Integrating Ethics in the NLP curriculum

ACL 2020
July 5

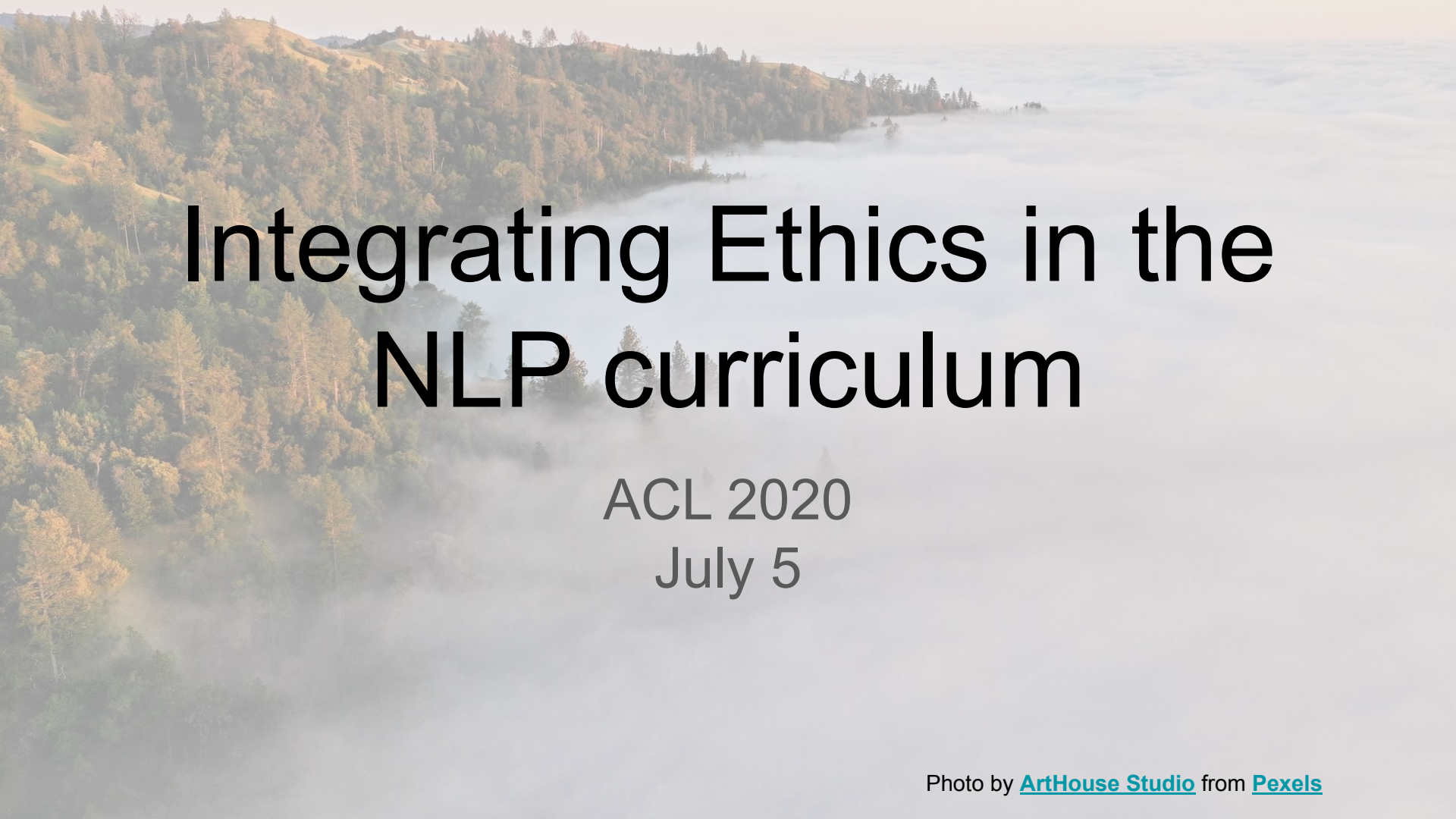


Photo by ArtHouse Studio from Pexels

About us

Emily M. Bender: UW Linguistics, working in Ethics & NLP space since late 2016, starting with pulling together a class for the CLMS program

Dirk Hovy: Bocconi University, Milan, working on ethics in NLP since 2016, interested in bias, dual use, and policy consequences

Xanda Schofield: Harvey Mudd College, working on making NLP tools available to domain experts in humanities and social science, interested in privacy, undergrad education

Schedule for Today

Plenary [60min]

Break [15min]

Break out sessions [30min] : ex1

Plenary [30min]

Break [15min]

Break out sessions [30min] : ex 2

Plenary [30min]

Requests:

1. Please stay for the whole tutorial
2. Please do not logout of Zoom during the breaks

Goals for today

- What are some of the ethical risks of NLP technology?
- What are our desired outcomes from incorporating this topic into the NLP curriculum?
- What are the challenges to integrating ethics into the NLP curriculum?
- Work through/develop sample exercises
  - How could these be modified to fit your instructional context?
- Share out results of discussion to ACL community at large (via wiki)

Ground rules

http://www.nonprofitinclusiveness.org/agreements-courageous-conversations-and-active-learning

- Stay engaged
- Experience discomfort
- Speak your truth
- Expect and accept non-closure
- Maintain confidentiality
- Listen with the intent to learn
- Suspend judgment

Goals in your curriculum: Your input

- What is your motivation for being here?
- What classes are you thinking about integrating ethics into? What are they like (grad/undergrad/other; class size; class format)?
- What are your goals for your students of including ethics in your NLP class?

Why is this hard? Your input

- What challenges do you see in getting students to engage with this material?
- Which learning goals are likely to be harder/easier to accomplish?

ACM Code of Ethics: https://www.acm.org/code-of-ethics

- Adopted by the ACL (March 2020)
- Key excerpt:

2.2 Maintain high standards of professional competence, conduct, and ethical practice.

High quality computing depends on individuals and teams who take personal and group responsibility for acquiring and maintaining professional competence. Professional competence starts with technical knowledge and with awareness of the social context in which their work may be deployed. Professional competence also requires skill in communication, in reflective analysis, and in recognizing and navigating ethical challenges. Upgrading skills should be an ongoing process and might include independent study, attending conferences or seminars, and other informal or formal education. Professional organizations and employers should encourage and facilitate these activities.

Section 1: Core Ethics Concepts

(More reading: https://aclweb.org/aclwiki/Ethics_in_NLP)

Bias

Has seemingly become the main topic in ethics in NLP (bias in models, embeddings, etc.).

However, bias is not necessarily bad: it's a preset (or Bayesian: prior), that can help us make decisions absent more information.

However, if this bias overwhelms the evidence, or if it influences predictive systems, it becomes problematic.

It is almost impossible to have an unbiased system.

Bias can arise from **data**, **annotations**, **representations**, **models**, or **research design**, among other things

Dual Use

Every technology has an **intended use**, and **unintended consequences** (nuclear power, knives, electricity could all be abused for things they were not originally designed to do).

Ethicist **Hans Jonas** commented on the inevitability of use: if a technology is available, it will be used.

Since we do not know how people will use it, we need to be aware of this duality.

Privacy

Privacy is often conflated with **anonymity**, but they are separate:

Privacy means nobody know I am doing *something*, **anonymity** means everyone knows *what* I am doing, but not that it is *me*.

GDPR requires us to **anonymize** data so that people cannot be identified "*without significant effort.*" It is unclear what that means given author attribute predictions: Given enough attributes, the set is small enough to identify people.

Theoretical and *differential privacy* are concepts to take into account.

Photo by Noelle Otto from Pexels

Normative vs. Descriptive Ethics

Useful concept to distinguish moral gradation: What we want the world to be vs. what it is.

A coreference system that cannot resolve female pronouns with the noun "doctor" is both normatively wrong (we want women to be doctors), and descriptively wrong (the sentence is actually referring to a female doctor).

Racially or gender-biased word embeddings are normatively wrong (we do not want the systems to proliferate stereotypes), but might be descriptively correct (they reflect how societies talk about gender and ethnicity).

Van de Poel, 2016: https://pure.tudelft.nl/portal/files/10265757/art_10.1007_s11948_015_9724_3.pdf

New Technology as a Large-Scale Experiment

New technology like NLP can be conceived as a social experiment (Van de Poel, 2016)

If we assume we are all participating in a large experiment, we need to make sure it meets certain criteria of responsible experimentation:

- Beneficence (no harm to subjects, maximize benefits, minimize risk)
- Respect for subjects' autonomy (informed consent)
- Justice (benefits vs. harms, protection of vulnerable subjects)

Section 2: A Few Pedagogy Concepts

Reflection Question:

What goals do you have for your students in including ethics in your NLP teaching?

Learning Outcomes: Statements of what **students will be able to do** on completion of your course, training, or program.

What are learning outcomes?

Statements of what **students will be able to do** on completion of your course/training/program.

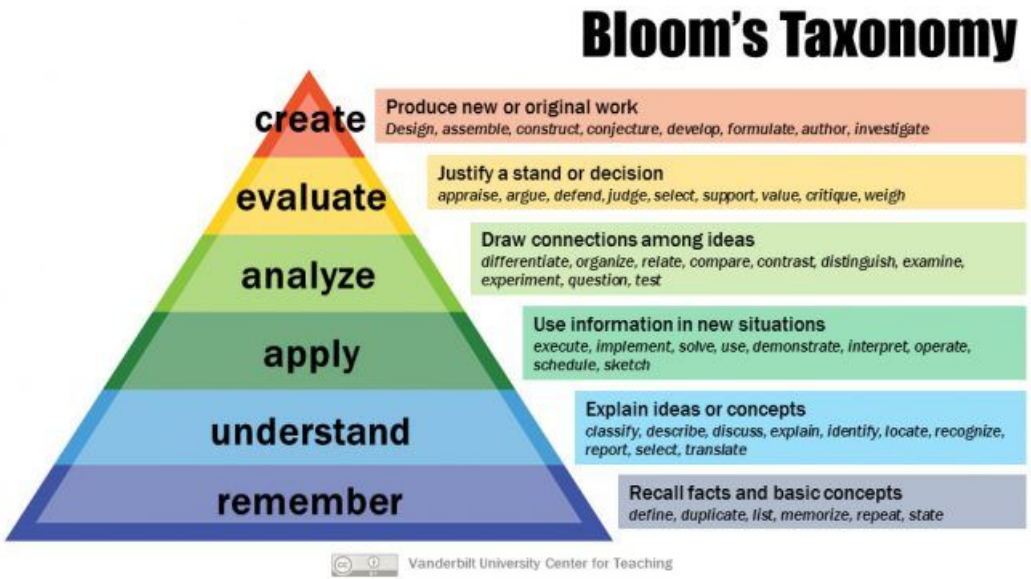
Realistic

Specific

Student-Centered

Measurable

Bloom's Taxonomy



Describes student activities that correspond to measurable skills.

Higher = more complex
(but not necessarily more valuable)

All layers can work for different types of knowledge:

- Factual
- Conceptual
- Procedural
- Metacognitive

Designing assessments for learning outcomes

Well-designed learning outcomes **should make it easier** to design meaningful learning assessments.

Example. Students should be able to...

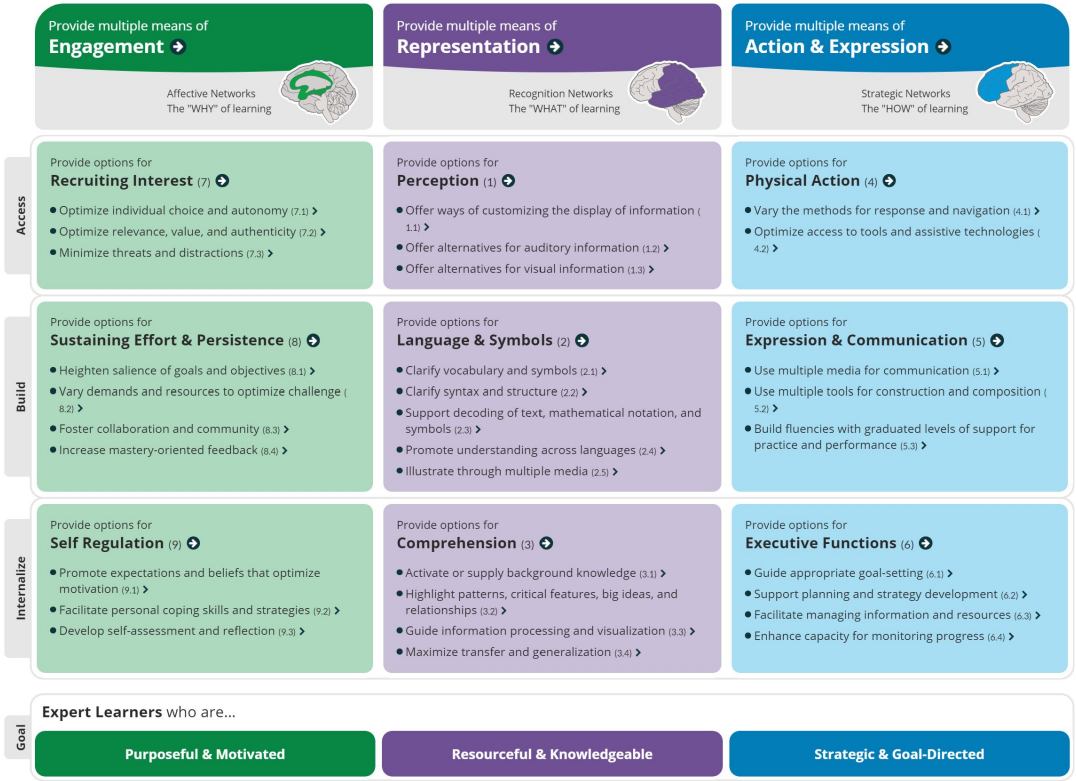
- Identify parts of speech of words in a sentence
- Describe how part of speech tags are useful in NLP tasks
- Compute ML parameter estimates for an HMM POS tagger
- Implement the Viterbi algorithm
- Evaluate the quality of a POS tagger

Access & universal design

How can we assume one learning strategy will work for everyone?

When possible, give **multiple paths** to success.

Access & universal design



Universal Design for Learning (UDL): udlguidelines.cast.org

Dimensions:

Why an outcome is important

What the material is to learn

How to practice and demonstrate knowledge

And how to **access**, **build**, and **internalize** learning.

Nobody expects perfection.

Communicating with your students can help you spend time on design where it counts.

Exercise 1: In-Class

Dual-Use

Learning Outcome: Students should be able to

- Recognize the dual nature of an issue
- Analyze the pros and cons

Exercise. A fellow student suggests a group project topic they want to explore: gendered language in the LGBTQ community. They are very engaged in the community themselves and have access to data. Their plan is to write a text classification tool that distinguishes LGBTQ from heterosexual language. What do you tell the student?

Bias: Sensitivity to language variation

Learning Outcome: Students should be able to

- Describe the potential impact of linguistic variation on the functioning of NLP/speech technology
- Reason about how differential performance for different social groups can lead to adverse impacts
- Articulate what kind of documentation should accompany NLP/speech technology to facilitate safe deployment

Bias: Sensitivity to language variation

- Pick an application of speech/language technology, determine what kind of training data is typically used for it (whose language? recorded when/where/how?).
- Next, imagine real world use cases for this technology. What speaker groups would come in contact with the system?
- If their language differs substantially from the training data, what would the failure mode of the system be and what would the real-world impacts of that failure be?
- How could systems, their training data or documentation be designed to be robust to this kind of problem?

Privacy

Learning Outcome: Students should be able to

- Select possible de-anonymizing features from documents
- Use statistics to argue about the effect of a document on a simple model

Exercise. Consider a simple Naive Bayes classifier trained on a subset of 20 Newsgroups using word frequencies as features. For five sample messages, could you tell whether or not they were included in the subset? How would you check? How certain could you be?

Follow-up questions

How does this adapt to your class format and composition?

What preparation would students need, and how would you provide it?

How would you assess student understanding from this exercise?

Get started here: **https://bit.ly/teachingnlpethics**

Exercise 2: Project Work

Dual-Use

Learning Outcome: Students should be able to

- Come up with arguments and explanations of unintended use
- Device responsible ways to address dual use issues

Exercise. One group develops a tool to detect personal attributes with high accuracy. Another group tries to "break" this tool.
Why, or why not, should you release it as an app?

Related discussion exercise. An ACL submission claims to be able to undo ciphers used by dissenters on social media. Who benefits from this? Is it better to release it in a peer-reviewed venue than to not know it?

Bias

Learning outcome. Students should be able to:

- Measure effect of bias in word vectors on a sentiment analysis system
- Discuss implications of treating found discourse as an objective representation of the world

Exercise: http://blog.conceptnet.io/posts/2017/how-to-make-a-racist-ai-without-really-trying/ (by Robyn Speer)

Exercise: Read and discuss growing literature on debiasing embeddings. What would make embeddings “safe” enough to use? How would we test that for given applications?

Privacy

Learning Outcome: Students should be able to

- Implement an inverted index for unigram searches
- Use the randomized mechanism of differential privacy

Exercise. Design a small search engine around an inverted index that uses random integer noise from a two-sided geometric distribution (Ghosh et al., 2012) to shape which queries are retrieved. Analyze how much this changes the search results with different noise levels. Are there systematic changes?

Follow-up questions

How does this adapt to your class format and composition?

What preparation would students need, and how would you provide it?

How would you assess student understanding from this exercise?

Get started here: **https://bit.ly/teachingnlpethics**

Wrap-up

Core ideas to walk away with

Teach students to ask questions, rather than treat ethics as a checklist.

- Who will this impact and how?
- Where are possible sources of ethical problems?
- What do I need to learn about in order to deploy this safely?

Present NLP as part of broader socio-technical systems, rather than just technical solutions.

Our students have future roles as technologists, informed consumers, informed readers of media reports and informed advocates for appropriate policy

How do we integrate this into a class?

Courses are hard to change overnight!

- Gather feedback as you test new exercises and assessments
- Center student experience and learning
- Pursue a goal of having ethics integrated with your class

Future directions

How can we continue the conversation we started here?

- Wiki?
- Website?
- Collection of case studies?
- …?

“What’s learned here, leaves here. What’s shared here, stays here.”

Thanks for participating!