Self-Portrayals of GI Junior Fellows

Daniel Braun*

# Natural Language Processing for Social Good: contributing to research and society

https://doi.org/10.1515/itit-2024-0095
Received November 5, 2024; accepted February 19, 2025;
published online April 24, 2025

**Abstract:** With great power comes great responsibility. As Artificial Intelligence in general and Natural Language Processing, in particular, continue to shape the world we live in, there is an increased need for Computer Scientists in academia and industry to acknowledge the pivotal role our discipline plays in shaping the future of our society. Accepting this role should come with the responsibility to make positive contributions, whether through research, teaching, application, or by contributing to the public discourse surrounding AI technology and its regulation. This text outlines how Natural Language Processing can contribute to shaping a better future in which society at large can benefit from technological advances.

**Keywords:** Natural Language Processing; Large Language Models; AI for Social Good

## 1 Introduction

The increased abilities and ubiquitous availability of Large Language Models (LLMs) have raised the urgency of practical and societal questions about Artificial Intelligence (AI) and its applications that have been discussed within the scientific community for many years and have brought them to the center of public debates. Deepfakes, disinformation, copyright concerns, threats to the integrity of assessments, and many other problems have transformed from scholarly debates to real-world problems in a matter of months.

The public debates that are fuled by these developments often seem to be dominated by extreme positions: On the one hand, techno-optimists that believe the chase for ever bigger and better AI models will somehow overcome all the problems both society and technology face today, and any limit put on the development or use of AI will just slow down our pathway to this better future. On the other hand, doomsday fears about uncontrollable AI systems that go rogue and have the potential to end civilization as we know it if we do not stop all AI research immediately.

Just like every other fundamental technology, AI has the potential to (and already does) improve our lives, but also comes with potential dangers that have to be taken seriously. One way to mitigate such risks is through legislation. The European Union's AI Act and initiatives from other lawmakers have sparked lively debates about the regulation of AI. Laws and regulations are important instruments for protecting individuals and society at large from potential dangers that could arise from the use of AI systems. However, it is debatable whether the approach taken by the AI Act is the correct one and even whether "Artificial Intelligence" as such is a useful regulatory category (see [1]).

Independent of a specific approach to regulation, we should strive for more than just mitigating risks and preventing harm. With this text and in extension also with my research, I want to advocate for a more proactive approach within AI research, and computer science more broadly, that acknowledges the responsibility we have as a discipline and individual researchers. I believe that, particularly in publicly funded research, we should not just chase metrics in order to be on top of an imaginary or real leaderboard. Instead, we should also pursue methods and applications that go beyond just preventing harm from AI systems and aim to actively contribute to a better society.

## 2 Natural Language Processing for Social Good

The idea of focusing on the positive social and societal impact that NLP, and AI more broadly, can have has gained traction in recent years thanks to initiatives such as "NLP for Social Good"[1] and a number of dedicated events like the "Workshop on NLP for Positive Impact" and the "Symposium on NLP for Social Good". While the impact AI [2], [3] and NLP [4]–[6] can potentially have has been discussed for many years, recent advances in technology

***Corresponding author: Daniel Braun**, Department of Mathematics and Computer Science, University of Marburg, Marburg, Germany, E-mail: daniel.braun@uni-marburg.de. https://orcid.org/0000-0001-8120-3368

[1] https://nlp4sg.vercel.app/.

have made practical applications more viable than ever before. Nevertheless, many challenges remain with regard to the underlying technologies as well as with regard to specific applications and their impact.

This section briefly outlines my research interests and contributions in the area of NLP for Social Good, starting with NLP models (Sections 2.1 and 2.2), followed by evaluation approaches (Section 2.3), and finally applications in the legal domain (Section 2.4).

## 2.1 Trustworthiness

One of the key issues in state-of-the-art NLP technology and particularly LLMs remains to be factual correctness. Despite all advances, LLMs still tend to "hallucinate", i.e. produce texts that are factually incorrect (see [7] for a more comprehensive definition of hallucination). It seems unlikely that we will be able to comprehensively address this issue on the current path of scale, which is purely based on ever larger data sets to train on and models with an ever larger number of parameters. Despite incremental improvements in factual correctness, mistakes will always happen. The fact that LLMs have no explicit notion of certainty (or even factuality) when generating texts makes it difficult to detect the unavoidable errors. This limits the unsupervised application of LLMs in domains in which they could have significant positive impact and remains one of the biggest challenges in NLP research.

Instead of relying solely on scaling, we need to investigate novel approaches like post-generation fact-checking or leveraging explicit representations of knowledge, e.g. in the form of knowledge graphs, in order to not just improve the factual correctness of the output generated by LLMs, but also the verifiability (and hence the trustworthiness) of LLMs and their outputs.

Another important factor to build trust are the processes in which the use of AI technology is embedded. Human oversight, in the form of human-in-the-loop approaches, is one of the most effective measures to build trust in AI systems, as we have shown in the context of higher education in [8].

## 2.2 Perspectives and disagreement

Most Machine Learning (ML) approaches rely on finding a single "truth" or "gold standard", e.g. by majority vote and removal of "outliers" during the annotation process. This assumed truth is subsequently used for the training and evaluation of ML models. Only recently, questioning this assumption of a single truth that can be found in an annotation process has gained attention within the NLP community (see e.g. [9]–[12]).

Particularly for inherently subjective tasks, like the detection of offensive language, it seems reasonable to assume that such an objective truth does not exist. In other domains, like politics, eradicating divergent positions is not only factually wrong but can become a thread for our pluralistic society as AI models become more deeply embedded in our daily lives. But even in domains that are traditionally seen as more "factual", like medicine or law, experts regularly disagree on decisions and conclusions. Instead of trying to remove these instances, we should try to leverage the information such a disagreement between expert annotators can provide.

In my research, I investigate how we can include different perspectives, opinions, and standpoints in all steps of the ML pipeline:

Starting with the annotation of data, where we have to separate noise, such as honest mistakes and attention slips during the annotation, from genuine disagreement that can provide valuable insights using, by using annotation processes that allow for such distinctions, as outlined in [9]. Followed by the training of ML models, where we have to develop new approaches that can be trained on a range of (potentially disagreeing) inputs rather than a single truth. E.g. by using probability-based multi-label methods, ensemble approaches, or instruction tuning, as exemplified in [13], [14].

And finally, the presentation of results from such models, which can no longer rely on just presenting the most likely option, but has to be a representation of the different opinions and annotations that went into the model. Presenting such a more nuanced and comprehensive results is also preferred by potential users, as we have shown in [13].

## 2.3 Evaluation

Evaluating Natural Language Generation (NLG) systems and their outputs is particularly challenging. It is rarely possible to classify such outputs in a binary correct/incorrect schema and the overall quality of a text depends on many factors, like fluency, readability, grammaticality, and (factual) correctness [15]. Conducting evaluations that consider all these factors is time-consuming and expensive. Therefore, researchers often rely on using automated evaluation metrics that are much cheaper to apply. Although these metrics are designed to correlate with human judgments on specific tasks and domains, they are often used in cases where their applicability and validity are questionable [16]. While these problems have existed since the first NLG systems were built, LLMs have added a further challenge to current evaluation practices due to the fact that they are trained on such vast amounts of data that often we cannot be sure anymore that they have not been trained on the test data included in benchmark datasets (and sometimes we can even be reasonably sure that they have been) [17].

Independent of such problems, purely metric-based evaluations can have limited significance with regard to the benefits a system can provide in practical application. Even if a system is very good at performing a certain task, it does not necessarily add significant value in practice, e.g. if the task has limited relevance or if the system is not properly integrated into existing workflows. To evaluate the potential benefits a system can have in a real-world application, task-based evaluations in realistic contexts are indispensable. In [18], for example, we conducted a task-based evaluation with real drivers of a behavior-change support systems that supports driving safety. In [19], we conducted a task-based evaluation of a system that can automatically detect potentially void clauses in consumer contracts with experts from consumer protection NGOs.

From a commercial perspective, the current practices in NLP evaluation often provide little insight into the economic viability of a system. We need frameworks that are able to quantify whether an AI system is beneficial from an economical perspective by taking into account its benefits (savings, service improvements, reduced waiting time, …) but also its potential costs (error frequency, costs of specific error classes, …) in order to allow companies to make informed decisions about the use of AI systems in production.

### 2.4 Access to justice

I believe that one of the most important contributions research can make to ensure AI is benefiting our society as a whole is to actively seek and work on applications that focus on positive impact, e.g., along the lines of the UN Sustainable Development Goals. One focus of my work is therefore on applications of NLP that support access to justice. Legal matters are often complicated to assess for ordinary citizens and seeking legal advice from trained professionals is expensive. NLP technology has the potential to lower the barrier to access to justice and legal advice. However, most existing so-called “Legal Tech” solutions focus on supporting companies rather than consumers, contributing to an already existing imbalance of power between both groups.

In my research, I work with legal scholars and practitioners, as well as NGOs, like the German “Verbraucherzentralen”, in order to support consumer protection. In the AGB-Check project, for example, which was funded by the German Federal Ministry of Justice and Consumer Protection, we developed a tool that is able to support consumer protection lawyers in the assessment of clauses from Terms and Conditions from online shops by automatically identifying the topic of a clause and clauses that are potentially void [20]–[22].

Many of the issues described above, like trustworthiness, but also other aspects that are currently at the center of ongoing NLP research like explainability or multimodality are particularly important for applications in the legal domain and other domains that are relevant for our society, like medicine or education. Working on such use cases is therefore not only relevant from a societal perspective but also from a scientific one.

## 3 The role of computer science in society

AI will continue to shape how we work, learn, and live in the future. Computer Science, as a discipline, is at the core of these developments, and with this position comes a responsibility that is still too often neglected. Hiding behind the alleged neutrality of technology instead of acknowledging the influence we have. Interdisciplinary collaboration, particularly with experts from the legal domain, can be eye-opening with regard to the different approaches different disciplines take with regard to their own responsibility.

Lawyers are usually liable for legal advice they provide. That is one of the reasons why it is sometimes difficult to get an unambiguous statement in legal matters. Rumor has it that many answers from legal experts start with “it depends”. Computer science, on the other hand, is often based on the principle of “move fast and break things”. In order to keep up with the stunning pace of innovation, one sometimes has to compromise on caution. It seems, for example, to be perfectly acceptable to put an “autonomous” car with a software labeled as “beta” in the hands of ordinary customers. An approach that is diametral to how legal professionals act.

Innovative technology almost always comes with risks. It is neither realistic nor necessary to try to eliminate all risks associated with AI and other innovative technology. However, as AI systems are deployed across many different parts of our lives, including in critical settings like medicine, the stakes get higher, and it might be time to re-evaluate how much we are willing to “break” in order to accelerate the speed of development. We should not wait for answers from regulators but take an active role in public debates, explain technical backgrounds to a wider audience, and actively shape and influence the discussion. While associations like the German Informatics Society (GI) or the Association for Computing Machinery (ACM) already play such a role in our society and inform legal decision-making processes, they rely on an active and engaged Computer Science community to which we can all contribute.

While regulation currently takes the spotlight in many debates, it should only represent the lowest bar we agree on as a society to prevent existential harm. While it is important to have these safeguards in place, we should strive for positive change and technology that goes beyond the bare minimum that is required by law and be a force of positive change in the world.

## 4 Outlook

Exciting but challenging times lay ahead of us. There is little doubt that we will continue to see rapid improvements in AI capabilities and the applications they fuel. While the EU already presented the first comprehensive legal framework for AI regulation, it remains to be seen how it will be implemented and enforced in practice. Many of the demands towards AI systems outlined in the AI Act, e.g. with regard to explainability, are technically impossible to be satisfied by state-of-the-art LLMs. However, it is hard to imagine that any regulation will be able or even willing to turn the clock back on the innovation of the last years in a highly competitive and international market.

The path taken by companies like *Open*AI has emphasized the importance of truly open, publicly funded research and AI infrastructure. With my research, I want to continue to raise awareness for NLP research that is oriented towards the common good and foster interdisciplinary collaborations to follow this path.

Moreover, I believe that one key task for academia will be to ensure that the next generation of computer scientists will leave universities with an increased awareness of their professional role and the role of their discipline within our society. We can no longer afford to hide behind our technology and pretend that the way in which it is applied and the consequences it has on people and their lives is none of our business.

**Research ethics:** Not applicable.
**Informed consent:** Not applicable.
**Author contributions:** The author has accepted responsibility for the entire content of this manuscript and approved its submission.
**Use of Large Language Models, AI and Machine Learning Tools:** None declared.
**Conflict of interest:** Author states no conflict of interest.
**Research funding:** None declared.
**Data availability:** Not applicable.

## References

[1] D. Braun, “Why “artificial intelligence” should not be regulated,” *Digital Gov. Res. Pract.*, 2024.

[2] B. Berendt, “Ai for the common good?! pitfalls, challenges, and ethics pen-testing,” *Paladyn, Journal of Behavioral Robotics*, vol. 10, no. 1, pp. 44–65, 2019.

[3] L. Floridi, *et al*., “Ai4people—an ethical framework for a good ai society: opportunities, risks, principles, and recommendations,” *Minds Mach.*, vol. 28, no. 4, pp. 689–707, 2018.

[4] D. Hovy and S. L. Spruit, “The social impact of natural language processing,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, K. Erk, and N. A. Smith, Eds., Berlin, Germany, Association for Computational Linguistics, 2016, pp. 591–598.

[5] Z. Jin, G. Chauhan, B. Tse, M. Sachan, and R. Mihalcea, “How good is NLP? a sober look at NLP tasks through the lens of social impact,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. [Online], C. Zong, F. Xia, W. Li, and R. Navigli, Eds., Association for Computational Linguistics, 2021, pp. 3099–3113.

[6] N. Tomašev, *et al*., “Ai for social good: unlocking the opportunity for positive impact,” *Nat. Commun.*, vol. 11, no. 1, p. 2468, 2020.

[7] K. van Deemter, “The pitfalls of defining hallucination,” *Comput. Linguist.*, vol. 50, no. 2, pp. 807–816, 2024.

[8] D. Braun, P. Rogetzer, E. Stoica, and H. Kurzhals, “Students’ perspective on ai-supported assessment of open-ended questions in higher education,” in *15th International Conference on Computer Supported Education, CSEDU 2023*, 2023, pp. 73–79.

[9] D. Braun, “I beg to differ: how disagreement is handled in the annotation of legal machine learning data sets,” *Artif. Intell. Law*, vol. 32, no. 3, pp. 839–862, 2023.

[10] F. Cabitza, A. Campagner, and V. Basile, “Toward a perspectivist turn in ground truthing for predictive computing,” *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 6, pp. 6860–6868, 2023.

[11] B. Plank, “The “problem” of human label variation: on ground truth in data, modeling and evaluation,” in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds., Abu Dhabi, United Arab Emirates, Association for Computational Linguistics, 2022, pp. 10671–10682.

[12] A. N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, and M. Poesio, “Learning from disagreement: a survey,” *J. Artif. Intell. Res.*, vol. 72, pp. 1385–1470, 2021.

[13] J. Xu, M. Theune, and D. Braun, “Leveraging annotator disagreement for text classification,” in *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, M. Abbas, and A. A. Freihat, Eds., Association for Computational Linguistics, 2024.

[14] L. Zhang and D. Braun, “Twente-BMS-NLP at PerspectiveArg 2024: combining bi-encoder and cross-encoder for argument retrieval,” in *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, Y. Ajjour, R. Bar-Haim, R. El Baff, Z. Liu, and G. Skitalinskaya,

Eds., Bangkok, Thailand, Association for Computational Linguistics, 2024, pp. 164—168.

[15] C. van der Lee, A. Gatt, E. van Miltenburg, and E. Krahmer, "Human evaluation of automatically generated text: current trends and best practice guidelines," *Comput. Speech Lang.*, vol. 67, 2021, Art. no. 101151. https://doi.org/10.1016/j.csl.2020.101151.

[16] E. Reiter, "A structured review of the validity of BLEU," *Comput. Linguist.*, vol. 44, no. 3, pp. 393—401, 2018.

[17] S. Balloccu, P. Schmidtová, M. Lango, and O. Dusek, "Leak, cheat, repeat: data contamination and evaluation malpractices in closed-source LLMs," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Y. Graham, and M. Purver, Eds., St. Julian's, Malta, Association for Computational Linguistics, 2024, pp. 67—93.

[18] D. Braun, E. Reiter, and A. Siddharthan, "Saferdrive: an nlg-based behaviour change support system for drivers," *Nat. Lang. Eng.*, vol. 24, no. 4, pp. 551—588, 2018.

[19] D. Braun, "Automated semantic analysis, legal assessment, and summarization of standard form contracts," Ph.D. thesis, Technische Universität München, 2021.

[20] D. Braun and F. Matthes, "NLP for consumer protection: battling illegal clauses in German terms and conditions in online shopping," in *Proceedings of the 1st Workshop on NLP for Positive Impact*. [Online], Association for Computational Linguistics, 2021, pp. 93—99.

[21] D. Braun and F. Matthes, "AGB-DE: a corpus for the automated legal assessment of clauses in German consumer contracts," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds., Bangkok, Thailand, Association for Computational Linguistics, 2024, pp. 10389—10405.

[22] D. Braun, E. Scepankova, P. Holl, and F. Matthes, "Consumer protection in the digital era: the potential of customer-centered legaltech," in *INFORMATIK 2019: 50 Jahre Gesellschaft für Informatik — Informatik für Gesellschaft*, K. David, K. Geihs, M. Lange, and G. Stumme, Eds., Bonn, Gesellschaft für Informatik e.V, 2019, pp. 407—420.

# Bionote

**Daniel Braun**
Department of Mathematics and Computer Science, University of Marburg, Marburg, Germany
**daniel.braun@uni-marburg.de**
**https://orcid.org/0000-0001-8120-3368**

Daniel Braun is a Professor of Computer Science and Head of the Natural Language Processing Group at the University of Marburg. He was awarded a Junior-Fellowship by the German Informatics Society (GI) in 2024.