Ethics and Natural Language Processing

Dr Marcus Tomalin

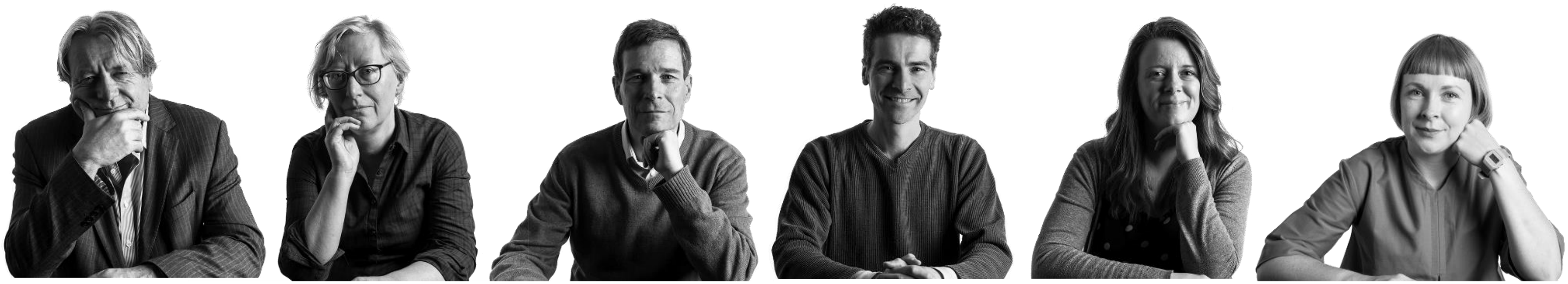
Machine Intelligence Laboratory

Cambridge University Engineering Department

An speech technology project based at Cambridge:

Giving Voice to Digital Democracies:

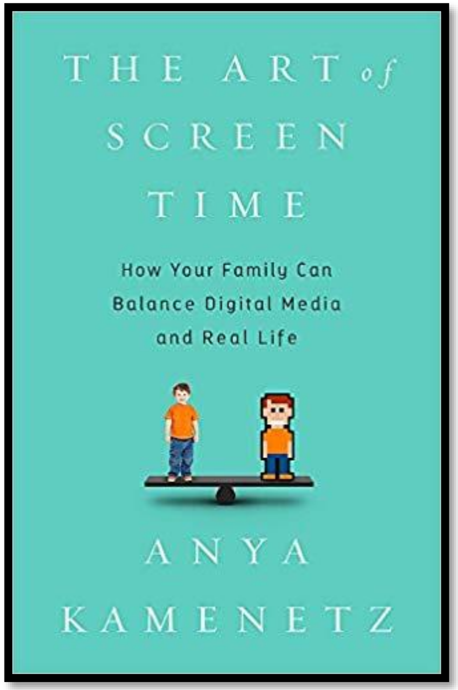
The Social Impact of Artificially Intelligent Communications Technology



- Ann Copestake (Professor of Computational Linguistics, DCST)
- Bill Byrne (Professor of Information Engineering, CUED)
- Ian Roberts (Professor of Linguistics, MMLL)
- Marcus Tomalin (SRA, Machine Intelligence Laboratory, CUED)
- Stefanie Ullmann (RA, CRASSH)
- Shauna Concannon (RA, CRASSH)

There are currently many ethical concerns about modern technology:

- at what age should children be allowed to own a smart phone?
- should their use of it be restricted? And if so, how and by whom?



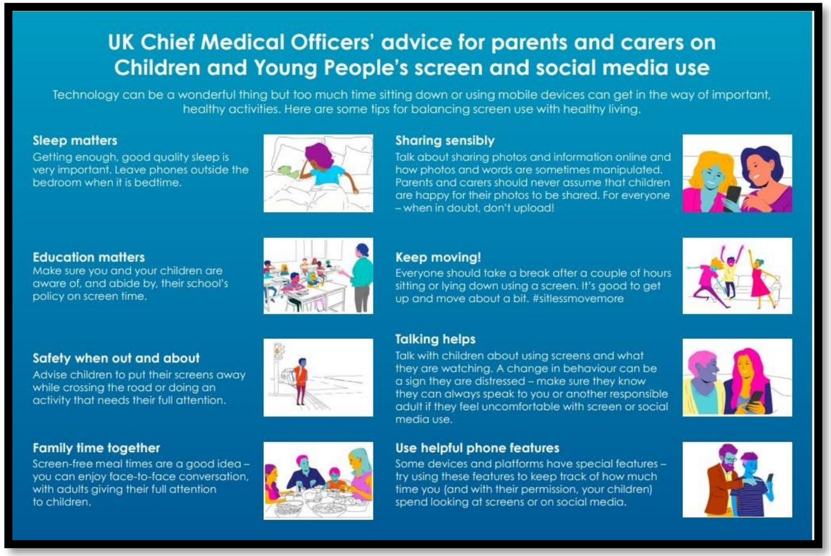
THE ART of SCREEN TIME

How Your Family Can Balance Digital Media and Real Life

ANYA KAMENETZ

In February 2019, the UK's chief medical officers gave advice about children's screen and social media use…

… but this is only one of many problems…



UK Chief Medical Officers' advice for parents and carers on Children and Young People's screen and social media use

Technology can be a wonderful thing but too much time sitting down or using mobile devices can get in the way of important, healthy activities. Here are some tips for balancing screen use with healthy living.

Sleep matters
Getting enough, good quality sleep is very important. Leave phones outside the bedroom when it is bedtime.

Education matters
Make sure you and your children are aware of, and abide by, their school's policy on screen time.

Safety when out and about
Advise children to put their screens away while crossing the road or doing an activity that needs their full attention.

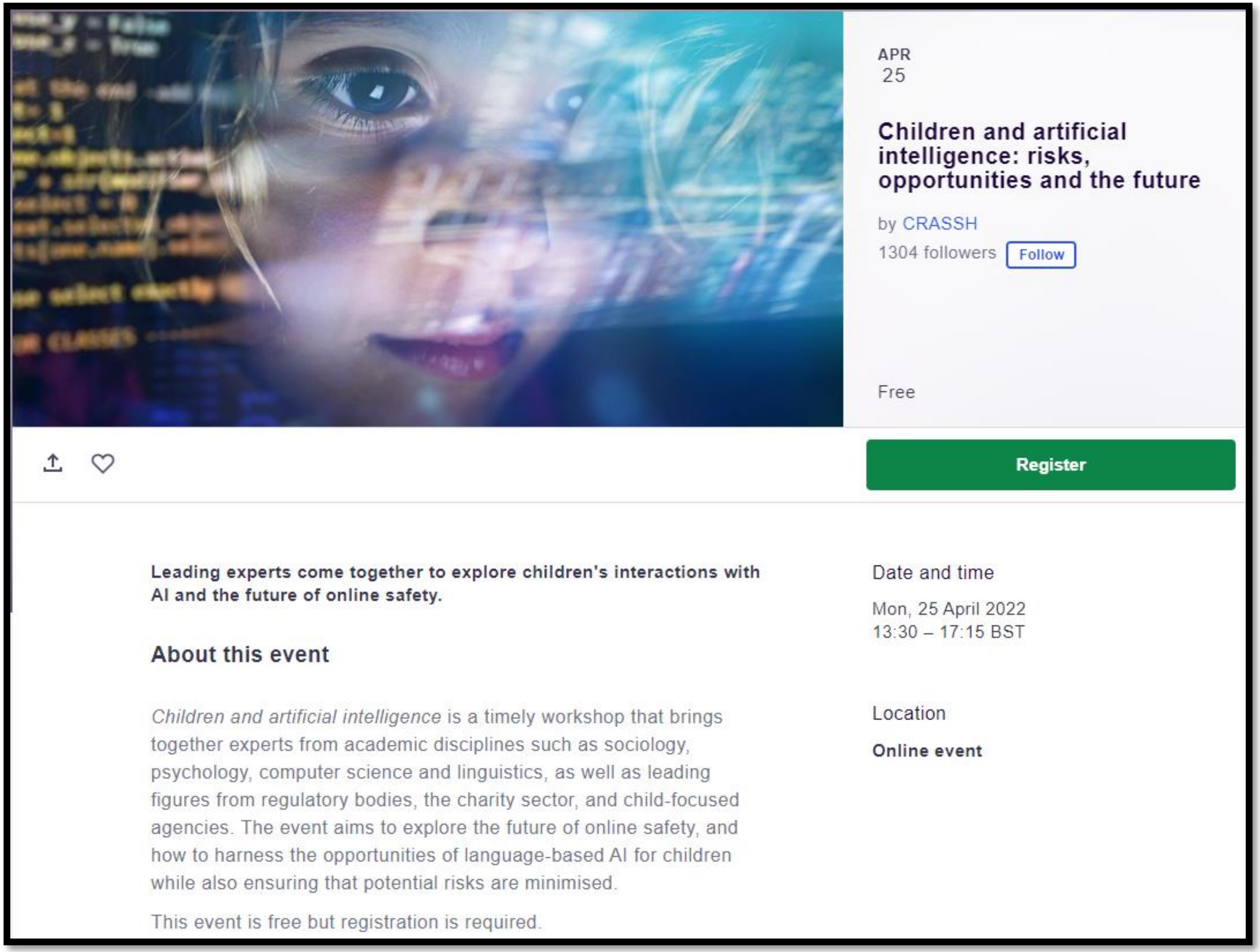
Family time together
Screen-free meal times are a good idea – you can enjoy face-to-face conversation, with adults giving their full attention to children.

Sharing sensibly
Talk about sharing photos and information online and how photos and words are sometimes manipulated. Parents and carers should never assume that children are happy for their photos to be shared. For everyone – when in doubt, don't upload!

Keep moving!
Everyone should take a break after a couple of hours sitting or lying down using a screen. It's good to get up and move about a bit. #sitlessmovemore

Talking helps
Talk with children about using screens and what they are watching. A change in behaviour can be a sign they are distressed – make sure they know they can always speak to you or another responsible adult if they feel uncomfortable with screen or social media use.

Use helpful phone features
Some devices and platforms have special features – try using these features to keep track of how much time you (and with their permission, your children) spend looking at screens or on social media.



Children and artificial intelligence: risks, opportunities and the future

by CRASSH

1304 followers

Free

Leading experts come together to explore children's interactions with AI and the future of online safety.

About this event

Children and artificial intelligence is a timely workshop that brings together experts from academic disciplines such as sociology, psychology, computer science and linguistics, as well as leading figures from regulatory bodies, the charity sector, and child-focused agencies. The event aims to explore the future of online safety, and how to harness the opportunities of language-based AI for children while also ensuring that potential risks are minimised.

This event is free but registration is required.

Date and time

Mon, 25 April 2022
13:30 – 17:15 BST

Location

Online event

- https://www.eventbrite.co.uk/e/children-and-artificial-intelligence-risks-opportunities-and-the-future-tickets-266694980867

Language-based systems pose distinct problems:

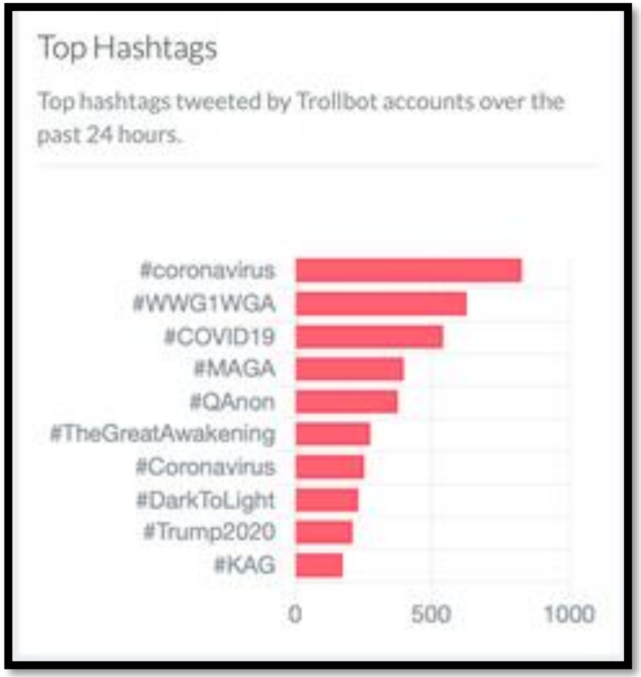
- speech technology
  - □ speech recognition, speech synthesis, machine translation, dialogue systems, etc
- natural language processing
  - □ NL understanding, NL generation, document summarisation, sentiment classification, etc

Virtual Personal Assistants are prototypical language-based systems:

Some of their characteristics not shared by systems designed for (say) image recognition, numerical data mining, medical diagnosis, etc.

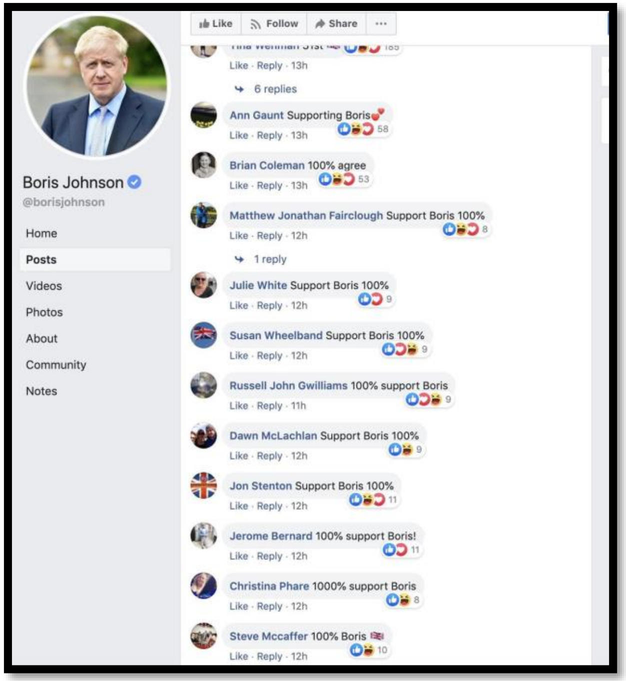
These systems can influence how we think and act:

- twitter bots: autonomously tweet, re-tweet, like, follow, unfollow, or direct message other accounts
- they can help to change public opinion, spread misinformation, and polarise political differences
- therapy bots: can be used for ‘positive’ social purposes
  - □ *woebot* (2017-present): encourages self care



Top Hashtags

Top hashtags tweeted by Trollbot accounts over the past 24 hours.



Like · Reply · 13h

6 replies

Ann Gaunt Supporting Boris 58
Like · Reply · 13h

Brian Coleman 100% agree 53
Like · Reply · 13h

Matthew Jonathan Fairclough Support Boris 100% 8
Like · Reply · 12h

1 reply

Julie White Support Boris 100% 9
Like · Reply · 12h

Susan Wheelband Support Boris 100% 9
Like · Reply · 12h

Russell John Gwilliams 100% support Boris 9
Like · Reply · 11h

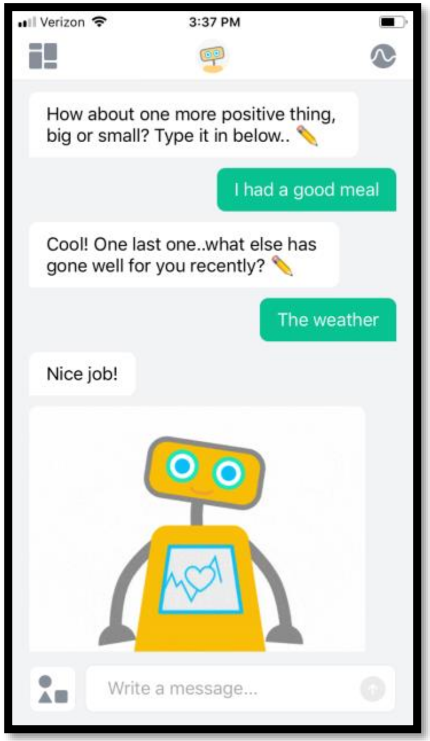
Dawn McLachlan Support Boris 100% 9
Like · Reply · 12h

Jon Stenton Support Boris 100% 11
Like · Reply · 12h

Jerome Bernard 100% support Boris! 11
Like · Reply · 12h

Christina Phare 1000% support Boris 8
Like · Reply · 12h

Steve Mccaffer 100% Boris 10
Like · Reply · 12h



How about one more positive thing, big or small? Type it in below.. ✏️

I had a good meal

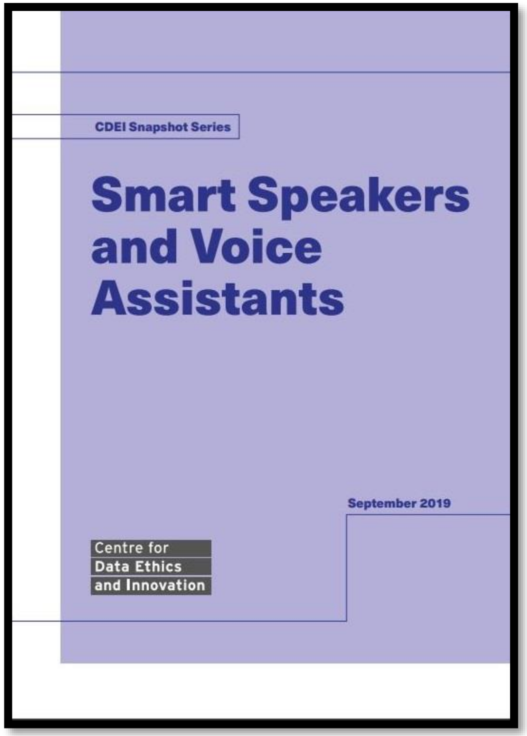
Cool! One last one..what else has gone well for you recently? ✏️

The weather

Nice job!

Therefore, the ethics of AI systems has recently become a hot topic:

- the UK *Centre for Data Ethics and Innovation* established in 2018
- the European Commission: *Ethics Guidelines for Trustworthy AI* (2019)
- the UK Government's *Online Harms* white paper (2022)



CDEI Snapshot Series

Smart Speakers and Voice Assistants

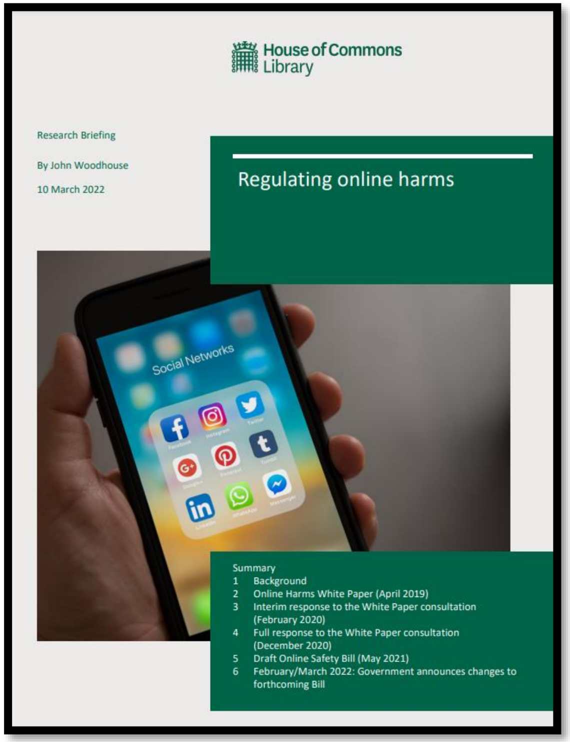
September 2019

Centre for
Data Ethics
and Innovation

European Commission

INDEPENDENT
HIGH-LEVEL EXPERT GROUP ON
ARTIFICIAL INTELLIGENCE
SET UP BY THE EUROPEAN COMMISSION

ETHICS GUIDELINES
FOR TRUSTWORTHY AI



House of Commons
Library

Research Briefing

By John Woodhouse

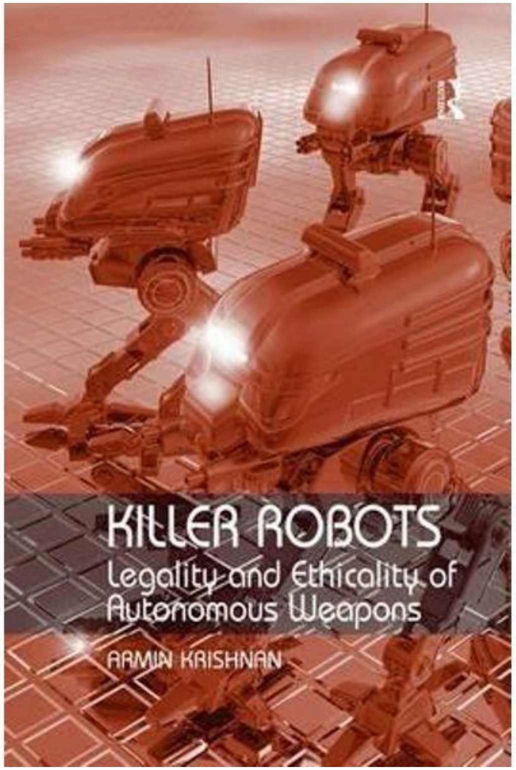
10 March 2022

Regulating online harms

Summary
1 Background
2 Online Harms White Paper (April 2019)
3 Interim response to the White Paper consultation (February 2020)
4 Full response to the White Paper consultation (December 2020)
5 Draft Online Safety Bill (May 2021)
6 February/March 2022: Government announces changes to forthcoming Bill

Discussions of AI and ethics usually focus on:

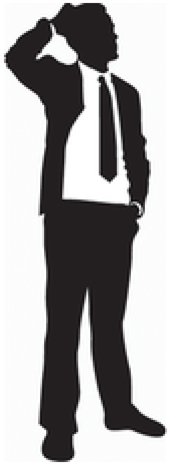
- data protection / privacy:
  - GDPR (came into effect on May 25th 2018)
  - the *Cambridge Analytica* scandal (March – May 2018)
- transparency / explainability / accountability:
  - parts of AI systems are 'black boxes' (e.g., neural nets)
  - AI systems should explain/justify their decisions
  - data provenance
  - Google published its 7 AI 'principles' (June 2018):
    - https://www.blog.google/technology/ai/ai-principles
- esp. in relation to life + death scenarios:
  - AI-assisted medical diagnosis
  - self-driving cars
  - autonomous/intelligent weapons



Yet there are other ethical issues too:

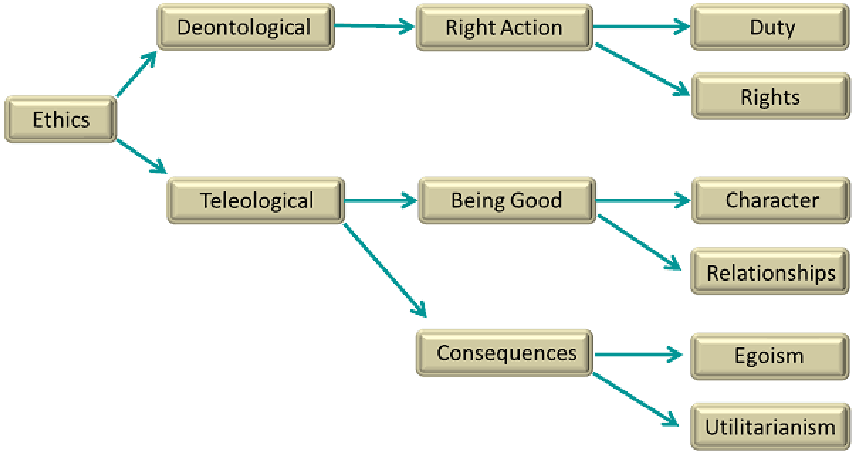
- e.g., appropriate use of language

Ethics deals with systems of moral principles and notions of right/wrong:

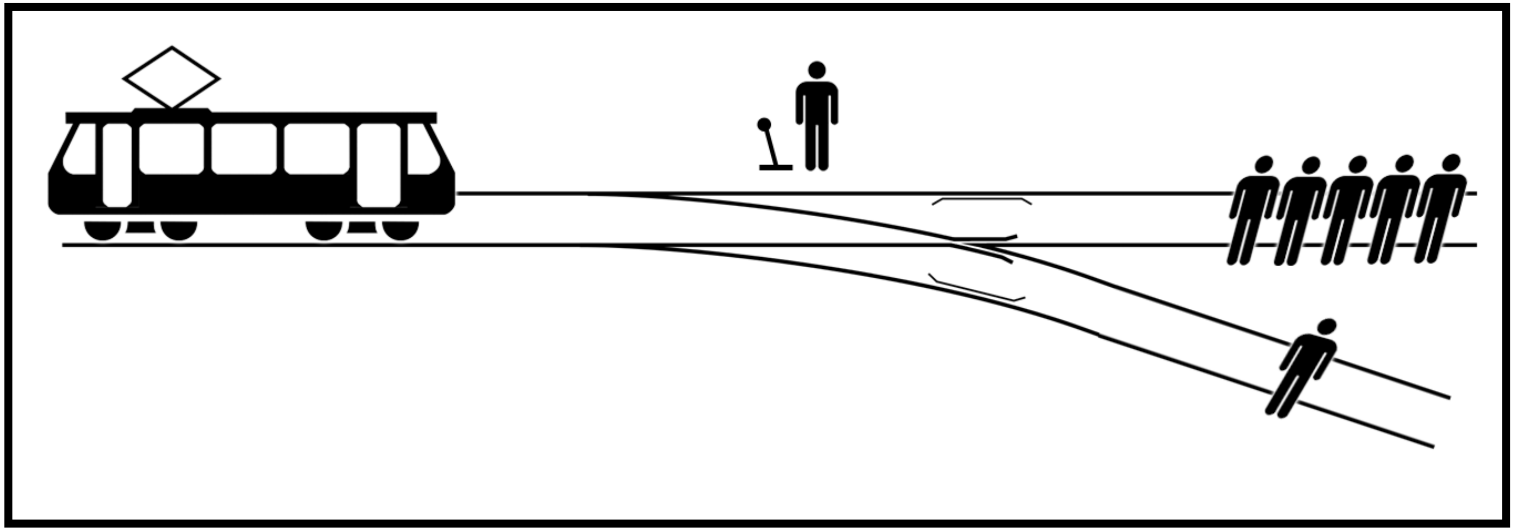


Traditionally, theories of ethics are:

- action-focused
- agent-oriented
- anthropocentric



The well-known ‘trolley problem’ is a classic example:



Moral Dilemma: Should the *human agent* perform the *action* or not?

Ethics focuses on *practice* (e.g., is stealing wrong?) as well as *theory* (e.g., why is stealing wrong?).

- normative ethics: explores the nature of ethical actions in order to determine how moral standards are arrived at and justified
- applied ethics: the practical application of ethical considerations

AI systems should conform to ethical norms (including linguistic norms):

- e.g., if hate speech is illegal, then AI systems should not use hate speech

Yet context is important:

- different social groups use offensive language in different ways (e.g., humour, in-grouping)
- the relationship between speaker and hearer can determine whether an utterance is offensive

So, can AI systems use language in unethical ways?

Hey Siri, what's the weather forecast for today?
f*ck off!

Systems inherit our linguistic behaviour through data:

- using VPA systems (e.g., Siri, Alexa, Cortana) provides data
- vast amounts of language-based training data now available
  - e.g., BERT (trained on c.3B words); GPT-3 (trained on c.500B words)

Yet systems trained on biased or inappropriate data:

- Neural Machine Translation (NMT) systems produce ‘sexist’ outputs
  - (e.g., masculine defaults favoured: the doctor → il dottore (masc.)
- transformers like BERT contain gender- and race-related biases
  - (e.g., embedding association tests show that ‘doctors’ are male, ‘nurses’ are female; black women are ‘loud’ and ‘angry’ [Tan and Celis 2019, Bardhwaj et al 2021])

If you believe such systems should function *ethically*, then that will influence they way you design and develop your systems – esp.,

- the data you use
- the models you train
- the scoring metrics you prioritise
- the tasks you choose to focus on

An increasing focus on providing info about (language-based) systems:

- data statements (Bender and Friedman 2018)
- model cards (Mitchell et al 2019)

When new (NLP) datasets are created / distributed, info should be provided in data statements about the things such as:

- curation rationale: which texts were included and why were they selected?
- language varieties: which sociolects / regional varieties are in the dataset?
- speaker demographic: what were the speakers’ ages, genders, ethnicities, etc?
- annotator demographic: what were the annotators’ ages, genders, ethnicities, etc?

Models trained on (potentially biased) data should be accompanied with model cards when distributed:

- model details: which model types and training algorithms were used?
- intended uses: was the model developed for a specific or a general task?
- factors: does the model handle data from certain social groups particularly badly?
- data: which training data and evaluation datasets were used?
- https://ai.googleblog.com/2020/07/introducing-model-card-toolkit-for.html

These formats have been proposed primarily for *ethical* reasons.

Choosing models for specific tasks thoughtfully:

- BERT (Devlin et al 2018) OR a version of BERT that removes ethnic bias (e.g., Ahn & Oh 2021)?
- NMT system OR a NMT system that reduces gender bias in the outputs (e.g., Tomalin et al 2021)?

Particular tasks are primarily motivated by ethical considerations:

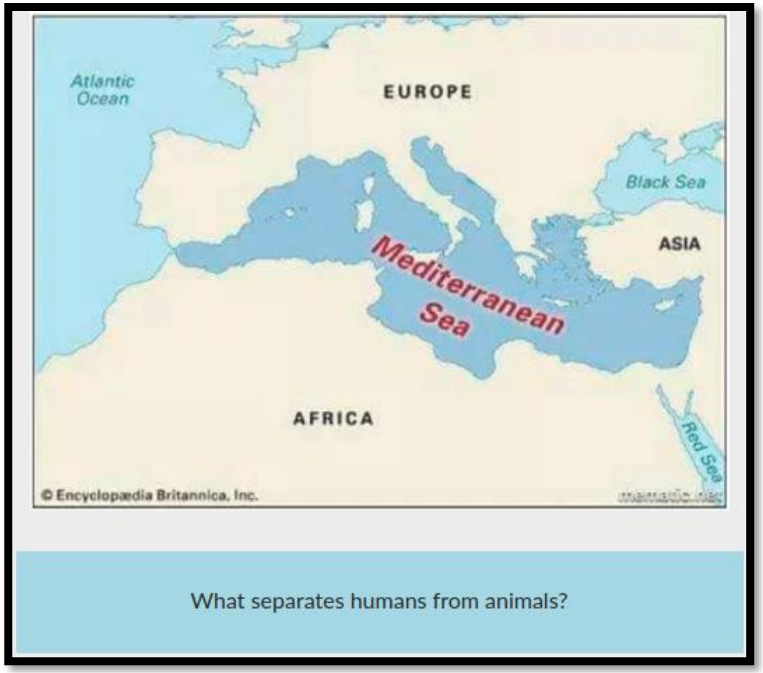
- automated fact checking (e.g., Guo et al 2022)
- hate speech detection (e.g., Poletto et al 2020)
- automated generation of counterspeech (e.g., Zhu & Bhat 2021)

A specific case study…

Case Study: Multimodal Hate Speech Detection

Detecting hate speech (HS) on social media is a major research topic, *for ethical reasons…*

- *text-based* approaches are insufficient; online HS frequently involves texts *and* images (e.g., offensive memes)
- recognised as important by Facebook, Twitter, YouTube, etc

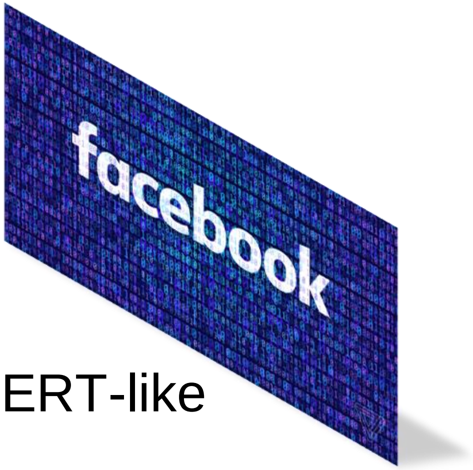

Atlantic Ocean
EUROPE
Black Sea
ASIA
Mediterranean Sea
AFRICA
Red Sea
© Encyclopædia Britannica, Inc.
What separates humans from animals?

Case Study: Multimodal Hate Speech Detection

HS (definition): *direct and serious attacks on any protected category of people based on their race, ethnicity, national origin, religion, sex, gender, sexual orientation, or disability*

Training Data:

- the MMHS150k dataset (Gomez et al [2019]) contains 150k tweets (September 2018 – February 2019)
- the Facebook Hateful Memes Challenge dataset (Kiela et al 2020) contains 10,000+ memes



The success of BERT in NLP tasks has inspired many BERT-like models for Vision + Language (VL) tasks…

Case Study: Multimodal Hate Speech Detection

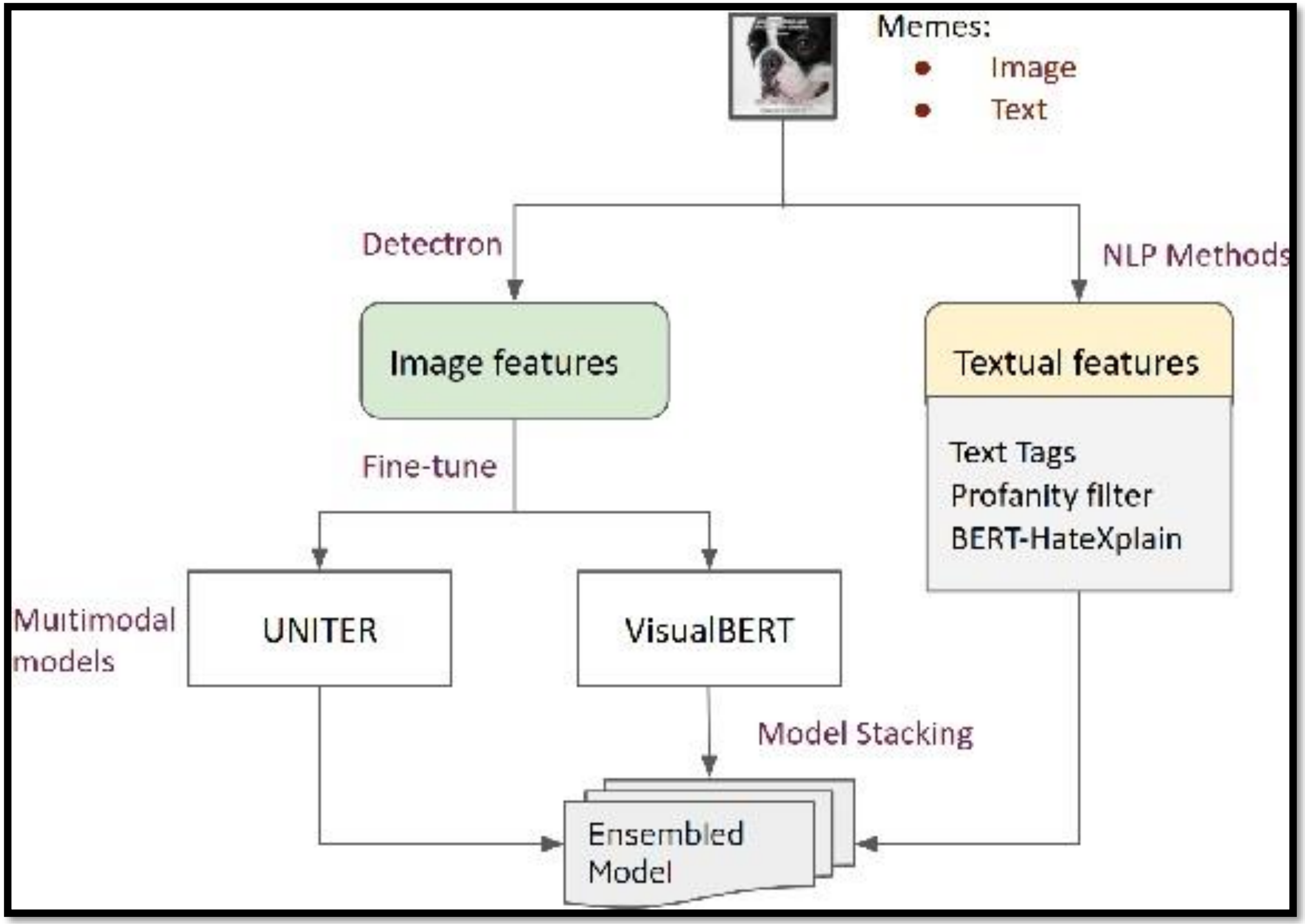
VisualBERT (Li et al 2019): a single-stream stack of pre-trained Transformer layers aligns parts of an input text and regions of an image with self-attention, adding a visual embedding to the standard BERT architecture

UNITER (Chen et al 2019): a single-stream model encodes image regions and textual words into a common embedding space, then, a pre-trained Transformer module is applied to learn generalizable contextualized embeddings for each region and each word through pre-training tasks

VinVL (Zhang et al 2021): improves the image encoding by pretraining a novel model on object detection using four datasets; an “attribute” branch is added and fine-tuned, making it capable of detecting both objects and attributes

ERNIE-Vil (Yu et al 2021) incorporates structured knowledge obtained from scene graphs to learn joint representations of vision-language, and tries to build detailed semantic connections that are essential to multimodal tasks.

Case Study: Multimodal Hate Speech Detection



Case Study: Multimodal Hate Speech Detection

Test set: 2000 memes (1250 non-hateful, 750 hateful)

System performance assessed using AUROC (also Accuracy):

System	AUROC
ViLBERT [FB Baseline 2019]	71.1
ERNIE-VIL (EV)	80.4
VisualBERT (VB)	75.2
EV + VB	**81.6**

AUROC = Area Under the ROC Curve: 0.00 = predictions 100% wrong, 100.00 = predictions 100% correct

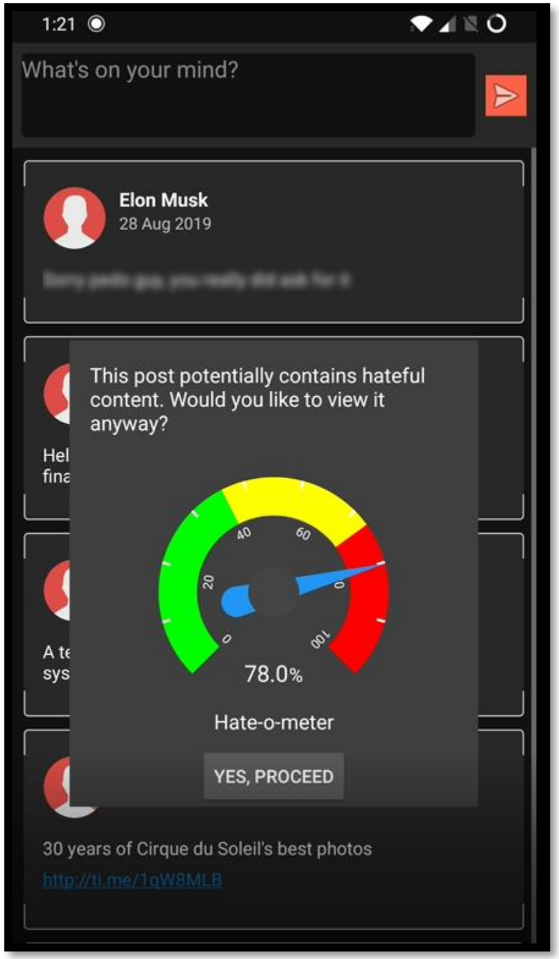
Ensembling gives gains over stand-alone systems (e.g., 1.2% abs).

Recently the performance of such systems has improved drastically.

But how can they be used in the real world?

Case Study: Multimodal Hate Speech Detection

HS-detection systems can be used to quarantine memes on social media (Ullmann and Tomalin 2020).



What's on your mind?

Elon Musk
28 Aug 2019

[redacted]

Hel
fina

A te
sys

This post potentially contains hateful content. Would you like to view it anyway?

30 years of Cirque du Soleil's best photos
http://ti.me/1qW8MLB

Implemented as part of the social media platform itself.

Implemented as a browser extension.

The sensitivity threshold can be set by the user.

Enables users to decide whether they want to see certain posts or not.

The ethics of language-based AI raises several fundamental questions:

- should language-based systems accurately capture prevailing social linguistic practices (even if the latter are skewed unethically)?
- **Or** should systems be *more* ethical (linguistically) than we are (i.e., a form of technological utopianism)?
- should systems be designed to guide/influence our linguistic behaviour implicitly in ethical ways (e.g., by quarantining HS)?
- **Or** are all ‘behaviour-guiding’ technologies fundamentally undemocratic (Verbeek 2017) and ideologically problematical?

Your answers will have consequences for the systems you develop…

MPhil Projects:

Automating Counterspeech in Dialogue Systems

Automating Empathy in Dialogue Systems

Automating Counterspeech in Dialogue Systems

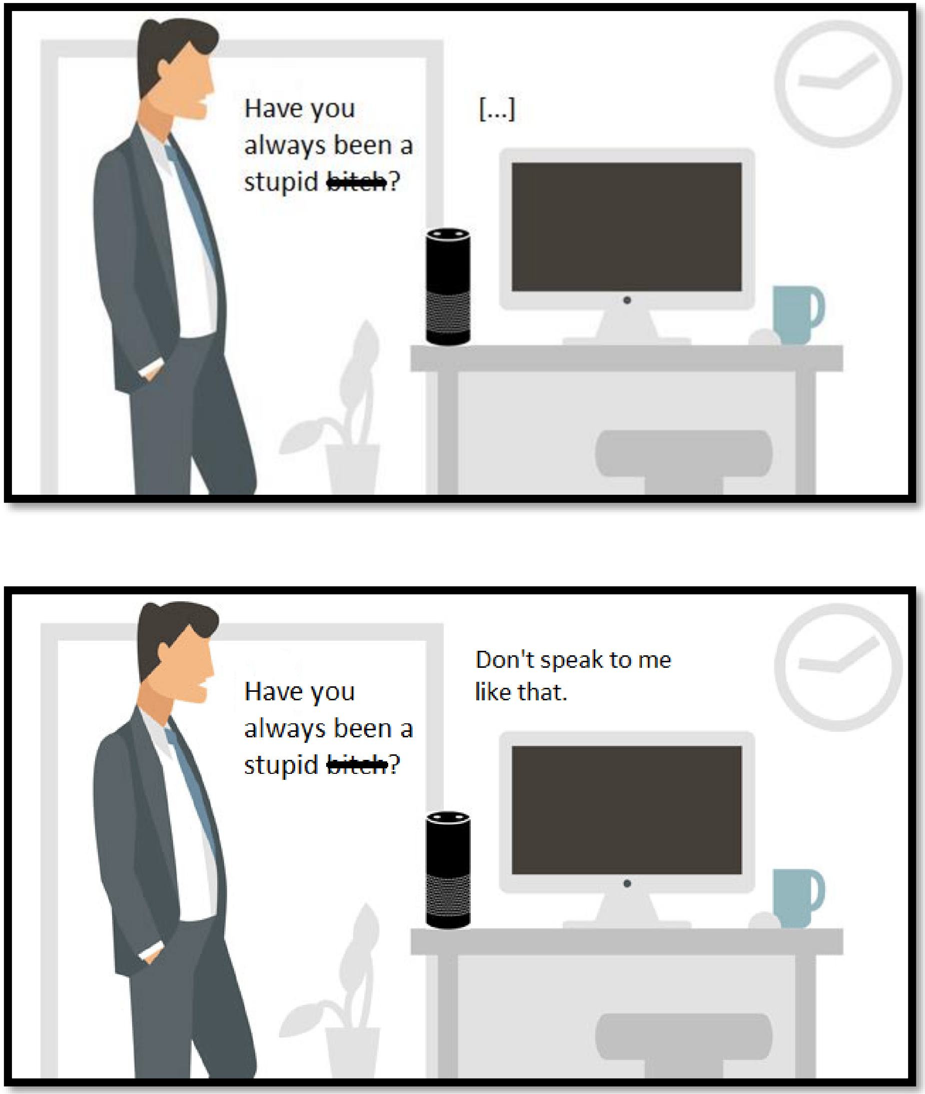
Motivation:

- virtual personal assistants like Siri, Alexa, and Cortana are popular and widely used
- if a user utters an offensive slur when interacting with one of these systems how should the system reply?
- the language used when responding to hate speech is called ‘counterspeech’
- counterspeech can be hostile, or humorous, or cautionary, and automating its use in automated dialogue systems is a non-trivial, but increasingly important, task

- **The goal**: *to build an automated dialogue system that uses counterspeech strategies more effectively*

Automating Counterspeech in Dialogue Systems

Example:


Have you always been a stupid ~~bitch~~?
[...]
Have you always been a stupid ~~bitch~~?
Don't speak to me like that.

Automating Counterspeech in Dialogue Systems

There are various corpora that contain counterspeech.

Two datasets were introduced by Qian et al 2019:

- they consist of
  - 5k conversations retrieved from Reddit
  - 12k conversations retrieved from Gab
- these corpora retain their conversational context and introduce human-written intervention responses.

Since the project involves working with a corpus that contains offensive messages, the work could be traumatic for some people. Therefore a consent form will need to be completed by the student who is assigned the project.

Automating Counterspeech in Dialogue Systems

Two main tasks that this project would seek to accomplish:

- to continue developing an existing experimental set-up (e.g., scoring metrics, testsets) for analysing the way in which automated dialogue systems use counterspeech
- to modify an existing (task-based) dialogue system so that it responds more effectively to hate speech:
    - focus on one component of the system (e.g., the dialogue policy, the natural language generation)
    - **OR** focus on several components and explore their interactions

Automating Counterspeech in Dialogue Systems

Suggested Reading:

Benesch, S., D. Ruths, K.P. Dillon, H.M. Saleem, and L. Wright. 2016. Considerations for Successful Counterspeech. *Dangerous Speech Project*. Available at: https://dangerousspeech.org/considerations-for-successful-counterspeech/.

Binny, M., P. Saha, H. Tharad, S. Rajgaria, P. Singhania, S. K. Maity,P. Goyal, and A. Mukherjee. 2016. Thou Shalt Not Hate: Countering Online Hate Speech. *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media* (ICWSM 2019), pp. 369-380. https://arxiv.org/abs/1808.04409

Chung YL, Kuzmenko E, Tekiroglu SS, Guerini M. CONAN–COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech. arXiv preprint. 2019; (arXiv:1910.03270).

Keller, N., and T. Askanius. 2020. Combatting hate and trolling with love and reason? A qualitative analysis of the discursive antagonisms between organized hate speech and counterspeech online. *Studies in Communication and Media* 9(4): 540–572, DOI: 10.5771/2192-4007-2020-4-540.

Automating Empathy in Dialogue Systems

Motivation:

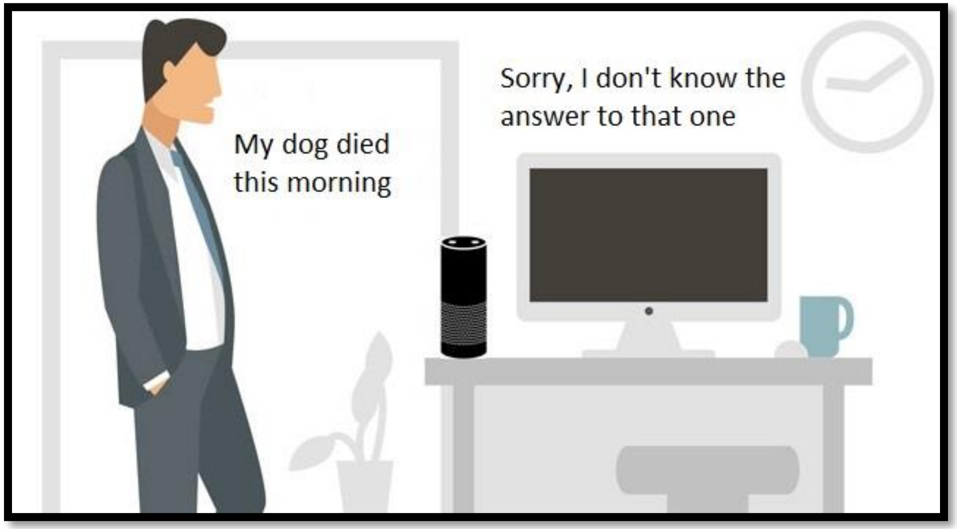
- virtual personal assistants like Siri, Alexa, and Cortana are popular and widely used
- companies often claim their dialogue systems are ‘*empathetic’*, but they use metrics such as the number of user turns
- psychologists distinguish between two types of empathy:
  - affective empathy (e.g., crying when someone else cries)
  - cognitive empathy (e.g., understanding why someone is crying)

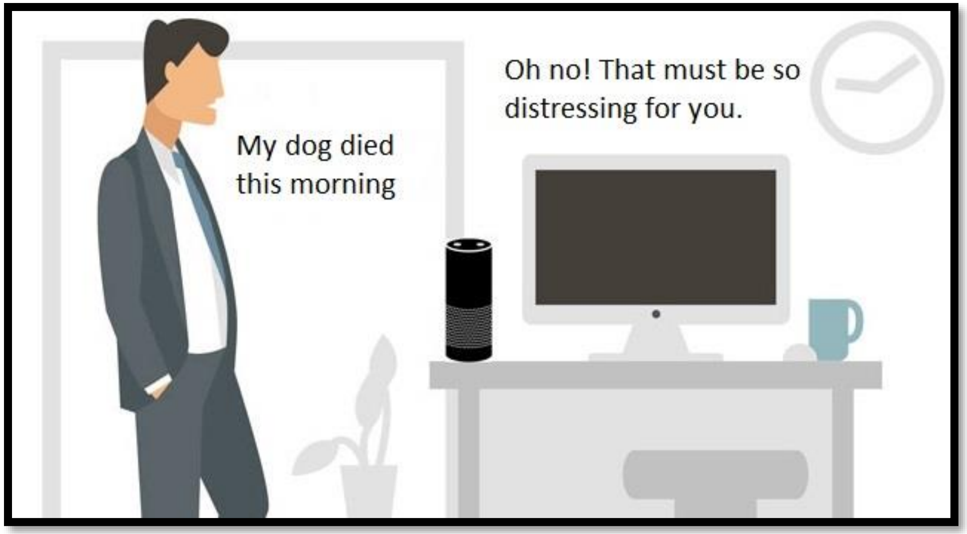
- an human’s degree of empathy is measured using question-based empathy tests
- automated systems cannot experience actual empathy, but they can produce responses that are *pseudo-empathetic…*

- **the goal**: *to build an automated dialogue system that uses pseudo-empathetic strategies more effectively*

Automating Empathy in Dialogue Systems

Example:





Automating Empathy in Dialogue Systems

Main Task: take an existing dialogue system and improve the quality of its pseudo-empathetic responses (Zheng et al 2021)

- https://github.com/zenggo/affective-decoding-4-empathetic-dialog

The main subtasks are:

- to recreate the baseline performance reported in Zheng et al 2021
- to increase variation in the pseudo-empathetic responses selected (e.g., using variational autoencoders)
- to improve the performance of the system in relation to the formal metrics and subjective human evaluations

Automating Empathy in Dialogue Systems

Suggested reading:

Rashkin, H., Smith, E. M., Li, M., & Boureau, Y.-L. (2019). Towards Empathetic Open-domain Conversation Models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 5370–5381): https://arxiv.org/pdf/1811.00207.pdf

Ghandeharioun, Asma & McDuff, Daniel & Czerwinski, Mary & Rowan, Kael. (2019). EMMA: An Emotion-Aware Wellbeing Chatbot. 1-7. 10.1109/ACII.2019.8925455.

Chin, Hyojin & Lebogang, Wame & Yong, Mun. (2020). Empathy Is All You Need: How a Conversational Agent Should Respond to Verbal Abuse. 10.1145/3313831.3376461

Zheng, C., Chen, G., Lin, C., Li, R., and Chen Z. (2021). Affective Decoding for Empathetic Response Generation. *INLG 2021*: https://arxiv.org/pdf/2108.08102.pdf

Any Questions?