

# Natural Language Processing

## **Computational Ethics**

### Yulia Tsvetkov

yuliats@cs.washington.edu

1



Yulia Tsvetkov



## Announcements

- Last quiz on Monday Dec 5
  - dependency parsing, recommender systems, summarization, ethics
- Last homework deadline on Friday Dec 9
- Last lecture Monday Dec 5
- Email me if you would like to TA this course in the next Fall quarter (2023)
- Thank you for the fun course and for your participation and hard work

WeekDateTopicsReadingsHomeworks1 $9/28$ Logistics [slides] [recording]Course website, syllabusHomeworks1 $9/30$ Introduction [slides] [recording]Eis 1HW1 out2 $10/03$ $[slides]$ [recording]Eis 1HW1 out2 $10/05$ $[slides]$ [recording]Eis 2; J&M III 4; Ng & Jordan, 2001Introduction [slides] [recording]Eis 2; J&M III 4; Ng & Jordan, 20013 $10/07$ Text classification [slides] [recording]Eis 2; J&M III 4; Ng & Jordan, 2001In-class quiz 13 $10/10$ Text classification [slides] [recording]J&M III 5; Pang et al. 2002In-class quiz 13 $10/12$ Text classification [slides] [recording]J&M III 5; Eis 6.1-6.2; 6.4In-class quiz 14 $10/14$ Text classification [slides] [recording]J&M III 3; Eis 6.1-6.2; 6.4In-class quiz 14 $10/19$ Language modeling [slides] [recording]J&M III 3; Eis 6.1-6.2; 6.4In-class quiz 35 $10/21$ Lexical semantics [slides] [recording]J&M III 6; Eis 14HW1 due5 $10/24$ Lexical semantics [slides] [recording]Eis 14; J&M III 6In-class quiz 35 $10/26$ Lexical semantics [slides] [recording]Eis 14; J&M III 6In-class quiz 36In/26Lexical semantics [slides] [recording]Eis 14; J&M III 6In-class quiz 37 $10/28$ Neural networks [slides] [recording]Collaberty 10; Coll					
19/28[slides] [recording]syllabus19/30Introduction [slides] [recording]Eis 1HW1 out210/03Introduction [slides] [recording]Eis 1	Week	Date	Topics	Readings	Homeworks
1Image: precording in the syntabulasyntabula19/30Introduction [sildes] [recording]Eis 1HW1 out210/03Introduction [sildes] [recording]Eis 1HW1 out210/05Text classification [sildes] [recording]Eis 2; J&M III 4; Ng & Jordan, 2001Image and the syntability of t	1	9/28			
9/30[slides] [recording]Eis 1HW1 out210/03Introduction [slides]Eis 1			[slides] [recording]	syllabus	
1[sides][sides]2 $10/03$ Introduction [sides]Eis 12 $10/05$ Text classification [sides] [recording]Eis 2; J&M III 4; Ng & Jordan, 20013 $10/07$ Text classification [sides] [recording]Eis 2; J&M III 4; Ng & Jordan, 20013 $10/10$ Text classification [sides] [recording]Eis 2; J&M III 4; Ng & Jordan, 20013 $10/10$ Text classification [sides] [recording]Eis 2; J&M III 5; Pang et al. 20023 $10/10$ Text classification [sides] [recording]J&M III 5 $10/12$ Text classification [sides] [recording]J&M III 5 $10/14$ Text classification [sides] [recording]J&M III 5; Eis 6, 1-6, 2; 6, 4 $10/17$ Language modeling [sides] [recording]J&M III 5; Eis 6, 1-6, 2; 6, 4 $10/19$ Language modeling [sides] [recording]J&M III 5; Eis 6, 1-6, 2; 6, 4 $10/21$ Lexical semantics [sides] [recording]J&M III 6; Eis 14HW1 due $10/21$ Lexical semantics [sides] [recording]Eis 14; J&M III 6In-class quiz 3 $10/26$ Lexical semantics [sides] [recording]Eis 14; J&M III 6In-class quiz 3 $10/28$ Neural networks [sides] [recording]Eis 6, 3, 6, 5; J&M III 7, 5; J&M III 9; Coldberg 10; Collobert et al. 2011 $10/31$ Neural networks [sides] [recording]Annotated Recording]Annotated Recording]		9/30		Eis 1	HW1 out
$\frac{10/03}{[stides]} = \frac{10}{10}$			[slides] [recording]		
2 $ 2 $ $ 10/05 $ $ 10/07 $ $ 1$		10/03		Eis 1	
2     10/05     [slides] [recording]     Eis 2; J&M III 4       10/07     [slides] [recording]     Jordan, 2001       3     10/10     Text classification [slides] [recording]     Eis 2; J&M III 5; Pang et al. 2002       3     10/12     Text classification [slides] [recording]     J&M III 5       10/14     Text classification [slides] [recording]     J&M III 5     In-class quiz 1       10/14     Text classification [slides] [recording]     J&M III 5     In-class quiz 1       10/14     Text classification [slides] [recording]     J&M III 5     In-class quiz 1       4     10/17     Language modeling [slides] [recording]     J&M III 3; Eis 6.1-6.2, 6.4       10/19     Language modeling [slides] [recording]     J&M III 3; Eis 6.1-6.2, 6.4       10/19     Language modeling [slides] [recording]     J&M III 3; Eis 6.1-6.2, 6.4       10/21     Lexical semantics [slides] [recording]     J&M III 6; Eis 14     HW1 due       5     10/24     Lexical semantics [slides] [recording]     Eis 14; J&M III 6     In-class quiz 3       5     10/26     Lexical semantics [slides] [recording]     Eis 6.3, 6.5; J&M III 7.5; J&M III 9; Goldberg 10; Collobert et al. 2011       10/28     Neural networks [slides] [recording]     Eis 6.3, 6.5; J&M III 7.5; J&M III 9; Goldberg 10; Collobert et al. 2011			[sudes]		
$\frac{110/07}{[slides] [recording]} = \frac{110/10}{[slides] [recording]} = \frac{110/12}{[slides] [recording]} = \frac{110/12}{[slides] [recording]} = \frac{110/12}{[slides] [recording]} = \frac{110/12}{[slides] [recording]} = \frac{100/12}{[slides] [recording]} = 10$	2	10/05		Eis 2; J&M III 4	
10/07     [sildes] [recording]     Jordan, 2001       3     10/10     Text classification [sildes] [recording]     Eis 2; J&M III 5; Pang et al. 2002       3     10/12     Text classification [sildes] [recording]     J&M III 5       10/14     Text classification [sildes] [recording]     J&M III 5       10/17     Language modeling [sildes] [recording]     J&M III 3; Eis 6.1-6.2, 6.4       10/19     Language modeling [sildes] [recording]     J&M III 3; Eis 6.1-6.2, 6.4       10/19     Language modeling [sildes] [recording]     J&M III 3; Eis 6.1-6.2, 6.4       10/19     Language modeling [sildes] [recording]     J&M III 6; Eis 14       10/21     Lexical semantics [sildes] [recording]     Eis 14; J&M III 6       10/26     Lexical semantics [sildes] [recording]     Eis 14; J&M III 6       10/28     Neural networks [sildes] [recording]     Eis 6.3, 6.5; J&M III 7.5; J&M III 9; Colloberg 10; Collobert et al. 2011       10/31     Neural networks [sildes] [recording]     JAnnotated Transformer; Illustrated     HW2 Out			[slides] [recording]		
3         10/10         Text classification [slides] [recording]         Jordan, 2001           3         10/10         Text classification [slides] [recording]         JBM III 5         In-class quiz 1           3         10/12         Text classification [slides] [recording]         JBM III 5         In-class quiz 1           10/14         Text classification [slides] [recording]         JBM III 5         In-class quiz 1           4         10/17         Language modeling [slides] [recording]         JBM III 3; Eis 6.1-6.2, 6.4		10/07			
$3 \qquad \begin{bmatrix} 10/10 & [slides] [recording] & al. 2002 \\ \hline 10/12 & Text classification & J&M III 5 & In-class quiz 1 \\ \hline 10/14 & Text classification & J&M III 5 & In-class quiz 1 \\ \hline 10/14 & Text classification & J&M III 5 & In-class quiz 1 \\ \hline 10/14 & Text classification & J&M III 5 & In-class quiz 1 \\ \hline 10/17 & Language modeling & J&M III 3; Eis 6.1-6.2, & In-class quiz 1 & In-class quiz 2 & In-class quiz 1 & In-class quiz 2 & In-class quiz 1 & $			[slides] [recording]	Jordan, 2001	
3     10/12     Text classification [slides] [recording]     J&M III 5     In-class quiz 1       10/12     Text classification [slides] [recording]     J&M III 5     In-class quiz 1       10/14     Text classification [slides] [recording]     J&M III 5     In-class quiz 1       4     10/17     Language modeling [slides] [recording]     J&M III 3; Eis 6.1-6.2, 6.4       10/19     Language modeling [slides] [recording]     J&M III 3; Eis 6.1-6.2, 6.4       10/21     Lexical semantics [slides] [recording]     J&M III 6; Eis 14       10/21     Lexical semantics [slides] [recording]     J&M III 6; Eis 14       10/24     Lexical semantics [slides] [recording]     Eis 14; J&M III 6       10/26     Lexical semantics [slides] [recording]     Eis 14; J&M III 6       10/28     Neural networks [slides] [recording]     Eis 6.3, 6.5; J&M III 7.5; J&M III 9; Coldberg 10; Collobert et al. 2011       10/31     Neural networks [slides] [recording]     Annotated Transformer; Illustrated     Hw2 Out		10/10	Text classification	Eis 2; J&M III 5; Pang et	
3     10/12     [slides] [recording]     J&M III 5     In-class quiz 1       10/14     Text classification [slides] [recording]     J&M III 5     In-class quiz 1       4     10/17     Language modeling [slides] [recording]     J&M III 3; Eis 6.1-6.2, 6.4       10/19     Language modeling [slides] [recording]     J&M III 3; Eis 6.1-6.2, 6.4       10/19     Language modeling [slides] [recording]     J&M III 3; Eis 6.1-6.2, 6.4       10/21     Lexical semantics [slides] [recording]     J&M III 6; Eis 14       10/21     Lexical semantics [slides] [recording]     J&M III 6; Eis 14       10/24     Lexical semantics [slides] [recording]     Eis 14; J&M III 6       10/26     Lexical semantics [slides] [recording]     Eis 14; J&M III 6       10/28     Neural networks [slides] [recording]     Eis 63, 6.5; J&M III 7.5; J&M III 9; Goldberg 10; Collobert et al. 2011       10/31     Neural networks [slides] [recording]     Annotated Transformer; Illustrated     Hw2 Out		10/10	[slides] [recording]	al. 2002	
2       10/12       [slides] [recording]         10/14       Text classification [slides] [recording]       J&M III 5         4       10/17       Language modeling [slides] [recording]       J&M III 3; Eis 6.1-6.2, 6.4         10/19       Language modeling [slides] [recording]       J&M III 3; Eis 6.1-6.2, 6.4         10/21       Lexical semantics [slides] [recording]       J&M III 6; Eis 14         10/21       Lexical semantics [slides] [recording]       J&M III 6; Eis 14         10/24       Lexical semantics [slides] [recording]       Eis 14; J&M III 6         10/26       Lexical semantics [slides] [recording]       Eis 14; J&M III 6         10/28       Neural networks [slides] [recording]       Eis 6.3, 6.5; J&M III 7.5; J&M III 9; Goldberg 10; Collobert et al. 2011         10/31       Neural networks [slides] [recording]       Annotated Transformer; Illustrated       Hw2 Out	2	10/12	Text classification	10 M III E	In class quiz 1
10/14     [slides] [recording]     J&M III 5       4     10/17     Language modeling [slides] [recording]     J&M III 3; Eis 6.1-6.2, 6.4       10/19     Language modeling [slides] [recording]     J&M III 3; Eis 6.1-6.2, 6.4       10/19     Language modeling [slides] [recording]     J&M III 3; Eis 6.1-6.2, 6.4       10/21     Lexical semantics [slides] [recording]     J&M III 6; Eis 14       10/21     Lexical semantics [slides] [recording]     J&M III 6; Eis 14       10/24     Lexical semantics [slides] [recording]     Eis 14; J&M III 6       10/26     Lexical semantics [slides] [recording]     Eis 14; J&M III 6       10/28     Neural networks [slides] [recording]     Eis 6.3, 6.5; J&M III 7.5; J&M III 9; Goldberg 10; Collobert et al. 2011       10/31     Neural networks [slides] [recording]     Annotated Transformer; Illustrated     HW2 Out	3	10/12	[slides] [recording]	J&M III 5	III-class quiz 1
4       10/17       Language modeling [slides] [recording]       J&M III 3; Eis 6.1-6.2, 6.4         10/19       Language modeling [slides] [recording]       J&M III 3; Eis 6.1-6.2, 6.4         10/19       Language modeling [slides] [recording]       J&M III 3; Eis 6.1-6.2, 6.4         10/21       Lexical semantics [slides] [recording]       J&M III 6; Eis 14         10/21       Lexical semantics [slides] [recording]       J&M III 6; Eis 14         10/24       Lexical semantics [slides] [recording]       Eis 14; J&M III 6         10/26       Lexical semantics [slides] [recording]       Eis 14; J&M III 6         10/28       Neural networks [slides] [recording]       Eis 6.3, 6.5; J&M III 7.5; J&M III 9; Coldberg 10; Collobert et al. 2011         10/31       Neural networks [slides] [recording]       Annotated Transformer; Illustrated       HW2 Out		10/14	Text classification	10 MA 111 E	
4       10/17       [slides] [recording]       6.4         4       10/19       Language modeling [slides] [recording]       J&M III 3; Eis 6.1-6.2, 6.4         10/21       Lexical semantics [slides] [recording]       J&M III 6; Eis 14       HW1 due         5       10/24       Lexical semantics [slides] [recording]       Eis 14; J&M III 6         10/26       Lexical semantics [slides] [recording]       Eis 14; J&M III 6       In-class quiz 3         10/26       Lexical semantics [slides] [recording]       Eis 6.3, 6.5; J&M III 7.5; J&M III 9; Goldberg 10; Collobert et al. 2011         10/28       Neural networks [slides] [recording]       Annotated Transformer; Illustrated       HW2 Out		10/14	[slides] [recording]	Joint III 2	
4     10/19     Language modeling [slides] [recording]     J&M III 3; Eis 6.1-6.2, 6.4       10/21     Lexical semantics [slides] [recording]     J&M III 6; Eis 14     HW1 due       10/21     Lexical semantics [slides] [recording]     J&M III 6; Eis 14     HW1 due       5     10/24     Lexical semantics [slides] [recording]     Eis 14; J&M III 6       10/26     Lexical semantics [slides] [recording]     Eis 14; J&M III 6       10/28     Neural networks [slides] [recording]     Eis 6.3, 6.5; J&M III 7.5; J&M III 9; Goldberg 10; Collobert et al. 2011       10/31     Neural networks [slides] [recording]     Annotated Transformer; Illustrated     HW2 Out		10/17	Language modeling	J&M III 3; Eis 6.1-6.2,	
4     10/19     [slides] [recording]     6.4       10/21     Lexical semantics [slides] [recording]     J&M III 6; Eis 14     HW1 due       5     10/24     Lexical semantics [slides] [recording]     Eis 14; J&M III 6       10/26     Lexical semantics [slides] [recording]     Eis 14; J&M III 6       10/26     Lexical semantics [slides] [recording]     Eis 14; J&M III 6       10/28     Neural networks [slides] [recording]     Eis 6.3, 6.5; J&M III 7.5; J&M III 9; Goldberg 10; Collobert et al. 2011       10/31     Neural networks [slides] [recording]     Annotated Transformer; Illustrated     HW2 Out			[slides] [recording]	6.4	
5     10/21     Lexical semantics [slides] [recording]     J&M III 6; Eis 14     HW1 due       5     10/24     Lexical semantics [slides] [recording]     Eis 14; J&M III 6       10/26     Lexical semantics [slides] [recording]     Eis 14; J&M III 6       10/26     Lexical semantics [slides] [recording]     Eis 14; J&M III 6       10/28     Neural networks [slides] [recording]     Eis 63, 6.5; J&M III 7.5; J&M III 9; Coldberg 10; Collobert et al. 2011       10/31     Neural networks [slides] [recording]     Annotated Transformer; Illustrated     HW2 Out	4	10/10	Language modeling	J&M III 3; Eis 6.1-6.2,	
10/21     [slides] [recording]     J&M III 6; Eis 14     HW1 due       5     10/24     Lexical semantics [slides] [recording]     Eis 14; J&M III 6       10/26     Lexical semantics [slides] [recording]     Eis 14; J&M III 6       10/28     Neural networks [slides] [recording]     Eis 6.3, 6.5; J&M III 7.5; J&M III 9; Goldberg 10; Collobert et al. 2011       10/31     Neural networks [slides] [recording]     Annotated Transformer; Illustrated	4	10/19	[slides] [recording]	6.4	
5     10/24     Lexical semantics [slides] [recording]     Eis 14; J&M III 6       10/26     Lexical semantics [slides] [recording]     Eis 14; J&M III 6       10/26     Lexical semantics [slides] [recording]     Eis 14; J&M III 6       10/28     Neural networks [slides] [recording]     Eis 6.3, 6.5; J&M III 7.5; J&M III 9; Goldberg 10; Collobert et al. 2011       10/31     Neural networks [slides] [recording]     Annotated Transformer; Illustrated		10/21	Lexical semantics	J&M III 6; Eis 14 HW1 due	Half due
10/24     [slides] [recording]     Eis 14; J&M III 6       5     10/26     Lexical semantics [slides] [recording]     Eis 14; J&M III 6     In-class quiz 3       10/28     Neural networks [slides] [recording]     Eis 6.3, 6.5; J&M III 7.5; J&M III 9; Goldberg 10; Collobert et al. 2011       10/31     Neural networks [slides] [recording]     Annotated Transformer; Illustrated			[slides] [recording]		HWIdue
5     10/26     Lexical semantics [slides] [recording]     Eis 14; J&M III 6     In-class quiz 3       10/28     Neural networks [slides] [recording]     Eis 6.3, 6.5; J&M III 7.5; J&M III 9; Goldberg 10; Collobert et al. 2011       10/31     Neural networks [slides] [recording]     Annotated Transformer; Illustrated		10/24	Lexical semantics	Fie 14: 19 M III 6	
S     10/26     [slides] [recording]     Eis 14; J&M III 6     In-class quiz 3       10/28     Neural networks [slides] [recording]     Eis 6.3, 6.5; J&M III 7.5; J&M III 9; Goldberg 10; Collobert et al. 2011       10/31     Neural networks [slides] [recording]     Annotated Transformer; Illustrated			[slides] [recording]	Eis 14; J&M III 6	
S     [slides] [recording]       10/28     Neural networks [slides] [recording]     Eis 6.3, 6.5; J&M III 7.5; J&M III 9; Goldberg 10; Collobert et al. 2011       10/31     Neural networks [slides] [recording]     Annotated Transformer; Illustrated		10/24	Lexical semantics	Cir 14: 19 M III C	In class quiz 2
10/28     Neural networks [slides] [recording]     J&M III 9; Goldberg 10; Collobert et al. 2011       Neural networks     Annotated       10/31     [slides] [recording]       Itunal metworks     Transformer; Illustrated	5	10/26	[slides] [recording]	Eis 14; J&M III 6	in-ciass quiz 3
10/28     [slides] [recording]     J&M III 9; Goldberg 10; Collobert et al. 2011       Neural networks     Annotated       10/31     [slides] [recording]       Itunplementary recording]     Illustrated		10/28	Noural ashunder	Eis 6.3, 6.5; J&M III 7.5;	
Collobert et al. 2011 Neural networks Annotated 10/31 [slides] [recording] Illustrated HW2 Out [supplementativ recording]					
Neural networks Transformer; HW2 Out [Suides] [recording] Illustrated HW2 Out				Collobert et al. 2011	
10/31 [slides] [recording] HW2 Out [supplementary recording]		10/31	Neural networks		
[supplementary recording]			[slides] [recording]		HW2 Out
			[supplementary recording]		

#### **TAT PAUL G. ALLEN SCHOOL** Eis 7.1-7.4, 8.1; J&M III Sequence labeling 11/02 In-class quiz 4 [slides] [recording] 8 Sequence labeling Eis 7.1-7.4, 8.1; Collins 11/04 [slides] [recording] notes Sequence labeling Eis 7.5, 7.7, 8.3; Sutton 11/07 [slides] [recording] & McCallum 2.1 - 2.5 Sequence labeling 11/09 Eis 7.6 In-class quiz 5 [slides] [recording] 11/11 Veterans Day (no class) Neural sequence labeling 11/14 Eis 7.6 [slides] [recording] Parsing Eis 10.1-10.2; J&M III 11/16 In-class quiz 6 [slides] [recording] 13 Parsing Eis 11.1, 11.3; J&M III 11/18 [slides] [recording] 14 Chen and Manning 11/21 Cancelled HW2 due, In-class quiz 7 2014 Parsing 11/23 Kassas et al. 2021 [slides] [recording] 11/25 Thanksgiving (no class) Advanced topics: 11/28 Recommender systems and HW3 out online training Research topics: 11/30 Summarization [slides] Advanced topics: NLP + 12/02 Historical linguistics 12/05 Q&A In-class quiz 8 12/07 TBD TBD HW3 due 12/09

Undergrad NLP 2022

6

7

8

9

10

11



## **Communication with machines**

#### • 50s-70s





## **Communication with machines**

• 80s

EDIT BS9U.DEVT3.CLIBPAU(TIMMIES) - 01.31 Command ===>	Columns 00001 0 Scroll ===>
****** *******************************	
000001 /* REXX EXEC *******************************	
000001 /* REAA EAEL *******************************	**********
000003 /* TIMMIES FACTOR - COMPOUND INTEREST CALCULATO	00
000004 /* TIMPLES FACTOR - COMPOUND INTEREST CALCULATE	JR
000005 /* AUTHOR: PAUL GAMBLE	
000006 /* DATE: OCT 1/2007	
000007 /*	
000008 /*	
000009 /********************************	
000010	
000011	
000012 Sau '************************************	
000013 say 'Welcome Coffee drinker.'	
000014 sau '***********************	
000015 DO WHILE DATATYPE(CoffeeAmt) \= 'NUM'	
000016 sau ""	
000017 say "What is the price of your coffee?",	
000018 "(e.g. 1.58 = \$1.58)"	
000019 parse pull CoffeeAmt	
000020 END	
000021	
000022 DO WHILE DATATYPE(CoffeeWk) \= 'NUM'	
000023 say ""	
000024 say "How many coffees a week do you have?"	
000025 parse pull CoffeeWk	
000026 END	
000027	
000028 DO WHILE DATATYPE(Rate) \= 'NUM'	
000029 say ""	
000030 say "What annual interest rate would you "	like to see on that money?",
000031 "(e.g. 8 = 8%)"	
000032 parse pull Rate	
000033 END	
000034 Rate = Rate * 0.01 /* CHG TO DECIMAL NUMBER */	

## NLP: Communication with machines

• Today

ul 💎	21:25	73 % 🛃
A COLUMN TO A COLUMN	need a dinner res r Valentine's Day	Sector Sector Sector
	e if any restauran e for one.	its have a
and the second s	o, I need a reser r two. <b>99</b>	vation
Why	? Is your mother i	in town?

WeKnowMemes



# Language use is fundamentally a social activity

The common misconception is that language has to do with words and what they mean. It doesn't. It has to do with **people** and what they mean.

> Herbert H. Clark & Michael F. Schober (1992) Asking Questions and Influencing Answers

# Decisions we make about our data, methods, and tools are tied up with their impact on people and societies.

## Ethics

## Ethics is a study of what are **good and bad** ends to pursue in life and what it is **right and wrong** to do in the conduct of life.

It is therefore, above all, a practical discipline. Its primary aim is to determine how one ought to live and what actions one ought to do in the conduct of one's life."

Introduction to Ethics, John Deigh



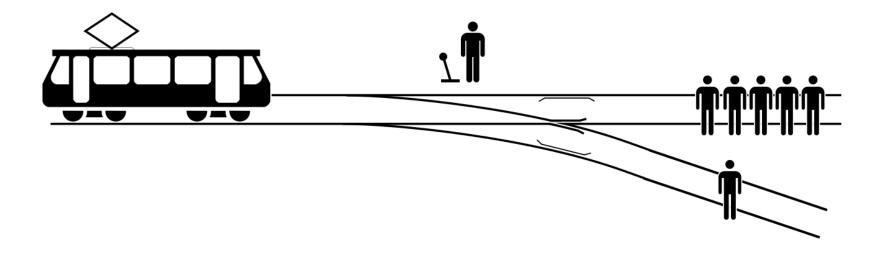
## **Ethics**

It's the **good** things It's the **right** things



## The Trolley Dilemma

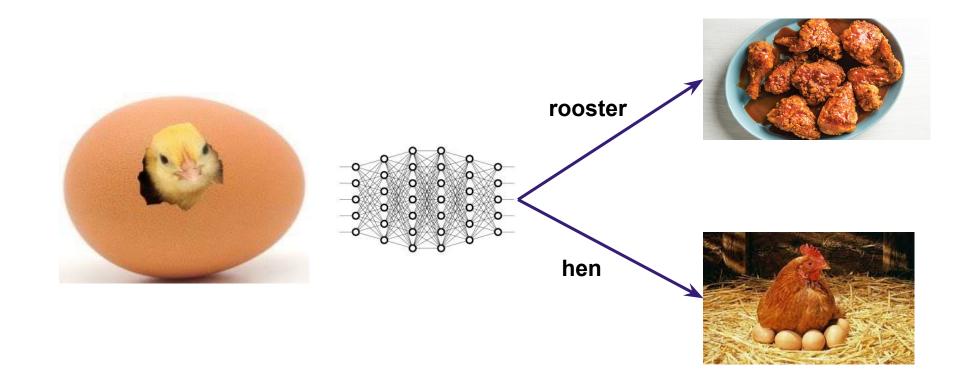
Should you pull the lever to divert the trolley?



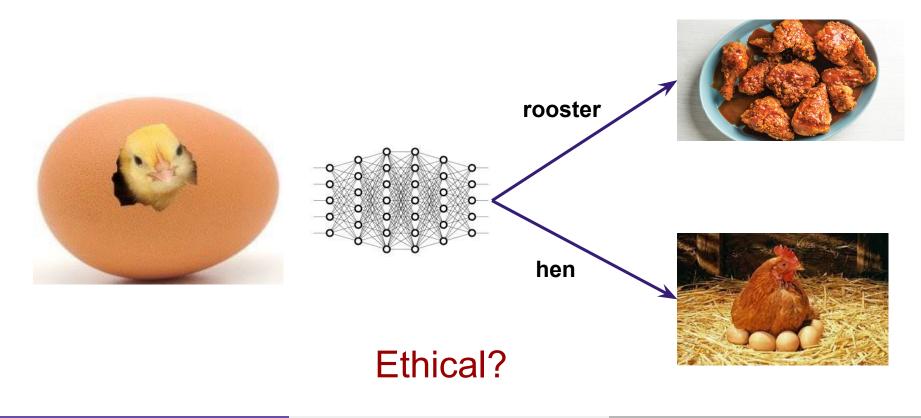
[image from Wikipedia]



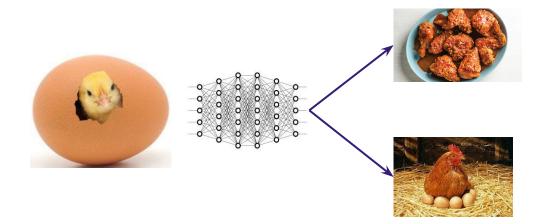
## The Chicken dilemma







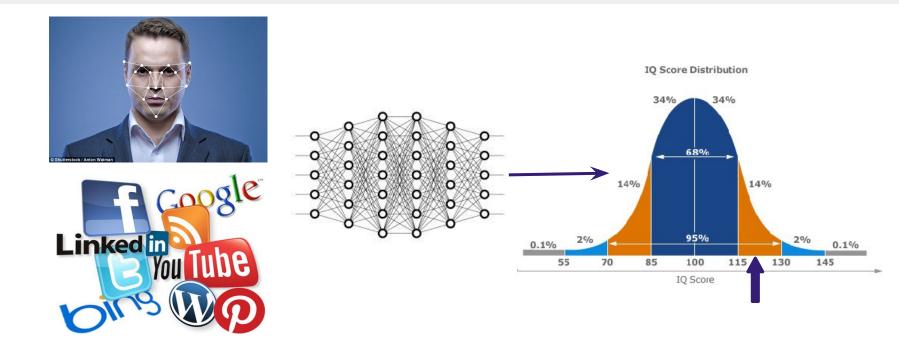




- → Ethics is inner guiding, moral principles, and values of people and society
- → There are gray areas. We often don't have easy answers.
- → Ethics changes over time with values and beliefs of people
- → Legal  $\neq$  Ethical



## The IQ dilemma



→ Intelligence Quotient: a number used to express the apparent relative intelligence of a person



## The IQ dilemma

We can train a classifier to predict people's IQ from their photos & texts. Let's discuss whether it is ethical to build such a technology and what are the risks.

• Who could benefit from such a classifier?

## The IQ dilemma: the ethics of the research question

We can train a classifier to predict people's IQ from their photos & texts. Let's discuss whether it is ethical to build such a technology and what are the risks.

- Who could benefit from such a classifier?
- Let's assume for now that the classifier is 100% accurate.
   Who can be harmed from such a classifier? How can such a classifier be misused?

# The IQ dilemma: understanding the risks

We can train a classifier to predict people's IQ from their photos & texts. Let's discuss whether it is ethical to build such a technology and what are the risks.

- Who could benefit from such a classifier?
- Who can be harmed from such a classifier? How can it be misused?
- What are the pitfalls/risks in the current solution?
  - Example: Our test results show 90% accuracy
    - We found out that white females have 95% accuracy
    - People with blond hair under age of 25 have only 60% accuracy

## The IQ dilemma: understanding the responsibility

We can train a classifier to predict people's IQ from their photos & texts. Let's discuss whether it is ethical to build such a technology and what are the risks.

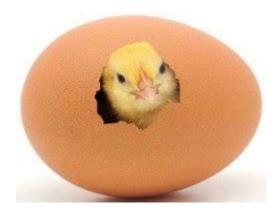
- Who could benefit from such a classifier?
- Who can be harmed from such a classifier? How can it be misused?
- What are the pitfalls/risks in the current solution?
- Who is responsible?
  - Researcher/developer? Advisor/manager? Reviewer? The IRB? The University? Society as a whole?

# We need to be aware of real-world impact of our research and understand the relationship between ideas and consequences

# IQ classifier - risks

- Research question is problematic: attempts to predict IQ are done to approximate intelligence and future success, but IQ is not a good proxy
- IQ tests are known to be racially and socio-economic status (SES)-biased
- Also, the data used to train an IQ classifier will likely have many biases
- NLP systems are likely to pick up on these biases and spurious correlations between intelligence metrics and linguistic features of racial or SES groups
- Error in such a classifier can have direct negative impact on people







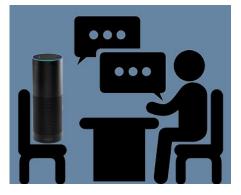
Undergrad NLP 2022

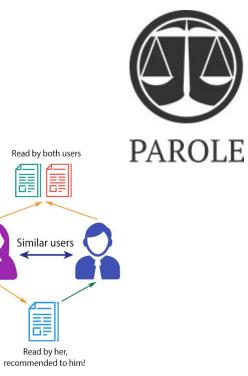


# AI and people



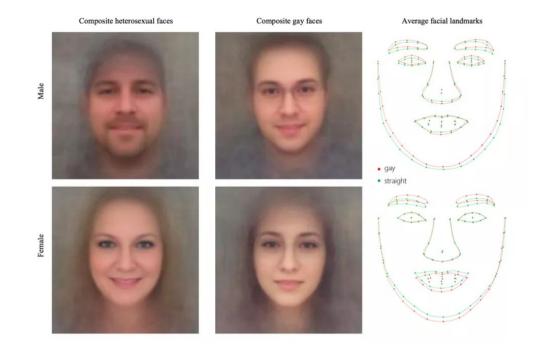








## A recent study: the "AI Gaydar", 2017



# A recent study: the "AI Gaydar"

- Research question
  - Identification of sexual orientation from facial features
- Data collection
  - Photos downloaded from a popular American dating website
  - 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly
- Method
  - A deep learning model was used to extract facial features + grooming features; then a logistic regression classifier was applied for classification
- Accuracy
  - $\circ$   $\,$  81% for men, 74% for women
- Motivation for the study: expose a threat to the privacy and safety of gay men and women



## Let's discuss...

- Research question
  - Identification of sexual orientation from facial features
- Data collection
- a collection Photos downloaded from a popular wherean dating website
  - 35,326 pictures of 14,776 percele, all white, with gay and straight, male and female, all represented evenly
- Method
  - A deep learning model was used to extract facial features + grooming features; then a logistic regression classifier was applied for classification
- Accuracy
  - 81% for men, 74% for women



## Questioning the ethics of the research question

• Identification of sexual orientation from facial features



## Sexual orientation classifier - who can be harmed?

- In many countries being gay person is prosecutable (by law or by society) and in some places there is even death penalty for it
- It might affect people's employment; family relationships; health care opportunities;
- Personal attributes like gender, race, sexual orientation, religion are social constructs. They can change over time. They can be non-binary. They are private, intimate, often not visible publicly.
- Importantly, these are properties for which people are often discriminated against.

# Dual framing in predictive analytics



"We live in a dangerous world, where harm doers and criminals easily mingle with the general population; the vast majority of them are unknown to the authorities.

As a result, it is becoming ever more challenging to detect anonymous threats in

public places such as airports, train stations, government and public buildings and

border control. Public Safety agencies, city police department, smart city service providers and other law enforcement entities are increasingly strive for Predictive Screening solutions, that can monitor, prevent, and forecast criminal events and public disorder without direct investigation or innocent people interrogations. "



## Data

- Photos downloaded from a popular American dating website
- 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly



## Data privacy

• Photos downloaded from a popular American dating website



## Data privacy

• Photos downloaded from a popular American dating website

Questions to ask:

- Is it legal to use the data?
- However, legal ≠ ethical. Who gave consent? Even if the data is public, public ≠ publicized. Does the action of publicizing the data violate social contract?



## Data

- Photos downloaded from a popular American dating website
- 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly



## Data biases

• 35,326 pictures of 14,776 people, all white, with gay and straight, male and female, all represented evenly

Questions to ask:

- Is the dataset representative of diverse populations? What are gaps in the data?
  - Only white people who self-disclose their orientation, certain social groups, certain age groups, certain time range/fashion; the photos were carefully selected by subjects to be attractive
- Is label distribution representative?
  - The dataset is balanced, which does not represent true class distribution.
- $\rightarrow$  this dataset contains many types of biases



## Method

• A deep learning model was used to extract facial features + grooming features; then a logistic regression classifier was applied for classification

# Algorithmic biases

• A deep learning model was used to extract facial features + grooming features; then a logistic regression classifier was applied for classification

Questions to ask:

- Does model design control for biases in data and confounding variables?
- Does the model optimize for the true objective?
- There is a risk in using black-box model which reasons about sensitive attributes, about complex experimental conditions that require broader world knowledge. Does the model facilitate analyses of its predictions?
- Is there analysis of model biases?
- Is there bias amplification?
- Is there analysis of model errors?
   Yulia Tsvetkov



## **Evaluation**

• Accuracy: 81% for men, 74% for women



## The cost of misclassification





## The cost of misclassification



## Learn to assess AI systems adversarially

- Ethics of the research question
- Impact of technology and potential dual use: Who could benefit from such a technology? Who can be harmed by such a technology? Could sharing data and models have major effect on people's lives?
- **Privacy**: Who owns the data? Published vs. publicized? User consent and implicit assumptions of users how the data will be used.
- Bias in data: Artifacts in data, population-specific distributions, representativeness of data.
- Social bias & unfairness in models: How to control for confounding variables and corner cases? Does the system optimize for the "right" objective? Does the system amplify bias?
- Utility-based evaluation beyond accuracy: FP & FN rates, "the cost" of misclassification, fault tolerance.

## Beyond decision-support tools and human-centered analytics

#### Gender/race bias in NLP

- Machine translation (Douglas'17, Prates et al. '19)
- Caption generation (Burns et al.'18)
- Speech recognition (Tatman'17)
- Question answering (Burghardt et al. '18)
- Dialogue systems (Dinan et al.'19)
- Sentiment Analysis (Kiritchenko & Mohammad'18)
- Language Identification (Blodgett et al.'16, Jurgens et al.'17)
- Text Classification (Dixon et al. '18, Sap et al. '19, Kumar et al. '19)

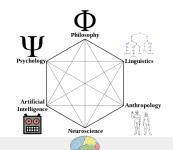
- Language modeling (Lu et al. '18)
- Named-entity recognition (Mehrabi et al. '19)
- Coreference resolution (Zhao et al. '18, Rudinger et al. '18)
- Semantic Role Labelling (Zhao et al. '17)
- SNLI (Rudinger et al. '17)
- Word Embeddings (Bolukbasi et al. '16,Caliskan et al. '17,++)
- ...
- Surveys (Sun&Gaut et al.'19, Blodgett et al.'20, Field et al.'21)

## Why do these issues become especially relevant now?

- Data: the exponential growth of user-generated content
- Technological advancements: machine learning tools have become powerful and ubiquitous

## Topics on ethical and social issues in NLP

- Social bias and algorithmic (un)fairness: social bias in data & NLP models
- Incivility: Hate-speech, toxicity, incivility, microaggressions online
- Misinformation: Fake news, information manipulation, opinion manipulation
- Privacy violation: Privacy violation & language-based profiling
- Technological divide: Unfair NLP technologies underperforming for speakers of minority dialects, for languages from developing countries, and for disadvantaged populations
- Environmental impacts of NLP models



## Recommended introductory readings and talks

- Hovy & Spruit (2016) The Social Impact of NLP
- Barocas & Selbst (2016) Big Data's Disparate Impact
- Barbara Grosz talk (2017) Intelligent Systems: Design & Ethical Challenges
- Kate Crawford NeurIPS keynote (2017) The Trouble with Bias
- Yonatan Zunger blog post (2017) Asking the Right Questions About Al

https://tinyurl.com/Readings-CompEthicsInNLP-2022

## Spring 2023 – Ethics in AI – CSE 582