Repository Entry
Embedded EthiCS @ Harvard Teaching Lab

Overview		
Course:	CS 187: Introduction to Computational Linguistic	
Course Level:	Undergraduate	
Course Description:	"Natural-language-processing applications are ubiquitous: Alexa can set a reminder if you ask; Google Translate can make emails readable across languages; Watson outplays world Jeopardy champions; Grover can generate fake news, and recognize it as well. How do such systems work? This course provides an introduction to the field of computational linguistics, the study of human language using the tools and techniques of computer science, with applications to a variety of natural-language-processing problems such as these. You will work with ideas from linguistics, statistical modeling, and machine learning, with emphasis on their application, limitations, and implications. The course is lab- and project-based, primarily in small teams, and culminates in the building and testing of a question-answering system."[1]	
Module Topic:	Defending Against Neural Fake News	
Module Author:	Samuel Dishaw	
Semesters Taught:	Fall 2020	
Tags:	E.g. natural language processing [CS], GPT-3 [CS], fake news [both], bystander immunity [phil], necessity condition [phil]	
Module Overview:	The module identifies a recent threat of computer-generated fake news (GPT-3), and a recent proposal for how to defend against that threat (GROVER). After introducing concepts from the ethics of defending against threats, students come up with different proposals for how we should use GROVER.	
Connection to Course Material:	In class, students learn how to create natural language processing algorithms as well as how to use them to either generate text or identify patterns in authorship. One potential use of natural-language processing is to identify computer-generated text. The risks and benefits of tapping this potential are the focus of this module.	The module was chosen partly for its timeliness (GPT-3, and the threats of fake news that it poses, were being discussed in popular news outlets at the time of the module), and because it served well to highlight some of the real-world threats posed by natural language processing programs more powerful than those the students would have had the chance to work with in class.

Goals

[1] https://courses.my.harvard.edu/psp/courses/EMPLOYEE/EMPL/h/?tab=HU_CLASS_SEARCH&SearchReqJSON=%7B%22SearchText%22:%22117372%22%7D

Module Goals:	1. Introduce a framework for thinking about the ethics of defending against threats. 2. Apply this framework to a proposed defense against neural fake news ("GROVER"). 3. Identify and discuss ethical problems with that proposal. 4. Consider alternative solutions (ethical design).	
Key Philosophical Questions:	1. What harms is it permissible to impose, and on whom, in defending against a threat? 2. In what ways is online censorship harmful, and who is harmed by it? 3. What are the ethical considerations when it comes to using the outputs of GROVER to defend against computer-generated fake news?	The questions under heading "1" are discussed in the context of introducing two principles from the ethics of defending against threat (the principle of bystander immunity and the principle of necessity) illustrated by examples from the ethics of just conduct in war. The questions under heading "2" bring out some ethical concerns with the proposal that the right way of using a detector of computer-generated fake news (viz. GROVER) on social media platforms is to prevent the news item identified as computer-generated fake news from being posted at all. The third set of questions invites students to consider alternative uses of GROVER's output beyond censorship.

Materials		
Key Philosophical Concepts:	● Bystander Immunity ● Liability to Harm ● Necessity Doctrine ● Censorship	These concepts provide a framework for thinking about false positives (i.e. cases where a news post is incorrectly identified by GROVER as being computer-generated fake news and subsequently prevented from being posted).
Assigned Readings:	● Zellers, R. et al. (2019), "Defending Against Neural Fake News" ● Arneson, R. (2006), "Just Warfare Theory and Noncombatant Immunity", *Cornell International Law Journal* (excerpt, pp. 666-68) ● Lazar, S. (2012), "Necessity in Self-Defense and War", *Philosophy and Public Affairs*, (excerpt, pp. 3-5)	Zellers et al. (2019) introduces GROVER and makes a positive proposal about how best to use it. Arneson (2006) discusses the concept of noncombatant immunity, which was used to illustrate a broader point about liability to defensive action. Lazar (2012) provides a tidy summary of

		the Necessity Condition in defending against threats.

	Implementation	
Class Agenda:	1. Introduce computer-generated fake news 2. Quiz: differentiating real news from computer-generated fake news 3. Two Principles from the ethics of defending against threats 4. Harms of Online Censorship 5. Problems with using GROVER to filter out fake news 6. Morally better uses of GROVER	The point of the quiz was to get students to take seriously the threat of computer-generated fake news. On a quiz of eight news items (four written by humans, four by either GROVER or GPT-3), the class on average scored below 50% in accuracy.
Sample Class Activity:	The main active learning exercise in the module was a discussion of better alternative uses of GROVER as a defense against computer-generated fake news. While a short list of possible alternative uses was provided, students were encouraged to come up with proposals of their own (which many of them did). Students were first divided into small groups and then reconvened to discuss their solutions.	Leading up to this activity, students were given a few examples of alternative uses of GROVER. One of these is the already common practice of flagging content deemed untrustworthy rather than removing it. Another was to give individuals more autonomy by allowing them to override the verdict given by GROVER. An additional idea that came up in discussion is that GROVER should label *people* rather than *posts* as untrustworthy. This proposal is also put forward by Rini (2017), although the reasoning for it is different (the rationale here was to minimize harm to "bystanders"; the rationale for Rini is efficiency in response time, which is not an issue where GROVER is concerned). The overlapping conclusions makes Rini (2017) a good reading to assign for this module, although it would be best used as a follow-up after the module, so as to allow students to work out the user-based proposal for themselves.
Module Assignment:	Write a 300-400 word essay responding to the following prompt: Would it be ethical to use Grover to defend against neural fake news?	The essays were peer-evaluated. Each student received three essays from other students. They then had to paraphrase the main thesis of the essay and grade it along a rubric we provided them with. Students thus learn not only to express their views using

	If so, focus on one use of the output of Grover, and explain why you think that use is to be preferred over others. If not, explain why you think using Grover is not justified. Your essay should concern whether using Grover would be *ethical*. Although you may certainly discuss whether using Grover would be *effective*, you should also discuss the ethics of using it.	argument, but also to evaluate the arguments of others, and respond to them in a helpful way (feedback on essays was also peer-graded) .
Lessons Learned:	1. Reactions to the module were positive on the whole. The same module could probably be given with a little less theoretical machinery. It seems likely that a version of this module that did not discuss the necessity condition but only focused on the principle of bystander immunity would be equally successful. This is because the simplest version of the necessity condition ("Do no unnecessary harm") is so straightforward as to almost go without saying. The principle of bystander immunity is what does most of the heavy lifting in this module. 2. One topic that was left underexplored is the issue of 'innocent' threats and to what extent they are liable for harm. (An innocent threat is someone who poses a threat to others but is not at fault for doing so, perhaps because they are ignorant of the fact that they are posing a threat, and faultlessly so.) This issue might be worth highlighting in future iterations of this module, in part because one of the most sensible alternative uses of GROVER (flagging individuals rather than posts) seems to address this worry.	