Ethical Issues in Language Resources and Language Technology – New Challenges, New Perspectives

Paweł Kamocki, Andreas Witt
Leibniz-Institut für Deutsche Sprache
R5 6-13, 68161 Mannheim, Germany
kamocki | witt@ids-mannheim.de

Abstract

This article elaborates on the author's contribution to the previous edition of the LREC conference, in which they proposed a tentative taxonomy of ethical issues that affect Language Resources (LRs) and Language Technology (LT) at the various stages of their lifecycle (conception, creation, use and evaluation). The proposed taxonomy was built around the following ethical principles: Privacy, Property, Equality, Transparency and Freedom.
In this article, the authors would like to: 1) examine whether and how this taxonomy stood the test of time, in light of the recent developments in the legal framework and popularisation of Large Language Models (LLMs); 2) provide some details and a tentative checklist on how the taxonomy can be applied in practice; and 3) develop the taxonomy by adding new principles (Accountability; Risk Anticipation and Limitation; Reliability and Limited Confidence), to address the technological developments in LLMs and the upcoming Artificial Intelligence Act.

Keywords: ethics, generative AI, privacy

1. Introduction

In our contribution to the previous edition of the LREC Conference (Kamocki, Witt, 2022), we proposed a tentative taxonomy of ethical issues affecting Language Resources (LRs) and Language Technology (LT) tools throughout their entire lifecycle, built around the principles of Privacy, Property, Equality, Transparency and Freedom. In this article, we would like to elaborate on this idea.

1.1 Ethical Norms over Time

It is a tempting perspective to think of ethical norms as something universal and perfectly static, i.e. not changing over time. The proponents of this view on ethics would use the Decalogue as an example: formulated millenia ago (probably around 7th century BC), the Ten Commandments are still the cornerstone of ethics and a foundation of (not only Western) civilisation. The argument, however, is inherently flawed, as the biblical version of the 10th commandment ('*Thou shalt not covet thy neighbour's house, thou shalt not covet thy neighbour's wife, nor his manservant, nor his maidservant, nor his ox, nor his ass, nor any thing that is thy neighbor's*'. – Exodus 20:17) is nowadays more generally known in its simplified version: You shall not covet. The reasons behind this are certainly not purely mnemonic; rather, in today's world wives are not considered property, slavery had been abolished, and oxen and donkeys are not generally seen as particularly desirable items (compared to, for example, high-end laptops, luxury watches or electric cars). The world has changed, and ethical norms, even as fundamental as the Ten Commandments, had to be adapted to the new reality.

It is therefore not surprising that ethical guidelines concerning something as dynamic as LT and LRs also need to change, and quite often. The taxonomy proposed in our previous contribution should be revised and, if necessary, completed.

1.2 Changes since 2022

The two years that passed since our original proposal seem to be a very short period of time, even in the evolving field of LT and LRs. However, some important developments have taken place during that time.

Most importantly, since the launch of Chat GPT in late 2022, LLMs have attracted a lot of public attention. Before that date, LLMs were mostly regarded as a useful tool in applications such as Machine Translation or Speech Recognition, but few were able to predict that LLMs will become independent tools, and almost ontologically independent beings. The debate on ethical implications of LLMs is now present in mainstream media (e.g. Metz, Weise, 2023), with reports on individuals having romantic relations with language models, or even committing suicide under their influence (Xiang, 2023).

In this unprecedented context, the EU Artificial Intelligence (AI) Act is taking shape (European Commission, 2021); it is expected to be soon, and become applicable in 2024. The AI Act already existed as a draft in 2022, but seemed, we admit it, rather far removed from LT and LRs, focused instead on such applications of AI systems as biometric identification, law enforcement and administration of justice. However, ChatGPT has revolutionised the perception of LLMs, which now may seem qualifiable as high-impact AI systems. Such systems are heavily regulated by the draft AI Act; before adoption, the draft is expected to be substantially modified in such a way as to regulate LLMs (and other foundation models) even more (Volpicelli, 2023; Zenner, 2023).

As explained in our previous contribution, while we agree that ethical norms are distinct in nature from legal norms, we also believe that the two systems – law and ethics – affect each other. This mutual influence is particularly visible in the field of new technologies, as laypersons, usually unable to comprehend the functioning of these technologies and their underlying principles, are more inclined to

LEGAL2024 @LREC-COLING-2024, pages 19–23

perceive them as 'evil' or 'immoral' when they are prohibited or restricted by law. This is why, in our opinion, the AI Act, even *in statu nascendi*, has an impact on LR and LT ethics.

1.3 Continous Relevance of the Proposed Taxonomy

All the above does not mean that the taxonomy we proposed in 2022 is now outdated. Quite the contrary, we stand by all the principles we formulated in our previous contribution, i.e. Privacy, Property, Equality, Transparency and Freedom. At the same time, we admit that some of these principles, and especially the principle of Equality, is increasingly unclear and difficult to apply in the context of LLMs and generative AI. Nevertheless, and perhaps all the more so, it should remain in sight throughout the entire lifecycle of an LR or an LT tool.

2. Overview of Ethical Issues throughout the LR or LT tool Lifecycle

In providing an overview of the ethical issues affecting LRs and LT, we will follow the same structure as in our previous contribution, dividing their lifecycle into four stages: conception, construction, use and evaluation. In this contribution, we would include a tentative checklist with questions that need to be addressed at every stage.

2.1 Conception Phase

Already at the earliest stage, i.e. the conception phase (before any data collection), certain ethical considerations need to be addressed. The questions that should be asked include:

1. **Under whose responsibility is the tool or resource developed?**

This fundamental question may be overlooked in joint research projects or public-private partnerships, as it is not always well-integrated in data-intensive technology research (Wagner, 2020). Although it may seem as a legal (more than ethical) question (cf. Article 5(2) of the GDPR), it should, in our view, precede any legal analysis. The legal situation (regarding responsibility, intellectual property, warranties…) should be modelled by a contract, or a series thereof, based on the answer to this question, which can also be formulated as: if things go wrong, who is to take the blame? This person (or, more often, this entity), would then be morally obliged to minimise the associated risks, and should be given the organisational (and legal) tools to do so. With great power comes great responsibility, and, ideally, *vice versa*.

This question is also essential to address the principle of accountability, which is a basic principle of the GDPR, but also of the AI Act, and one of the OECD Principles for responsible stewardship of trustworthy AI (OECD, 2019).

In practice, responsibility can be limited to specific tasks: for example, one entity can be responsible for training a model, another one for developing an application based on that model, and yet another one for commercialising it. However, situations where responsibility is thinly spread should generally be avoided, and whenever possible, responsibility should be concentrated in as small a number of entities as possible.

2. **What is the intended purpose of the tool or resource? What are its potential uses and foreseeable misuses?**

Although sometimes the purpose might be difficult to grasp (some resources or tools can initially be general-purpose, and then shift to a more specific application, or vice versa), this question still helps to anticipate the associated risks. A resource intended for researchers (e.g. a corpus of 17th century theatrical plays) is held to a lower standard of risk anticipation than, e.g., a chatbot intended to assist air traffic controllers, as the potential harm caused by malfunctioning or misuse is much higher in case of the latter.

Defining the intended use is also instrumental in assessing the reliability of the tool or resource – it is reliable if it performs its main task in a way that is both reasonably accurate and proportionate. Accuracy in this context means that the output does not contain false information; proportionality: that the output does not contain more or less information than reasonably needed or expected. For example, a chatbot intended to provide passengers with information about train schedules is accurate if it provides information about existing trains only (not information that is out-of-date, or, worse, that is 'hallucinated', like an imaginary direct train connection between Prague and Oxford). It is proportionate if it provides relevant information such as time of departure, expected perturbations and a crowd forecast; it is disproportionate if it omits to provide essential information (e.g., departure time) or if it provides irrelevant information (e.g., the number of the seat next to an unaccompanied teenager, or the name of the conductor).

As per the AI Act, it is also necessary to consider 'foreseeable misuses', as they play a prominent role in risk assessment. For example, the abovementioned corpus of 17th century theatrical plays can hardly be misused, whereas a speech synthesis tool may potentially be misused by minors to circumvent age restrictions (by making them sound as adults on the phone). This brings us to the next question, i.e.,:

3. **What are the intended user groups?**

The intended users are, of course, closely linked to the intended use, at least *prima facie*. In particular, it should be considered here whether the tool or resource can be made available to minors, or other groups that deserve special protection (people with disabilities, elderly people, refugees), in which case a higher standard of risk anticipation and management should be applied.

Attention needs to be paid to tools and resources that, although primarily intended for a narrow group of users (e.g. university researchers), are going to be made openly available, to satisfy the requirements of the Open Science movement. Open availability means that the tool may also be used by groups like minors or convicted criminals. This should be taken into account on the one hand in risk anticipation and management (see above about foreseeable misuses), and on the other hand in the decision whether the resource or tool should actually be made openly available or not. We believe that in many cases ethical considerations related to risk mitigation should prevail over Open Science ideals. After all, FAIR data should be as open as possible, but also as closed as necessary (Landi et al., 2020).

4. **What is the potential impact of the tool or resource on the users?**

This question seems closely related to the two previous ones. However, one should consider the impact not only of the intended use by the target user groups, but also more broadly: what is the worst possible scenario involving the tool? Can it harm the user in any way, by its normal functioning or malfunctioning? How likely is this worst scenario to happen? How can the risk of it happening be further reduced?

Of course, we are aware of the fact that few human activities are completely risk-free – one can be killed or seriously injured while turning the light on, if several factors coincide (e.g. wet hands, bare feet and faulty electric installation). This does not mean that light switches should be banned, or only made available to trained professionals, or that an average user should be constantly reminded about the risk of being electrocuted while operating a switch. However, while in the presence of a relatively new technology, such as a very large language model (as opposed to a known and tested piece technology, as a light switch), any non-negligible risks should be carefully considered, anticipated and mitigated, and the users warned about their existence.

The 'impact on users', as discussed here, includes impact on their privacy, understood both as *'freedom from unauthorised intrusion'* and as a *'right to keep one's personal matters and relationships secret'*. This is related to the GDPR principle of privacy by design (Kamocki, Witt 2020), and discussed at length in our previous contribution (Kamocki, Witt 2022).

Moreover, a tool or resource that is ill-balanced since the conception phase would disregard the principle of Equality (also described in our previous contribution), and have a negative impact on some users.

The answers to all the above questions should be thoroughly documented and made available to users in an appropriate form, in the spirit of the fundamental principle of Transparency (OECD, 2019; Kamocki, Witt 2022).

2.2 Construction Phase

The construction phase contains for the most part of data collection and preparation (annotation, etc.). At this stage, the following questions should be taken into account:

1. **Are the data (and other material) subject to (intellectual) property rights?**

This question is directly related to the principle of Property, which we elaborated upon in our previous contribution (Kamocki, Witt, 2022). Even though today many researchers decide to openly share their data and code, in the spirit of Open Science, this does not change the fact that language data and software code are (almost always) protected by copyright, which means they can only be copied and shared with the right holder's permission or under a statutory exception. In the EU, such exceptions for Text and Data Mining (whether carried out for research purposes or for any other purpose) were introduced in the 2019 Directive on copyright in the Digital Single Market. In the US, to the best of our knowledge, the fair use doctrine allows for a wide range of uses related to research and new technologies in general.

In any case, it is important to decide whether the data can be used (e.g. scraped from the Internet) on the basis of an exception, or whether a permission (licence) should be obtained. Needless to say, the decision should be grounded in a thorough legal analysis, and properly documented.

2. **Are privacy-sensitive data used? If so, is the concerned individual allowed to opt-out?**

We use the term 'privacy-sensitive data' instead of 'personal data' for a good reason: as explained in our previous contribution, we want to examine the issue of privacy not from the legal perspective, but from a broader, ethical one.

In general, it seems to us that in most cases providing the person whose interests (privacy or others) are affected by the data the right to opt-out by withdrawing 'their data' from the processing – even if such an opt-out is not required by law – is the best way to address 'data sensitivity'. However, in certain situations an opt-out may require a delicate balance of interests, as opt-out by one individual can negatively affect another one. In a somewhat simplistic example, if Team A loses in competition with Team B, Team A can argue that this information negatively affects its members, who might be seen as less competent. Withdrawal of this information from the processing, however, would negatively impact the interests of the winning Team B. The opt-out request, in such circumstances, should not be acted upon.

Specific consideration should be given to the re-use of user input data for further development of the LR or LT tool (e.g., for training an underlying language model). If users are given complete freedom as to the types of data they can input, their data should not by default be re-used for such purposes – rather, they

should be given a possibility to opt-in (and then opt-out at a later stage, if they change their mind). In some specific applications, however, this optic can be reversed, if the balance of interests justifies it. For example, if the user input is of low sensitivity (e.g. a query history in a corpus of 17th century theatrical plays), and it can be used to significantly improve the resource or tool, the re-use should be a default, and an opt-out request should only be acted upon if it is well-justified.

2.3 Use Phase

To comply with ethical requirements, LRs and LT tools should also be used responsibly. The questions that any user should ask themself include:

1. **Is the resource or tool suitable for the envisaged use?**

This question is particularly important in the context of generative AI tools, such as Chat GPT. A responsible user should be aware of the drawbacks of AI-generated data, such as the fact that they under-represent dialects and other local specificities of a given language. Furthermore, AI-generated data, if used for training AI tools (which is not uncommon in practice, as a sizable portion of texts on the Internet is likely to be machine-generated) can only reinforce their own shortcomings and create a negative feedback loop. Finally, the use of AI-generated data comes with a risk of overlooking the 'human factor', which in certain scenarios is particularly undesirable (e.g., in tools intended for some form of emotional support).

Furthermore, it is important that the user be transparent about the use of AI-generated data.

Regarding all sorts of LR and LT outputs, the user should maintain some 'healthy scepticism', rather than blindly rely on the results. LT, even the most advanced (or: especially the most advanced) is known for occasionally 'hallucinating' (Alkaissi, McFarlane, 2023), i.e. producing a credible output that is not based on any real-world input. For example, when asked to generate a bibliography for a scientific article, Chat GPT is likely to provide references that look credible *prima facie*, but do not correspond to any real publications (Walters, Wilder, 2023). Such outputs, before they can be used, should be manually verified or cross-referenced with a credible source (e.g., Google Scholar).

2. **Do I know the conditions (terms) of use of the resource or tool?**

It is easy to lose sight of the fact that some LR and LT tools come with conditions (or terms) of use. This is the case of ChatGPT, available only via a dedicated API. These terms of use may prohibit certain uses of the tool or resource, or related outputs. For example, the Open AI's Terms of use prohibit the use of outputs from their services (like Chat GPT) to develop competing models (Open AI, 2023).

We believe that the respect of such conditions is also an ethical requirement, as they are rooted in the principle of property, even if not based directly on an existing Intellectual Property right.

2.4 Evaluation Phase

In the evaluation phase, reliability of the tool or resource, as well as continuous risk management, seem to be primary concerns. Therefore, the following questions, similar to the one asked at the conception phase, should be answered here:

1. **Who is the person or entity responsible for the tool? Has it changed since the conception phase?**
2. **Is the purpose for which the tool or resource is being or can be used different from its initial purpose?**
3. **Are the actual users of the tool the same as those for whom the tool was initially intended?**
4. **Given the answers to the questions above, what is the potential impact of the tool or resource, as it is used now, on its current users?**

All these questions reflect one idea: the context in which the tool is used can evolve, which requires a new risk assessment. Such a review should be carried out periodically.

3. Ethical Principles for LR and LT

Based on the analysis above, we would like to propose the following list of ethical principles for LR and LT:

1. **Privacy**: stakeholders should be protected against disproportionate intrusion and allowed to keep certain information secret;
2. **Property**: intellectual and cultural property should be handled with respect, in compliance with applicable law, ensuring that any potential harm (evaluated from the owner's perspective) is outweighed by collective benefit;
3. **Equality**: no group of stakeholders or contributors should be directly or indirectly discriminated against;
4. **Transparency**: LT outputs should be clearly marked as such; stakeholders should be informed about the main principles of, and given a possibility to learn the details about the functioning of LT;
5. **Freedom**: data providers should be free to contribute their data to LR&LT, and, to a reasonably practicable extent, to change their mind at any later stage; human intervention should be necessary and decisive in any process involving the use of LT the outcome of which may seriously impact the user;
6. **Accountability:** the person(s) or entity(-ies) responsible for the resource or tool at different stages of its creation should be clearly identified. The accountability should not be unnecessarily distributed across too many stakeholders;

7. **Risk Anticipation and Mitigation:** any risks related to the use or foreseeable misuse of a LR or LT tool, taking into account its actual use and actual user groups, should be anticipated and, if necessary, mitigated by employing appropriate measures;
8. **Reliability and Limited Confidence:** these principles are like two sides of one coin: a) LRs and LT tools should be built in such a way as to be fit for their intended purpose (Reliability) and b) any results produced with LRs and LT tools should be met with limited confidence and, if appropriate, verified (Limited confidence).

Principles 1-5 restate those that we proposed in our previous contribution. Principles 6-8, which are of more over-arching nature, constitute an original input of this article.

4. Conclusion

Since the last edition of the LREC conference, the debate on ethical issues affecting LRs and LT tools has intensified. Since ethical norms are not a static system, the guiding ethical principles for our field should be periodically revised, to ensure that they maintain their validity and do not become detached from the reality of the field.

In this contribution, we proposed a "checklist", a list of questions that should be examined at various stages of development of an LR or an LT tool. We do hope it will help in ethics assessments by the data providers, the developers, the users, the evaluators, and potentially even the funders. We also proposed three new guiding principles, which are not intended to replace the principles we previously proposed, but rather to reinforce them by introducing a larger, more overarching perspective on ethics. These new principles are: Accountability, Risk Anticipation and Mitigation, and Reliability and Limited Confidence.

The debate on ethics in our field is bound to continue, and we do hope that this contribution will help structure it, at the very least by proposing a common terminology.

5. Bibliographical References

Alkaissi H., McFarlane S.I. (2023). Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*. 2023 Feb 19;15(2):e35179

European Commission (2021). *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*. COM/2021/206 final.

Landi, A., Thompson, M., Giannuzzi, V., Bonifazi, F., Labastida, I., Bonino da Silva Santos, L. O., Roos, M. (2020). The "A" of FAIR – As Open as Possible, as Closed as Necessary. *Data Intelligence* 2020; 2 (1-2): 47–55.

Metz, C. and Weise, K. (2023). How 'A.I. Agents' That Roam the Internet Could One Day Replace Workers. *The New York Times*, 16 October 2023.

OECD (2019). *Recommendation of the Council on Artificial Intelligence*. OECD/LEGAL/0449.

Open AI (2023). *Terms of Use*. Updated 13 March 2023. Available at: https://openai.com/policies/terms-of-use (access: 20.10.2023)

Volpicelli, G. (2023). ChatGPT broke the EU plan to regulate AI. *Politico*, 3 March 2023.

Wagner, B. (2023). Accountability by design in technology research. *Computer Law & Security Review*, vol. 37, July 2020.

Walters, W.H., Wilder, E.I (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports* 13, 14045 (2023).

Xiang, Ch. (2023). 'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says. *Vice*, 30 March 2023.

Kamocki, P. and Witt, A. (2020). Privacy by Design and Language Resources. In *Proceedings of the LREC 2020*.

Kamocki, P. and Witt, A. (2022). Ethical Issues in Language Resources and Language Technology – Tentative Categorisation. In *Proceedings of the LREC 2022*.

Zenner, K. (2023). A law for foundation models: the EU AI Act can improve regulation for fairer competition. *OECD AI Policy Observatory*, 20 July 2023.