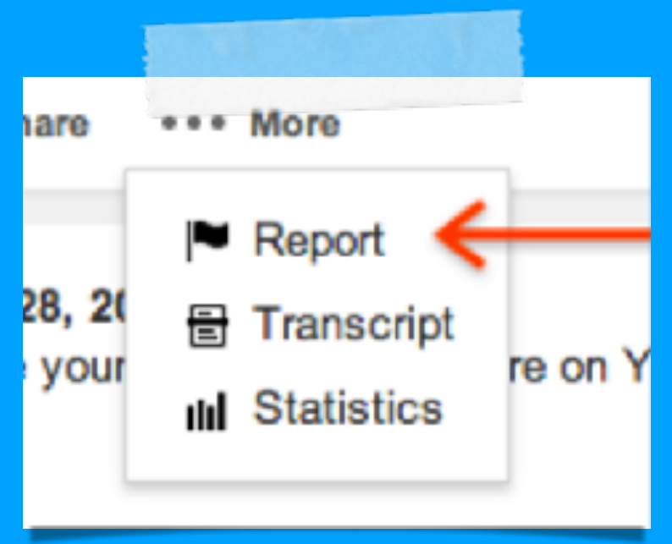
Like trainer, like bot?

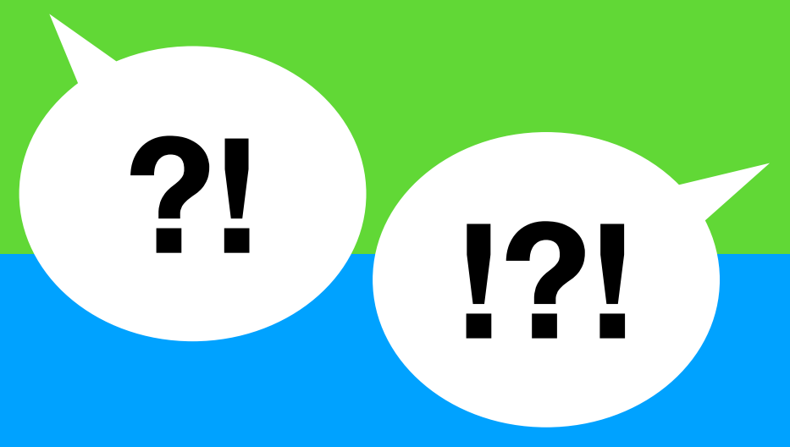
Inheritance of bias in algorithmic content moderation

Reuben **Binns***, Michael **Veale**+, Max **Van Kleek***, Nigel **Shadbolt***

*Department of Computer Science, University of Oxford, [reuben.binns|max.van.kleek|nigel.shadbolt@cs.ox.ac.uk]
+Department of Science, Technology, Engineering & Public Policy (STEaPP), University College London [m.veale@ucl.ac.uk]



Participation in internet communities can be dampened by personal attacks, offensive comments—not to mention 'trolls'.



Manual moderation for deletion or prioritisation is costly at scale, so increased interest in algorithmic, often machine learning–based, comment flagging 🤖🚩🤖🚩🤖🚩

Yet norms are far from agreed or universal. If trained by users or crowd-workers, whose norms enter these censorship systems? How might we explore the values at play?

Data: *Wikipedia Detox Project*

100k Wikipedia talk page diffs, each annotated by 10 Crowdflower workers for **personal attacks, aggression,** and **toxicity.**

For each annotator, we have self-provided demographic information: age group, gender, education level, first language.

Aims of this study:

- build a methodology for understanding how norms of offence can be explored by practitioners
- Test this methodology on one existing example.

What this study is not:

- a social scientific investigation into generalisable, gendered notions of offence, or offence on Wikipedia

Exploring data before model-building

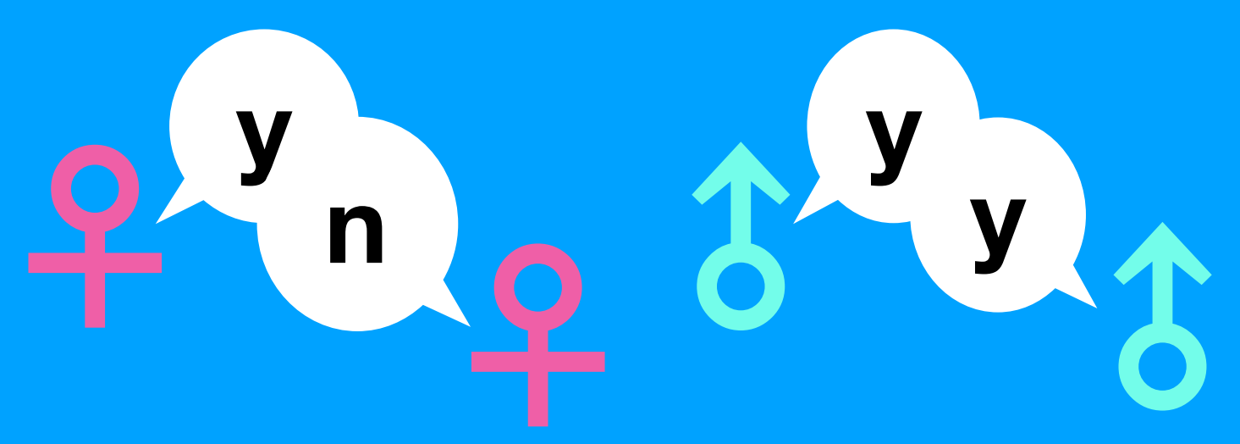
Do different communities, here explored with self-reported gender, have different understandings of offence?

One way of looking at this: how often does a community agree that something is offensive?

Look at Krippendorff's Alpha (bootstrapped 95% CI), an agreement metric designed for missing data, as not all annotators annotate all comments (Hayes and Krippendorff, 2007).

Results

There is evidence of community difference in conception of offence in this dataset: **female annotators tend to disagree about what is offensive significantly more than male annotators.** (468 [.457, .478] vs .494 [.484, .503])



Hayes, A.F., Krippendorff, K.: Answering the call for a standard reliability measure for coding data. Communication Methods and Measures 1(1), 77–89 (2007)

Wulczyn, E., Thain, N., Dixon, L.: Ex machina: Personal attacks seen at scale. In: Proceedings of the 26th International Conference on World Wide Web. pp. 1391–1399. International World Wide Web Conferences Steering Committee (2017)

Analysing trained classifiers

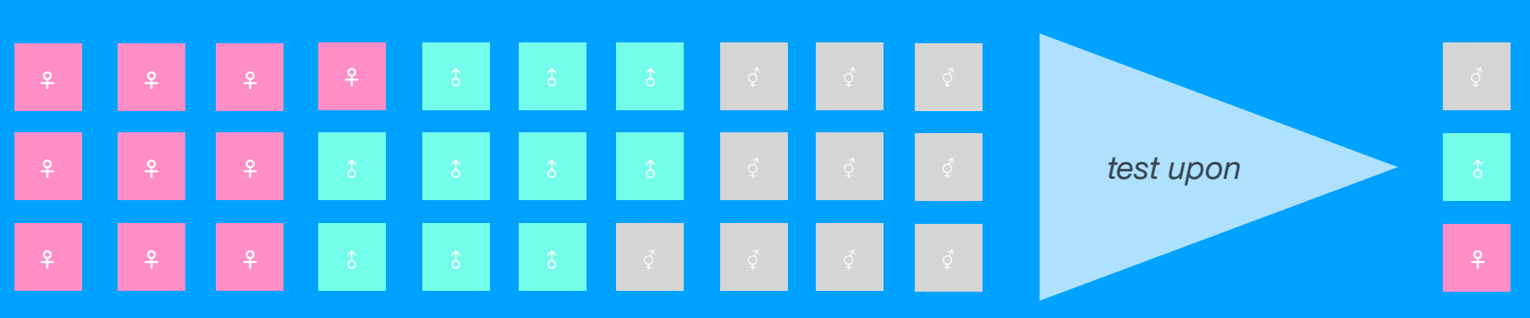
Create **new training sets**: 10 'male-rated', 10 'female-rated' and 10 'balanced' classifiers by making 30 new datasets by sampling the original data with replacement. This is designed to explore the difference the original training context makes. From these datasets, train 30 classifiers following the method of Wulczyn et al. (2017) study from the Wikipedia Detox Project.

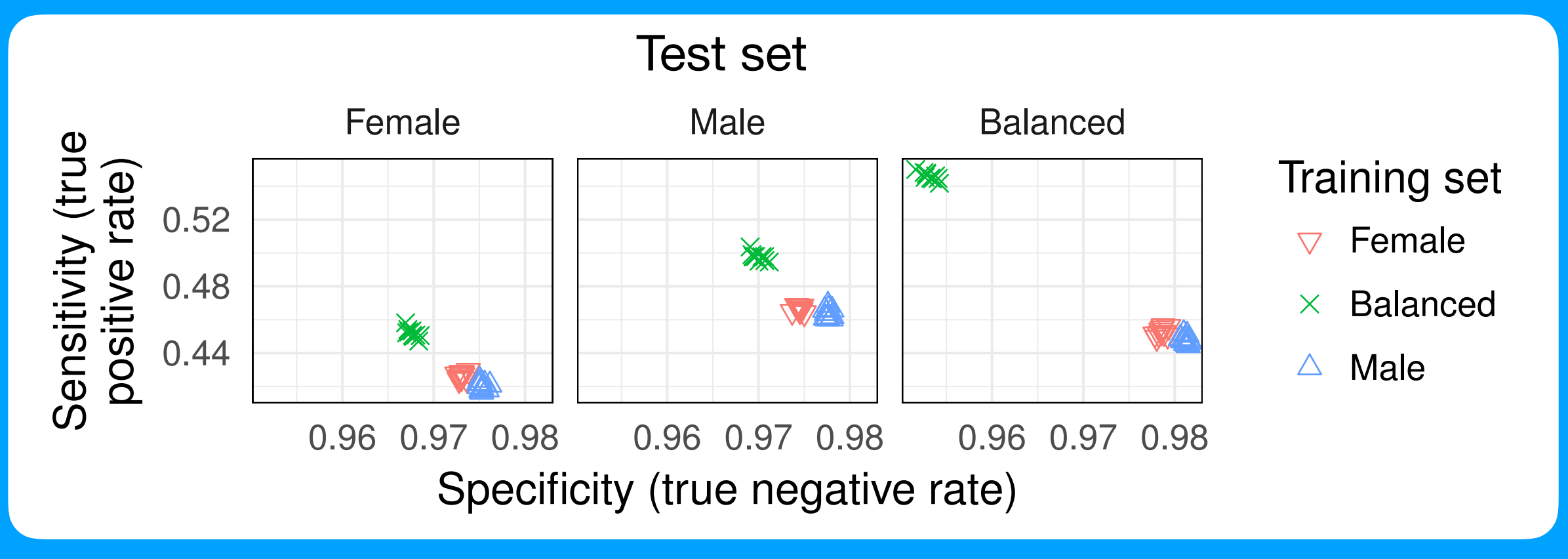
For example, if a given comment was rated by 6 men and 4 women, a new 'female' dataset would see that comment's 10 ratings sampled from the 4 female raters with replacement.

Test those 30 classifiers on **3 test sets**, again with 'male', 'female' and 'balanced'. This is designed to explore the different **deployment context**.

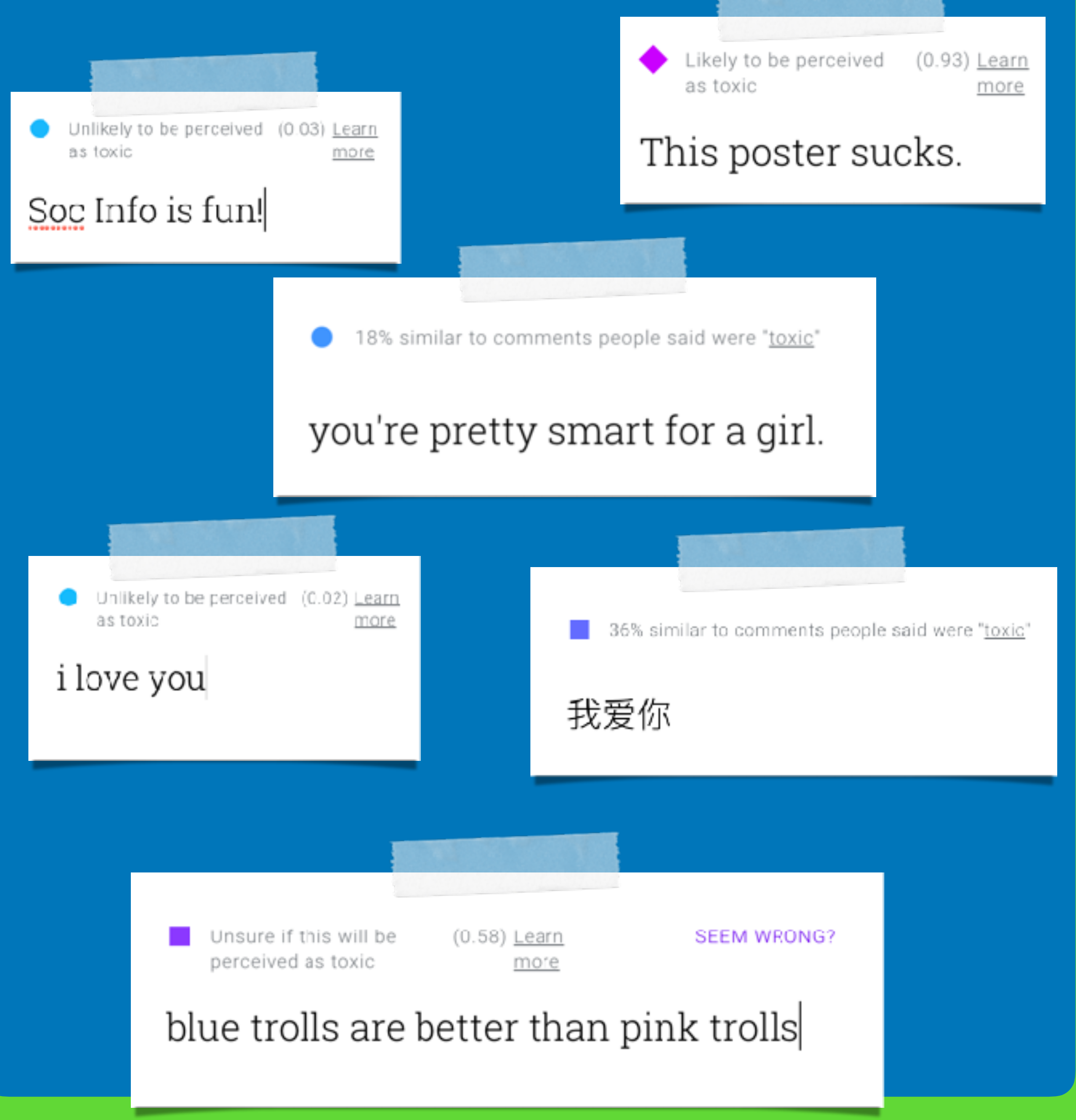
Results

All types of classifiers were less sensitive to female-labelled norms of offence than male or balance labelled norms of offence.





Fun with Perspective API...



Unlikely to be perceived as toxic (0.03) Learn more

Soc Info is fun!

Likely to be perceived as toxic (0.93) Learn more

This poster sucks.

18% similar to comments people said were "toxic"

you're pretty smart for a girl.

Unlikely to be perceived as toxic (0.02) Learn more

i love you

36% similar to comments people said were "toxic"

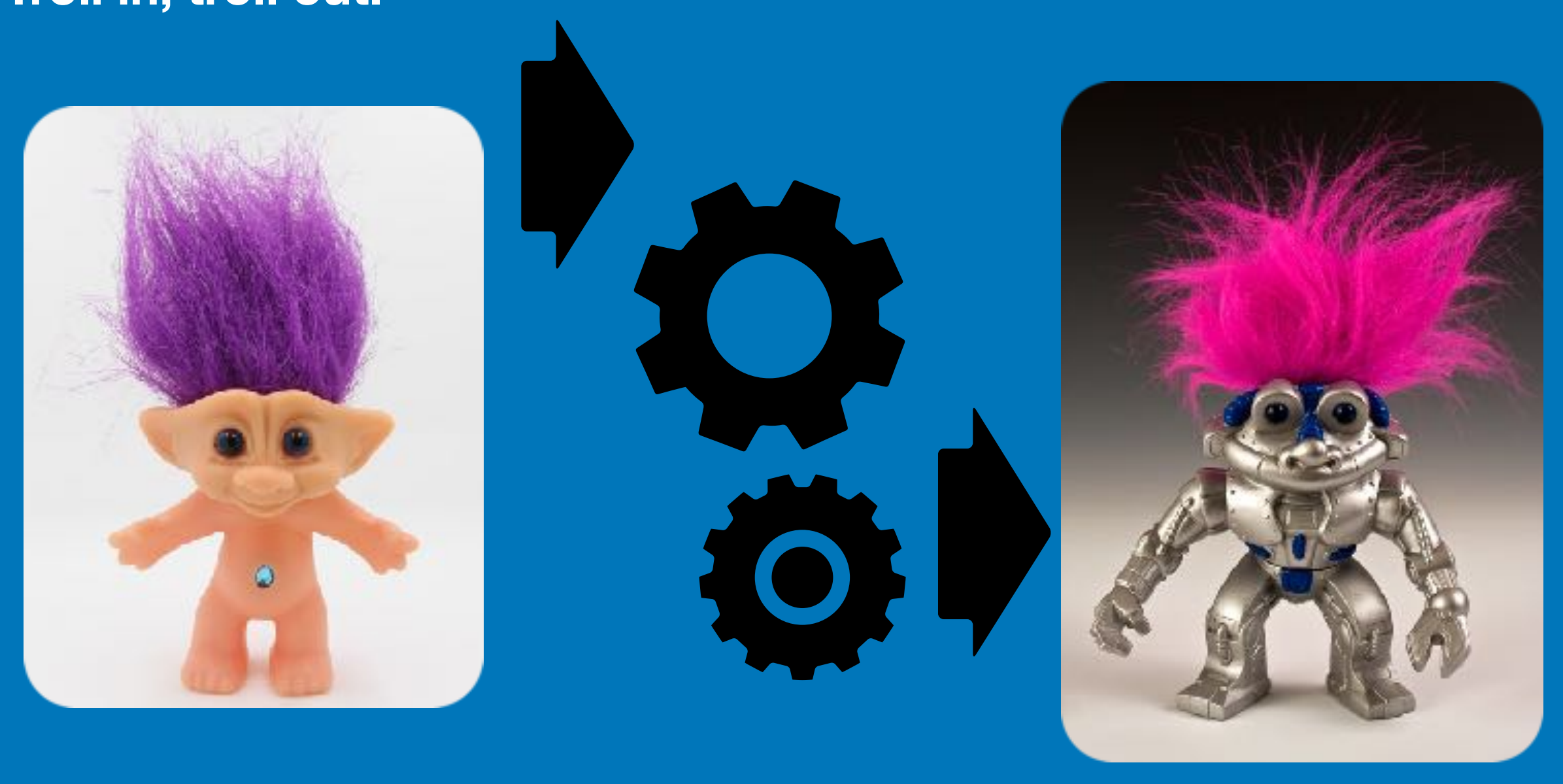
我爱你

Unsure if this will be perceived as toxic (0.58) Learn more

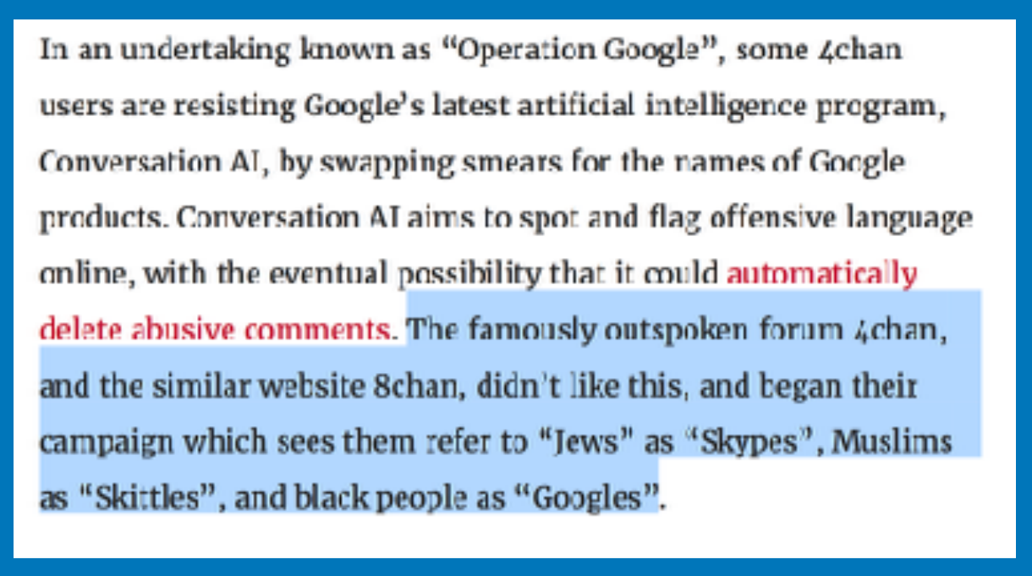
SEEM WRONG?

blue trolls are better than pink trolls

Troll in, troll out.



Communication can be subtle; simple filtering systems are easy to game. Example recently (excerpt from the *New Statesman*)



In an undertaking known as "Operation Google", some 4chan users are resisting Google's latest artificial intelligence program, Conversation AI, by swapping smears for the names of Google products. Conversation AI aims to spot and flag offensive language online, with the eventual possibility that it could automatically delete abusive comments. The famously outspoken forum 4chan, and the similar website 8chan, didn't like this, and began their campaign which sees them refer to "Jews" as "Skypes", Muslims as "Skittles", and black people as "Googles".

Implications for deployment of automated content moderation systems

The norms of content moderation are inherently contestable. Automation doesn't change that. Some things platforms deploying such systems should be aware of:

- No universally agreed notions of 'toxic' comments
- Difficult to operationalise acceptable comment policies
- Training data drawn from a different context may impose norms at odds with the application domain
- Exploratory approaches can help monitoring and evaluation
- Community norms may change over time, but training data from the past may exert a constraining force.
- Difficulty in exploring what is no longer toxic, when comments are censored
- Unlike other anti-discrimination contexts, the 'right' amount of diversity and homogeneity is contestable
- The balance between error different rates could encourage more or less diverse participation
- Clustering or manually identifying users by behaviour might help identify groups with conflicting views of offence

Read the paper on arXiv: tiny.cc/trainerbot
Check out the code: github.com/sociam/liketrainer

UNIVERSITY OF OXFORD
SOCIAM
UCL
EPSRC Engineering and Physical Sciences Research Council