Say the Right Thing Right: Ethics Issues in Natural Language Generation Systems

Charese Smiley & Frank Schilder
Thomson Reuters R&D
610 Opperman Drive
Eagan, MN 55123
USA
FirstName.LastName@tr.com

Vassilis Plachouras & Jochen L. Leidner
Thomson Reuters R&D
30 South Colonnade
London E14 5EP
United Kingdom
FirstName.LastName@tr.com

Abstract

We discuss the ethical implications of Natural Language Generation systems. We use one particular system as a case study to identify and classify issues, and we provide an ethics checklist, in the hope that future system designers may benefit from conducting their own ethics reviews based on our checklist.

1 Introduction

With the advent of big data, there is increasingly a need to distill information computed from these datasets into automated summaries and reports that users can quickly digest without the need for time-consuming data munging and analysis. However, with automated summaries comes not only the added benefit of easy access to the findings of large datasets but the need for ethical considerations in ensuring that these reports accurately reflect the true nature of the underlying data and do not make any misleading statements.

This is especially vital from a Natural Language Generation (NLG) perspective because with large datasets, it may be impossible to read every generation and reasonable-sounding, but misleading, generations may slip through without proper validation. As users read the automatically generated summaries, any misleading information can affect their subsequent actions, having a real-world impact. Such summaries may also be consumed by other automated processes, which extract information or calculate sentiment for example, potentially amplifying any misrepresented information. Ideally, the research community and industry should be building NLG systems which avoid altogether behaviors that promote ethical violations. However, given the difficulty of such a task, before we reach this goal, it is necessary to have a list of best practices for building NLG systems.

This paper presents a checklist of ethics issues arising when developing NLG systems in general and more specifically from the development of an NLG system to generate descriptive text for macro-economic indicators as well as insights gleaned from our experiences with other NLG projects. While not meant to be comprehensive, it provides high and low-level views of the types of considerations that should be taken when generating directly from data to text.

The remainder of the paper is organized as follows: Section 2 covers related work in ethics for NLG systems. Section 3 introduces an ethics checklist for guiding the design of NLG systems. Section 4 describes a variety of issues we have encountered. Section 5 outlines ways to address these issues emphasizing various methods we propose should be applied while developing an NLG system. We present our conclusions in Section 6.

2 Related work

Many of the ethical issues of NLG systems have been discussed in the context of algorithmic journalism (Dörr and Hollnbuchner, 2016). They outline a general framework of moral theories following Weischenberg *et al.* (2006) that should be applied to algorithmic journalism in general and especially when NLG systems are used.

We are building on their framework by providing concrete issues we encounter while creating actual NLG systems.

Kent (2015) proposes a concrete checklist for robot journalism[1] that lists various guidelines for utilizing NLG systems in journalism. He also points out that a link back to the source data is

[1] http://mediashift.org/2015/03/an-ethical-checklist-for-robot-journalism/

Proceedings of the First Workshop on Ethics in Natural Language Processing, pages 103–108,
Valencia, Spain, April 4th, 2017.

QUESTION	EXAMPLE RESPONSE	SECTION
Human consequences		
Are there ethical objections to building the application?	No objections anticipated	4.3
How could a user be disadvantaged by the system?	No anticipated disadvantages to user	4.4-4.7
Does the system use any Personally Identifiable Information?	No PII collected or used	4.5
Data issues		
How accurate is the underlying data?*	Data is drawn from trusted source	4
Are there any misleading rankings given?	Yes, detected via data validation	4.1
Are there (automatic) checks for missing data?	Yes, detected via data validation	4.2
Does the data contain any outliers?	Yes, detected via data validation	4.2
Generation issues		
Can you defend how the story is written?*	Yes via presupposition checks and disclosure	5
Does the style of the automated report match your style?*	Yes, generations reviewed by domain experts	5
Who is watching the machines?*	Conducted internal evaluation and quality control	5
Provenance		
Will you disclose your methods?*	Disclosure text	4.4
Will you disclose the underlying data sources?	Provide link to open data & source for proprietary data	4.4

Table 1: An ethics checklist for NLG systems. There is an overlap with questions from the checklist Thomas Kent proposed and they are indicated by $*$.

essential and that such systems should at least in the beginning go through rigorous quality checks.

A comprehensive overview of ethical issues on designing computer systems can be found in (IEEE, 2016). More specifically, Amodei et al. (2016) propose an array of machine learning-based strategies for ensuring safety in general AI systems, mostly focussing on autonomous system interacting with a real world environment. Their research questions encompass avoiding negative side effects, robustness to distributional shift (i.e. the machine's situational awareness) and scalable oversight (i.e. autonomy of the machine in decision-making). The last question is clearly relevant to defining safeguards for NLG systems as well. Ethical questions addressing the impact of specifically NLP systems are addressed by Hovy and Spruit (2016).

To ensure oversight of an AI system, they draw inspiration from semi-supervised reinforcement learning and suggest to learn a reward function either based on supervised or semi-supervised active learning. We follow this suggestion and propose creating such a reward-based model for NLG systems in order to learn whether the generated texts may lay outside of the normal parameters.

Actual NLG systems are faced with word choice problem and possible data problems. Such systems, however, normally do not address the ethical consequences of the choices taken, but see Joshi et al. (1984) for an exception. Choosing the appropriate word in an NLG system was already addressed by (Ward, 1988; Barzilay and Lee, 2002), among others. More recently, Smiley et al. (2016), for example, derive the word choice of verbs describing the trend between two data points from an extensive corpus analysis. Grounding the verb choice in data helps to correctly describe the intensity of a change.

The problem of missing data can taint every data analysis and lead to misleading conclusions if not handled appropriately. Equally important as the way one imputes missing data points in the analysis is the transparent description of how data is handled. NLG system designers, in particular, have to be very careful about which kind of data their generated text is based on. To our knowledge, this problem has not been systematically addressed in the literature on creating NLG systems.

At the application level, Mahamood and Reiter (2011) present an NLG system for the neonatal care domain, which arguably is particularly sensitive as far as medical sub-domains are concerned. They generate summaries about the health status of young babies, including affective elements to calm down potentially worried parents to an appropriate degree. If a critically ill baby has seen dramatic deterioration or has died, the system appropriately does not generate any output, but refers to a human medic.[2]

3 Ethics Checklist

While there is a large body of work on metrics and methodologies for improving data quality (Batini et al., 2008), reaching a state where an NLG system could automatically determine edge cases (problems that occur at the extremes or outside of normal data ranges) or issues in the data, is a dif-

[2]Ehud Reiter, personal communication

	2009	2010	2011	2012	2013	2014
Curaçao	76.15609756	..	77.47317073	..	..	77.82439024

Table 2: Life expectancy at birth, total (years) for Curaçao.

	2006	2007	2008	2009	2010	2011
South Sudan	..	..	15,550,136,279	12,231,362,023	15,727,363,443	17,826,697,892

Table 3: GDP (current US$) for South Sudan.

ficult task. Until such systems are built, we believe it could be helpful to have some guidance in the form of an ethics checklist, which could be integrated in any existing project management process.

In Table 1, we propose such a checklist, with the aim to aid the developers of NLG systems on how to address the ethical issues arising from the use of an NLG system, and to provide a starting point for outlining mechanisms and processes to address these issues. We divided the checklist up into 4 areas starting with questions on developing NLP systems in general. The table also contains the response for a system we designed and developed and pointers to sections of the paper which discuss methods that could be deployed to make sure the issues raised by the questions are adequately addressed. The checklist was derived from our own experience with NLG systems as well as informed by the literature. We do not assert its completion, but rather offer it as a starting point that may be extended by others; also, other kinds of NLP systems may lead to specific checklists following the same methodology.

4 Current issues

This section consists of issues encountered when developing an NLG system for generating summaries for macro-economic data (Plachouras et al., 2016). To illustrate these issues we use World Bank Open Data,[3] an open access repository of global development indicator data. While this repository contains a wealth of data that can be used for generating automatic summaries, it also contains a variety of edge cases that are typical of large datasets. Managing edge cases is essential not only due to issues of grammaticality (e.g. noun-number agreement, subject-verb agreement), but because they can lead to misstatements and misrepresentations of the data that a user might act on. These issues are discussed in turn in this section.

4.1 Ranking

It is common to provide a ranking among entities with values that can be ordered. However, when there are a small number of entities, ranking may not be informative especially if the size of the set is not also given. For example, if there is only one country reporting in a region for a particular indicator an NLG engine could claim that the country is either the highest or lowest in the region. A region like North America, for which World Bank lists Bermuda, Canada, and the United States will sometimes only have data for 2 countries as Bermuda is dramatically smaller, so clarity in which countries are being compared for a given indicator and timespan is essential.

4.2 Time series

Missing Data: Enterprise applications will usually contain Terms of Use of products stating that data may be incomplete and calculations may include missing points. However, users may still assume that content shown by an application is authoritative leading to a wrong impression about the accuracy of the data. Table 3 shows the life expectancy for Curaçao from 2009-2015. Here we see that 2010, 2012, and 2013 are missing. NLG systems should check for missing values and should be informed if calculations are performed on data with missing values or if values presented to the user have been imputed.

Leading/trailing empty cells: Similar to issues with missing data, leading/trailing zeros and missing values in the data may be accurate or may signal that data was not recorded during that time period or that the phenomena started/ended when the first or last values were reported. For example, Table 3 shows empty leading values for South Sudan, a country that only recently became independent.

Small Changes: The reported life expectancy of St. Lucia was very stable in the late 1990s. In 1996, World Bank gives a life expectancy of

[3] http://data.worldbank.org

71.1574878 and in 1997, 71.15529268. Depending on our algorithm, one generation would say that there was *no change* in St. Lucia's life expectancy between 1996 and 1997 if the number was rounded to 2 decimal places. If the difference is calculated without rounding then the generation would say that there was *virtually no change*. Using the second wording allows for a more precise accounting of the slight difference seen from one year to the next.

Temporal scope: It is common to report activity occurring from a starting from the current time and extending to some fixed point in the past (e.g. *over the past 10 years*). While this is also a frequent occurrence in human written texts and dialogues, it is quite ambiguous and could refer to the start of the first year, the start of the fiscal calendar on the first year, a precise number of days extending from today to 10 years ago, or a myriad of other interpretations. Likewise, what it meant by the current time period is also ambiguous as data may or may not be reported for the current time period. If, for example, the Gross Domestic Product (GDP) for the current year is not available the generation should inform the user that the data is current as of the earliest year available.

4.3 Ethical Objections

Before beginning any NLG project, it is important to consider whether there are any reasons why the system should not be built. A system that would cause harm to the user by producing generations that are offensive should not be built without appropriate safeguards. For example, in 2016, Microsoft released Tay, a chatbot which unwittingly began to generate hate speech due to lack of filtering for racist content in its training data and output.[4]

4.4 Provenance

In the computer medium, authority is ascribed based on number of factors (Conrad et al., 2008): the user may have a prior trust distribution into humans and machines (on the "species" and individual level), they may ascribe credibility based on the generated message itself. Only being transparent about where data originated permits humans to apply their prior beliefs, whereas hiding whether generated text originated from a machine or a human leaves the user in the dark about how to use their prior beliefs to ascribe trust (or not). Once users are informed about the provenance of the information, they are enabled to decide for themselves whether or how much they trust a piece of information output by a system, such as a natural language summary.

As pointed out by Kent (2015) disclaimers on the completeness and correctness of the data should be added to the generation, or website where it's shown. Ideally, a link to the actual data source should also be provided and in general a description of how the generation is carried out in order to provide full transparency to the user. For example, such description should state whether the generated texts are personalized to match the profile of each user.

4.5 Personalization

One of the advantages of NLG systems is the capability to produce text customized to the profile of individual users. Instead of writing one text for all users, the NLG system can incorporate the background and context of a user to increase the communication effectiveness of the text. However, users are not always aware of personalization. Hence, insights they may obtain from the text can be aligned with their profile and history, but may also be missing alternative insights that are weighed down by the personalization algorithm. One way to address this limitation is to make users aware of the use of personalization, similar to how provenance can be addressed.

4.6 Fraud Prevention

In sensitive financial systems, in theory a rogue developer could introduce fraudulent code that generates overly positive or negative-sounding sentiment for a company, for their financial gain. A code audit can bring attempts to manipulate any code base to light, and pair programming may make any attempts less likely.

4.7 Accessibility

In addition to providing misleading texts, the accessibility of the texts generated automatically is an additional way in which users may be put in a disadvantaged position by the use of an NLG system. First, the readability of the generated text may not match the expectations of the target users, limiting their understanding due to the use of specialized terminology, or complex structure. Second, the quality of the user experience may be af-

[4] http://read.bi/2ljdvww

fected if the generated text has been constructed without considering the requirements of how users access the text. For example, delivering a text through a text-to-speech synthesizer may require to expand numerical expressions or to construct shorter texts because of the time required for the articulation of speech.

5 Discussion

The research community and the industry should aim to design NLG systems that do not promote unethical behavior, by detecting issues in the data and automatically identifying cases where the automated summaries do not reflect the true nature of the data.

There are a couple of methods we want to highlight because they address the problems of solving ethical issues from two different angles. The first method we called *presupposition check* draws principled way of describing pragmatic issues in language by adding semantic and pragmatic constraints informed by Grice's Cooperative Principles and presupposition (Grice, 1975; Beaver, 1997): Adding formal constraints to the generation process will make NLG more transparent, and less potentially misleading (Joshi, 1982).

If an NLG system, for example, is asked to generate a phrase expressing the minimum, average or maximum of a group of numbers ("The smallest/average/largest (Property) of (Group) is (Value)"), an automatic check should be installed that determines whether the cardinality of the set comprising that group is greater than one. If this check only finds one entity, the generation should be licensed and the system avoids that user is misled into believing the very notion of calculating a minimum, average or maximum actually makes sense. Instead, in such a situation a better response may be "There is only one (Property) in (Group), and it is (Value)." (cf. work on the NLG of gradable properties by van Deemter (2006)).

A second method to ensure that the output of the generated system is valid involves evaluating and monitoring the quality of the text. A model can be trained to identify problematic generations based on an active learning approach. For example, interquartile ranges can be computed for numerical data used for the generation determining outliers in the data. In addition, the fraction of missing data points and the number of input elements in aggregate functions can be estimated from the respective data. Then, domain experts can rate whether the generated text is acceptable or not as a description of the respective data. The judgements can be used to train a classifier that can be applied to future data sets and generations.

6 Conclusions

We analyzed how the development of an NLG system can have ethical implications considering in particular data problems and how the meaning of the generated text can be potentially misleading. We also introduced best practice guidelines for creating an NLP system in general and transparency in interaction with a user.

Based on the checklist for the NLG systems we proposed various methods for ensuring that the right utterance is generated. We discussed in particular two methods that future research should focus on: (a) the validation of utterances via a presupposition checker and (b) a better evaluation framework that may be able to learn from feedback and improve upon that feedback.

Checklists can be collected as project management artifacts for each completed NLP project in order to create a learning organization, and they are a useful resource that inform Ethics Review Boards, as introduced by Leidner and Plachouras (2017).

Acknowledgments

We would like to thank Khalid Al-Kofahi for supporting this work.

References

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Regina Barzilay and Lillian Lee. 2002. Bootstrapping lexical choice via multiple-sequence alignment. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 164–171. Association for Computational Linguistics, July.

Carlo Batini, Federico Cabitza, Cinzia Cappiello, and Chiara Francalanci. 2008. A comprehensive data quality methodology for Web and structured data. *Int. J. Innov. Comput. Appl.*, 1(3):205–218, July.

David Beaver. 1997. Presupposition. In Johan van Benthem and Alice ter Meulen, editors, *The Hand-*

book of Logic and Language, pages 939–1008. Elsevier, Amsterdam.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of nlg systems. In *In Proc. EACL'06*, pages 313–320.

Jack G. Conrad, Jochen L. Leidner, and Frank Schilder. 2008. Professional credibility: Authority on the Web. In *Proceedings of the 2nd ACM Workshop on Information Credibility on the Web*, WICOW 2008, pages 85–88, New York, NY, USA. ACM.

Konstantin Nicholas Dörr and Katharina Hollnbuchner. 2016. Ethical challenges of algorithmic journalism. *Digital Journalism*, pages 1–16.

Paul Grice. 1975. Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax and Semantics III: Speech Acts*, pages 41–58. Academic Press, New York, NY, USA.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 591–598.

IEEE, editor. 2016. *Ethically Aligned Design: A Vision For Prioritizing Wellbeing With Artificial Intelligence And Autonomous Systems*. IEEE - advanced Technology for Humanity.

Aravind Joshi, Bonnie Webber, and Ralph M. Weischede. 1984. Preventing false inferences. In *Proceedings of the 10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*, pages 134–138, Stanford, California, USA, July. Association for Computational Linguistics.

Aravind Joshi. 1982. Mutual beliefs in question-answering systems. In Neil S. Smith, editor, *Mutual Knowledge*, pages 181–197. Academic Press, London.

Thomas Kent. 2015. "an ethical checklist for robot journalism. Online, cited 2017-01-25, http://mediashift.org/2015/03/an-ethical-checklist-for-robot-journalism/.

Jochen L. Leidner and Vassilis Plachouras. 2017. Ethical by design: Ethics best practices for natural language processing. In *Proceedings of the Workshop on Ethics & NLP held at the EACL Conference, April 3-7, 2017*, Valencia, Spain. ACL.

Saad Mahamood and Ehud Reiter. 2011. Generating affective natural language for parents of neonatal infants. In *Proceedings of the 13th European Workshop on Natural Language Generation*, ENLG 2011, pages 12–21, Nancy, France. Association for Computational Linguistics.

Leo L. Pipino, Yang W. Lee, and Richard Y. Wang. 2002. Data quality assessment. *Commun. ACM*, 45(4):211–218, April.

Vassilis Plachouras, Charese Smiley, Hiroko Bretz, Ola Taylor, Jochen L. Leidner, Dezhao Song, and Frank Schilder. 2016. Interacting with financial data using natural language. In Raffaele Perego, Fabrizio Sebastiani, Javed A. Aslam, Ian Ruthven, and Justin Zobel, editors, *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, Pisa, Italy, July 17-21, 2016*, SIGIR 2016, pages 1121–1124. ACM.

Ehud Reiter. 2007. An architecture for data-to-text systems. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*, ENLG '07, pages 97–104, Stroudsburg, PA, USA. Association for Computational Linguistics.

Frank Schilder, Blake Howald, and Ravi Kondadadi. 2013. Gennext: A consolidated domain adaptable nlg system. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 178–182, Sofia, Bulgaria, August. Association for Computational Linguistics.

Charese Smiley, Vassilis Plachouras, Frank Schilder, Hiroko Bretz, Jochen L. Leidner, and Dezhao Song. 2016. When to plummet and when to soar: Corpus based verb selection for natural language generation. In *The 9th International Natural Language Generation Conference*, page 36.

Kees van Deemter. 2006. Generating referring expressions that involve gradable properties. *Computational Linguistics*, 32(2):195–222.

Nigel Ward. 1988. Issues in word choice. In *Proceedings of the 12th Conference on Computational Linguistics-Volume 2*, pages 726–731. Association for Computational Linguistics.

Siegfried Weischenberg, Maja Malik, and Armin Scholl. 2006. Die Souffleure der Mediengesellschaft. *Report über die Journalisten in Deutschland. Konstanz: UVK*, page 204.