2Kindness

Exploring computer-mediated communication interfaces that reduce toxic interactions

Dr Matthew Higgs - Lecturer in Business Analytics, University of Bristol

(PI: Dr Mark Warner, Northumbria University)

(Dr Angelika Strohmayer, Northumbria University)

(Prof Lynne Coventry, Northumbria University)

(Dr Biju Issac, Northumbria University)

Northumbria University NEWCASTLE
University of BRISTOL
REPHRAIN Protecting citizens online

The problem

In the UK:

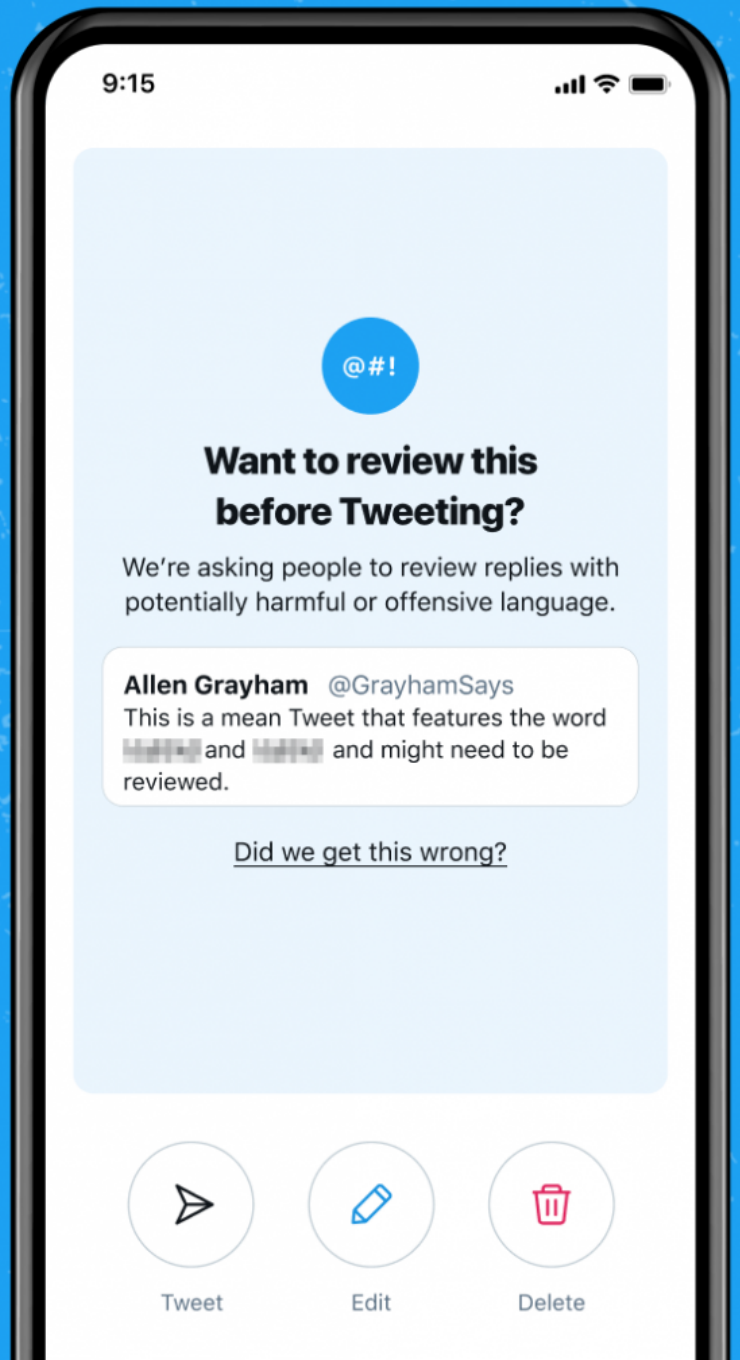
- 30 - 40% of people had been exposed to online abuse
- 10 – 20% have been a direct target of abuse

(Vidgen, Margetts, Harris, 2019)

Exposure to online abuse can significantly impact wellbeing, leading to:

- depression
- anxiety
- suicidal ideation

(Stevens, Nurse, Arief, 2021)



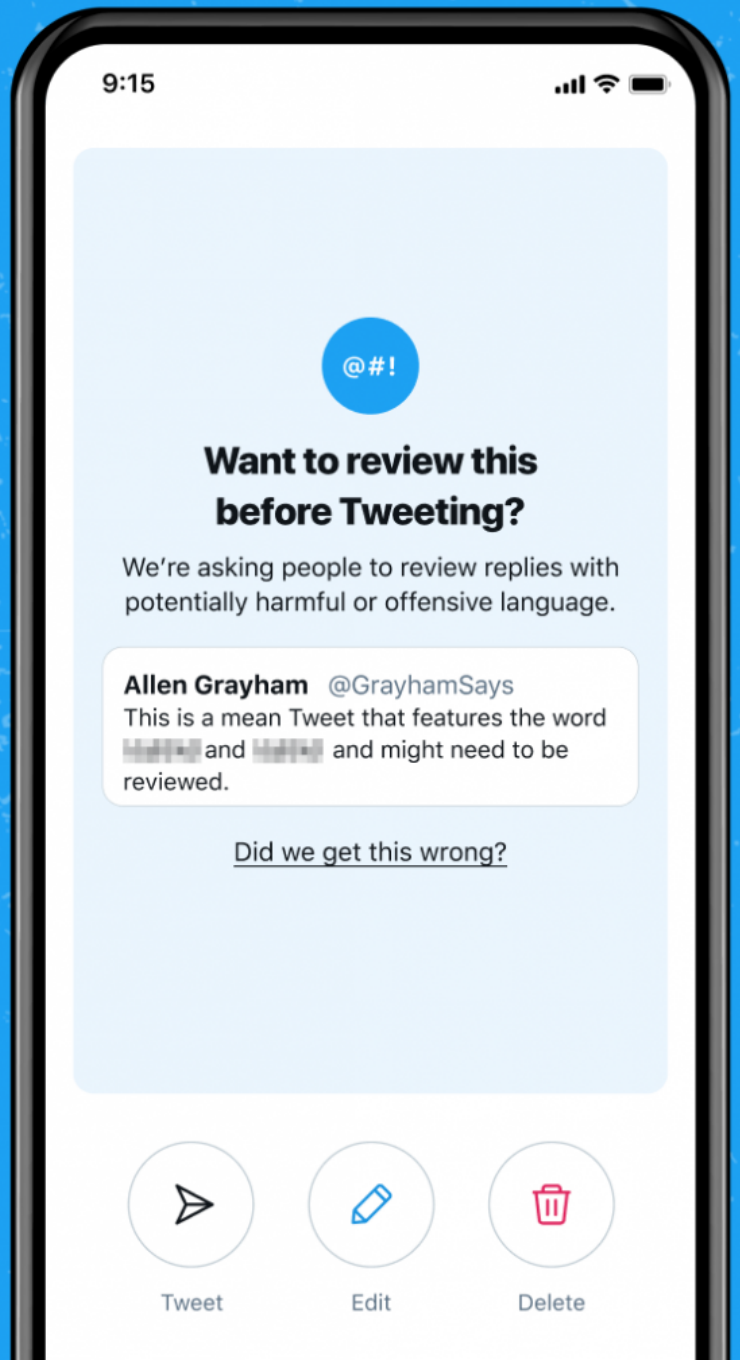
What's being done?

Moderation

- Social networks using ML models to help detect online abuse, to help with post-creation moderation.

Limitations / Issues

- Concerns related to freedom of speech due to automated approach to moderation
- Concerns related to fairness and discrimination due to the way ML models are trained. *See: Sap et al., (2021) bias evaluation of Perspective API.*



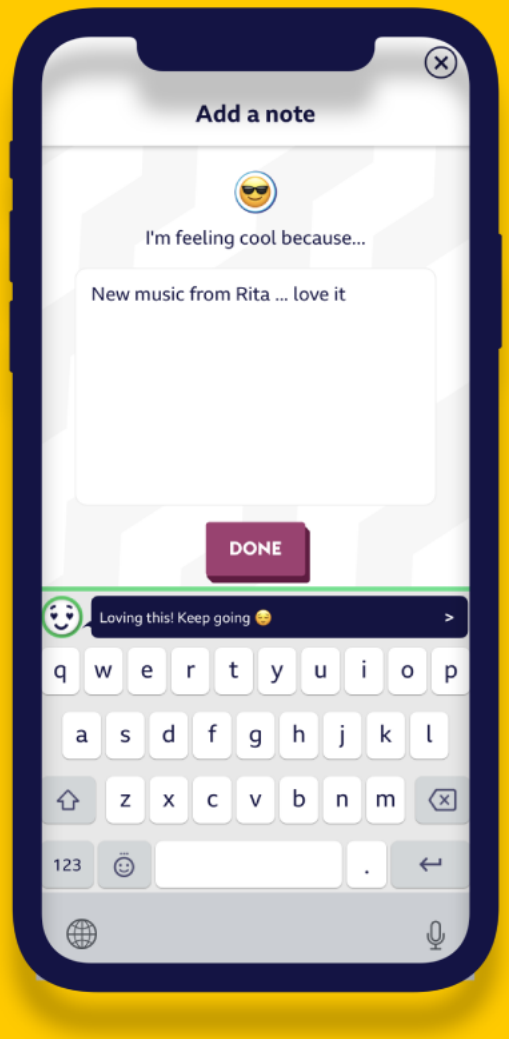
What's being done?

Embedded "in the moment prompts"

- Platforms experimenting/deploying ML models to prompt users prior to sending
- Twitter prompt shows some evidence of effectiveness. See: Katsaros et al., (2021)
- Tinder deployed offensive language detection in IM (no evaluation available)

Limitations / Issues

- Platform dependent (i.e. doesn't learn from users behaviour across platforms)
- Little research on effectiveness of different approaches



What's being done?

Keyboard based interventions

- BBC's Ownit App detects tone of language and provides educational prompts

- No evaluation available, and is designed specifically for younger children

Reflective Interfaces

- ”In the moment” approaches aim to cause moments of reflection
- Jones (2021) and Van Royen et al., (2017) both identified reflective interfaces reducing instances of harassment. They explored:
  - Text-based prompts
  - Time delays
- Prompts can result in self-corrective behaviour towards positive behavioural norms.

Behavioural “nudge”

Negative → Positive

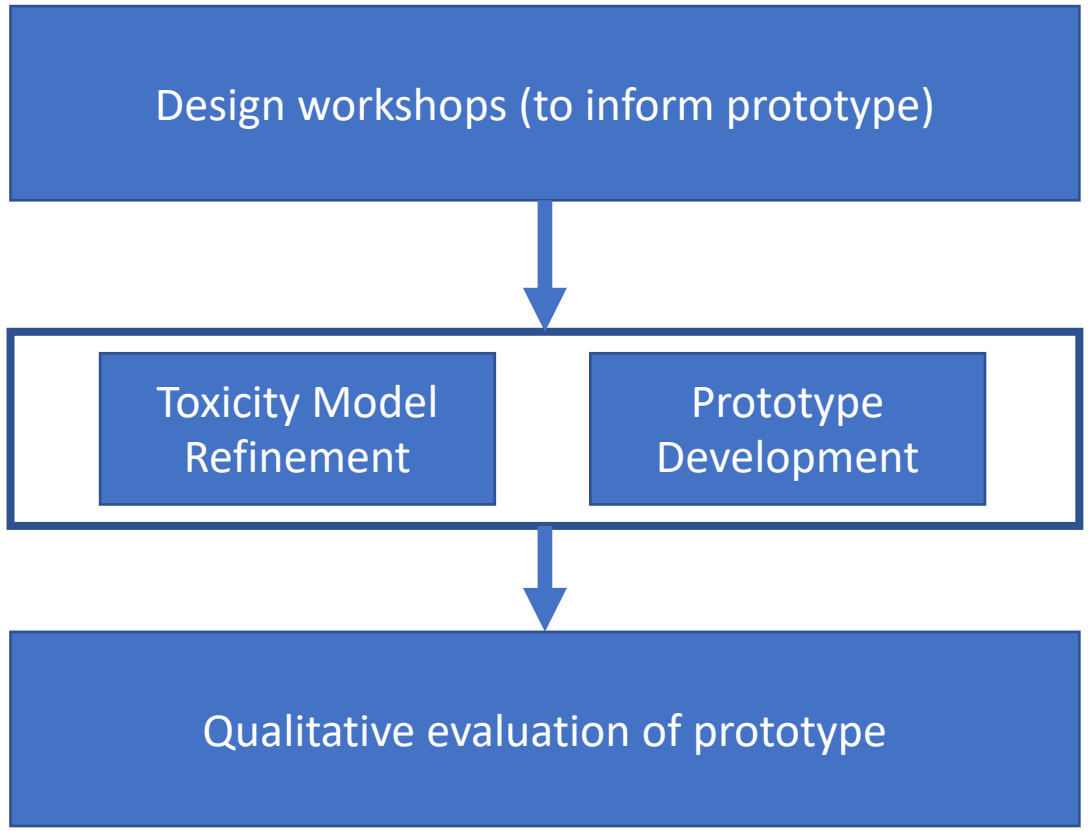
Northumbria University NEWCASTLE

University of BRISTOL

REPHRAIN Protecting citizens online

Could an “in the moment” intervention built into a mobile keyboard, reduce toxic interactions?

Project approach



Design Workshops

Four design workshops conducted with:

1. Postgraduate students working in social justice / HCI
2. Experts in cyber in cyberbullying, online hate, and other forms of toxic content
3. Online moderations (dealing with and receiving abuse)
4. People who have previous sent harmful content, but regret it

What participants did

Participants were asked to:

1. Discuss where the boundaries are between different types of content (e.g., hate and harmful messages)
2. Discuss the pro's and con's of this form of intervention
3. Design an imagined keyboard to respond to toxic content being written

Data was transcribed and analysed using a reflective thematic analysis approach

Codes and themes developed collaboratively

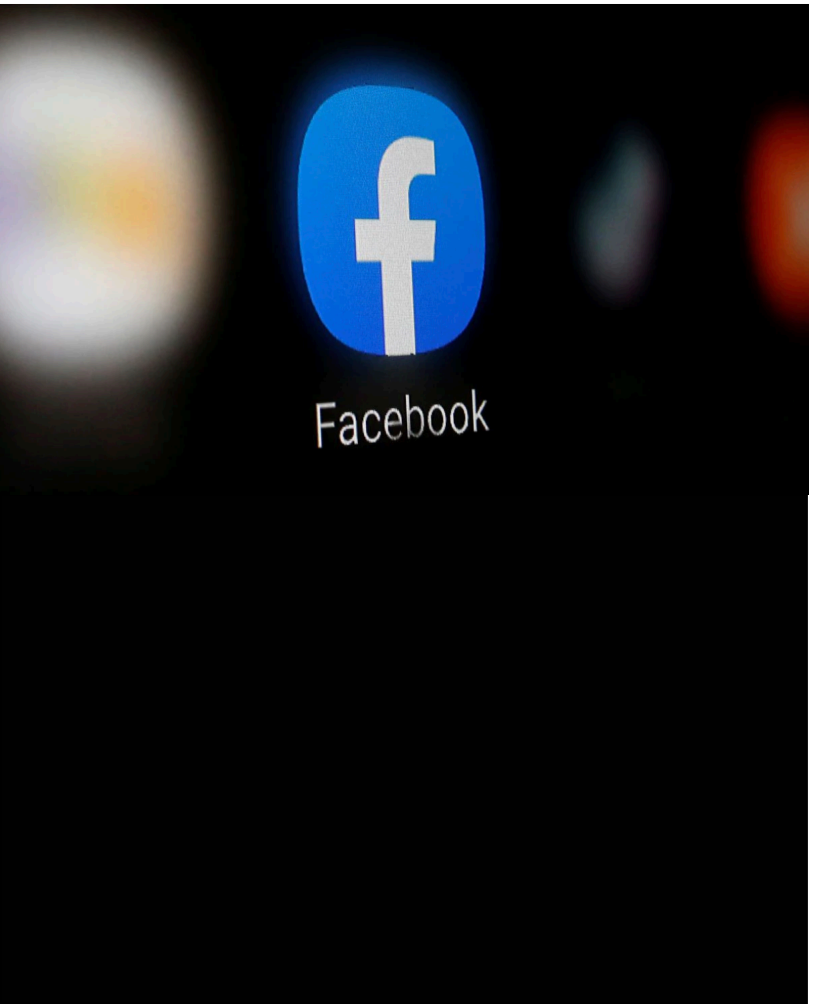
Initial themes:

1. A layered understanding of ‘context’
2. An audience continuum
3. Abusability of applications which embed toxicity models

Facebook allows war posts urging violence against Russian invaders

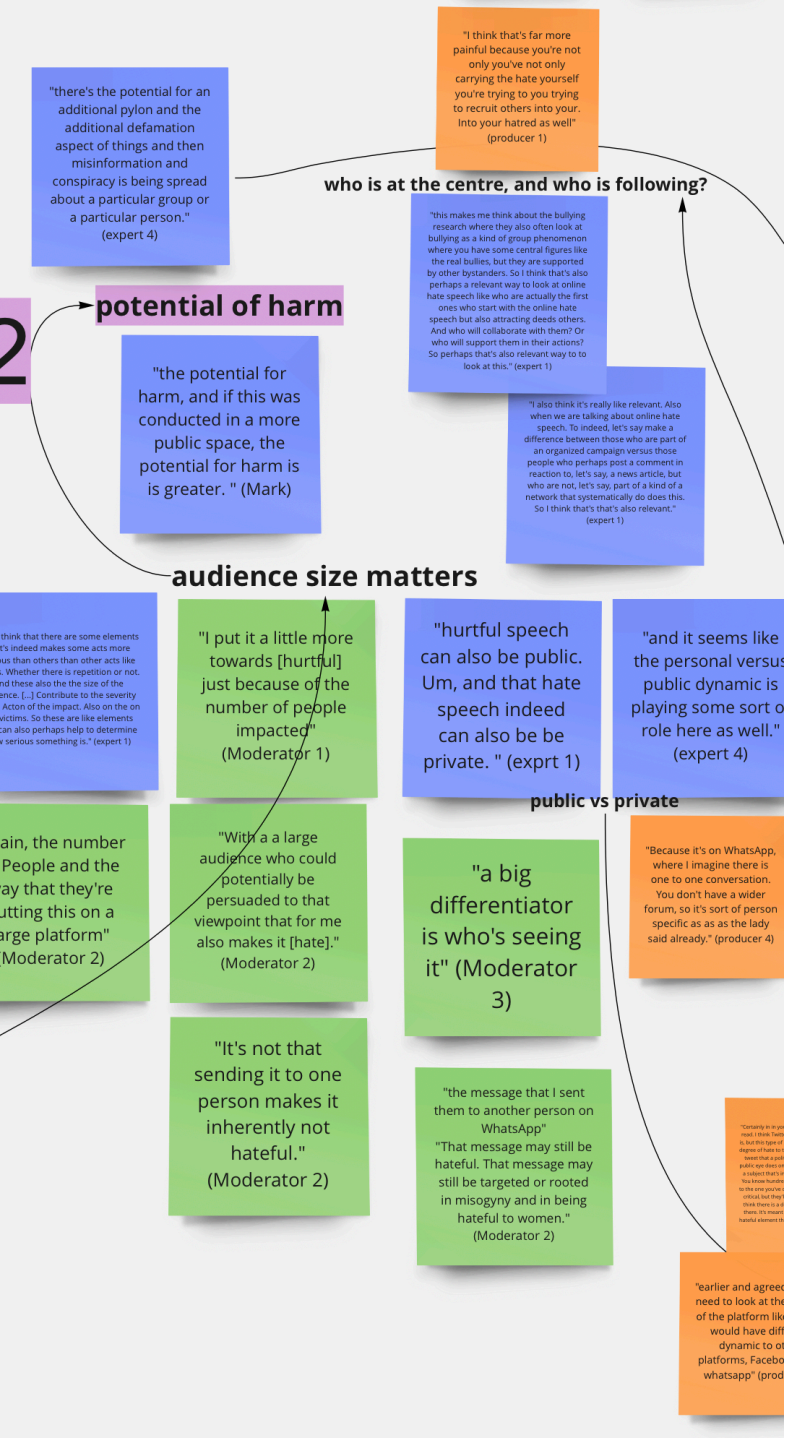
By Munsif Vengattil and Elizabeth Culliford

3 minute read



Layered ‘context’

- Type of platform
- **In vs out-group conversations**
- Public vs private
- **Message frequency**
- Audience size
- Recruiting others into targeted attacks
- **Conversation history**
- Social histories of oppression and power structures



Northumbria University NEWCASTLE

University of BRISTOL

REPHRAIN Protecting citizens online

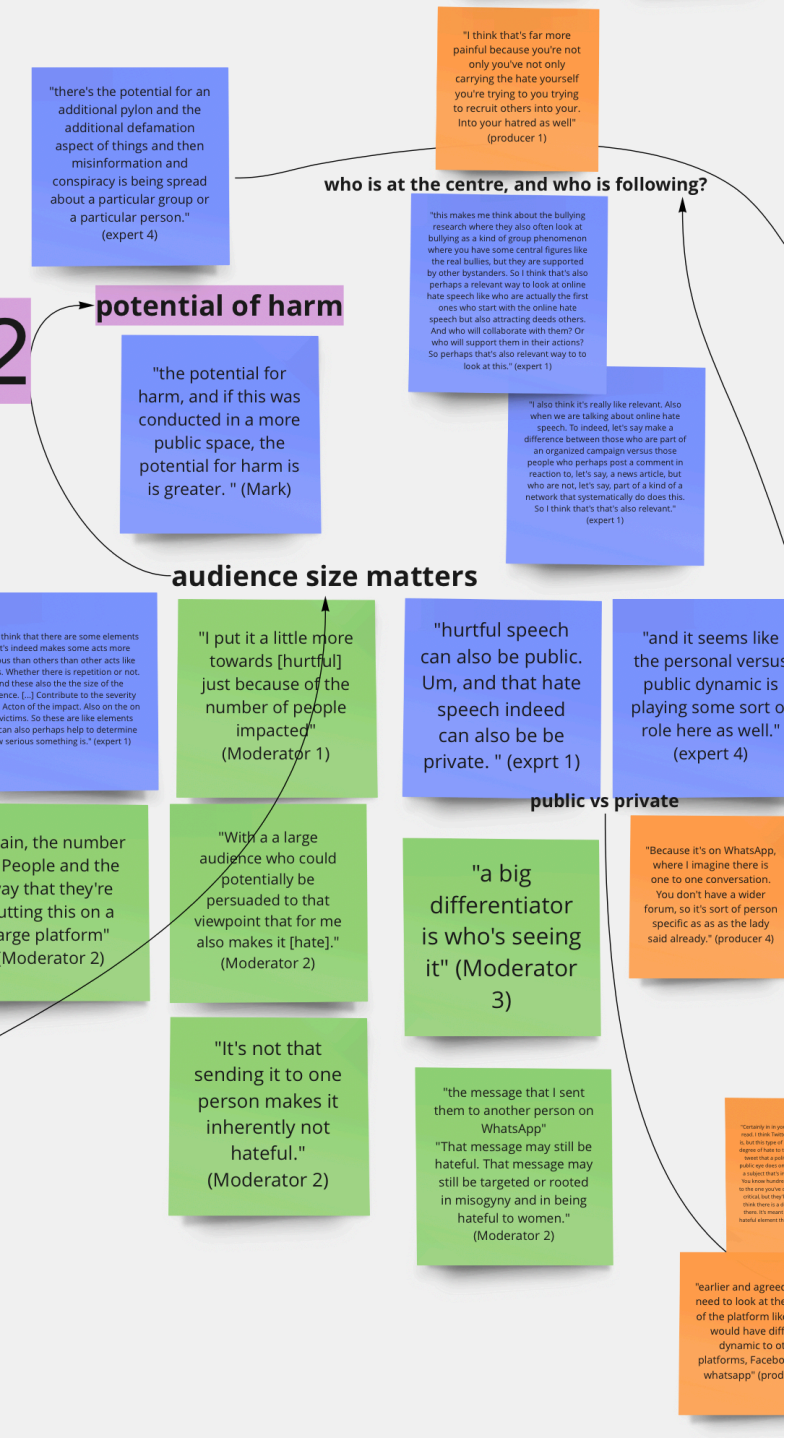
Audience continuum

- **The unaware**
- Those wanting to learn
- Emotionally triggered
- Hate as an emotional arousal
- Playing to an audience
- **Determined and organised**

Low intention

High intention

Wang et al.(2011); Sleeper et al. (2013); Warner et al., (2021)



Northumbria University NEWCASTLE | University of BRISTOL | REPHRAIN Protecting citizens online

Abusability

Abusing for validation

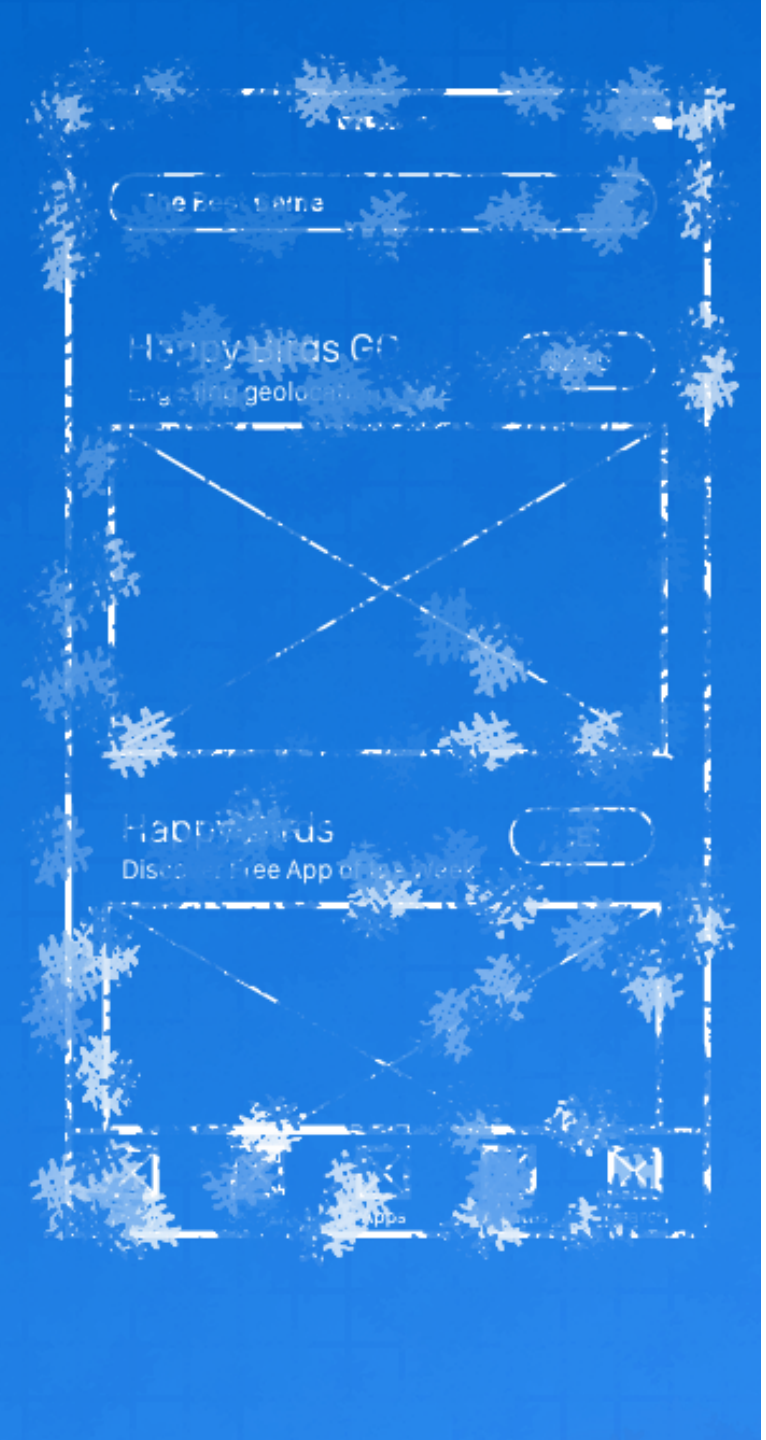
"I got validated, I'm being hateful" (Moderator 2)

Gamification of the system

"Some people would actually relish that needle go up" (Producer 2)

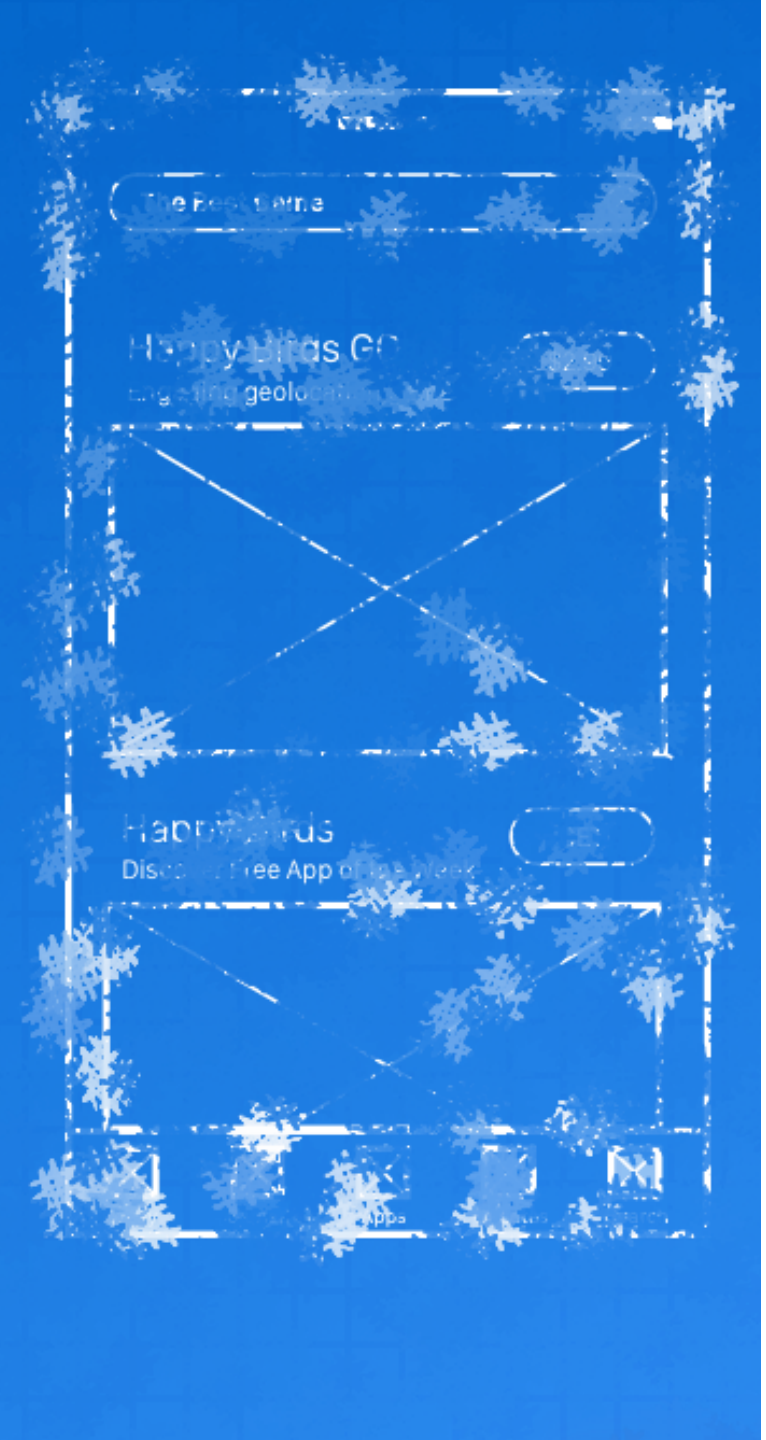
System circumvention

"How can I just […] find another way to say something hurtful in a way that the system doesn't recognise" (Expert 2)



How to design?

- Learning and with machine co-learning
- Designing with adaptability
- Designing for direct and indirect effects

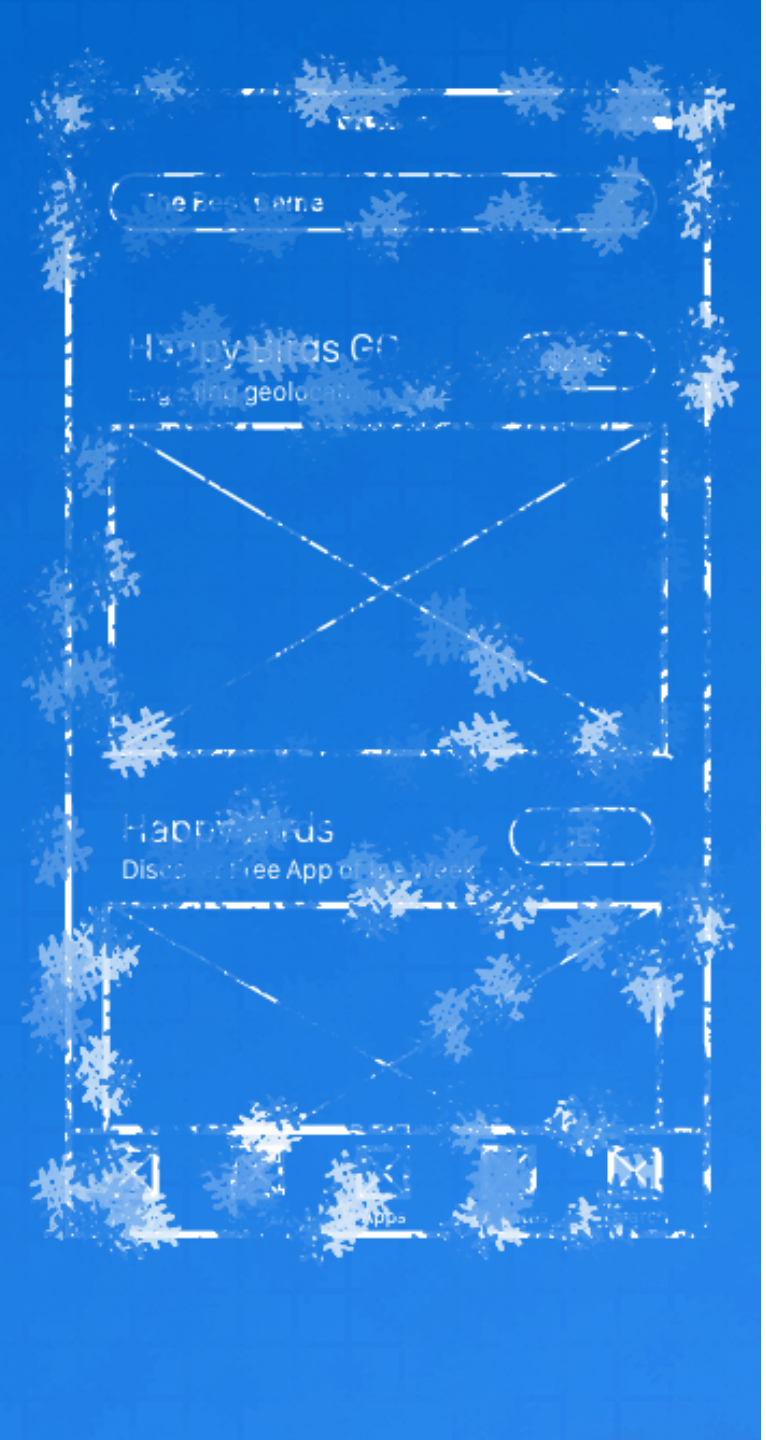


How to design?

Manage the tension between:

- Usability and friction
- Transparency (which could be open to abuse) and abstraction.

- **Currently exploring designs at different stages of the interaction:**
  - Long before writing
  - Immediately prior to writing
  - During writing
  - At the point of sending
  - Immediately after sending
  - Long after sending



Next steps…

- Integrating design insights into a working prototype keyboard
- Deployment and qualitative evaluation

Future future work:

- Iteration on the keyboard design
- Long-term deployment with control group
- Evaluation of different ML models within this sociotechnical system