

Key

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)



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.*



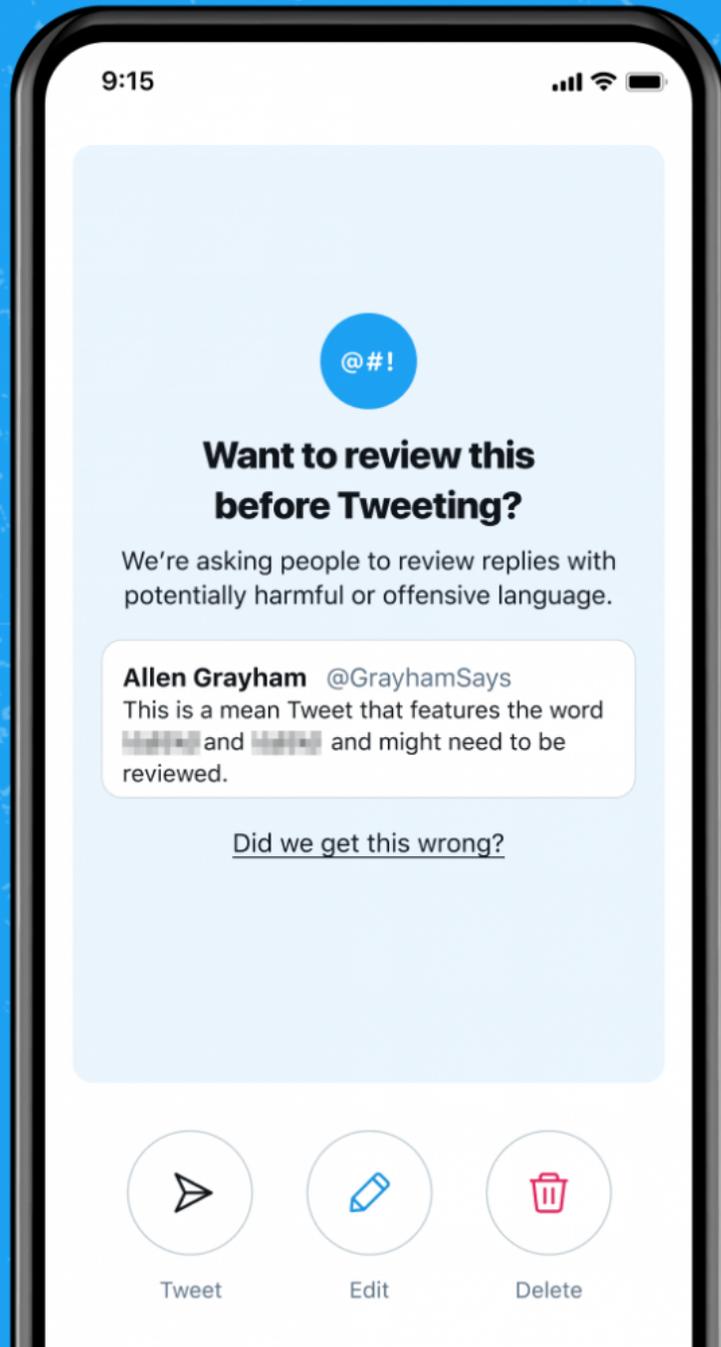
What's being done?

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

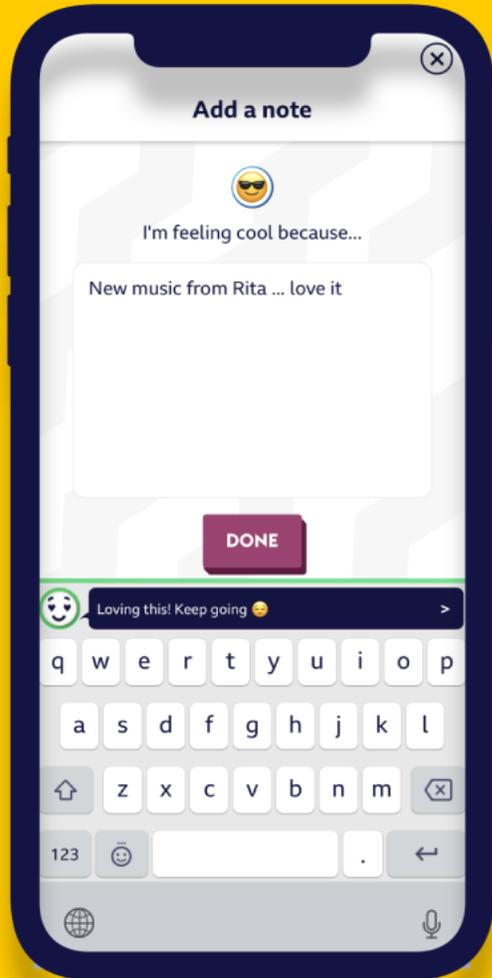
- Platform dependent (i.e. doesn't learn from users behaviour across platforms)
- Little research on effectiveness of different approaches



What's being done?

Keyboard based interventions

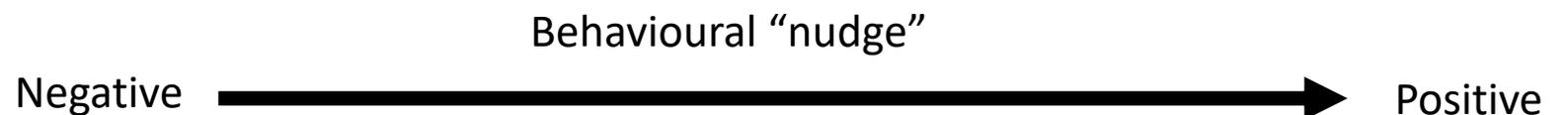
- BBC's Ownit App detects tone of language and provides educational prompts
- No evaluation available, and is designed specifically for younger children





Reflective Interfaces

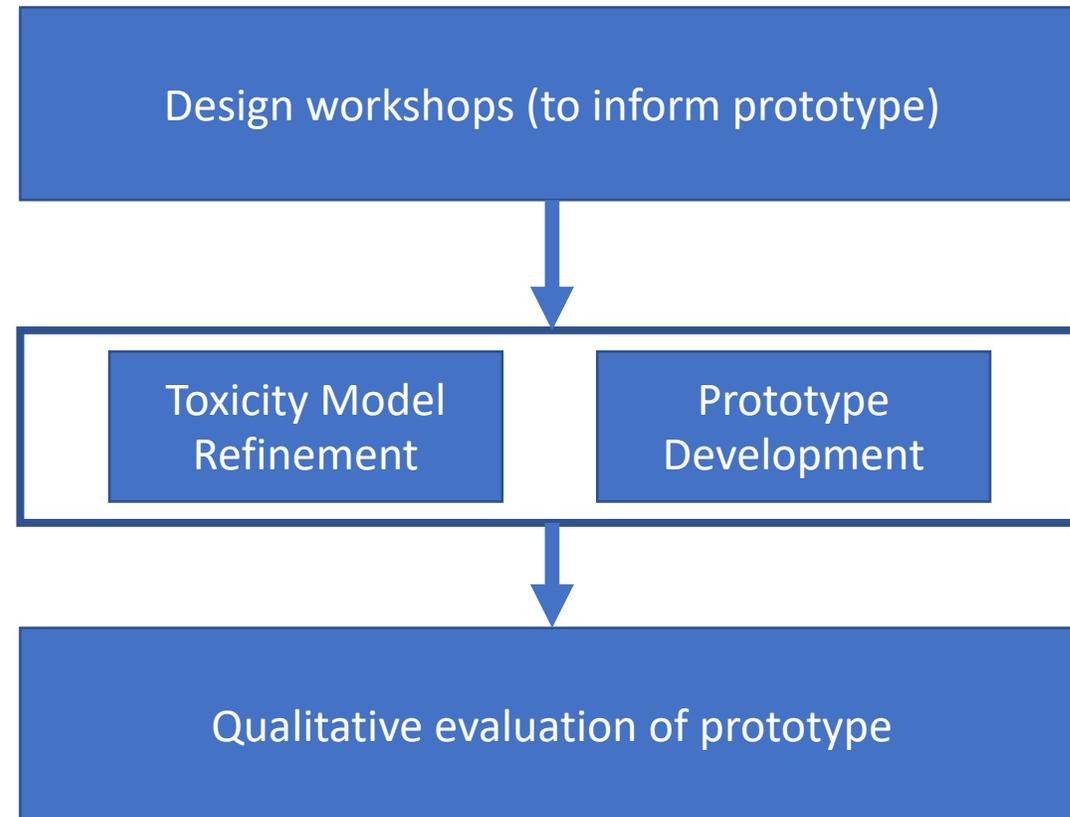
- “In the moment” approaches aim to cause moments of reflection
- Jones (2021) and Van Royen et al., (2017) both identified reflective interfaces reducing instances of harassment. They explored:
 - Text-based prompts
 - Time delays
- Prompts can result in self-corrective behaviour towards positive behavioural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



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

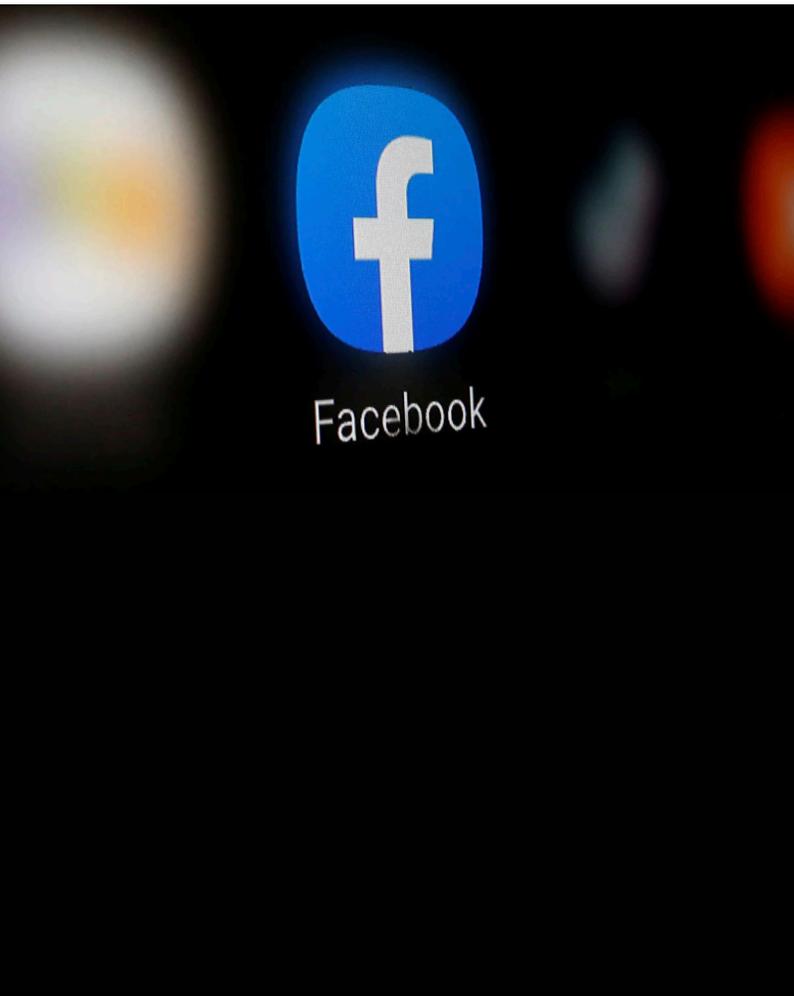
Data was transcribed and analysed using a reflective thematic analysis approach



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

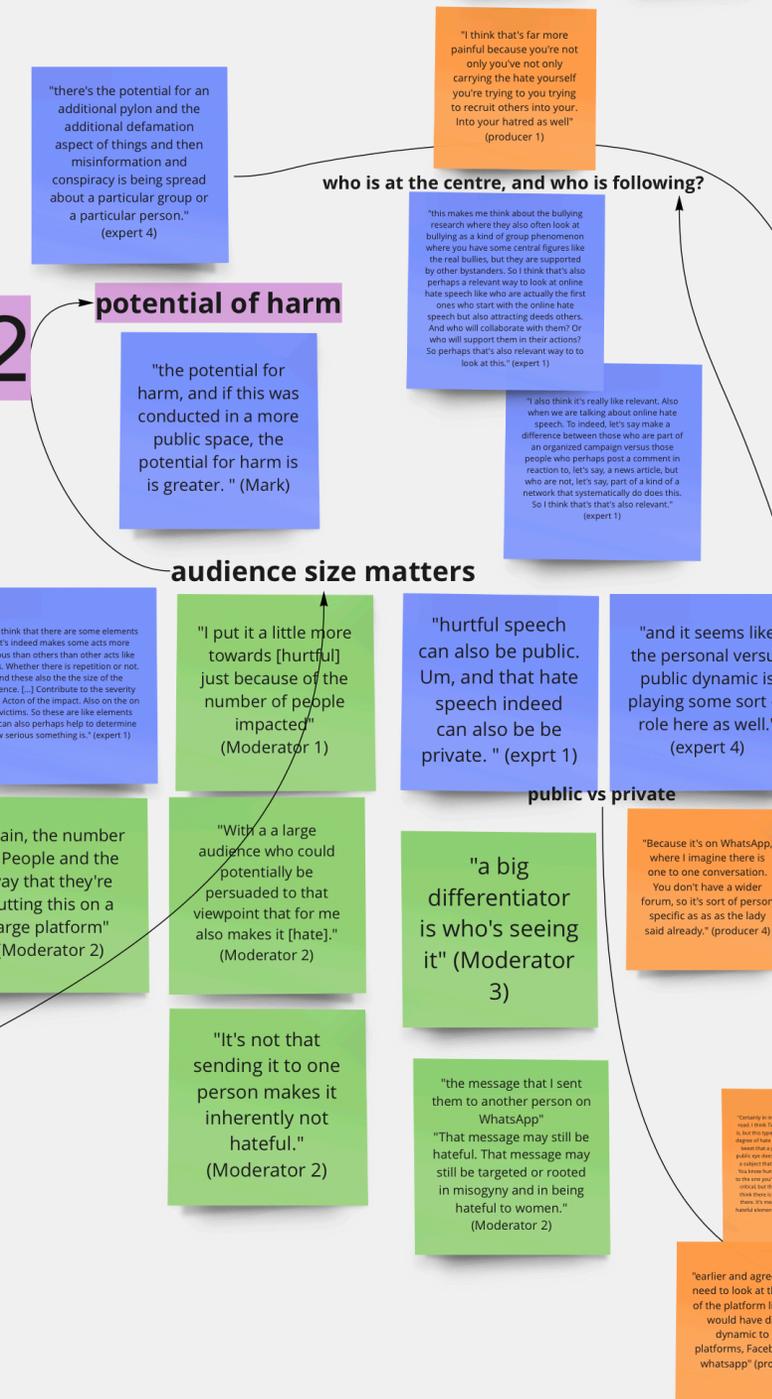
Audience continuum

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

Low intention



High intention



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)

who is at the centre, and who is following?

"there's the potential for an additional defamation aspect of things and then misinformation and conspiracy is being spread about a particular group or a particular person." (expert 4)

"I think that's far more painful because you're not only you've not only carrying the hate yourself you're trying to you trying to recruit others into your. Into your hatred as well" (producer 1)

"this makes me think about the bullying research where they also often look at bullying as a kind of group phenomenon where you have some central figures like the real bullies, but they are supported by other bystanders. So I think that's also perhaps a relevant way to look at online hate speech like who are actually the first ones who start with the online hate speech but also attracting deeds others. And who will collaborate with them? Or who will support them in their actions? So perhaps that's also relevant way to look at this." (expert 1)

"I also think it's really like relevant. Also when we are talking about online hate speech. To indeed, let's say make a difference between those who are part of an organized campaign versus those people who perhaps post a comment in reaction to, let's say, a news article, but who are not, let's say, part of a kind of a network that systematically do does this. So I think that's that's also relevant." (expert 1)

potential of harm

"the potential for harm, and if this was conducted in a more public space, the potential for harm is greater." (Mark)

audience size matters

"I think that there are some elements that indeed makes some acts more serious than others than other acts like [...] Whether there is repetition or not, and these also the size of the audience. [...] Contribute to the severity of the action of the impact. Also on the on the victims. So these are like elements that can also perhaps help to determine how serious something is." (expert 1)

"I put it a little more towards [hurtful] just because of the number of people impacted" (Moderator 1)

"hurtful speech can also be public. Um, and that hate speech indeed can also be private." (exprt 1)

"and it seems like the personal versus public dynamic is playing some sort of role here as well." (expert 4)

public vs private

"...ain, the number of people and the way that they're putting this on a large platform" (Moderator 2)

"With a large audience who could potentially be persuaded to that viewpoint that for me also makes it [hate]." (Moderator 2)

"a big differentiator is who's seeing it" (Moderator 3)

"Because it's on WhatsApp, where I imagine there is one to one conversation. You don't have a wider forum, so it's sort of person specific as as the lady said already." (producer 4)

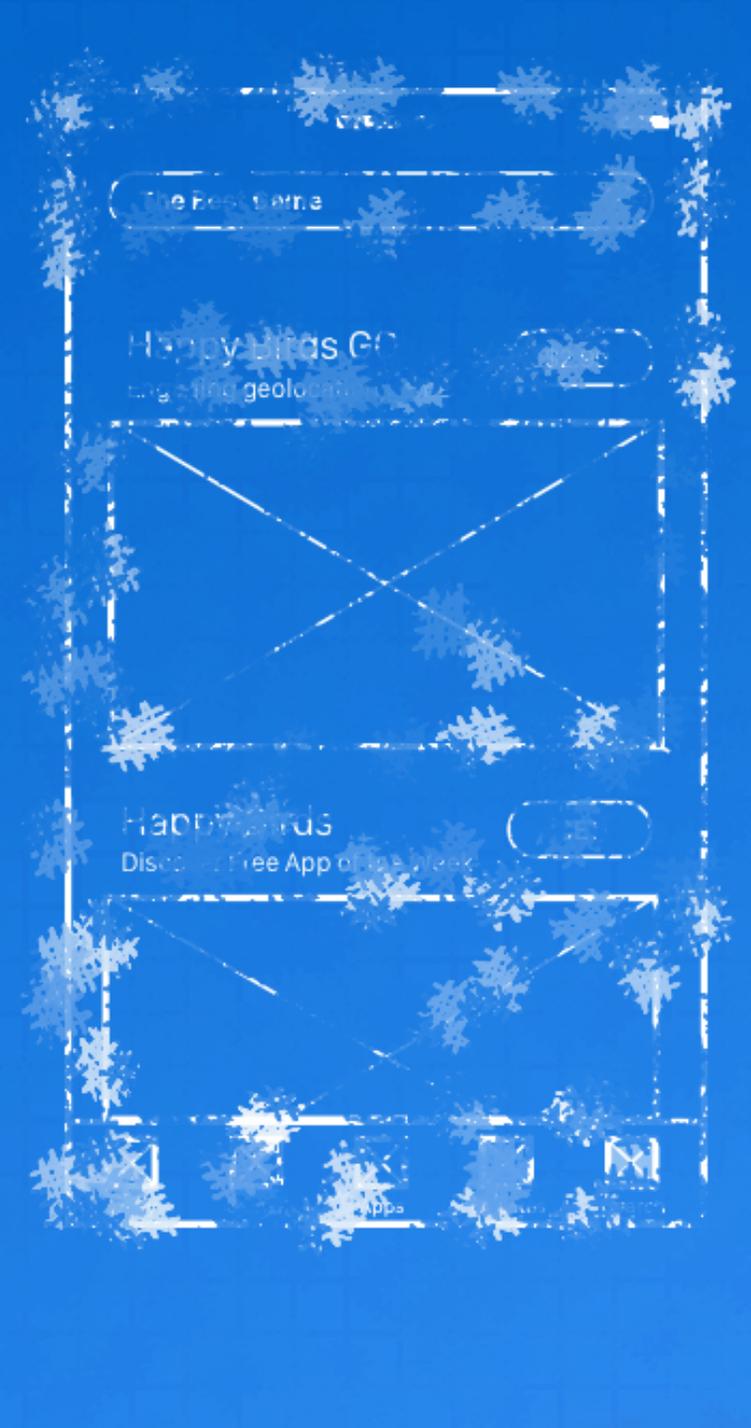
"It's not that sending it to one person makes it inherently not hateful." (Moderator 2)

"the message that I sent them to another person on WhatsApp" "That message may still be hateful. That message may still be targeted or rooted in misogyny and in being hateful to women." (Moderator 2)

"...earlier and agree need to look at the of the platform like would have diff dynamic to of platforms, Facebook whatsapp" (prod

How to design?

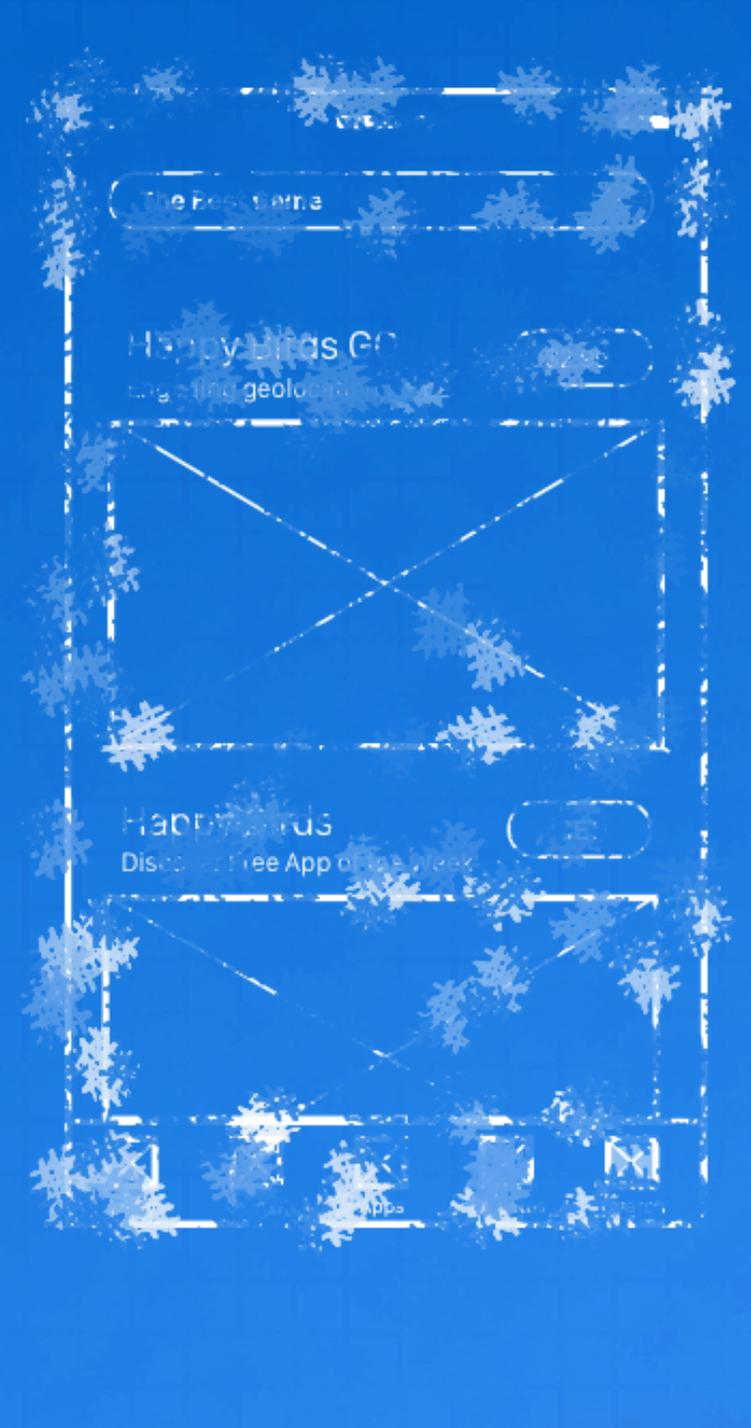
- Learning and with machine co-learning
- Designing with adaptability
- Designing for direct and indirect effects



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

