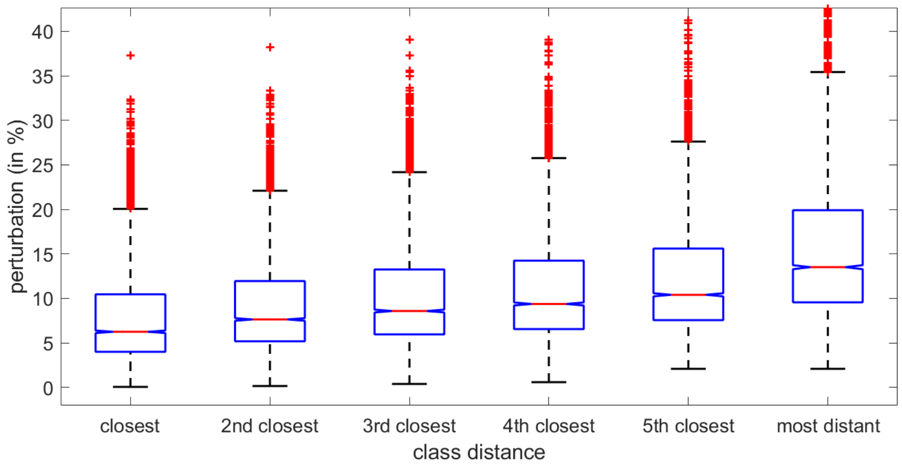


Figure 5. MLP classifier: minimum article perturbation in function of the class distance.