UNIVERSITY OF CAMBRIDGE
Department of Engineering

Participant Information Sheet 11/07/2022

Title

Automating Counterspeech in Dialogue Systems

Invitation paragraph

Before you decide to take part in this study it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Please ask a member of the team if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part.

Purpose of the study

Online hate speech[1] is a major social problem. Whilst the amount of online hate speech continues to increase, *counterspeech* offers a promising strategy for countering the hate speech without invoking freedom of speech concerns. Counterspeech refers to any direct response to hate speech that seeks to undermine it and challenge the hate narratives.

The primary question we are investigating in this research is whether we can use AI to build *automated dialogue systems* that make appropriate use of counterspeech. There are many ways this could be useful. For example, the system could be used for generating counterspeech suggestion prompts for social media users when they encounter online hate speech, thus making it easier for the public to speak up against online hatred.

The aim of this study in particular is to help with *evaluating the quality of the counterspeech responses* produced by various systems in response to hate speech comments. This will help our research team draw conclusions about the effectiveness of the dialogue systems' counterspeech and how to improve them.

Do I have to take part?

Taking part is entirely voluntary. Refusal or withdrawal will involve no penalty or loss, now or in the future.

[1] The United Nations considers hate speech to be *any form of communication that attacks or uses pejorative or discriminatory language towards a person or group on the basis of an identity factor (such as race, gender, religion etc.).*

What will happen to me if I take part?

You will be sent an electronic form where you will be asked to evaluate the counterspeech produced by the dialogue systems according to specific criteria. This will include, for a set of sample hate-speech/counterspeech interactions, ranking the quality of the counterspeech response on a scale from 1 to 5. More specific details will be provided on the form.

Completing the form is expected to take approximately 20 minutes in total, but does not have to be completed in one sitting.

Are there possible disadvantages and/or risks in taking part?

Because you will be evaluating counterspeech responses to comments that directly contain hate speech and/or offensive language (e.g. racist, sexist or homophobic comments), there is a risk of psychological distress should you find the comments offensive or traumatic. Should you at any point experience any psychological distress due to reading these hate speech comments, you can stop and withdraw from the study without giving any reasons for why you no longer want to take part.

Will my taking part in this project be kept confidential?

All identifiable information collected (name and contact details) will be kept strictly confidential on a secure computer with access only by the immediate research team. Your responses and evaluations will be aggregated amongst all participants in the study.

Ethical review of the study

The project has received ethical approval from the Engineering Research Ethics Committee of the University of Cambridge.

Contact for further information

Please contact Shane Weisz (sw984@cam.ac.uk) for further information.