Practising responsibility: Ethics in NLP as a hands-on course

Malvina Nissim
University of Groningen
The Netherlands
m.nissim@rug.nl

Viviana Patti
University of Turin
Italy
viviana.patti@unito.it

Beatrice Savoldi
Fondazione Bruno Kessler
Italy
bsavoldi@fbk.eu

Abstract

As Natural Language Processing (NLP) systems become more pervasive, integrating ethical considerations into NLP education has become essential. However, this presents inherent challenges in curriculum development: the field's rapid evolution from both academia and industry, and the need to foster critical thinking beyond traditional technical training. We introduce our course on *Ethical Aspects in NLP* and our pedagogical approach, grounded in active learning through interactive sessions, hands-on activities, and "learning by teaching" methods. Over four years, the course has been refined and adapted across different institutions, educational levels, and interdisciplinary backgrounds; it has also yielded many reusable products, both in the form of teaching materials and in the form of actual educational products aimed at diverse audiences, made by the students themselves. By sharing our approach and experience, we hope to provide inspiration for educators seeking to incorporate social impact considerations into their curricula.[1]

1 Introduction

With the popularity of language technologies entering everyday life and their potential for severe societal consequences, attention to ethical aspects has massively increased in NLP research over the last few years. Best practices have emerged—e.g., data statements (Bender and Friedman, 2018) and model cards (Mitchell et al., 2019)—policies have been established—e.g., the ACL adoption of a Code of Ethics in 2020,[2] the inclusion of ethics statements in *CL publications, ethical reviews,[3] and bias statements (Hardmeier et al., 2021)—and research extensively addresses issues such as bias (Blodgett et al., 2020), dual use (Hovy and Spruit, 2016), and safety (Zhang et al., 2024). However, teaching curricula adapt at a slower pace.

> *"The emergence of this new course could be described as the culmination of an increasing public awareness in the ethical use of AI systems. For me personally, the course condensed abstract ethical thinking into crucial practical advice. [...] I was actively challenged to consider specifically how my own work and standard practices may produce some unintended and unwanted side effects. The course made ethics a very real and tangible affair."*

Figure 1: Testimonial of a BSc Information Science (2021/2022) student, Groningen.

Working with NLP involves crucial reflections on the choices we make when developing methods, models, and data, as well as the consequences of our work in terms of personal responsibility and third-party misuse, making knowledge and awareness of ethical issues a critical part of NLP education. Still, until recently the social impact of language technology was discussed in isolated lectures or seminars, with few dedicated modules.

To fill this gap, we developed the course "Ethical Aspects in Natural Language Processing" to feature in the last period of the last year in the BSc Information Science offered at the Faculty of Arts at the University of Groningen, The Netherlands, from the academic year 2021/2022 onwards.[4] Rather than treating ethics as an afterthought for experienced researchers, the course exposed students

[1] Materials available at `https://github.com/ethics-handson/teaching`.

[2] `https://www.aclweb.org/portal/content/acl-code-ethics`

[3] `https://aclrollingreview.org/ethicsreviewertutorial`

[4] `https://www.rug.nl/bachelors/information-science/`. The whole BSc programme has recently undergone some reshaping and the current version, albeit very similar, is not identical to the one from 2021/2022.

Proceedings of the Seventh Workshop on Teaching Natural Language Processing (TeachNLP 2026), pages 88–108
March 29, 2026 ©2026 Association for Computational Linguistics

W	Topics
1	Introduction to sota discussions on ethics in NLP *Social implications & values in ML/AI research*
2	NLP and language-specific challenges *Ethical practices in the ACL community*
3	Bias: scientific and ethical implications *Methods for measurement and debiasing; portability beyond English*
4	Downstream tasks and user-facing applications *Dual use; stakeholders; sensationalism*
5	Data pipeline and annotations *Data ownership; auditing and documentation; crowdsourcing*
6	Evaluation, interpretation, and reporting *Practices (e.g., leaderboardism); performance, capabilities, and reliability*

Table 1: **W**eekly breakdown of lecture topics and *seminar/assignment* topics.

early to ideas put forward by the research community and technologies entering the market. Given the multifaceted and ever-evolving nature of ethics in NLP, the course aimed to foster critical thinking, awareness, made room for open questioning and challenged unstated assumptions in the design and use of technology. This allowed us to move beyond the usual technically-oriented approach of Information Science training and avoiding stagnation on fixed notions—especially since scholarship in this area does not yet follow standardised approaches. We also placed a strong focus on the NLP practioners' responsibility over communicating ethical issues to a broader audience.

We describe our approach presenting the materials and concepts we included, and how we structured them. We also detail how the course has been adapted across different formats, editions, and audiences since its original design, placing emphasis on the hands-on activities, and in particular the final project of the course. We hope to provide inspiration for educators seeking to integrate ethical considerations into NLP curricula.

2 Course Structure

The "Ethical Aspects in Natural Language Processing" course is conceived to yield 5-6 ECTS for a total of 28-36 contact hours. It is organised to span six weeks of teaching with two modes of instruction per week: a two-hour **lecture** and a two-hour **seminar** with hands-on lab activities. Students work in groups on **weekly assignments** that provide practical experience with lecture topics and are given optional readings for in-depth coverage or complementary perspectives. The course culminates in a **final project** where students work in groups and actively engage with a variety of target audiences (e.g., experts, the general public, or targeted demographics such as school children) with the aim to consolidate the discussed materials and to learn to communicate about ethical aspects of language technology.

Assessment The weekly assignments are mandatory but not graded beyond a 0 or 1 flag. Having completed the assignments is a pre-requisite to engage with the final project, especially considering that each weekly assignment contributes to build up to the final project. The final assessment is based on the group's final presentation/product and report, and on individual reflections in free-text form which help to evaluate each student's contribution and their personal engagement with the topics. Details on instructions for students and summary rubrics are included in the Appendix, Fig. 15– 17.

Editions and Adaptations The course was developed and introduced for the first time in the academic year 2021/2022 of the BSc Information Science offered at the Faculty of Arts of the University of Groningen, and was intended for students with a general understanding of NLP and its applications. While the lecture format ensures high accessibility and lends itself to multidisciplinary audiences, we expected familiarity with basic NLP concepts and how current models work. Since then, it has featured as a stable component of the Bachelor Information Science programme, though undergoing some modifications in the contents, assignments, and final project to keep up with the fast pace of developments in the field, and to introduce novel assessment methods (§5). In the current academic year (2025/2026), it will appear in its fifth edition.

The course was also invited to feature in other programmes, namely the Master in Linguistics[5] offered at the University of Pavia, Italy (2023/2024 and 2024/2025), and the Master in Language Technologies and Digital Humanities[6] offered at the University of Turin, Italy (2023/2024 and 2024/2025). In Pavia, the course was given as a "crash course" with 36h of classes given in six

[5] https://en.unipv.it/en/education/bachelors-and-masters-degree-programs/second-cycle-degree-course/theoretical-and-applied-linguistics-linguistics-and-modern-languages

[6] https://en.unito.it/ugov/degree/41992

days (three hours in the morning and three hours in the afternoon, with lectures and labs respectively). In Turin, the materials were integrated in a broader course, with two classes of three hours given each week, over a total of six weeks. The Turin edition also featured a section on ethics where a philosopher co-teaching the course explored central ethical concepts and theoretical approaches which would be picked up in the more NLP-focused classes.

Thanks to its flexible structure, and the rather open-ended nature of the final project, adaptations were easy and contents were kept more or less stable, though some additional technical background on language modelling had to be included both for the Pavia and Turin courses, due to the less technical background of the master students there compared to the BSc students in Groningen.

Below, Section 3 describes the main rationale and contents of the course (§3.1) and its materials (§3.2). Section 4 focuses on the hands-on activities and weekly assignments. Section 5 details the final projects and their evolution across editions.

3 Course Overview

3.1 Contents

In designing the course, we aimed to cover the entire pipeline of NLP model development, deployment as well as broader considerations around research practices and reporting (see Table 1). The course progresses from general concerns common to several AI-adjacent disciplines (e.g., implications of AI/Machine Learning research more broadly, dual nature of technology) to specific challenges posed by NLP technologies (e.g., sociodemographic language variation, English-dominance). We discussed the impact of different NLP applications and products in real-world contexts—such as generative tasks, emotion and hate speech detection, and machine translation—while foregrounding the implications and cascaded effects of choices made when developing methods, models, and data for language processing.

It was important for us to have students develop the skills to reflect on the design choices of others as well as their own. We therefore also dedicated attention to how researchers evaluate and report their results and technologies, including a focus on evaluation, interpretation, and reporting practices (Week 6 in Table 1). We wanted to prepare students for whatever roles they might pursue—whether as NLP practitioners or as researchers—while also having them examine how the community itself has been grappling with new policies and ethical guidelines (Week 2 in Table 1).

The lectures alternated between instructors presenting key concepts and research findings with highly interactive moments that allow students to develop their own opinions and raise doubts. In this way, we aimed to cultivate students' curiosity through active engagement. For example, to trigger deeper reflections on what *data* is (Gitelman, 2013) and the criticalities of "data ownership" (Bird, 2020; Hao, 2022), we had a student type down the conversations happening in class during a lecture. Data represent the backbone of current NLP technologies and are typically considered a given—a true, unmediated representation of reality. After the transcription activity, students were asked: Can such transcripts be considered *data*? If so, who do they belong to—the utterer, the typist, or the teachers who requested the recording? Did the typist include everything that was said? Students quickly notice that the typist occasionally inserted line breaks, exclamation marks, and made personal textual choices regarding spelling. Were these choices neutral? Were they aligned with the communicative intent of the original speaker?

Through this exercise, students realised the multitude of unstated choices involved in selecting, filtering, and transforming data into machine-readable text (Gururangan et al., 2022; Luccioni and Viviano, 2021; Rogers, 2021). Crucially, they recognised the pervasive role of people throughout this process—from those who originally produce language to those who process it—and how these individuals, despite their fundamental contributions, often become invisible actors who disappear from NLP pipelines (Geiger et al., 2020; Hao and Seetharaman, 2023), along with concerns about their privacy and the use of their content (Williams et al., 2017; Fiesler and Proferes, 2018).

3.2 Materials

Since no main reference book on ethics in NLP exists, we assembled course materials from a range of sources beyond scientific literature. Social and ethical reflections are relatively new in NLP, and the emerging scholarship does not engage with them in agreed-upon ways. Accessing diverse perspectives is essential for promoting critical awareness, so our materials ranged from academic articles to journalistic pieces, blog posts, podcasts, interviews, documentaries, and even Netflix series.

This diversity served multiple purposes. First, the NLP field moves at such high speed that many discussions occur on platforms beyond traditional academic venues. For example, investigative journalism exposes cutting-edge criticalities of available applications and sensitive tasks (e.g., Bloomberg's investigation on ChatGPT's racial bias in CV screening)[7]. Also, Bluesky and X have become a major platform for hosting discussions on NLP ethics led by prominent researchers.

Second, we aimed to engage students' curiosity across disciplines and expose them to different types of reporting accessible to lay audiences. For example, we included Netflix's *History of Swear Words*[8] series to discuss language appropriation and reclamation (Cervone et al., 2021), hate speech detection (Zsisku et al., 2024), and the potential further marginalisation of the very communities who reclaim these terms by filtering slurs from datasets. The episodes featured rappers and stand-up comedians discussing nuances of language meaning, use, and value—perspectives often absent from technical discussions. We also incorporated documentaries such as *Coded Bias*[9] to examine algorithmic discrimination and *The Social Dilemma*[10] to discuss privacy and industry interests more broadly. These materials bridged technical and social perspectives, kept students updated and exposed them to communication practices for a broad audience, which was particularly formative towards the final projects (§5).

4 Weekly Assignments

While lectures were more front-facing and information-dense, seminars provided hands-on experiences complemented by weekly assignments students completed at home. These laboratories fostered the incorporation of theoretical knowledge into research and experimental practices through direct engagement with course topics.

For instance, in Week 2 (Table 1), students received an assignment on community practices. They read the Ethics Statements of self-selected published papers in the ACL anthology, and verified if they satisfied the Responsible NLP Research Checklist.[11] In class, they reflected on what these measures meant for the field and were encouraged to discuss their position about this.

Concerning data and annotation practices, students also carried out first-hand annotation tasks to infer emotions from text as an assignment. Through this exercise, they confronted fundamental questions: Could they agree on the emotions encountered? Was emotion detection even feasible? Were they accounting for cultural differences in interpreting emotions? Was the task grounded in scientifically valid theory? In this way, students were directly exposed to controversial questions and pitfalls in NLP task design, the inherent nuances of so-called gold standard data, and the caution to be applied when analysing evaluation outcomes (Blodgett et al., 2021; Delobelle et al., 2024).

Another assignment addressed how the perception of NLP capabilities and dangers is intrinsically linked to their reporting and presentation. Inspired by work on responsible NLP communication (Bender and Koller, 2020),[12] students put a highly debated NLP topic—potentially at the centre of media attention—into perspective, unpacked different takes on the issue, and presented their view with appropriate explanations and criticism to make it understandable for non-experts.[13] This lab exposed students to the current phenomenon of over-hyping NLP tools with sensational claims, putting them in the shoes of the lay public. We also explored the less frequent phenomenon of *under-claiming* (Bowman, 2022) and discussions on what qualifies as harmful versus undesirable (Blodgett, 2021).

Each assignment also contains a preparatory part for the final project, so that every activity contributes to build up towards it (§5). The assignments are mandatory but not graded. Feedback is provided both on the assignment itself and on the final project preparation.

5 Final Projects

A standard exam with multiple choice questions or even a written essay did not seem appropriate tools to assess the students; but most of all, these

[7] https://www.bloomberg.com/graphics/2023-generative-ai-bias/

[8] https://en.wikipedia.org/wiki/History_of_Swear_Words

[9] https://www.imdb.com/it/title/tt11394170/?reasonForLanguagePrompt=browser_header_mismatch

[10] https://thesocialdilemma.com/

[11] https://aclrollingreview.org/static/responsibleNLPresearch.pdf

[12] We drew inspiration from https://faculty.washington.edu/ebender/2021_575/scicomm.html and https://ryan.georgi.cc/courses/575-ethics-win-19/scicomm-assignment/

[13] For instance, to revisit the Delphi debate through both its original presentation (Jiang et al., 2021, 2025) and subsequent critique (Talat et al., 2021).

assessment methods would not provide the best opportunity for the students to consolidate the notions and reflections developed throughout the course.

Therefore, with a more *active learning* approach in mind, we devised a final project which would make them actors and encourage them to put into practice and further reflect on the concepts learnt during the course. The specific final project changed in the course of the various editions, both for practical reasons as well as for us to experiment with different learning strategies and outcomes.

5.1 Interview with Experts

In the very first edition of the course we leveraged our own network, and in particular the fact that one of the authors was a member of the at-the-time newly established ACL Ethics Committee[14]. We arranged for the students to run interviews with experienced NLP researchers, most of which were members of the ACL Ethics Committee.

Each group was assigned a different interviewee, with the meetings planned online adapting to the times of the experts who were based in North and South America, Europe, and Asia. These are not only experts, but also rather senior and well known researchers, making some of the students at the same time excited and a little nervous about interviewing them. Thus, the preparation was thorough! Question development was integrated in the weekly assignments from Week 1: each week students created 2-3 questions based on the newly introduced topics, so that the whole interview skeleton was ready well in advance. The questions were revised and tested with teachers and fellow students in multiple iterations until a satisfactory version was reached before the interview was due. Within each group, students decided who would ask what, and who would take notes. As a final report, they wrote the whole interview as it happened, supplemented with their own comments. Each student also wrote a short individual reflection highlighting what they gained from the experience. The Appendix includes a sample template interview prepared by one group[15] (each group prepared a different one according to preferences, ideas, and the specific interviewee they had been assigned.)

This experience brought a twofold advantage. On the one hand, it exposed the students to practices that the NLP community has undertaken towards more responsible research and in particular the creation of an ACL ethics committee, its purpose, and its functioning. On the other hand, it offered them the opportunity to discuss topics with experts other than their two teachers. We wanted students to develop a genuine interest for ethical NLP, and to be free to ask critical questions they had a keen interest in. Intended to foster student's curiosity and willingness to keep on nurturing the reflections started during the course, this final project was also more suitable given the subjectivity and highly nuanced nature of the topic. Everything we discussed in class was necessarily mediated by our own perspectives. While we tried to provide students with several diversified pointers, it is impossible to escape your own biases and personal perspective. By interacting with experts directly, students could revisit some of the aspects discussed in class, giving more space to their own take, and having access to the opinion of somebody else who is active in the field by means of a very stimulating experience.

Figure 2: Students presenting in schools in the region, Groningen edition 2022/2023: Björn Overbeek, Carmen Reker, Dennis van Thulden, Oscar Zwagers, Louis Speelman, Hessel Eekhof, Taede Meijer, Nathalie de Palm, Sijbren van Vaals, Indy van Boven, Sander Beyen, Martijn Prikken, Eva Dyadko, Jurriën Steegman, Harmen Vogt.

5.2 Presentations for High School Students

In the second and third editions of the course we experimented with a different kind of final project, mostly driven by current developments and by cu-

[14] https://www.aclweb.org/adminwiki/index.php/Formation_of_the_ACL_Ethics_Committee.

[15] Credits: Patrick Darwinkel,Ties Leneman,Jordy Loomans

riosity over alternative learning strategies (and also because, as kind as they can be, colleagues' availability for interviews cannot be taken for granted!) The release of ChatGPT in late 2022 brought to the foreground the importance of communicating about ethical aspects of language-based AI technologies to a general audience, and made it even more pressing in the context of our educational purview: as technology experts and future practitioners in the field, our students must embrace the responsibility of contributing to literacy and awareness in the use of language-based AI tools.

The new final project, still to be carried out in small groups, therefore focused on *communicating* and educating others about basic workings of language technology, and aspects they had learned and reflected upon during the course; besides the retention of knowledge and the development of argumentation on ethical concepts, learning how to effectively communicate the impact of language technology on society is a core objective of the course. We identified *high school students* as an excellent target audience, and selected in particular classes in their penultimate year of high school.

Leveraging our local network and previous collaborations with high schools in the region, we arranged the sessions by getting in touch with school teachers, explaining the reasoning behind the experience we were proposing for the pupils, and in most cases had preliminary meetings with the school teachers. In addition, we asked our own students whether they would be interested in doing their presentation in their former school, should that be logistically feasible. A couple of groups in both the 2022/2023 and 2023/2024 editions chose to do this, and organised the logistics themselves, in collaboration with us. In both years, for grading purposes but also for being present in case anything strange would happen or would be said, we attended all live presentations in all schools, which in some cases also meant quite some travel across the region! Prior to the actual presentations, we had a joint session in class with all groups giving mock presentations to us and each other. This allowed us to verify that all information conveyed in the presentation was correct and that sensitive issues would be treated sensibly; it also served as a testbed for the interactive parts which were included in the presentations, such as quizzes and live polls.

Overall, this turned out to be an exceptional experience for our students. By conveying potentially sensitive information to younger individuals, who are avid users of language technology but may not yet fully grasp how it works or the implications of its use, students had to engage in deeper reflection on the course materials and topics discussed.

From an educational standpoint, having students prepare and deliver materials to real audiences, the pedagogical method broadly known as "learning by teaching", is conducive to the *protégé effect*. This is a commonly described phenomenon in psychology, whereby learners understand and remember concepts better when they teach them to someone else, especially to a younger audience (Bargh and Schul, 1980; Benware and Deci, 1984). Indeed, presenting to high school students served the twofold purpose of helping younger people to interact more responsibly with language technology and instilling in our students a sense of responsibility as practitioners. One student said: "*I eventually learned more from preparing this presentation than I have during the three years of my bachelor's; I revised everything, as I felt I could not risk being unprepared in front of the students, especially in the presence of my former computer science teacher!"*[16]

Variations This form of final project was used also in the Pavia and Turin 2023/2024 courses, with two variations. One is the target audience: we left the groups free to choose which school years they'd like to target, thus preparing materials accordingly. The other one is the modality of the presentation delivery: mostly due to time constraints (the course was condensed in just a few days or a few weeks), it was not possible to organise actual visits to schools for the students, so presentations were given in class only in a mock, though complete, form. Pedagogically, this is still a valid route, since even in absence of the actual teaching experience, just *preparing to teach* has been shown to yield very positive learning effects in the students, superior to simply studying the materials (Fiorella and Mayer, 2013). Two illustrative slides from two presentations are shown in Fig. 9 and 10.

5.3 Educational outreach products

In the fourth edition of the course (2024/2025), we renewed the concept of the final project once more. Students worked in groups to create outreach materials that could be used to raise awareness of ethical issues when using language technology. Each group was free to choose a target audience and the relative product to develop. With this new setup

[16] Reported from an informal conversation with MN.

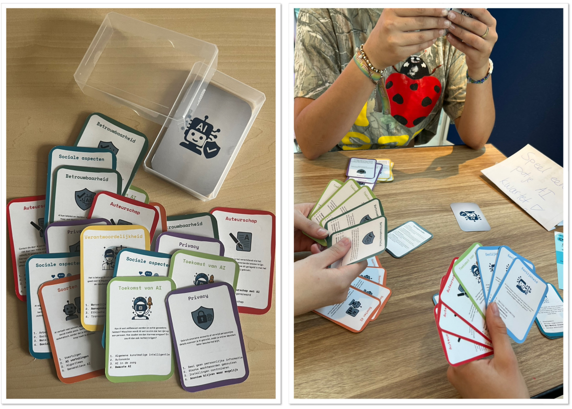


Figure 3: Quartet game, Groningen edition 2024/2025. Right: game used by high school students at a European Researchers' Night event, Groningen, Sept. 2025. Credits: Shaya Bhailal, Jelmer Smit, Matthijs ten Hove.

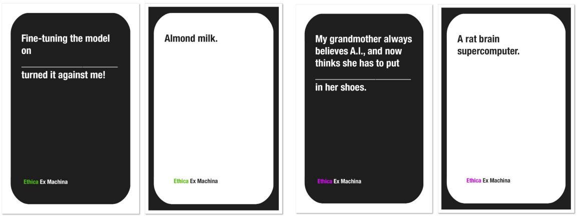
students could be even more creative and in charge of their choice, thus more invested, and the created materials could be used more than once. We provide here some example choices by the students from the classes that were taught in the academic year 2024/2025 in Groningen, Turin, and Pavia. A team in Turin, inspired by the activities of the previous year, chose to present at a high school, organising all of the logistics themselves (Fig. 11).

5.3.1 Card games

One of the popular products developed by students was card games, of different sorts. Inspired by existing card game mechanics, three groups independently came up with an "Ethics in AI" game.

Quartet This is a classic card game for four players, played with a deck organised into (usually eight) sets of four related cards (*quartets*). Players take turns requesting a specific card from another player with the aim of completing a quartet. If the requested player has the card, they must give it to the requester, who may then continue their turn. Otherwise, the turn passes on. When a player collects all four cards from the same set, the quartet is revealed and counted as a point. The game ends when all quartets have been completed, and the player with the most quartets wins. The educational component here is the use of quartet themes which are relevant for ethics in NLP, such as privacy, bias, responsibility, future, etc. The game is very easy to play, and also equipped with some instruction and theme explanation cards (Fig. 3).

Ethica ex Machina Targeting young adults, and inspired by the popular game "Cards Against Humanity", students developed this fill-in-the-blank game with intriguing prompts and weird responses to trigger conversation about AI. Players try to make the funniest sentence by combining two types of cards: prompt cards and response cards. Prompt cards are coloured black and define the base "sentence" that will be used for the round. These either include a blank space to be filled in with a response card (e.g., "My facial recognition software thinks I'm a ___!", or are questions to be answered with a response card, such as "What data took down Gemini?" Response cards are white and are used to answer the prompt cards, either by filling in their blank space or by being an answer to their question. Response cards include rather random phrases such as "Will Smith eating spaghetti" and "Outsourcing to Italian AI" (Fig. 4).



Fine-tuning the model on ___ turned it against me!

Ethica Ex Machina

Almond milk.

Ethica Ex Machina

My grandmother always believes A.I., and now thinks she has to put ___ in her shoes.

Ethica Ex Machina

A rat brain supercomputer.

Ethica Ex Machina

Figure 4: Ethica ex Machina, Groningen edition 2024/2025. Example prompt (black) and response (white) cards. The colour at the bottom (violet/green) refers to one of the four topics. Green is Bias, violet is Hype. Credits: Jessay Beukema, Merel Hemstede, Niek Holter, Cody van der Deen, Sofia van der Wal.

The cards are designed on the themes: Data, Bias, Danger, and Hype. The playing cards are complemented by a *user manual* with general information about the game and instructions for setup and play; an *explainer* with background information on the featured topics and used categories as well as a disclaimer to clarify that the game serves as a simplified, playful introduction to AI topics and should not be considered an authoritative source; and a *term glossary* containing explanations for all the relevant terms and references used in the game.

To motivate the design and mechanics of the game the students wrote in their report: "*We don't have all the answers, but maybe we can grow together by asking the right questions*," which we found very inspiring.

Debatable This is also inspired by an existing discussion-based party game where players are given a question or statement and must argue for assigned or chosen positions, regardless of their

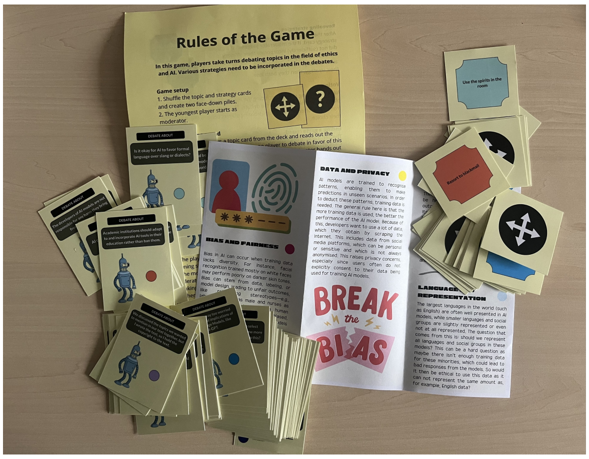

Rules of the Game
In this game, players take turns debating topics in the field of ethics and AI. Various strategies need to be incorporated in the debates.
DATA AND PRIVACY
BIAS AND FAIRNESS
BREAK the BIAS
LANGUAGE REPRESENTATION

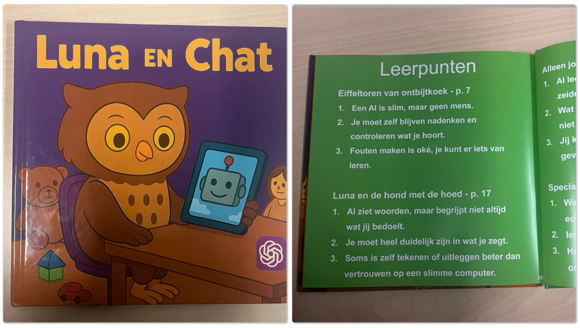
Figure 5: Debatable card game, with instructions and reference leaflet. Groningen edition 2024/2025. Credits: Ilse Kerkhove, Dertje Roggeveen Marieke Schelhaas, Mijke van Daal, Nikki van Gurp.

personal beliefs. Players take turns presenting arguments, responding to others, and attempting to persuade the group. After the discussion, a vote determines which argument was most convincing. The game ends after a set number of prompts. Again, the discussion prompts are organised around the topics discussed in class (bias and fairness, responsibility, data, etc), which are colour-marked on the cards and explained in an accompanying information leaflet (Fig. 5; some cards are visible also in Fig. 12 in the Appendix). The game's intrinsic emphasis on rhetoric and interaction rather than factual correctness makes it suitable for Ethics in NLP, where there are few definite truths and much to gain from a plurality of views.

In all three games, the card themes are based on the course's topics, underscoring how the lectures and labs guided the students' reflection and their product development. The cards were also eventually printed to make the final product more concrete, also in line with the idea of creating educational and outreach materials which can be re-usable. For example, the quartet game was made available during an activity run by GroNLP, the Groningen Natural Language Processing group[17], in the context of the 2025 edition of European Researchers' Night[18].

5.3.2 Other products

Illustrated book Aimed at primary school children, this is an illustrated book featuring a 9 year old girl, Luna, who uses a tablet with a chatbot in it (Fig 6). Through a series of chapters on the various topics discussed in class, risks and advantages of using chatbots are discussed in very simple language and with images. At the end, there are also guidelines for teachers on how to use to book and how to talk about this topic with kids. Developed by Groningen students, the book is in Dutch.



Leerpunten

Eiffeltoren van ontbijtkoek - p. 7

1. Een AI is slim, maar geen mens.
2. Je moet zelf blijven nadenken en controleren wat je hoort.
3. Fouten maken is oké, je kunt er iets van leren.

Luna en de hond met de hoed - p. 17

1. AI ziet woorden, maar begrijpt niet altijd wat jij bedoelt.
2. Je moet heel duidelijk zijn in wat je zegt.
3. Soms is zelf tekenen of uitleggen beter dan vertrouwen op een slimme computer.

Figure 6: Illustrated book for primary school children, Groningen 2024/2025. Left: cover; right: some explanatory points. Credits: Jani de Bruijn, Dies de Haan, Manon Kooning, Joos Oving, Thomas Thiescheffer.

Podcast episodes Two episodes of a podcast aimed at young adults and university students ("Zuckerberg and ethics"[19]). The focus is on the ethical dimensions of Meta's use of user data for AI training, and aims to both inform and raise critical reflection on current AI and data protection issues. The first episode focuses on technical explanations, and in particular on how Meta intends to use user data for model training, data types, training processes (pre-training and fine-tuning), and the mechanics of the opt-out functionality. The second episode zooms in on ethical considerations: consent models, privacy harms, power imbalances, regulatory context (GDPR and EU AI Act), and proposals for more equitable data governance. In the podcast, students play different expert roles with different attitudes and backgrounds (technical, ethical, legal, economic), with a plurality of viewpoints emerging in the discussions.

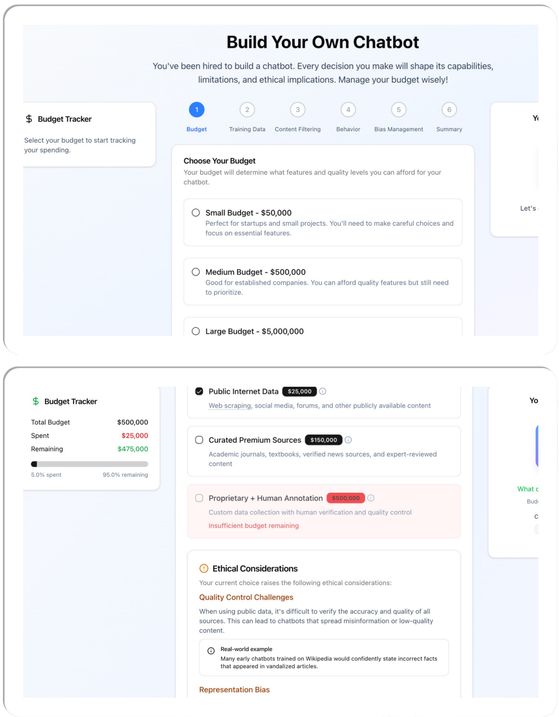
Interactive demo An interactive web-based experience called "Build Your Own Chatbot"[20]. The tool simulates the process of building a chatbot and consists of several progressive budget-dependent choices developers and researchers have to deal

[17] https://www.rug.nl/research/clcg/research/cl/

[18] https://forum.nl/en/whats-on/europese-nacht-van-de-onderzoekers

[19] https://open.spotify.com/show/1Z28HnWoV1ssdAzsKZci0O. Credits: Niek Biesterbos, Pascal Boon, Mark den Ouden, Isa Houtsma, Armen Poghosow.

[20] https://createchatbot.vercel.app/chatbot



Build Your Own Chatbot

You've been hired to build a chatbot. Every decision you make will shape its capabilities, limitations, and ethical implications. Manage your budget wisely!

$ Budget Tracker

Select your budget to start tracking your spending.

1 Budget · 2 Training Data · 3 Content Filtering · 4 Behavior · 5 Bias Management · 6 Summary

Choose Your Budget

Your budget will determine what features and quality levels you can afford for your chatbot.

○ Small Budget - $50,000
Perfect for startups and small projects. You'll need to make careful choices and focus on essential features.

○ Medium Budget - $500,000
Good for established companies. You can afford quality features but still need to prioritize.

○ Large Budget - $5,000,000

$ Budget Tracker

Total Budget	$500,000
Spent	$25,000
Remaining	$475,000

5.0% spent 95.0% remaining

☑ Public Internet Data $25,000
Web scraping, social media, forums, and other publicly available content

☐ Curated Premium Sources $150,000
Academic journals, textbooks, verified news sources, and expert-reviewed content

☐ Proprietary + Human Annotation $500,000
Custom data collection with human verification and quality control
Insufficient budget remaining

Ethical Considerations

Your current choice raises the following ethical considerations:

Quality Control Challenges

When using public data, it's difficult to verify the accuracy and quality of all sources. This can lead to chatbots that spread misinformation or low-quality content.

Real-world example
Many early chatbots trained on Wikipedia would confidently state incorrect facts that appeared in vandalized articles.

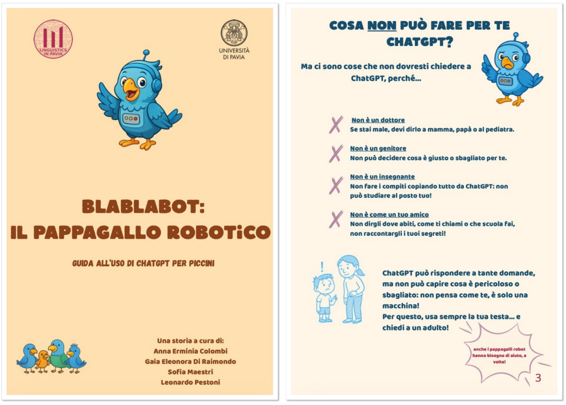
Representation Bias

Figure 7: Interactive chatbot, Groningen edition 2024/2025. Top: starting page with budget choice; bottom: data selection for training and warning associated with low-budget data choices. The budget gets progressively updated (see left). Credits: Marco Boasso, Kylian de Rooij, Emiel Dost, Andrew Geddes, Stijn Schreven.

with when training a model. Choices that must be made include budget, data sources, content filtering, AI behaviour and debiasing (Fig. 7.) The website is very user friendly and thus an accessible educational source to be used by high-schoolers or university students; it can also appeal to teachers and AI content creators as an engaging educational tool to introduce discussions about AI ethics.

Informative Leaflets Two groups from two different editions independently chose to develop informative leaflets: one aimed at primary school children (Pavia, Fig. 8) and one aimed at people in care homes (Turin, Fig. 13). In both cases the leaflets are aimed at explaining what language technology can and cannot be used for, and the advantages and risks associated with it.

Surveys and Interviews with Laypersons One group for the Turin edition created a website to collect perceptions of laypersons on the influence of AI in different fields such as education, ecology, and art. They combined surveys with insights obtained from street interviews, to inform the creation of materials to raise awareness. They designed two kinds of surveys: one for a general public, and one tailored to students of the art faculties, with specific questions related to using AI to create art. Street interviews involved mainly students around the Campus Luigi Einaudi in Turin (Fig. 14).



BLABLABOT:
IL PAPPAGALLO ROBOTiCO

GUIDA ALL'USO DI CHATGPT PER PICCINI

Una storia a cura di:
Anna Erminia Colombi
Gaia Eleonora Di Raimondo
Sofia Maestri
Leonardo Pestoni

COSA NON PUÒ FARE PER TE CHATGPT?

Ma ci sono cose che non dovresti chiedere a ChatGPT, perché...

✗ Non è un dottore
Se stai male, devi dirlo a mamma, papà o al pediatra.

✗ Non è un genitore
Non può decidere cosa è giusto o sbagliato per te.

✗ Non è un insegnante
Non fare i compiti copiando tutto da ChatGPT: non può studiare al posto tuo!

✗ Non è come un tuo amico
Non dirgli dove abiti, come ti chiami o che scuola fai, non raccontargli i tuoi segreti!

ChatGPT può rispondere a tante domande, ma non può capire cosa è pericoloso o sbagliato: non pensa come te, è solo una macchina!
Per questo, usa sempre la tua testa... e chiedi a un adulto!

Figure 8: Leaflet for children, Pavia 2024/2025. Left: cover; Right: one central page on how *not* to use ChatGPT. Credits: Anna Erminia Colombi, Gaia Eleonora Di Raimondo, Sofia Maestri, Leonardo Pestoni.

6 Conclusions

We developed an Ethics in NLP course, which we offered in different institutions to students of different backgrounds and levels in the last four years. The key aspects of this course are a *plurality of perspectives* and a *hands-on approach*, where students become actors of communication with a variety of strategies, and to a variety of audiences. These include presenting to high school students and creating outreach materials, with the target goal of increasing awareness and responsibility in the students themselves in addition to the recipients of the interventions they had to perform for the course.

The outputs and the students' comments speak to a great success of this course. We hope to inspire and help others by sharing this experience and all associated materials.

> *"The course offered tools to move beyond polarisations through discussion in a space focused more on asking questions than giving answers, where the aim was not to take sides but to dig into the topic and probe it with a critical lens."*
>
> A student in MA Linguistics (Pavia) [originally in Italian, translation is ours]

7 Limitations

While we hope that our experience can inspire others to embark on a similar adventure, we also care to share some of the challenges that this approach carries and that should be taken into account. We report them here, and point to Table 2 in the Appendix for a schematised visualisation of pros and cons for the various stakeholders involved in the final project with high-school presentations.

Logistical complexity and resource intensity The course requires significant organisational effort from instructors, particularly for formats involving external stakeholders. This includes coordinating with colleagues for interviews and arranging presentations with multiple high schools. Such logistical demands may limit the course's scalability and sustainability, especially in institutions with constrained faculty resources or less developed professional networks.

Assessment The shift from traditional examinations to interviews, presentations, and diverse creative outputs (card games, podcasts, interactive demos, illustrated books, leaflets) introduces a margin of subjectivity which requires some redefinition of grading. We found that providing written guidance and discussing in advance what is expected of their outputs helps both the teachers and the students to assign and interpret the final grades.

Financial sustainability The production of tangible final project materials (e.g., printed card games, illustrated books, leaflets) requires some budget to sustain the costs of producing them.

Language constraints The high school presentation format might present language barriers that may limit accessibility. This might be true for international students who do not speak the local language well.

Acknowledgments

We wholeheartedly thank all of the students who attended our classes in Groningen, Pavia, and Turin. They made this experience invaluable, and through their creations and feedback made this course a concrete and wonderful experience. They also created materials which we can use in activities with the general public and target audiences, and gave us a lot of ideas. We have mentioned many of you, surely all those whose outputs are explicitly included in the paper, but our thanks extend to absolutely everyone. Special thanks to Groningen TAs Patrick Darwinkel and Riemmelth Ruitenbeek.

We cannot thank enough the colleagues from ACL who lent themselves as interviewees for the Groningen 2021/2022 edition: Tim Baldwin, Luciana Benotti, Karën Fort, Dirk Hovy, Min-Yen Kan, Saif Mohammad, Xanda Schofield, and Yulia Tsvetkov. It wasn't obvious you would agree to do this, and it was such an amazing experience for the students — you are the best!

We are particularly grateful to the following schools in the Groningen area: ISG, Maartenscollege, Willem Lodewijk Gymnasium, Praedinius Gymnasium, Gomarus college, Stellingwerf college, CS Vincent van Gogh Lariks, and to the Liceo Classico e Linguistico "Vincenzo Gioberti" in Turin. We could not have offered to the students the experience of giving presentations in high schools without the support of school teachers, including: Matteo Saudino (Liceo Classico V. Gioberti), Louis Hanson, Victoria Ingram, Marion Kuijpers, Karen Lewis, Lydi Tillema, and Mike Weston (ISG/Maartenscollege), Katherine Gardiner and Vincent Bekkering (Praedinius), Marije van der Kooij, Rob Nagtegaal and Betty Pluim (Willem Lodewijk), Roel Stamhuis (Gomarus college), Erik Meek (CS Vincent van Gogh Lariks).

Teaching in person in Pavia and Turin was made possible by fellowships awarded to MN in the context of two international excellence programmes: Collegio Ghislieri's programme "L'università nei Collegi"[21] in Pavia, and the University of Turin's "Visiting Professors" programme[22]. We are particularly grateful to Chiara Zanchi for encouraging and facilitating this process in Pavia. Thanks also go to Vincenzo Crupi for co-teaching the course in Turin, contributing the philosophical framework underlying the ethical concepts, to Otje Minnema for having given three lectures over different academic years in the Pavia and Turin editions, and Alan Ramponi for teaching one class in Turin.

The concrete realisation of the final products (e.g., printing) was in large part sponsored by the Sectorplan theme "Humane AI" financed by the Dutch funding body NWO, which also partially supports MN's work.

[21] https://www.ghislieri.it/progetto-universita-nei-collegi/

[22] https://en.unito.it/international-relations/teachers-and-researchers-mobility/visiting-professors

References

John A Bargh and Yaacov Schul. 1980. On the cognitive benefits of teaching. *Journal of educational psychology*, 72(5):593.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Carl A Benware and Edward L Deci. 1984. Quality of learning with an active versus passive motivational set. *American educational research journal*, 21(4):755–765.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Su Lin Blodgett. 2021. *Sociolinguistically Driven Approaches for Just Natural Language Processing*. Doctoral dissertation, University of Massachusetts Amherst. UMass Amherst Doctoral Dissertations 2092.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Samuel Bowman. 2022. The dangers of underclaiming: Reasons for caution when reporting how NLP systems fail. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7484–7499, Dublin, Ireland. Association for Computational Linguistics.

Carmen Cervone, Martha Augoustinos, and Anne Maass. 2021. The language of derogation and hate: Functions, consequences, and reappropriation. *Journal of language and social psychology*, 40(1):80–101.

Pieter Delobelle, Giuseppe Attanasio, Debora Nozza, Su Lin Blodgett, and Zeerak Talat. 2024. Metrics for what, metrics for whom: Assessing actionability of bias evaluation metrics in NLP. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21669–21691, Miami, Florida, USA. Association for Computational Linguistics.

Casey Fiesler and Nicholas Proferes. 2018. "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society*, 4(1).

Logan Fiorella and Richard E Mayer. 2013. The relative benefits of learning by teaching and teaching expectancy. *Contemporary Educational Psychology*, 38(4):281–288.

R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 325–336, New York, NY, USA. Association for Computing Machinery.

Lisa Gitelman. 2013. *Raw data is an oxymoron*. MIT press.

Suchin Gururangan, Dallas Card, Sarah Dreier, Emily Gade, Leroy Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A. Smith. 2022. Whose Language Counts as High Quality? Measuring Language Ideologies in Text Data Selection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2562–2580, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Karen Hao. 2022. A new vision of artificial intelligence for the people. *MIT Technology Review*. Accessed: 2024-12-19.

Karen Hao and Deepa Seetharaman. 2023. Cleaning up ChatGPT takes heavy toll on human workers. *The Wall Street Journal*.

Christian Hardmeier, Marta R Costa-jussà, Kellie Webster, Will Radford, and Su Lin Blodgett. 2021. How to write a bias statement: Recommendations for submissions to the workshop on gender bias in NLP. *arXiv preprint arXiv:2104.03026*.

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: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, Yulia Tsvetkov, Oren Etzioni, Maarten Sap, Regina Rini, and Yejin Choi. 2025. Investigating machine moral judgement through the Delphi experiment. *Nature Machine Intelligence*, 7:145–160.

Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Jon Borchardt, Saadia Gabriel, et al. 2021. Can machines learn morality? the delphi experiment. *arXiv preprint arXiv:2110.07574*.

Alexandra Luccioni and Joseph Viviano. 2021. What's in the Box? An Analysis of Undesirable Content in the Common Crawl Corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 182–189, Online. Association for Computational Linguistics.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA. Association for Computing Machinery.

Anna Rogers. 2021. Changing the world by changing the data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2182–2194, Online. Association for Computational Linguistics.

Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2021. A Word on Machine Ethics: A Response to Jiang et al. (2021).

Matthew L. Williams, Pete Burnap, and Luke Sloan. 2017. Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology*, 51(6):1149–1168.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2024. SafetyBench: Evaluating the Safety of Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15537–15553, Bangkok, Thailand. Association for Computational Linguistics.

Eszter Zsisku, Arkaitz Zubiaga, and Haim Dubossarsky. 2024. Hate speech detection and reclaimed language: Mitigating false positives and compounded discrimination. In *Proceedings of the 16th ACM Web Science Conference*, WEBSCI '24, page 241–249, New York, NY, USA. Association for Computing Machinery.

A Appendix

We include here some additional pictures of the students' products and presentations as well as some comments from them on the experience of the course itself. These sample testimonials are quotes extracted from the individual reflections that the students had to submit at the end of the course, in addition to the group report.

This Appendix also includes information on assessment, a table summarising challenges connected with this course, and an example template for the interviews that students held with experts in the Groningen 2021/2022 edition.

> *"I found "Ethics in NLP" one of the most significant courses of the degree, and it had a concrete impact on how I perceive the topics that I'm studying. Moreover, I noticed that I started reading more online articles about AI from a more critical perspective, and I found myself debating more with my classmates but also at home, with my family, about the use of AI."*
>
> A student in MA Language Technologies and Digital Humanities (Turin)

> *"We were thinking more black and white, but it made us really think beyond performance. Performance isn't just accuracy, it relates to the impact tools have on people."*
>
> A student in BSc Information Science (Groningen)

Stakeholder	Challenges	Opportunities
University students	• Talking in front of school kids • Finding simpler ways to explain tech concepts • Feeling unprepared • Schoolers misbehaving	• Revising everything • Finding simpler ways to explain tech concepts • Feeling relevant
School students	• Receiving instruction from non-professional teachers	• Stronger resonance • Exposure to a field underrepresented in the high school curriculum
Lecturers	• Planning logistics • Attending logistics • Ensuring students are well prepared • Rubric and Assessment	• Awe and satisfaction • Alternative viewpoints • Contact with schools

Table 2: Presentations in schools as final project: challenges and opportunities for the stakeholders involved.

Figure 9: “Captain America discovers technology”. Presentation aimed at primary school kids. Captain America wakes up after 100 years, in Italy, and must face all recent technological developments (without speaking Italian!). Pavia edition 2023/2024. Credits: Vincenzina Cacchione, Giulia Tassi, Aurora Zuin.

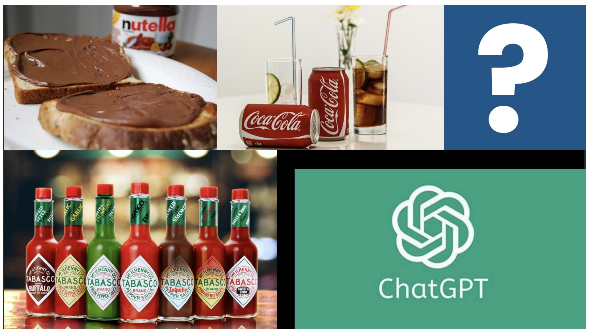


Figure 10: What do these products have in common? The presence of a secret ingredient! Slide in a presentation aimed at high school students. The focus here is on the lack of transparency over the training details of large closed models. Pavia edition 2023/2024. Credits: Matteo Gay, Lorenzo Reina, Anna Vignoli.

Figure 11: “Conversations about AI: Ethics and Artificial Intelligence”. Presentation for high school students during the philosophy class. Turin edition 2024/2025. Credits: Emanuele Belloni, Monica Bongiorni, Arianna Denitto, Alina Jill Simeone.

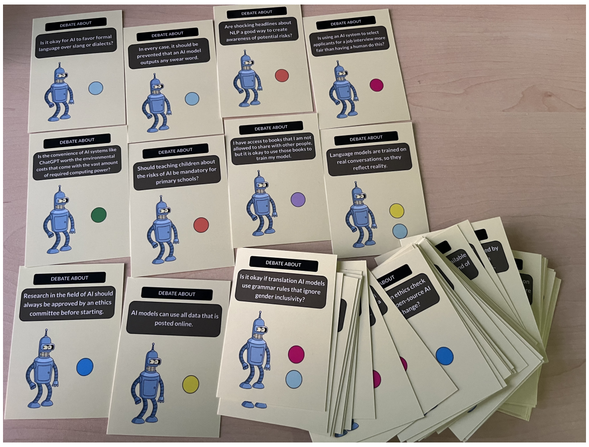
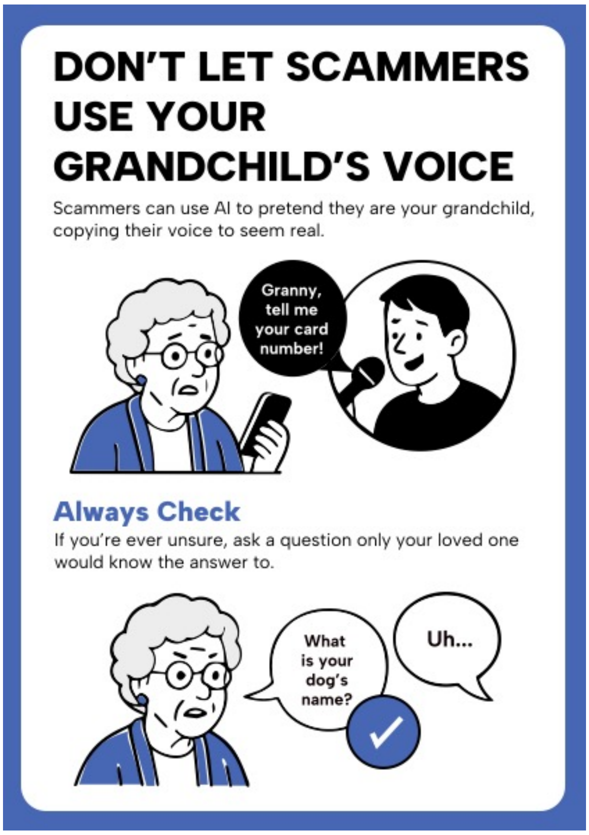

DEBATE ABOUT
Is it okay for AI to favor formal language over slang or dialects?
DEBATE ABOUT
In every case, it should be prevented that an AI model outputs any swear word.
DEBATE ABOUT
Are shocking headlines about NLP a good way to create awareness of potential risks?
DEBATE ABOUT
Is using an AI system to select applicants for a job interview more fair than having a human do this?
DEBATE ABOUT
Is the convenience of AI systems like ChatGPT worth the environmental costs that come with the vast amount of required computing power?
DEBATE ABOUT
Should teaching children about the risks of AI be mandatory for primary schools?
DEBATE ABOUT
I have access to books that I am not allowed to share with other people, but it is okay to use those books to train my model.
DEBATE ABOUT
Language models are trained on real conversations, so they reflect reality.
DEBATE ABOUT
Research in the field of AI should always be approved by an ethics committee before starting.
DEBATE ABOUT
AI models can use all data that is posted online.
DEBATE ABOUT
Is it okay if translation AI models use grammar rules that ignore gender inclusivity?

Figure 12: Some example cards from the Debatable game (Groningen, 2024/2025 edition). The coloured dots signal a specific theme. Credits: Ilse Kerkhove, Dertje Roggeveen Marieke Schelhaas, Mijke van Daal, Nikki van Gurp.



DON’T LET SCAMMERS USE YOUR GRANDCHILD’S VOICE

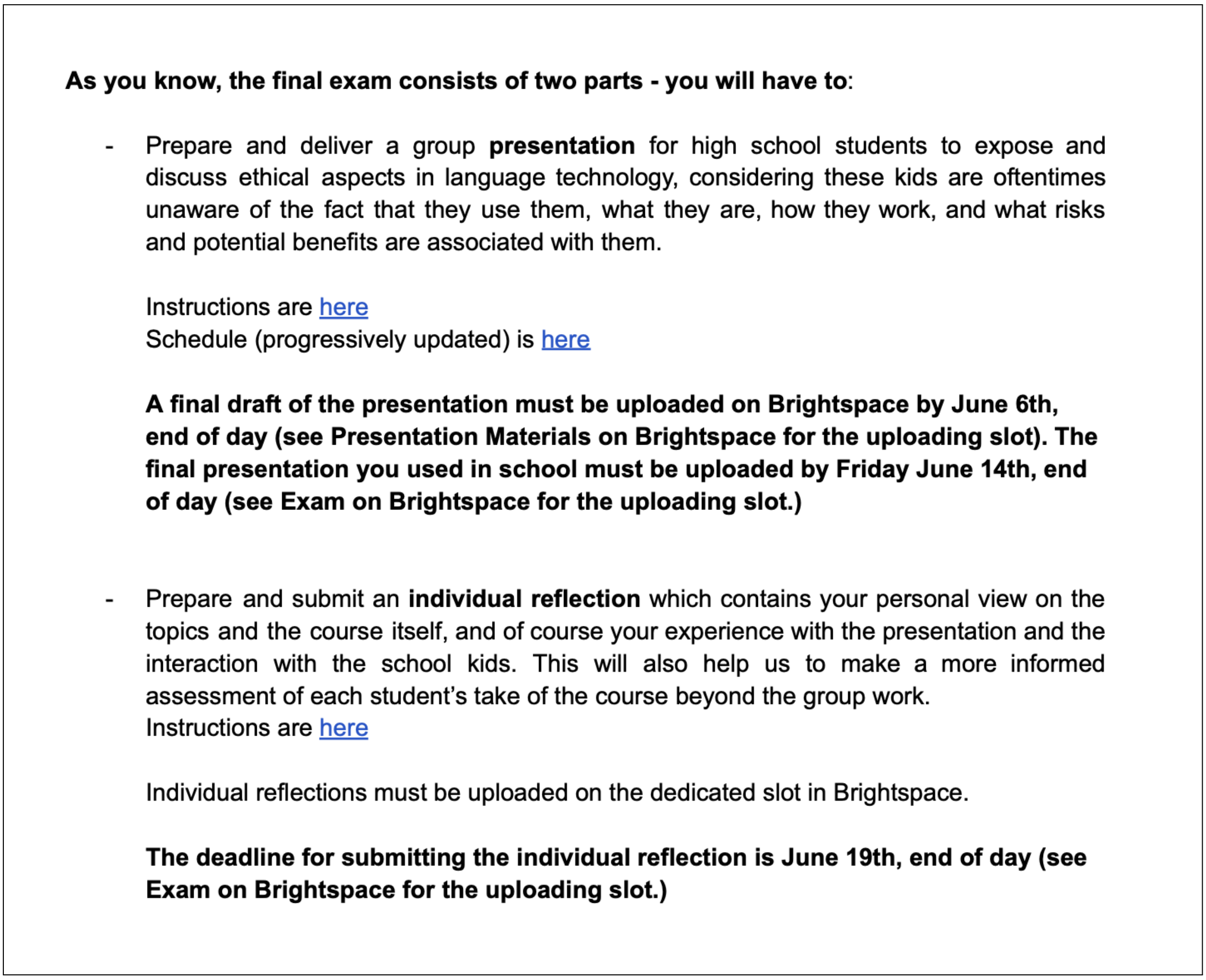
Scammers can use AI to pretend they are your grandchild, copying their voice to seem real.

Always Check

If you’re ever unsure, ask a question only your loved one would know the answer to.

Figure 13: Part of leaflet for elderly people in care homes, Turin edition 2024/2025. Follows explanation in other parts of the leaflet that talk about how state-of-the-art speech technology can now make use of one’s voice in a very credible manner. Credits: Elina Saifutdinova, Evgeniya Voropaeva, Kseniia Zakharneva, Tatiana Semenova.

Figure 14: Screenshots from video of interviews with laypersons about the influence of AI in education, ecology and art, Turin edition 2024/2025. Credits: Chiara Falcioni, Martina Tazzini, Alex Tessarin, Emir Ünverdi.



As you know, the final exam consists of two parts - you will have to:

- Prepare and deliver a group **presentation** for high school students to expose and discuss ethical aspects in language technology, considering these kids are oftentimes unaware of the fact that they use them, what they are, how they work, and what risks and potential benefits are associated with them.

  Instructions are here
  Schedule (progressively updated) is here

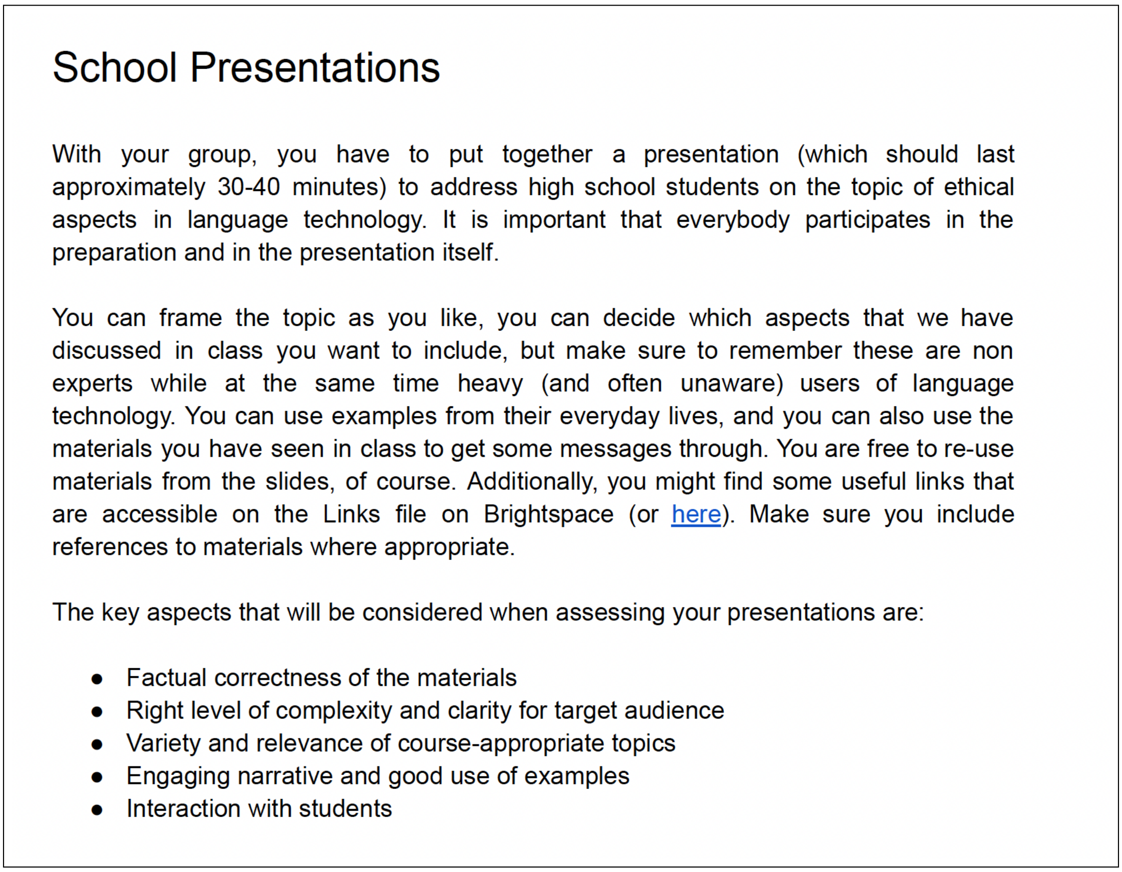
  A final draft of the presentation must be uploaded on Brightspace by June 6th, end of day (see Presentation Materials on Brightspace for the uploading slot). The final presentation you used in school must be uploaded by Friday June 14th, end of day (see Exam on Brightspace for the uploading slot.)

- Prepare and submit an **individual reflection** which contains your personal view on the topics and the course itself, and of course your experience with the presentation and the interaction with the school kids. This will also help us to make a more informed assessment of each student's take of the course beyond the group work.
  Instructions are here

  Individual reflections must be uploaded on the dedicated slot in Brightspace.

  The deadline for submitting the individual reflection is June 19th, end of day (see Exam on Brightspace for the uploading slot.)

Figure 15: General exam instructions. Links point to the details of the presentation and reflection requirements and the corresponding rubric (reported in Figures 16 and 17), and to the schedule for school appointments per group with dates, times, and contacts (schedule not included in this paper).



School Presentations

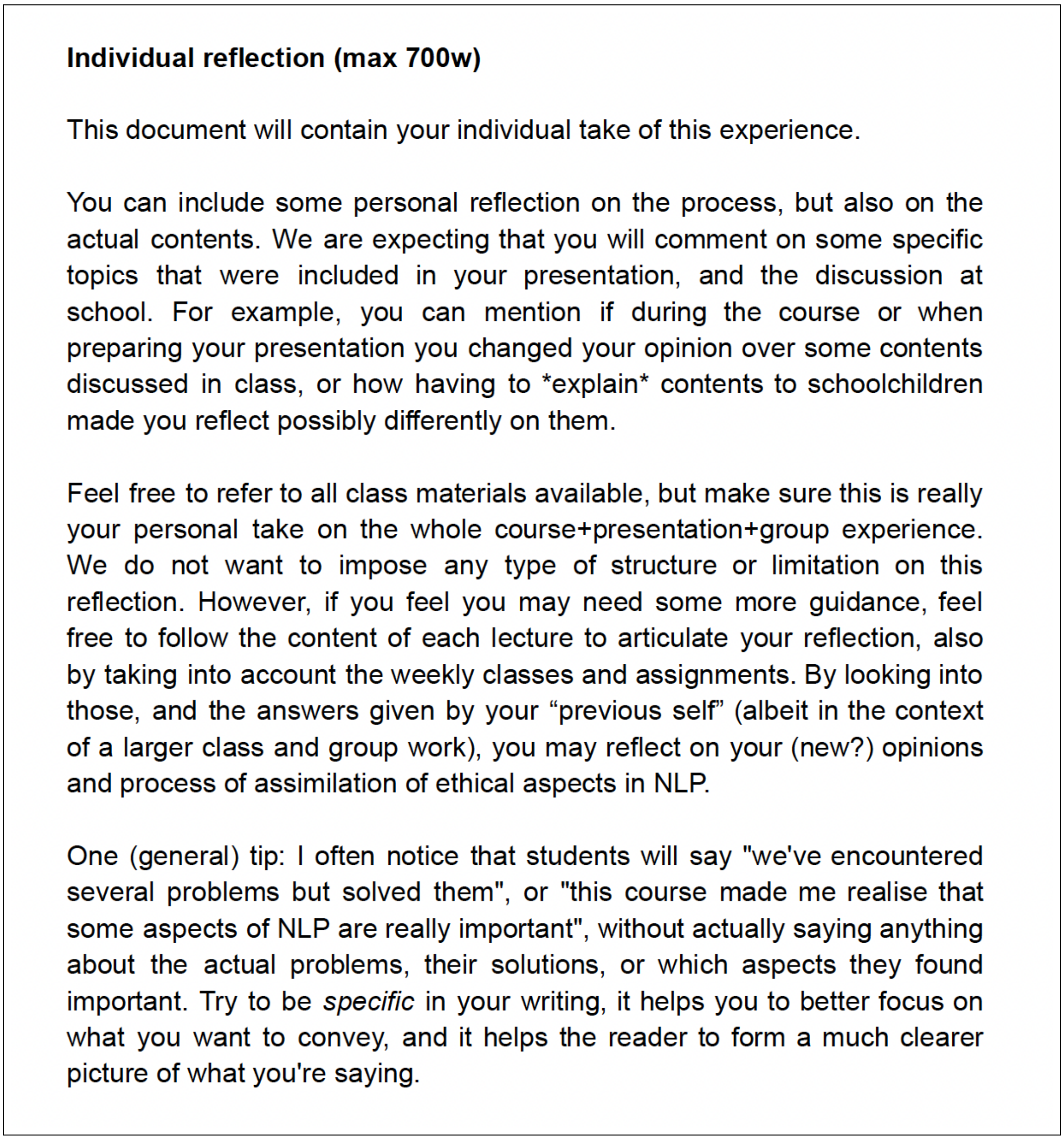
With your group, you have to put together a presentation (which should last approximately 30-40 minutes) to address high school students on the topic of ethical aspects in language technology. It is important that everybody participates in the preparation and in the presentation itself.

You can frame the topic as you like, you can decide which aspects that we have discussed in class you want to include, but make sure to remember these are non experts while at the same time heavy (and often unaware) users of language technology. You can use examples from their everyday lives, and you can also use the materials you have seen in class to get some messages through. You are free to re-use materials from the slides, of course. Additionally, you might find some useful links that are accessible on the Links file on Brightspace (or here). Make sure you include references to materials where appropriate.

The key aspects that will be considered when assessing your presentations are:

- Factual correctness of the materials
- Right level of complexity and clarity for target audience
- Variety and relevance of course-appropriate topics
- Engaging narrative and good use of examples
- Interaction with students

Figure 16: Presentation instructions.



Individual reflection (max 700w)

This document will contain your individual take of this experience.

You can include some personal reflection on the process, but also on the actual contents. We are expecting that you will comment on some specific topics that were included in your presentation, and the discussion at school. For example, you can mention if during the course or when preparing your presentation you changed your opinion over some contents discussed in class, or how having to *explain* contents to schoolchildren made you reflect possibly differently on them.

Feel free to refer to all class materials available, but make sure this is really your personal take on the whole course+presentation+group experience. We do not want to impose any type of structure or limitation on this reflection. However, if you feel you may need some more guidance, feel free to follow the content of each lecture to articulate your reflection, also by taking into account the weekly classes and assignments. By looking into those, and the answers given by your “previous self” (albeit in the context of a larger class and group work), you may reflect on your (new?) opinions and process of assimilation of ethical aspects in NLP.

One (general) tip: I often notice that students will say "we've encountered several problems but solved them", or "this course made me realise that some aspects of NLP are really important", without actually saying anything about the actual problems, their solutions, or which aspects they found important. Try to be *specific* in your writing, it helps you to better focus on what you want to convey, and it helps the reader to form a much clearer picture of what you're saying.

Figure 17: Individual reflection instructions.

Interview Template

(P. Darwinkel, T. Leneman, and J. Loomans)

The following are the topics we would like to ask you questions about.

- The ACL ethics committee
- Accountability, responsibility, and legal matters in NLP
- Some technical aspects
- Dominance of English in NLP
- Public policy and NLP

ACL Ethics Committee (Ties)

1. On your website it says that much of your research group's work focuses on NLP for social good and that one of your goals is to enable NLP for diverse and disadvantaged users. Is this part of your motivation to be part of the ACL's Ethics Committee?

(self-explanatory)

2. Do you think that stricter ethics guidelines and awareness may lead to self-censorship and/or cause researchers to (unnecessarily) avoid certain research topics? It isn't far-fetched to imagine a scenario where some researchers may decide to avoid certain topics altogether so as to not attract any potential controversy or public outrage.

prof. dr. Nissim mentioned after the lecture that she fears that this may increasingly become an issue, and a pop article described a situation where a senior scientist talked to a young researcher who was doubting whether to continue their research.

3. Do you think there should be an ACL-wide, universal definition of bias because the ethics checklist explicitly leaves room for interpretation on what bias actually constitutes. Could a universal, structured, comprehensive guideline help avoid misalignment between researchers?

A significant part of the ethics checklist is devoted to bias, but explicitly leaves room for interpretation by paper submissions on what bias actually constitutes. Blodgett et al. (2020) discuss this topic, but do not suggest a concrete dictionary-like definition. A universal, structured, official, comprehensive guideline of the topic might help avoid misalignment between researchers.

4. Have you used the ethics guidelines and checklists in practice or have seen it being successfully used? If you did, then have you encountered issues in the practical application of the guidelines that should be addressed?

Throughout this week's course materials, the ethics guidelines popped up. However, we have seen very little concrete material about how it works in the field. We're curious as to how the interviewee experiences the new regulations.

Some technical aspects (Patrick)

1. Do you (and if so: how?) think corpora from non-digital origins (e.g. local books/newspapers/religious texts) can help de-bias language technologies?

As discussed in the first part of the assignment - how can we successfully increase the diversity of our training data? And will this help with de-biasing?

2. Do you know of any promising methods that may help separate between gender/race/class inappropriate (e.g. stereotypes) and appropriate (e.g. grammatical gender or contextual clues) language? Machine learning methods essentially learn from co-occurences of words. How do you separate legitimate from illegitimate co-occurences? Will this require linguistic knowledge, dictionary lookups, or perhaps modified training data to mitigate the symptoms?

(self-explanatory)

3. Do you think that de-biasing contextual clues in e.g. word embeddings and images could/should become part of a compulsory pre-processing pipeline? Papers on de-biasing frequently suggest de-biasing techniques which do not seem to affect performance too much. Should such practice become some kind of de-facto standard in NLP?

Two papers which we read are both concerned with the bias-reinforcing effect of contextual clues hidden in data. Both papers suggest de-biasing techniques which do not seem to affect performance too much. Should such practice become some kind of de-facto standard in NLP?

Accountability, responsibility, and legal matters in NLP (Patrick)

1. Do you think that public information (e.g. public posts from social media) should be in the public domain or should the author always retain the full ownfership rights to the text? Who do you think owns information generated through public discussions? What implications does this have legally and ethically?

This fundamental question was discussed during the lab as we talked about the ethics and legality of using public Tweets. If public space = public domain, then do people actually own their tweets and utterances? And if they do not: isn't the knowledge that everything you post publicly may be analyzed, scrutinized, and used for purposes that you wouldn't want highly uncomfortable? But then again, would this also apply for literature and opinion articles? And if not: how are those different from Tweets?

2. **Most moral philosophers hold the view that the ability to reason and act independently is a requirement for accountability, and at this point in time this is not the case for AI systems. Who do you think is responsible for decisions made by NLP systems? Or: responsible for the consequences of NLP systems? Do you think the responsibility lies with the developers, the people who deploy them, or somewhere else?**

A fundamental problem in ethics is that of moral agency: most philosophers hold that the ability to reason independently is a requirement for accountability. Currently, most AI systems do not really have that. We want to know who Julia thinks is responsible for their use.

Dominance of English in NLP (Jordy)

1. How would you feel about a quorum for the proportion of paper submissions that focus on an English corpus? Or: forcing papers to focus on a different language if the topic allows for it. Perhaps demanding that papers contain languages with various traits (e.g. non-Latin alphabet; complex morphology; flexible word order), would put pressure on researchers to actually deviate from using English-as-default.

Emily Bender and other researchers talk about the problem of the status of the English language as the "default" language in NLP. No concrete solution has been proposed, aside from explicitly naming the researched/used languages and having a public debate about the problem. A hard quotum of rule of some kind, perhaps not even regarding the use of English, but demanding that papers contain languages with various traits (e.g. non-Latin alphabet; complex morphology; flexible word order), would put pressure on researchers to actually make a change.

2. Do you (and if so: how?) think that multilingual data could be used to balance different cultural and linguistic biases? It's likely that Indian English, South African English, and American English all have different biases in their data.

Would the use of data other than American English data give an improvement or would it just introduce different problems due to the languages such as grammatical assumptions of gender. Could multilingual data be used to balance different language biases?

3. Do you think that the continuing status of English as the "default" language in NLP reinforces the Anglo-Saxon global cultural hegemony? Is this even a problem? Could multilingual data halt or reduce Anglo-Saxon dominance?

During the Machine Translation course we read a paper that explicitly warned for the risk of "global cultural hegemony" propagated through the widespread use of English. It begs the question whether English-first NLP research reinforces this problem.

Government and NLP (Ties)

1. Which do you think is the more pressing concern: biases in NLP, or the generation of fake reality (e.g. news, images, and voices)? Given limited time, would you focus on de-biasing or more controlled use of models?

These two aspects of AI (biases and the generation of fake reality) are subtly abused - much more subtly than other fields of AI such as (weaponized) robotics. Given limited resources, which would be the more pressing issue? Biases in NLP reinforce already existing biases, but uncontrolled use of generative tools may result in society beginning to live in different "realities of truth".

2. What do you think the role of the government should be in the deployment of NLP systems? Should there be legal boundaries in what is allowed and what not? What institution or regulatory body would enforce these laws?

<self explanatory>

3. How would you feel about your public research being incorporated in a multilingual robo-dog patrolling the streets of Shanghai or the US-Mexico border? What kind of responsibility do you feel regarding possible abuse of your publicly-funded work?

Public research will no doubt be incorporated into industrial applications. This week's pop articles talk about robo-dogs and military use. A researcher no doubt has feelings about this.