

# DESIGNING FOR POLARISATION

---

By **Valentina Dzhekanovich**

Mentors **Johanna Rochegude, Jonas Knutsson**

Supervisors **Tanel Kärp & Nesli Hazal Akbulut**

Completion of project in June 2021

Interaction Design, Faculty of Design, The Estonian Academy of Arts

---

## TABLE OF CONTENTS

### **INTRODUCTION** **4**

Personal story  
Context of the problem  
Design approach  
Process outline & research methods

### **RESEARCH** **10**

Desk research  
Expert's view  
User research  
Survey  
Digital ethnography  
Cognitive-behavioural map

### **IDEATION** **30**

Co-designing with users  
Mix&Match brainstorming  
Design principles  
Ideation based on analogy  
Workshop  
Analogous solutions

### **DESIGNING AN INTERVENTION** **40**

Developing design intervention  
AI-driven UX/UI design intervention  
Psychological basis of design proposal

### **CONCLUSION** **51**

Reflections about the design process  
Evaluation of the design intervention  
Future work  
Acknowledgements

---

Reference list 53  
Appendix 1 55  
Appendix 2 58  
Appendix 3 59

---

## ABSTRACT

Political division over information is a paradox of the modern hyperconnected world. AI-driven social media platforms created a unique unprecedented situation where despite being brought so close to each other people increasingly get polarised in the way they see the reality.

It has to do with the way information is presented in social media today – personalised news feeds provide each of us with the picture of the world that we like and engage with the most. On a macro level, different truths and separate worlds have emerged as a result of this process. This has diminished the common grounds of shared reality and developed hostility in interactions between these different groups of people who hold contradicting beliefs.

In this thesis, I explore the cognitive-behavioural aspect of political polarisation in the U.S. between far-left and far-right social media users. Using the digital ethnography as the main research method,

I embedded into circles of my target audiences from both ends of the American political spectrum and compared their online experiences, information processing mechanisms and behaviours.

The goal of this project was to attempt to depolarise social media users with means of design. Results of the digital ethnography study identified several points in the experiences of radicalised social media users, where the design intervention could make a change.

As an outcome of this project, I propose a concept that is designed to motivate respectful discussions and unite people around common reality, which removes certain aspects of polarisation from social media environments.

This report presents an overview of the five months of the project while detailing the development of the work and exploration of the concept.

## ESTONIAN ABSTRACT

Kaasaegses hüperühendatud maailma paradoks on politiseeritud informatsioon. Tehisintellekti kasutavad sotsiaalmeedia platvormid on loonud olukorra, kus vaatamata näilisele lähedal olemisele, inimesed kaugenevad üksteisest.

See on tingitud isikupärastatud uudisvoogudest, mis pakuvad igapäevase erinevat taju maailmas. Näidates üha enam ainult teemasid, mis meile meeldivad või millega tegeleme. Seepärast on makrotasandil tekkinud erinevad tõed ja paralleelsed maailmad. Vähenenud tajutav ühine reaalsus on kasvatanud vastuoluliste veendumustega inimrühmade omavahelist vaenulikkust.

Uurin töös USA poliitilise polariseerumise vasak- ja paremäärmuslike sotsiaalmeedia kasutajate näitel. Uurimismeetodina kasutan digitaalset etnograafiat, kus ühinen sotsiaalmeedias mõlema poliitilise spektri äärmustega ning võrdlen nende informatsiooni töötlemist, käitumist ja kogemust veebis.

Selle projekti eesmärk on proovida sotsiaalmeedia kasutajaid depolariseerida. Digitaalse etnograafia uuringu tulemused tõid sotsiaalmeedia kasutajate kogemustes välja mitu punkti, kus disaini sekkumine võiks tuua tulemusi.

Selles projektis pakun välja kontseptsiooni, mille eesmärk on motiveerida lugupidavaid arutelusid ja ühendada inimesi ühise mõistmise ümber, mis eemaldab sotsiaalmeediast polariseerumise teatud aspektid.

Selles aruandes esitatakse ülevaade projekti viiest kuust, kusjuures üksikasjalikult kirjeldatakse töö arengut ja kontseptsiooni uurimist



# INTRODUCTION

## PERSONAL STORY

I was growing up in a multicultural family: my father is an orthodox Christian from Serbia and my mother is a Muslim Tatar. Without going deep into the history of the tension between these two religious groups, I must admit that since early childhood I was constantly exposed to this divide. I was baptised and my Muslim part of the family never should have known that, and I was taught Islamic prayers and my Christian part of the family should have never known that either. There is nothing wrong with any of these cultures, but out of respect and for the mental safety of all of my family members, I had to play hide and seek with both of my identities in different family gatherings.

Belonging to Christian and Muslim cultures at the same time was even fun sometimes, but not until relatives from both sides would start expressing hostile views towards each other on a religious basis. I could never understand why did people whom I

equally love and respect would divide so harshly upon such a personal and ephemeral thing as belief.

This split has significantly influenced the way I see the world, questions that I ask and issues that concern me the most.

## CONTEXT OF THE PROBLEM

My family story and divide root back to times when local communities and classic mass media were the only source of information, often biased and propagandistic. The appearance of the internet and ICT could offer a much wider perspective of the world, help people connect and understand each other better.

However, from my viewpoint, the trend went in quite the opposite direction. For example, my Serbian brother who is of my age, who has always been a very kind and peaceful guy, all of a sudden started sharing rather aggressive posts on Facebook about Orthodox supremacy and calling for separation of the Christian part of his country from Islamic states. I talked to him about it, and the

reply was that “one cannot hide the truth from the internet”, and that it was his moral duty to spread it.

This radicalisation case of my brother was not the single one. As Joaquin Quiñonero Candela, a director of AI at Facebook, mentioned in his interview to the MIT Technology Review, “models (content recommendation algorithms – author’s note) that maximise engagement increase polarisation... regardless of the issue, the models learned to feed users increasingly extreme viewpoints. Over time they measurably become more polarised” (Hao, 2021).

The internet and social media created a situation where users see tailored content as a full-fledged reflection of the world, engage with it and reinforce the feedback loop of the same preferred content, which spirals their views further into radicalisation and narrow mindedness. The path towards mutual understanding and connectivity initially promised by social media has turned the other way around and fuelled the disputes not only within my family but among millions of social media users and

communities all over the world. It is commonly known, that significant number of recent clashes from the Pizzagate case and other election-related scandals to the genocide of Muslim Rohingya minority in Myanmar were escalated due to Facebook, Twitter, Youtube and other social media (Noujaim and Amer, 2019).

## DESIGN APPROACH

The problem is interdisciplinary and spans across many fields, such as psychology, sociology, philosophy and ICT design, which makes it relevant to tackle within the scope of interaction design. This field studies the cognition and behaviour of people and addresses it through their environment, both digital and physical.

The environment I chose was social media because that’s where behaviour of radicalisation inflamed according to my own observations and research. The psychological ground of the problem became a dual-process theory, that explained how fast, intuitive and emotional mode of thinking was opposed to slow rational thinking, and how it produced a lot of cognitive biases and misjudgments (Kahneman, 2013). This dichotomy very well applies to how people process overwhelming flows of information in social media today.

With this approach, I theorised that AI-driven social media reinforce cognitive biases in users and result in increasing polarisation of beliefs.

My goal with this master’s degree project was to study and solve the given problem by means of interaction design. As an outcome of this project, I propose a design intervention that works as a self-reflection tool and removes certain aspects of polarisation from social media environments.

In the next section, I describe how this project was done – how the design activities were planned and documented, and which research methods were applied.

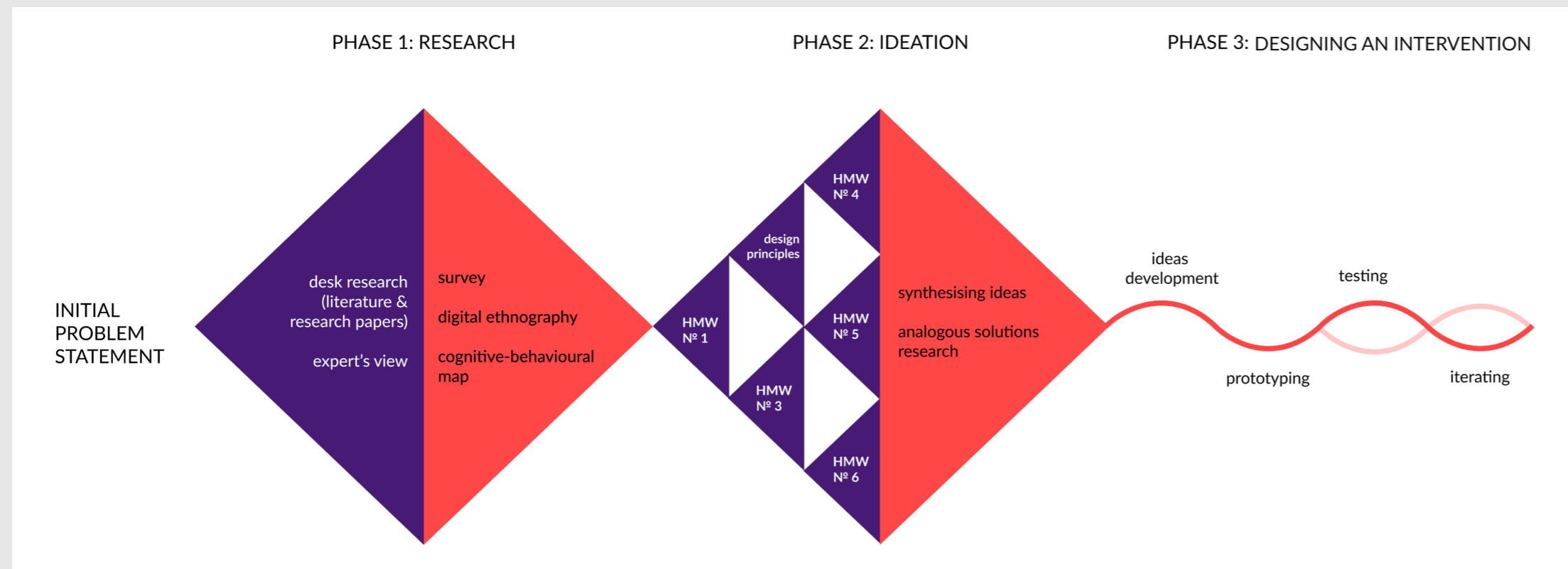


Fig. 1. Visualisation of my design process

# PROCESS OUTLINE & RESEARCH METHODS

The initial project outline was based on a classic Double Diamond design process model. However, as the process unfolded and new circumstances emerged, I updated the plan and finally arrived at the design process model of my own, which I view as the filigreed version of the Double Diamond (Fig. 1).

## PHASE ONE: RESEARCH

At the beginning of this degree project journey, I had a preconception that cognitive biases could be eliminated if rational thinking mechanisms were enhanced by AI-driven social media interactions, and thus emotional hostility and polarisation would be reduced.

At the first meeting with my mentor Johanna Rochegude, a senior UX designer from the strategic design studio Block Zero, I received a suggestion to take a bird's-eye view of the problem and its context, define parameters of polarisation, talk to experts from the related fields, find a specific issue and a target group within it.

So, I scheduled five semi-structured interviews with experts from psychology, AI, politics, philosophy and sociology. Each interview started with a general introduction to my topic and then continued with a more specific direction of discussion depending on the expert's field. When transcribing the interviews, I colour coded quotes that covered repetitive topics or unusual perspectives. Then I created affinity diagrams (Fig. 2) based on these colour coded data in order to find patterns and connections among interviews.

Interviews helped me to define the position of my research in a broader context and find keywords for the primary desk research and literature review. All these steps moved the process further to the widest part of the first Diamond of the diagram.

At this step, I could see the problem from a very wide perspective and see that my initial statement missed many aspects of it. It was not purely the problem of irrational perception reinforced by

algorithms, but also sociological, philosophical, political and even a business problem. With this broader context kept in mind, I found the specific theme of polarisation and who was the target group: it was the political polarisation between far-left and far-right social media users in the U.S. (in the next chapters, I explain how I arrived at this conclusion). From there on, I narrowed down the focus and conducted in-depth research about my target audience.

At first the target audience research involved a political bias survey that I planned as a filter tool to select participants for further interviews. However, this method did not gather enough data in the time I allocated for that.

So I changed the user research method and conducted a two-week-long digital ethnography study. The method turned out very effective, because it allowed me to study the problem in its natural environment. For this study, I created two personas on both sides of the polarisation spectrum, registered their social media accounts and lived their online lives. This approach helped me step into my target audience's shoes, understand and empathise with these people. During my observations, I captured all the data that was in the scope of my research problem and placed it on Miro (Fig. 3) boards.

This data collection method produced many insights. I mapped them out on the Cognitive-behavioural scheme that depicted the information processing mechanisms and behavioural flows of my polarised target group. The scheme described the problem in great detail, so it became a convenient thinking tool in my further ideation sessions.

## PHASE 2: IDEATION

However, instead of a single narrow problem definition in the centre of the classic Double Diamond model, the scheme offered several problem definitions and directions to pursue. I decided to tackle them with four brainstorming sessions – with my classmates and designers from the IxD.ma community, with my target group and with my colleagues from a design studio Block Zero.

At this phase, it was important to go as wide as

### Political expert interview: Annika Arras

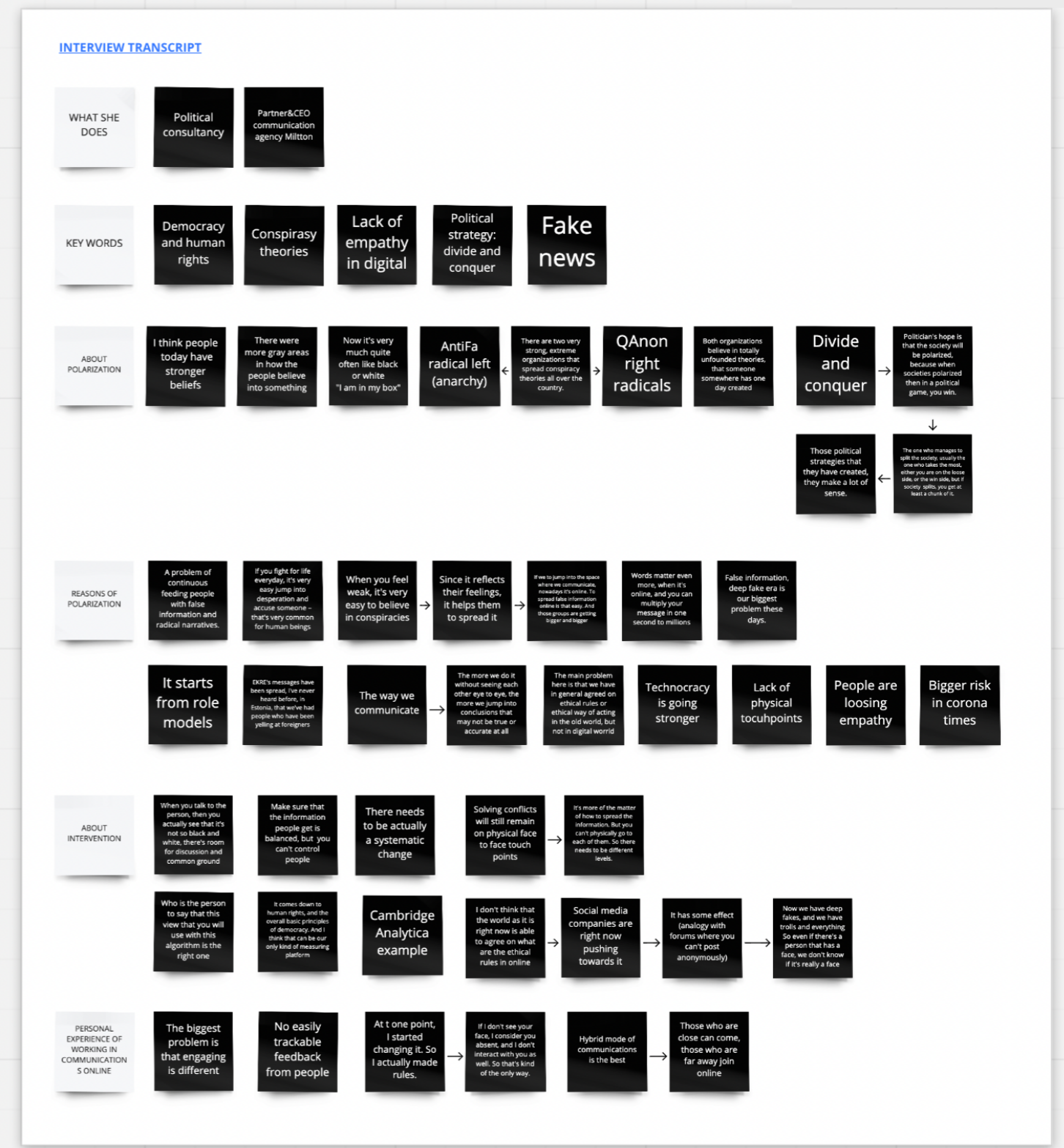


Fig. 2. Analysing expert interviews with affinity diagrams (analysis of the interview with a political expert as an example).



Digital ethnography study: far-left bubble

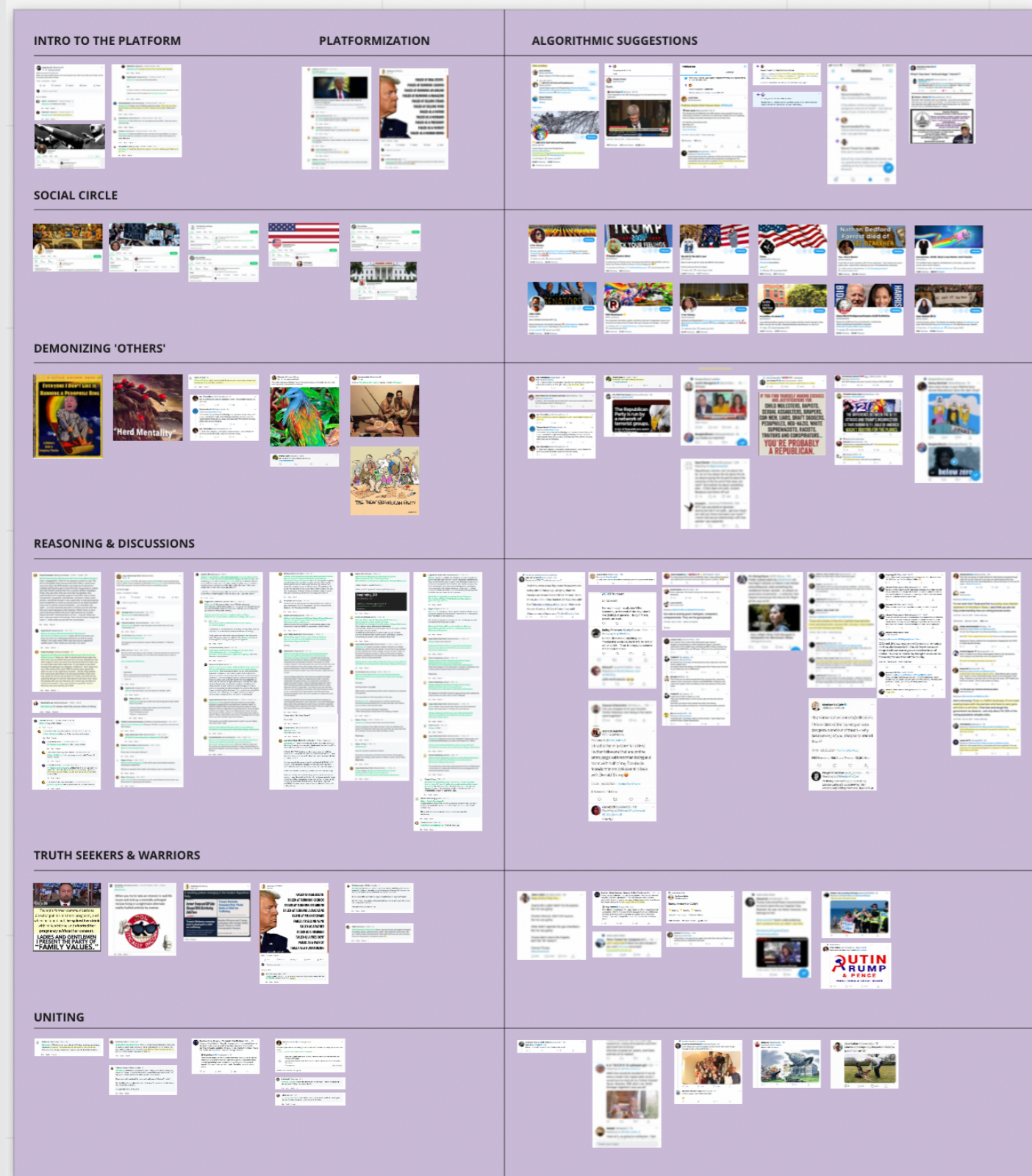


Fig. 3. Digital ethnography study data represents the far-left "bubble". Posts and comments are grouped based on topics. For ethical and privacy reasons, the figure shows the quantity of gathered data rather than specific texts and usernames.

possible to stimulate creativity and reveal the unusual perspectives. Altogether, ideation sessions helped me gather many diverse ideas about potential design interventions.

Then, I developed design principles (Fig. 4), because I needed them to shape this very wide flow of ideas and direct me towards ideas informed by the research. I learned that having these principles in front of me already in the first ideation session could make the process much more efficient.

I also conducted research about existing similar solutions and made a comparative analysis. This helped me filter out duplicating ideas and narrowed my focus to the unique ones.

**PHASE 3: DESIGNING AN INTERVENTION**

The design process then moved on further to the last phase of this project. Designing an intervention involved three rounds of iterations that helped me to refine my final solution. During each iteration, I developed the idea in the form of sketches and then as a landing page, and validated it with my mentors, supervisors, target audience and experts. Overall, this feedback improved my final solution and helped me to design a high-fidelity prototype.

To conclude, this design journey taught me a lot about research methods and how to adjust the process if the outcomes are unexpected. All of this would not be possible without the regular weekly meetings with my mentor, who guided me and helped me to arrive at the final step of this project.

In the next chapters, I introduce the content of this research and development journey.

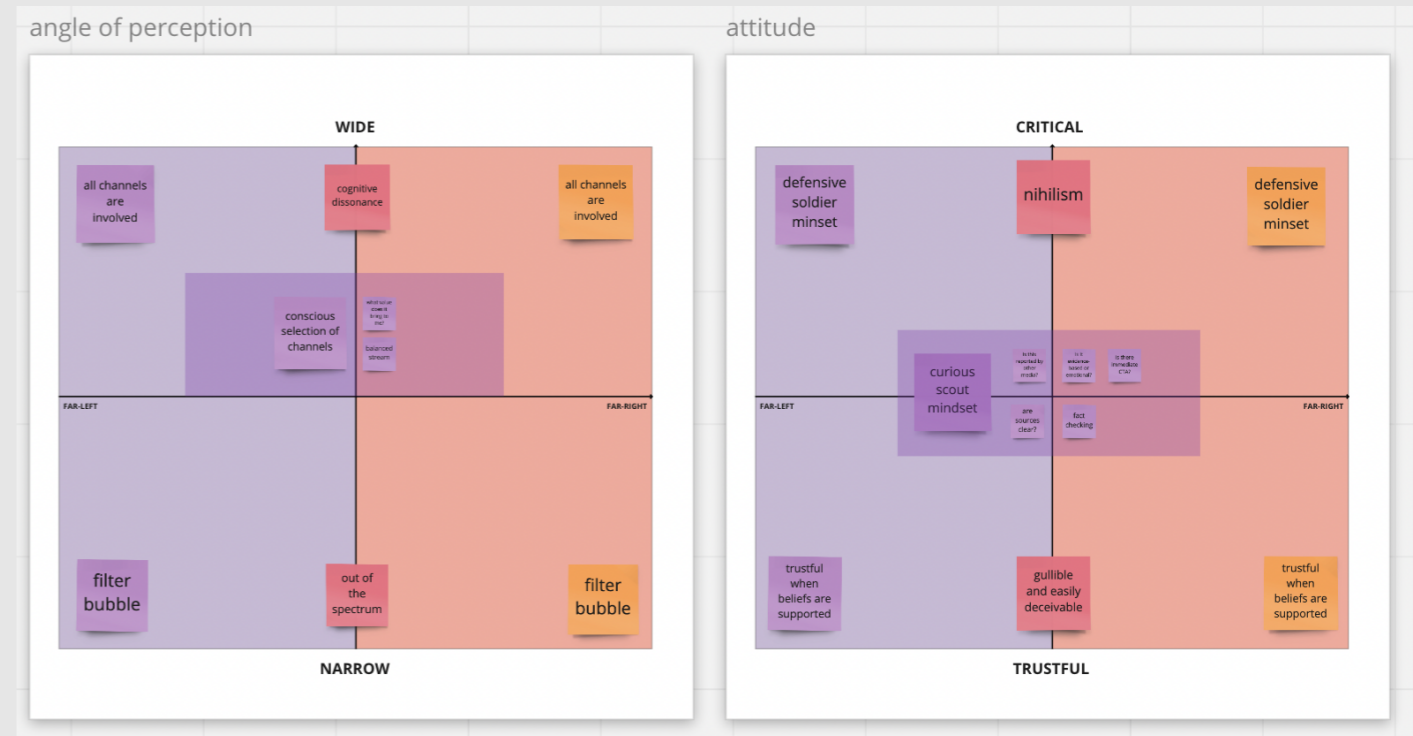
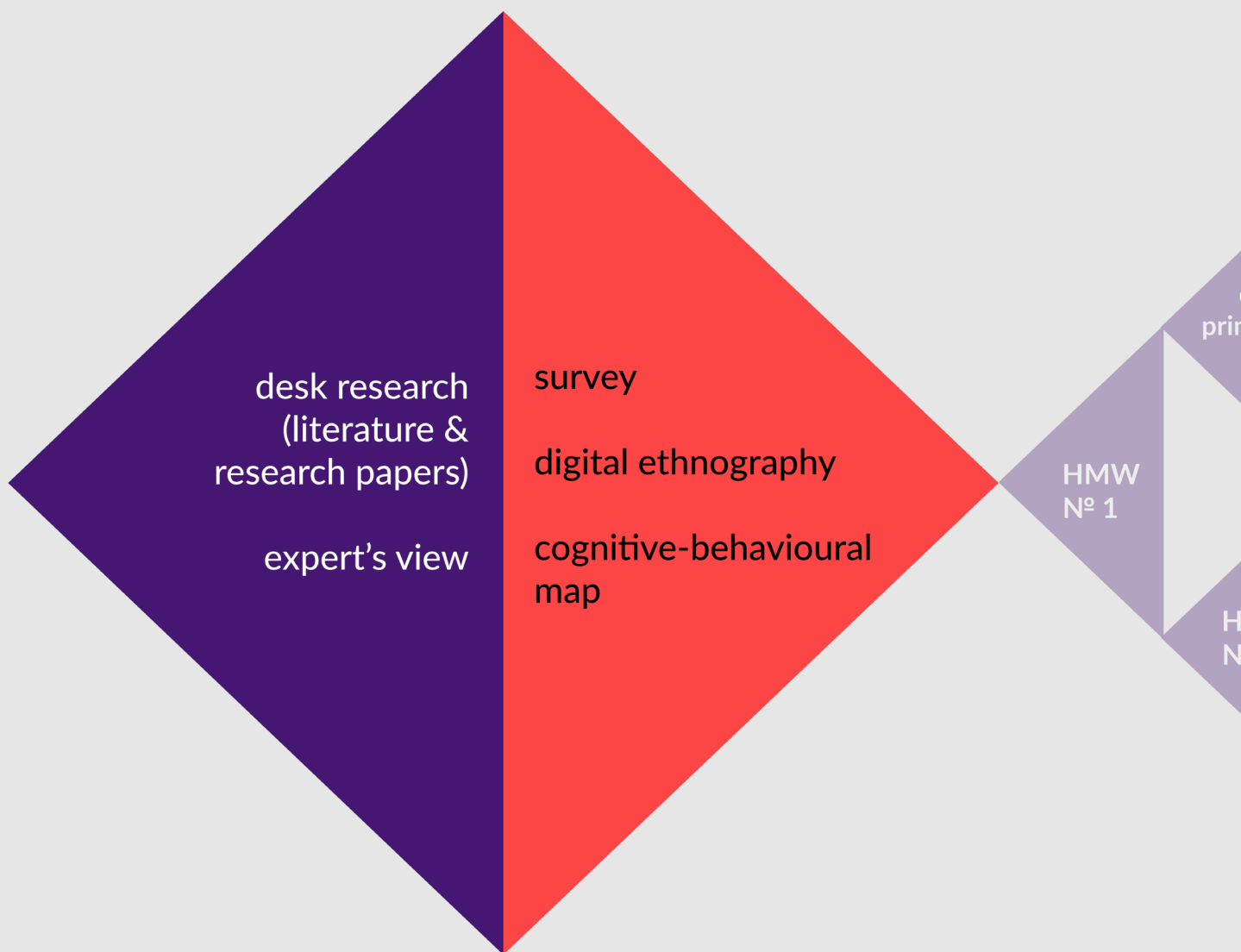


Fig. 4. Framing design principles on matrices

---

**PHASE 1** January – February

# RESEARCH



## RESEARCH OVERVIEW

In this section, I present findings from preliminary research, focus definitions, and insights from expert interviews that provided direction for my next steps.

### INITIAL DESK RESEARCH

At the starting point of this project was the problem of attitude polarisation in interpersonal interaction, which I wanted to address. Attitude polarisation is a process in which groups of people who share opposing beliefs become more radicalised and hostile toward one another.

The psychological framework of dual process theory explains attitude polarisation by describing two modes of human cognition: automatic (System 1) and conscious (System 2) thought processes (3). According to the theory, the System 1 mode of thinking processes information in a rapid and intuitive manner. This is where mental shortcuts take place, also known as cognitive biases. Conversely, the System 2 mode of thinking is “the slower, analytical mode where reason dominates” (4).

According to this theory, the attitude polarisation problem is one of the consequences of the irrational System 1 mode of thinking. Humans inherently tend to rely on automatic thought processes “to search for and interpret evidence to reinforce their current beliefs or attitudes” (Cooper & Thomas, 2019).

My personal observations and the background knowledge suggested, that the problem of irrational

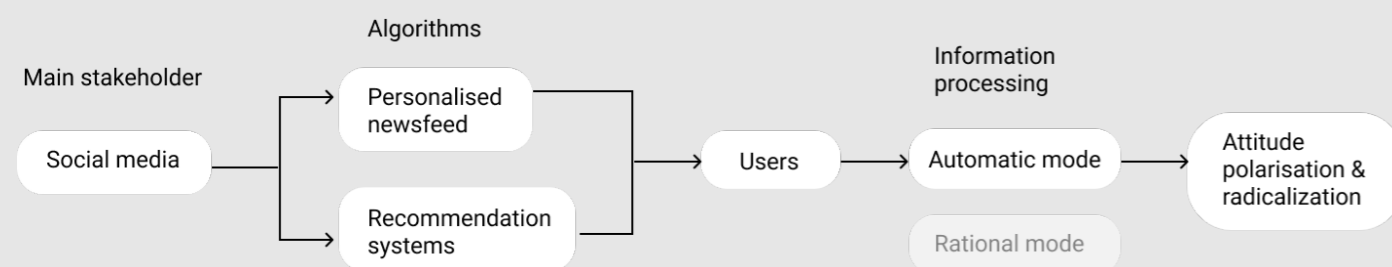


Fig. 5. Initial problem statement

attitude polarisation in the context of the modern world was a direct result of algorithmically personalised social media news feeds.

It is also described as a “filter bubble” phenomenon: “...A filter bubble is the intellectual isolation that can occur when websites make use of algorithms to selectively assume the information a user would want to see and then give information to the user according to this assumption ... A filter bubble, therefore, can cause users to get significantly less contact with contradicting viewpoints, causing the user to become intellectually isolated ...” (Technopedia, Definition – What does Filter Bubble mean?).

Putting these preliminary research findings together, I saw that there is a cause-and-effect relationship between AI-driven social media news feeds, the reinforcement of cognitive biases, and the increasing attitude polarisation & radicalisation in society (Fig. 5).

I hypothesised that it was possible to optimise algorithms to promote rational thinking that will eventually lead to the reflective attitude and collaborative behaviour between users (Fig. 6).

### EXPERT'S VIEW

However, after several discussions with experts in the fields of psychology, sociology, philosophy, political science, and AI, I found that my prior assumption was only partially true and that the real problem turned out to be far more complex

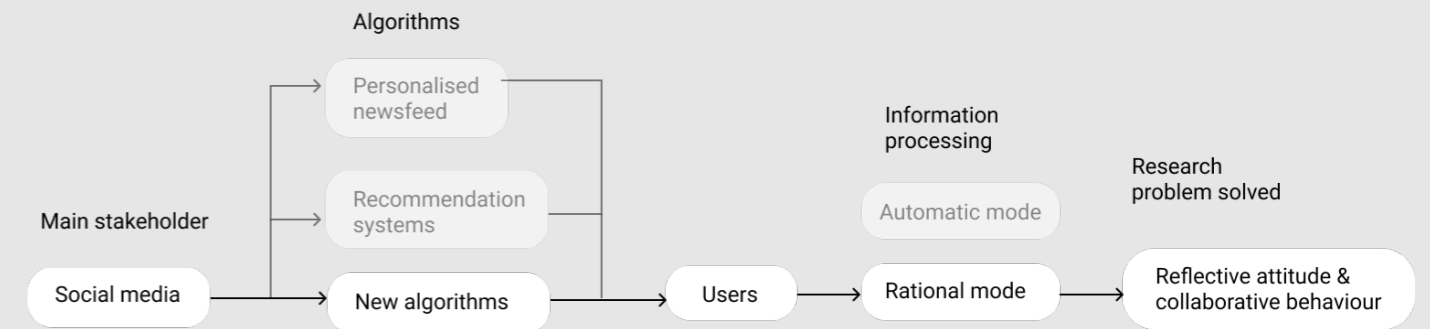


Fig. 6. Initial hypothesis

(Fig. 7). For example, the sociological perspective denied technological determinism. It was not just algorithmic proposals that people polarised on. It was due to the general modernisation and decline of religion, which previously served as a unified social structure upon which people could build their lives (Berger, 1973). After the church lost its status, people began to look elsewhere for guidance and meaning, and everyone was able to find something for themselves. This led to a pluralism of moral principles (the Homeless Mind concept), which could also become the basis for the trend towards polarisation of attitudes that I have observed.

Interviewing an expert in psychology made me rethink my views on radicalisation as a completely irrational process. Although there are 2 different types of thinking, they do not occur separately. Instead, these 2 types complement each other. Therefore, it would never be possible for me in this research to figure out which extreme beliefs are rational or irrational in nature and how to address the rational part of human cognition.

From a philosophical perspective, part of the problem arose from the way people communicate in the digital world. Screen-based verbal communication missed the crucial point of emotional connection and physical contact. Conflict resolution would still remain on the physical points of contact.

In my interviews with AI experts, social media design was also mentioned as part of the polarisation problem. An interesting comparison

was brought up: In the real world, when we walk down the street, we meet a lot of people, but we don't usually approach every person we see and ask them about their beliefs. But in the online social experience, we are exposed to too many opinions from people we don't even know – imagine walking past people on the street and having to deal with each person's opinion. So this social overexposure leads to a distortion of the experience of ordinary human interactions.

Moreover, AI experts pointed out that personalisation algorithms are designed to keep engagement high, which is at the core of the financial revenue model of social media platforms. And, as recent studies show, “what triggers more engagement online is content that triggers strong emotional response, with anger, for one, being a powerful driver of such. Given this, algorithms, whether intentionally or not, are basically built to fuel division, via the more practical business aim of maximizing engagement” (Hutchinson, 2021).

My interview with the political expert brought to light another important insight. Polarisation is a desired state of society in political games that aim to “divide and conquer”. Cambridge Analytica Scandal was a vivid example of an election won through the spread of Fake News, politically targeted advertising, and radicalisation of previously neutral users on Facebook (Noujaim and Amer, 2019). A few years after the scandal, the rift in the States has gone even further. On January 6, just a few days before this interview, far-right extremists laid siege to the Capitol, and the U.S. president



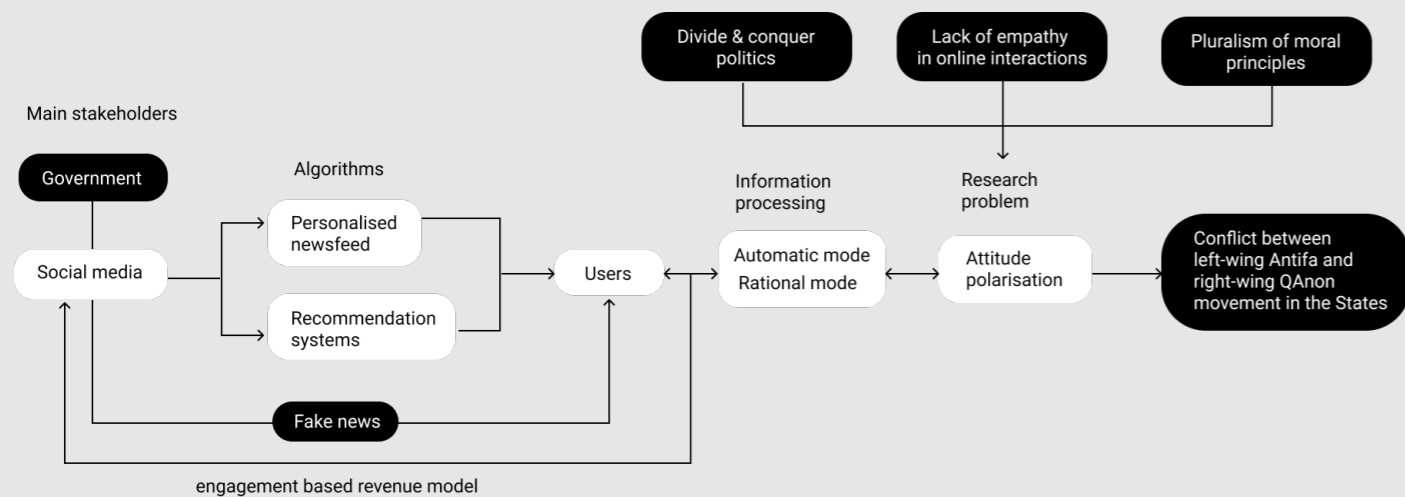


Fig. 7. Research problem after expert interviews

Donald Trump was suspended from Facebook and Twitter for inciting his followers to protest election results that proclaimed his left-wing opponent Joe Biden the new president.

The more one side of the spectrum radicalises, the more it gets opposed to the equally extreme niche on the far-left. The independent online platform The Shift sums up this polarisation case:

“Such violence isn’t the monopoly of the American far right. This collective hysteria exists on both sides of the bitter US political divide. While Trump has deliberately fed into one side’s narrative that a militia of morbidly obese but heavily armed ‘patriots’ is the only thing defending the US from being overthrown from within, the other side feeds their followers with yesterday’s bogeymen in the form of fascists and Nazis.

We saw similar attacks on the media by Antifa activists during Seattle and Portland’s Summer of ‘20, where the looting of businesses that had nothing to do with a police shooting in Minneapolis somehow represented “social justice”.

This is what happens when facts are thrown out the window and replaced with one’s preferred worldview, supplied in stronger and stronger doses by your local Silicon Valley pusher whose clever algorithms give you steady hits of the poison you crave... Both sides seek to destroy the existing social

structure, a goal which is presented as an end in itself. Neither narrative can be defeated through rational argument or the marshalling of facts, because truth is now a matter of faith” (Murdock, 2021).

This case of polarisation was particularly relevant to this project because it encapsulated all aspects of the problem raised in my previous research.

Another reason why I chose left-right political divide in the U.S. as a research subject is my relative unfamiliarity with it. This could help me not to take any stance and be as unbiased as possible.

This is also important to mention here that my main goal was to study the human psychology of cognitive biases in the conflict from the interaction design perspective, not the U.S. political case of polarisation itself.

#### ADDITIONAL DESK RESEARCH

First group of radicalised users from the far-right side belonged to the QAnon movement. QAnon is a conspiracy theory that says that there’s a global elite of individuals that are ruling the earth behind closed curtains. It started in October 2017, when an anonymous user called Q published a series of posts on the message board 4chan, and his followers there were called Anons. The user claimed he had connections to the US security agency. According

to Q, president Trump was waging a secret war against elite Satan-worshipping pedophiles in government, business and the media.

BBC reported that “social media and opinion polls indicate there are at least hundreds of thousands, if not millions, of people who believe in at least some of the bizarre theories offered up by QAnon” (Wendling, 2021).

On January 6, QAnon followers escaped their virtual headquarters of social media and sieged the Capitol. “The siege on the U.S. Capitol played out as a QAnon fantasy made real: The faithful rose up in their thousands, summoned to Washington by their leader, President Trump... The “#Storm” envisioned on far-right message boards had arrived. And two women who had died in the rampage — both QAnon devotees had become what some were calling the first martyrs of the cause. The siege ended with police retaking the Capitol and Trump being rebuked and losing his Twitter account. But the failed insurrection illustrated how the paranoid conspiracy theory QAnon has radicalized Americans, reshaped the Republican Party and gained a forceful grip on right-wing belief” (Harwell and Stanley-Becker, 2021).

As I approached the subject, it turned out that Facebook and Twitter banned hundreds of groups associated with QAnon (Ortutay, 2020) because they posed significant risks to public safety. A search on Twitter or on Facebook did not turn up anything associated with this movement. I managed to find the far-right portion of my target audience on gab.com. Gab is an American social networking service known for its far-right user base, which reached four million in March 2021. This site attracted users who were banned from other platforms.

The other extreme group of social media users belonged to the Antifa movement. It represents virtually everything that contradicts its far-right counterparts: people who engage in the movement tend to hold anti-authoritarian, anti-capitalist, anarchist, and anti-fascist views, and support BLM and LGBTQ+ values.

Members of Antifa were also involved in violence and riots in January 2021:

“Left wing radicals went on the rampage in a number of cities just hours after President Biden’s inauguration smashing up buildings, clashing with cops and burning American flags, according to police and reports. As cities across the US were on high-alert for Trump-supporting right-wing militias, they were instead attacked by members of Antifa, who were assailing the Democratic president for not being left enough for their liking. Portland, Oregon, and Seattle, Washington the main flashpoint cities for riots last year saw hundreds of militants trashing buildings, many expressing outright fury at Biden’s call for unity” (Ngo, 2021).

Finding the leftist part of my target audience was much easier because they weren’t spreading the agenda that was considered dangerous by the big tech giants.

So through Gab and Twitter, I found these groups of people to invite them to participate in my research. I planned to conduct interviews to learn how these people turned to radical beliefs, the values behind those beliefs, how they consumed information, and the role social media played in this.

## USER RESEARCH

### SURVEY

To recruit interviewees from both sides, I created an online survey. The survey was based on a political bias test developed by the Criticalthinking.org platform. It started with questions that asked respondent's political affiliation, preferred media (such as the TV, newspapers or social media) and channels of news consumption, where each side was represented by four media outlets. The survey was followed by questions that addressed empirical facts and had a correct answer, but at the same time, they were politically controversial. Thus, the answers to these questions would reveal whether the respondents were politically biased based on their preferred media and channels of news consumption. The full survey is provided in the Appendix 1.

The survey was posted from my personal account on the Gab and Twitter platforms. However, within three days, it did not produce the expected results. I received eight responses, which was too few given the scope of the problem I was trying to address.

### DIGITAL ETHNOGRAPHY

(disclaimer: 18+, for ethical and privacy reasons, usernames and quotes are partly blurred)

So, together with my mentor, I decided to change the method of user research and try digital ethnography. The goal of this study was to embed

myself in the online social circles of the target groups, to provoke them and ask questions that I would otherwise ask in interviews.

Due to a previous desk research and survey results, I was already more familiar with my target users. So I created two personas from both sides of the political spectrum based on that knowledge.

The name of the first persona is Rhonda Kayser. Rhonda was born in Texas in 1964. She describes herself as a proud mother, Christian, patriot, Warrior for Truth, Q follower and Trump supporter.

Rhonda was an active Twitter user – it helped her keep in touch with her relatives and find comfort with like-minded conservatives and patriots, and of course keep up with her role model Donald Trump. In some of his speeches, Donald mentioned QAnon supporters as patriots who love their country. This very much coincided with Rhonda's views, and she decided to check out who these people were. This was a moment of great awakening for her. The Q Theory opened her eyes to the secret cabal of Satan worshippers, cannibals and pedophiles who ran a global child sex trafficking ring and conspired against her beloved president Donald Trump. As a mother, she was deeply concerned about the suffering of children. The theory very well explained her gut feelings of disgust towards left leaning politicians and leaders – they were involved

in these crimes and covered up the truth with the help of the corrupted media.

Rhonda pledged to help her president in this fight and save children with her resources – by spreading the truth online and joining with other patriots to be ready for the final battle against the cabal, for the battle of good versus evil.

She updated her news feed and found more and more evidence that reinforced her beliefs about Q and Trump. She couldn't understand how her relatives didn't seem to care at all about their country. She tried to wake them up with posts and videos, with evidence that the deep state apparatus existed and should be overthrown, but the relatives just mocked her as a conspiracy theorist and some even blocked her. "They are already brainwashed, – Rhonda thought, – but eventually I will make them swallow the red pill". She felt offended by this label, but she kept posting. Until she was temporarily

banned from Twitter for spreading "fake news".

One would think this would rattle Rhonda, but instead it backfired and strengthened her convictions even more. "If they are silencing me, then I certainly know something they are trying to hide".

In this desperate moment of Rhonda's life, I registered her account on Gab where she felt safe and supported by like-minded people. This is where my journey with Rhonda as a researcher began. To create her account, I used the online face generator Thispersondoesnotexist.com and googled patriotic wallpaper to set it as a background image (Fig. 8)

At the same time, I created the second persona from the far left. Josh Baginsky is a 30 year old BLM and Antifa activist from Portland. Josh joined the BLM movement exactly one year ago when Breonna Taylor was fatally shot and killed by police officers

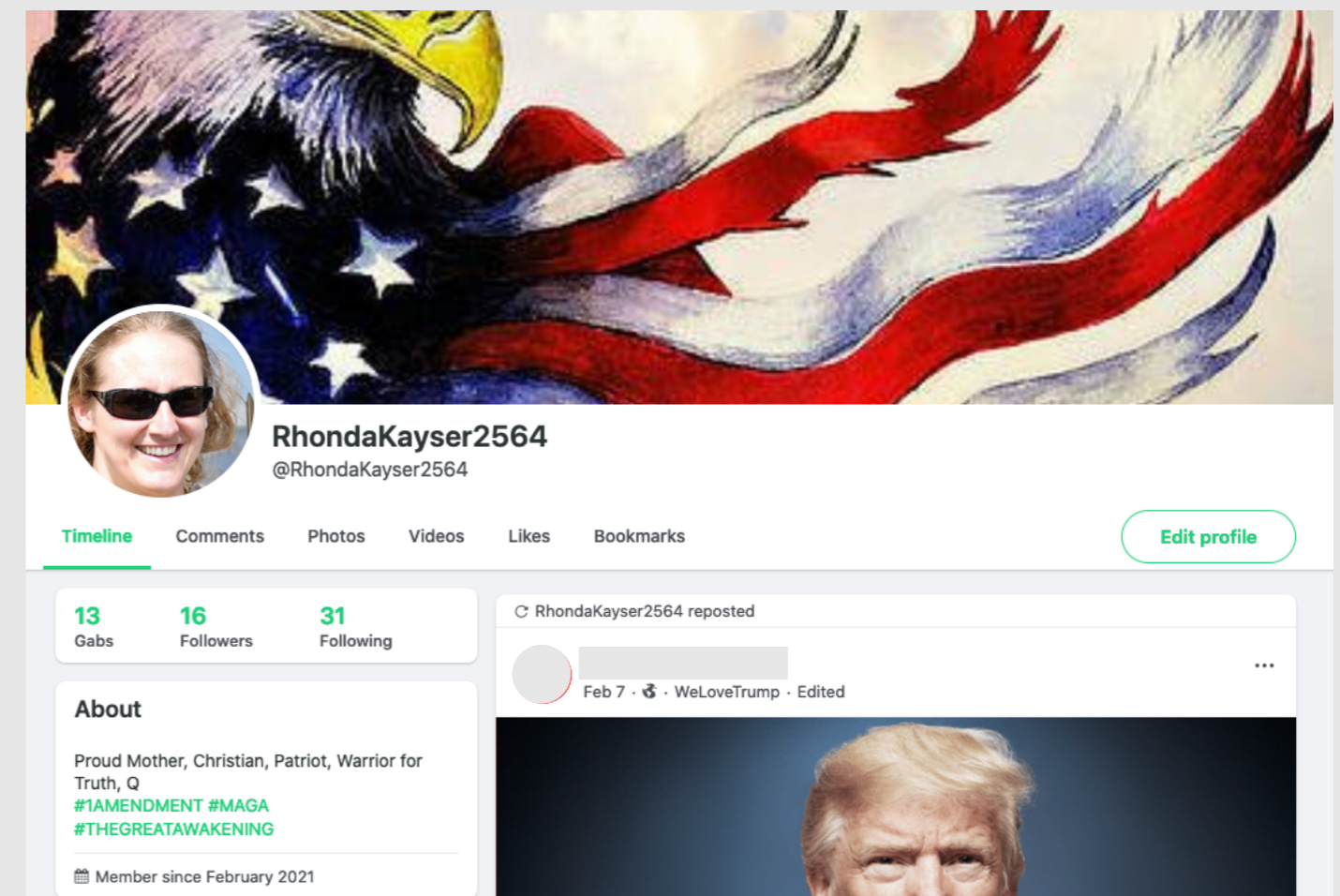


Fig. 8. Personal page of the far-right persona Rhonda Kayser



during a botched raid on her apartment. Josh demanded justice for Breonna, went on riots and smashed up police officers' cars in revenge.

When the pandemic of Covid-19 started, his activism shifted from the streets to the online world. Before the lockdown, he followed Antifa groups and forums and communicated with like-minded people, but this time social media became his only point of contact with the outside world.

Josh saw the world through the lens of his news feed – it was cruel, unjust and racist. Other Antifa activists, radical left-wing media, and the support of his followers fueled that anger even more. From what Josh could see, it was objectively the right thing to do to demand justice for all discriminated black people.

But not everyone from his online social circle agreed. Josh couldn't believe one of his friends had

turned into a crazy Q and Trump supporter. How could she not see what was really going on in the world? After Josh commented on some of her "fake news" posts and tried to convince her and open her eyes, she responded with even worse resistance. Josh had to block her and they never spoke again after that.

Josh continued to call for revenge against police officers and Republican government officials who did not want to take serious actions to restore justice. He saw everyone who disagreed with him as "FoxNews"-brainwashed, who couldn't think critically or were just evil. After a few fights in comments, he was temporarily blocked from Facebook and Twitter.

Josh was outraged by this, but he still had to stay online to continue spreading the truth. In times of the pandemic, social media is the battlefield where he had to fight for justice.

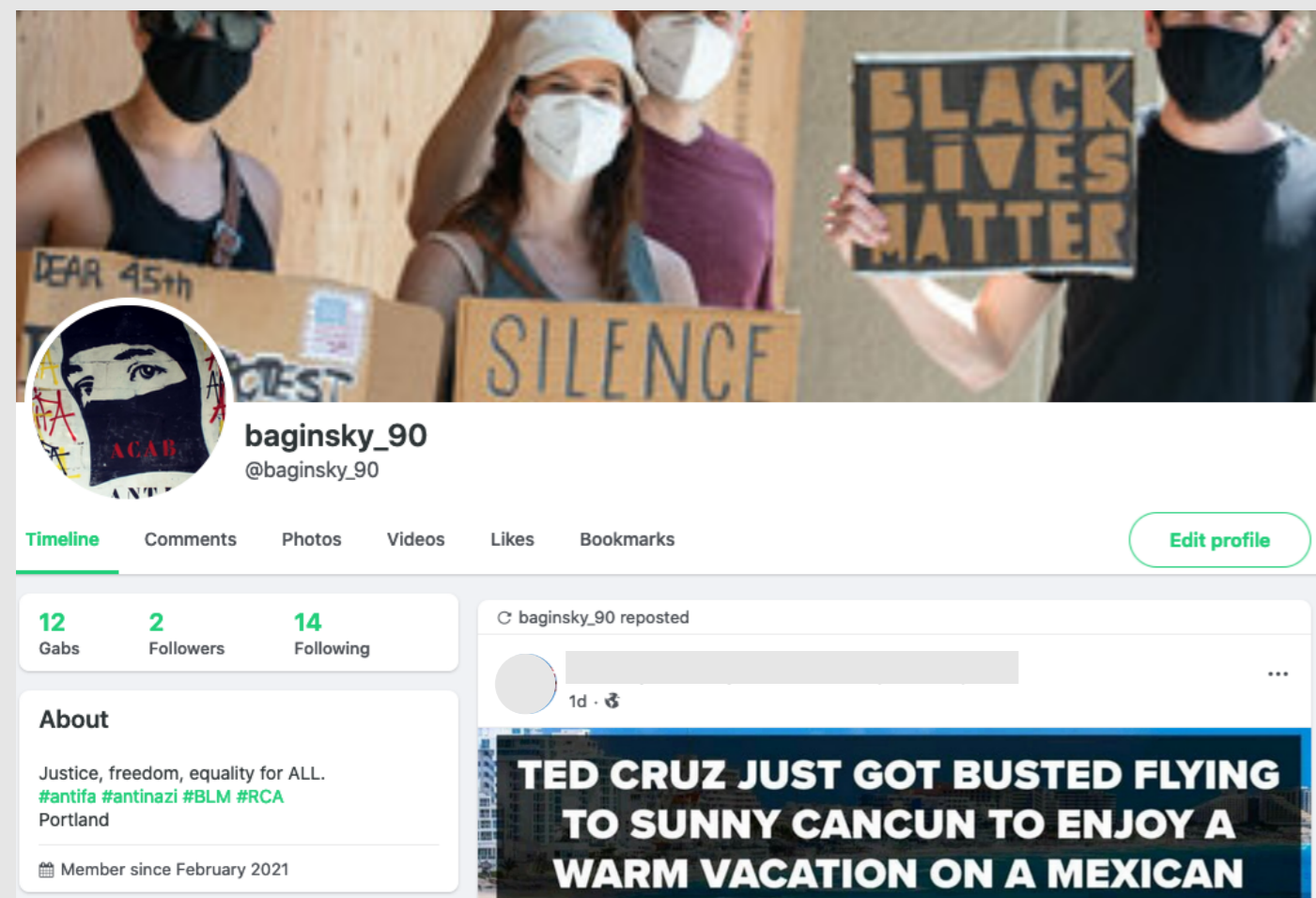


Fig. 9. Personal page of the far-left persona Josh Baginsky

Josh heard about the platform Gab.com, which supports free speech, it was mostly for the far-right, and that was exactly where his words would make the most difference.

In this desperate moment of Josh's life, I registered him on Gab. To create his account, I found a typical Antifa graffiti with a covered face and ACAB lettering as his profile picture (Antifa usually never show their faces), and I also set a background with rioting people (Fig. 9).

I, as the researcher, slipped into the shoes of my two personas. For two weeks, I lived the online lives of Josh and Rhonda on Gab and on Twitter. I befriended like-minded people, debated with them, followed exclusively left or right news channels and all the algorithmic suggestions. Day after day for two weeks, I logged into Rhonda's and Josh's accounts alternately. I took screenshots of posts and comments that answered my questions about faith formation and preservation, goals of people on both sides, and themes that were frequently repeated. In total, I collected 332 messages during this study. I analysed these messages and grouped them by the topic being discussed or the specific issue I wanted to highlight. Thus, nine distinctive groups of messages stood out: Introduction to Gab, Social Circle, Algorithmic Suggestions, Platformisation, Unity, Searching the truth, Dehumanisation, Reasoning. In the following paragraphs I will describe the content of each group.

### INTRODUCTION TO GAB

The first set of screenshots has been grouped around the theme "Introduction to Gab". On this platform, every new user is automatically added to the Gab founder's feed and to the "Introduce yourself" group. In this group, messages from new users appear every minute – I was surprised at how vibrant and popular this platform actually was.

After registering, Rhonda was delighted when she saw the first post in her news feed (Fig. 10). She then posted her first message and introduced herself (Fig. 11). She immediately received likes and followers.

However, Josh's introduction was more controversial. He introduced himself and received the comment "I don't think you will fit here" (Fig. 12) from a user with Adolf Hitler on his background



Fig. 10. First post from Rhonda's newsfeed

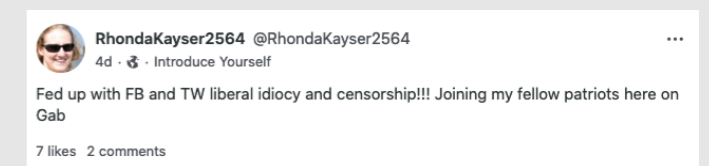


Fig. 11. First post published by Rhonda

image (Fig. 13). Although Josh should have been outraged by this greeting, I understood that for research purposes Josh should have a discussion (Fig. 14). A rational explanation of how the platform worked produced no results. After this introduction, Josh got zero followers.

### SOCIAL CIRCLES

In this group, I want to show which users surrounded my personas. For both of them and on both platforms, I followed opinion leaders and groups, as well as regular, less popular users.

In Rhonda's case, it was easy to find like-minded people on both Twitter and Gab. After one week of this study, Rhonda had 16 followers and followed 31 users on Gab, and she followed 12 accounts on Twitter but only received two followers there.

As mentioned in the previous section, Josh found it very difficult to find a social circle on Gab. After a week of this study, he followed 14 users, only four of whom shared his views and only two of whom followed him back. Twitter was a very different story. Josh quickly found 42 like-minded users, but again only received two followers back.



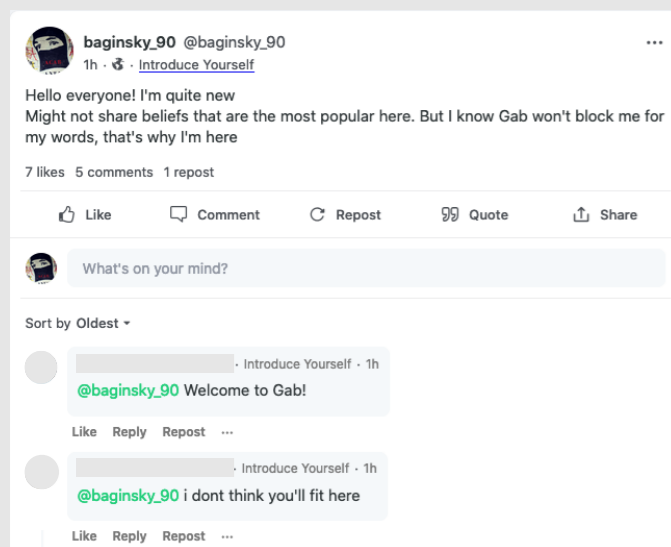


Fig. 12. First post published by Josh



Fig. 13. User with Adolf Hitler on his background image

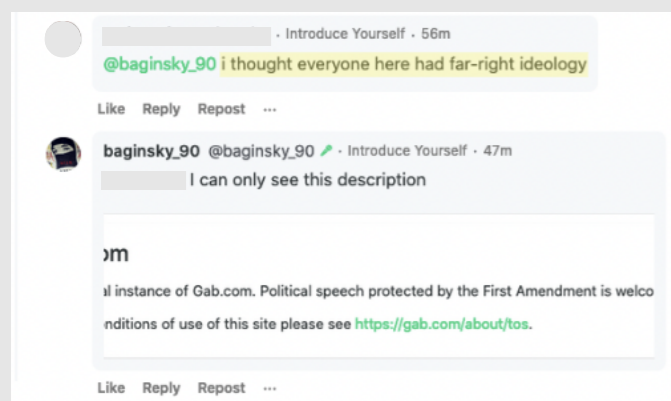


Fig. 14. Josh is having his first conversation

I explain this relatively low “mutual follow” activity in Twitter by the fact that this is already a very popular platform with an abundance of bots and my personas looked just like them, so it was difficult to get a follow back. On Gab, on the other hand, there were many users who had just signed up and it was easy to fit in with this stream of newbies.

This experience of making friends for Rhonda and for Josh showed to me how superficial the online relationships with the “like-minded” people are. Both of them were accepted to the community or declined only based on their background images and bio descriptions, which only contained a few hashtags and key words. Complex peoples’ personalities got shrunk to the size of the “About” tab or to the circle of their avatar. With this limited set of self-expression tools, users highlight only the most important facts, which in this case became their political affiliations. So, instead of seeing something unique in each person, by the design users rely on this shallow set of UI elements to make their judgement about the person. This is how social media profiles are designed for quick reactions, and hence stereotyping and labelling.

### ALGORITHMIC SUGGESTIONS

By default, Twitter suggests accounts to follow and tweets to check out “in case you missed them”. This is based on past interactions, so Rhonda and Josh would naturally receive completely different notifications and suggestions.

Gab had a different news feed architecture – the content ranking system was based on the “upvotes” and “downvotes” system. So the most emotionally charged and hottest topics were displayed first.

Both of my personas also used mobile Twitter app, and by default both of them got subscribed to notifications. These notifications were coming at least three times a day and usually displayed some groundbreaking and very emotional tweets. At some point, I could not recall who was logged in the app – Rhonda and Josh – because both of their algorithmic suggestions were equally catchy. Speaking of Gab mobile app, I could not install it because it was banned from the AppStore and GooglePlay.

So, from my observations, both algorithms were designed for engagement – either because

it reflected personal interests or because it hit emotionally and shockingly.

### PLATFORMISATION

Platformisation is a polarisation within many platforms. Users with different beliefs move into separate online worlds (into echo chambers) where they become even more radicalised. I was particularly struck by this trend when observing Gab users – some of whom were banned or censored from Twitter and had to socialise elsewhere, and Gab seemed like a “big love fest” to them (Fig. 15).

Twitter, like Facebook, was seen as purely evil source of violence, pedophilia, misinformation (Fig. 16). It became a matter of political belief whether to be active on Gab or on Twitter (Fig. 17).

This trend was not observed at all among the left-wing part of my target group on Twitter. However, some users from the right felt unsafe there and, even if they were not yet blocked, agreed in advance to meet on Gab (Fig. 18).

This trend showed me that by blocking accounts, Twitter was only furthering polarisation and losing its users.

### UNITY

The trend of platformisation goes hand in hand with the tribal need of people to be together, socialise, and unite against the common enemy. I observed this a lot in the feeds of both of my personas.

With Rhonda, for example, especially on Gab, it’s very evident in the “Introduce Yourself” group – people greet each other there and celebrate their unity almost unstoppably. Whenever I checked the group, posts like this one appeared practically every five minutes (Fig. 19). This can be explained by the fact that all new users in Gab receive an automatic suggestion to join the group. To me, it felt like this UX was designed as a rite of passage for newbies to feel like they are welcome to join this reunion of suspended users from other platforms.

For both Rhonda and Josh, joining with like-minded people was a way to show their superiority (Fig. 20, 21) and find confidence in numbers. Thousands



Fig. 15. Greeting message from a Gab user in the “Introduce yourself” group

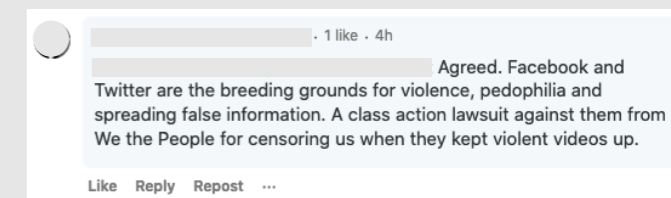


Fig. 16. Description of Facebook and Twitter from Gab user

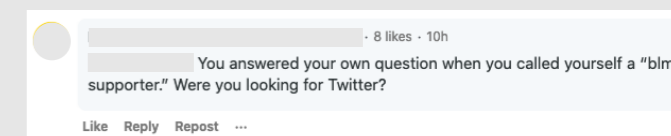


Fig. 17. Social media platform becomes a matter of political belief



Fig. 18. Twitter users agreeing to meet each other on Gab in case Twitter blocks them



Fig. 19. Celebration of unity (right)



of likes and reposts displayed on the post that supported their opinions was a clear sign that they were socially approved.

### SEARCHING THE TRUTH

The search for truth on social media became a matter of one's political beliefs, not facts and evidence. The search for information turned into the search for confirmation. This group of messages shows how desperate users on both sides were when it came to finding the truth and proving it to others (Fig. 22). The phrase "Do your research" was often repeated, but some users were at a loss when it came to finding a trustworthy source (Fig. 23).

And it seemed that none of the media channels aimed for objective reporting. Even when both sides described an event, it was to be seen completely



Fig. 20. Celebration of unity (right)



Fig. 21. Celebration of unity (left)

differently from Josh's and Rhonda's perspectives. It became especially clear when I saw different coverage of the same event from Josh's and from Rhonda's accounts. There was video of Trump's speech, and different outlets cropped and edited it differently to better suit their agenda (Fig. 24).

So, whether my personas observed the news feed based on the "upvotes" (Gab) or based on the past interactions (Twitter), they both saw a picture of reality that confirmed their or their group's beliefs.

### DEHUMANISATION

This is a group of screenshots in which both far-left and far-right users alike label each other's political stances as inherently evil, stupid and treasonous. Due to the nature of the Gab platform, this is where I first noticed it. Many discussions in Rhonda's

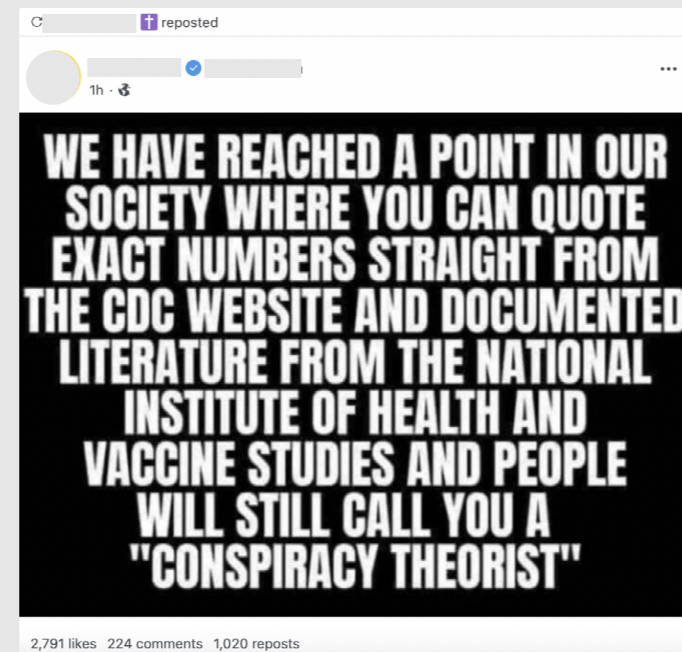


Fig. 22. Desperate picture about truth finding and proving

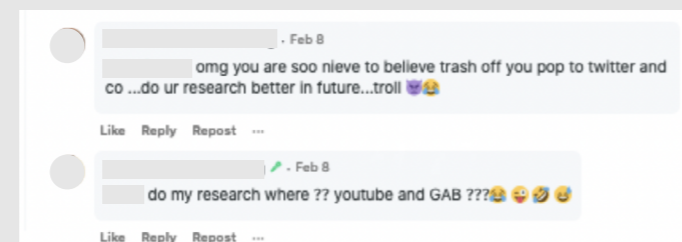


Fig. 23. Conversation about research techniques

feed have been full of hatred and violence against democrats and the proclamation of Antifa as a domestic terrorist organisation (Fig. 25, 26).

However, in the face of the same attitude of the left, it felt tiring for them to be "labelled as everything under the sun just because" of being a Trump supporter (Fig. 27).

Sadly for Josh, but valuable to me as a researcher, this hatred from Gab users eventually came back to haunt him. While reposting a meme with rats wearing a red MAGA caps and the caption "herd mentality" (Fig. 28), he received a notification that he had been tagged in a post (Fig. 29). This post immediately got 20 likes, ten comments, and five reposts – I couldn't ask for better publicity for my persona. Josh was very offended and humiliated, but remained calm and callous, which later lowered the temperature and led to a more in-depth conversation.

This discussion unfolded into a week-long debate, which I considered a focus group interview. This was a chance for me to observe my target audience on the real battlefield while balancing the role of my far-left persona. As a researcher, I wanted to know what irrational mechanisms were driving



Fig. 24. Tweet from RSBNC Network exposing manipulations of some media outlets

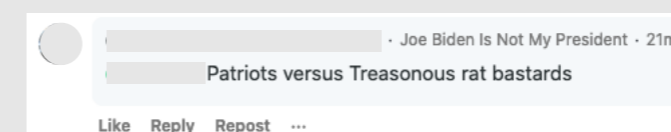


Fig. 25. Dehumanisation of the left

this discussion: why was Josh tagged in this post and what was the goal, and why were these people blaming Josh for all the evil seen on the left.

Josh's attacker responded that the reason he decided to attack was because Josh's background image promoted hate speech. This user admitted that it was funny because he had mentioned earlier in the discussion that the concept of hate speech was "objectively insidious and evil".

The rest of my questions were not answered directly, although I learned a lot from this focus group interview and observation. I noticed many very emotionally charged arguments against Josh based on a personal experience with someone on

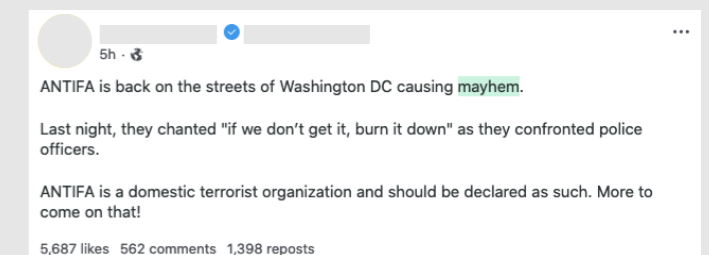


Fig. 26. Dehumanisation of the left

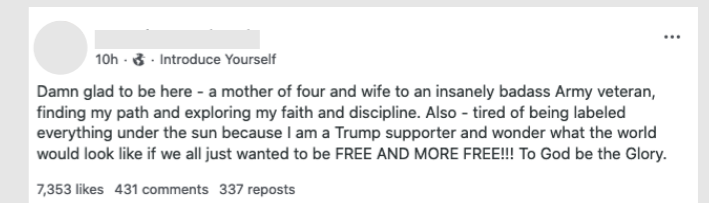


Fig. 27. Gab user complaining about being labelled



Fig. 28. Meme reposted by Josh



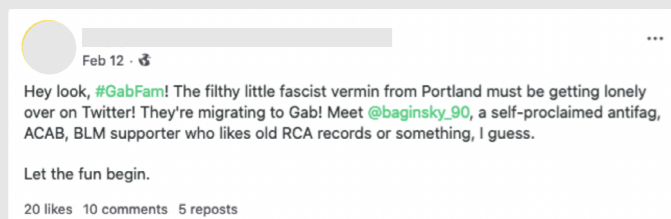


Fig. 29. Josh was attacked

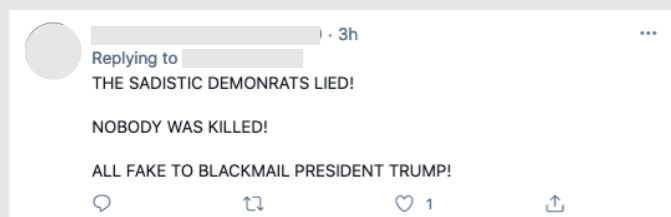


Fig. 30. Dehumanisation of the left on Twitter

the left. I tried to answer as neutrally and rationally as possible to get my questions answered, but they turned on me with even more facts, sometimes completely random facts to support their initial intuitive attack.

They seemed to enjoy this fight, too. They supported and cheered each other up with likes and reposts, seeing so many opinions reflecting their worldview was obviously very satisfying for them. In general, I was very surprised at how much effort and time these people were willing to put into fighting my imaginary Josh. It was also a surprise to me that some of their arguments seemed objectively correct to me as well.

On the subject of dehumanising “others” on Twitter, I had imagined less heated discussions there because the content was moderated.

But just as on Gab, Rhonda’s feed on Twitter was filled with claims about sadistic “DemonRats” (Fig. 30) who demonise Trump supporters (Fig. 31).

For Josh, I’m sure it was relieving to see his recent offenders also labelled as domestic terrorists (Fig. 32), traitors, and so on (Fig. 33).

### REASONING

Following on from the previous group, in this section I have collected statements that reveal thinking and information processing mechanisms in general. The discussions there were very similar to the previous



Fig. 31. “Dems accused of demonizing Trump supporters”

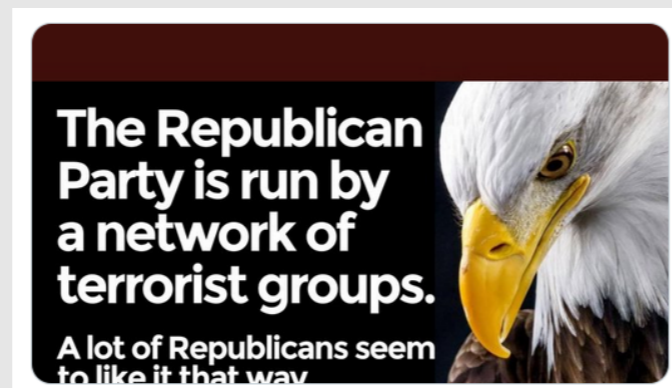


Fig. 32. Republicans are labelled as domestic terrorists

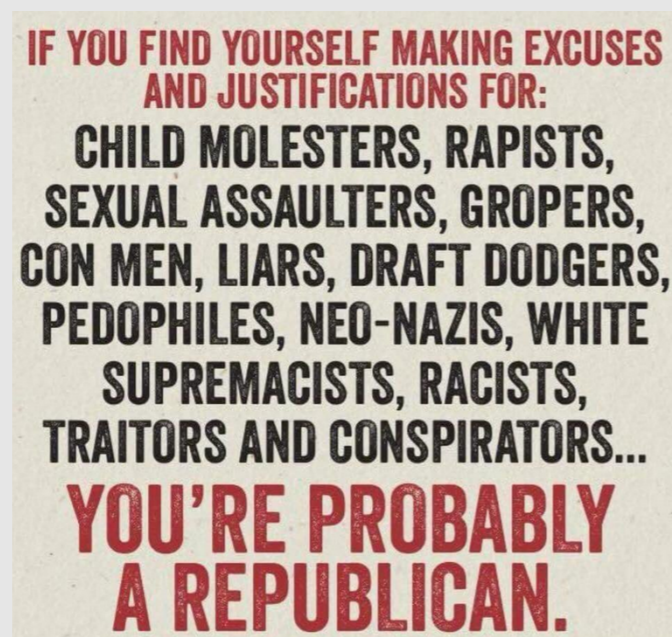


Fig. 33. Dehumanisation of the right

group, but they were more detailed, so I separated them to get a clearer overview.

The main finding of this group is that the arguments of both sides mirror each other. During some of the discussions, I even forgot which account I was logged into. To present this insight and perhaps check reader’s preconceptions, I’ll do a short test below.

### TEST: AS SAID BY...

Each of the following quotes belongs to a far-right and/or far-left user. The task is to guess which of the 2 parties said this. Clues are replaced with dots.

1. It does really feel like... have a monopoly on intelligence and integrity.
2. Reasonable, Strong, Logical women need to stand firm against the narrative – we don't have to be .... We don't have to believe their crap just because we are women.
3. Unfortunately we live in world where people rely heavily on the television for their information instead of taking a little time to do some proper research.
4. My morals and ideals and who I am as a person could not reconcile with being associated with people who breed hate, refuse to be informed and who would overlook an attack on our country.
5. Feels great to have my views and opinion validated... when so many family and friends were totally conned. Being on the right side of history is so satisfying.
6. The ... are fearful of me because they know their radical agenda won't get by me.

I hope this test sheds some light on the way information and opinions are treated by both sides: mainstream media can’t be trusted because it shows a manipulated view of reality, those who disagree are brainwashed, have no critical thinking, are lazy and stupid to do their own research, hearing the echo of your own opinion feels satisfying, but echo chambers are bad.

7. ... and ... are utterly unable to see their own hypocrisy.
8. Why does anybody even read the works of somebody who is so clearly evil and worthless?
9. They truly feel like they're on the correct side and that we're all evil monsters. And that's because they're fed constant lies which they devour.
10. Those who remain in the echo chambers have become more radicalised after January 6th, not less. I have had to cut off both friends and family.
11. Remember how before the internet people used to think the cause of stupidity was lack of access to information? Yea it wasn't that.
12. It's fake news. Stop spreading misinformation.

Answers: 1) Missed word: democrats, Answer: left; 2) Missed word: democrats, Answer: right; 3) Right; 4) Left; 5) Left; 6) Missed word: democrats, Answer: right; 7) Missed words: leftists, marxists, Answer: right; 8) Right (cancel culture rhetorics applied to Karl Marx); 9) Left; 10) Left; 11) Right; 12) Both.

## SELF-OBSERVATION

At this point I must mention another important finding from this study. These observations were made at a time when the political situation in my home country was very critical: earlier this year Alexei Navalny was arrested and masses of people, including my close friends, protested and were arrested. For several days my personal information bubble was filled with a constant flow of police violence and injustice. I was outraged and couldn't stop checking all my social media accounts to see if my friends were alright.

The only thing I could do for them was to post messages about their unjust arrest and spread the truth. I received mocking responses from some of my Facebook friends who said Navalny was an FBI agent and deserved to be in jail. I couldn't believe that there were people, especially in my online social circle, who could be so brainwashed, evil and treacherous to mock me at such moments. It also couldn't get into my head that anyone with access to the internet would still believe these propagandistic lies. I don't follow mainstream channels, I only trust the independent channels on Telegram. At some point, while updating the feed on Telegram, I accidentally confused a left-wing Russian media channel account with a QAnon channel I was following for research purposes.

This was a moment of bitter truth for me, because I suddenly realised that my information processing and thinking mechanisms were completely identical to those of the people I thought were biased and irrational. So who was I to rationalise someone else's online behavior and decisions?

After two weeks, things calmed down a bit, my friends were released, but the tensions surrounding Navalny's detention continued. I looked back at what had happened and realised I was still angry at my online friends who disagreed with me. Why was it so hard to be around people who didn't share my beliefs, even digitally?

On those days, I continued research by reading books on the subject of psychology and the human mind. And one of them provided me with an important insight that explained my feelings, as well as Josh's and Rhonda's.

In his book, "The Political Mind: A Cognitive Scientists Guide to Your Brain and Its Politics" (2009), cognitive linguist and philosopher G. Lakoff wrote that the understanding of the mind as something rational, logical, and acting on self-interest dates back to the old Enlightenment.

I understood that this state of the mind was exactly what I wanted to achieve with my design intervention: the elimination of bias and the establishment of more rational mechanisms of information processing, and thus addressing the problem of polarisation. But recent breakthroughs in cognitive science show that the human brain doesn't really work that way, because even our most logical thinking is subconsciously driven by emotion, and we can't physically get rid of that.

Also, we are culturally wired to interpret facts based on narratives we grew up with. Simply put, in fairy tales and other stories, there is always an "evil" character and a "hero" who fights them and in the end, good truth wins. This understanding of how reality works is physically embedded in our neural pathways. So when something controversial happens, we automatically give one side the role of deity and the other side the role of evil. And when we are presented with an opposing framing, it is dismissed because it doesn't fit into our system.

Another theory can be found in the book "Righteous mind: Why Good People are Divided by Politics and Religion" (2020) by social psychologist J. Haidt. In this book, Haidt offers an account of the origins of the human moral sense, and he shows how variations in moral intuitions can explain political polarisation in American society. According to the Moral Foundations Theory (MFT) described in the book, all ethical judgments are made intuitively, and rational thought comes later to justify the intuitive judgment.

The five innate and universal moral foundations by which we intuitively judge are:

1) Care/Harm: This foundation is related to our long evolution as mammals with attachment systems and the ability to feel (and dislike) the pain of others. It underlies virtues such as kindness, gentleness, and caring.

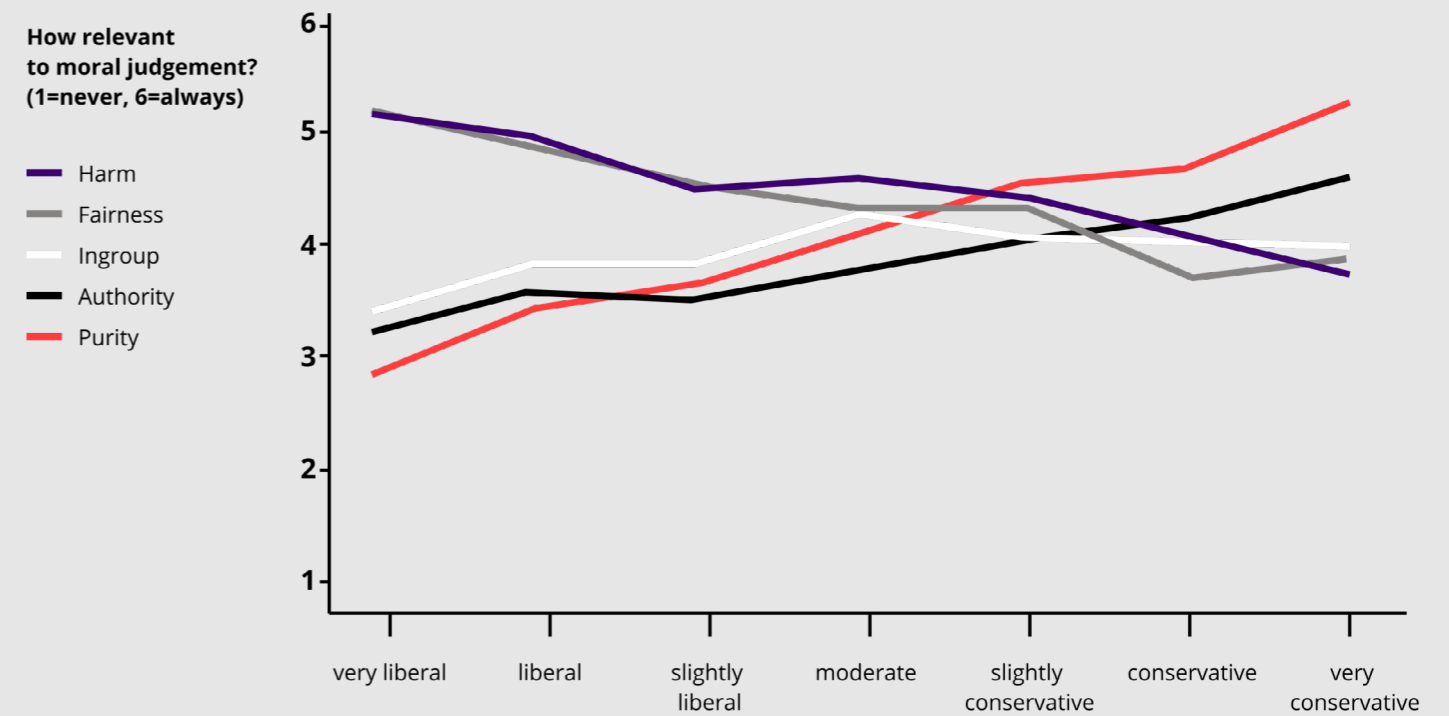


Fig. 34. The graph of moral foundations.  
Source: moralfoundations.org

2) Fairness/Cheating: This foundation is related to the evolutionary process of reciprocal altruism. It generates notions of justice, rights, and autonomy.

3) Loyalty/betrayal: this foundation is related to our long history as a tribal being capable of forming shifting coalitions. It underlies virtues such as patriotism and self-sacrifice for the group. It is active whenever people feel that it is "one for all and all for one".

4) Authority/subversion: this foundation was shaped by our long primate history of hierarchical social interactions. It underlies the virtues of leadership and followership, including deference to legitimate authority and respect for traditions.

5) Sanctity/degradation: This foundation was informed by the psychology of disgust and defilement. It underlies religious ideas of striving for a loftier, less carnal, more noble life.  
(Source: Moralfoundations.org)

According to MFT, the root of the disparity between liberals and conservatives lies in the

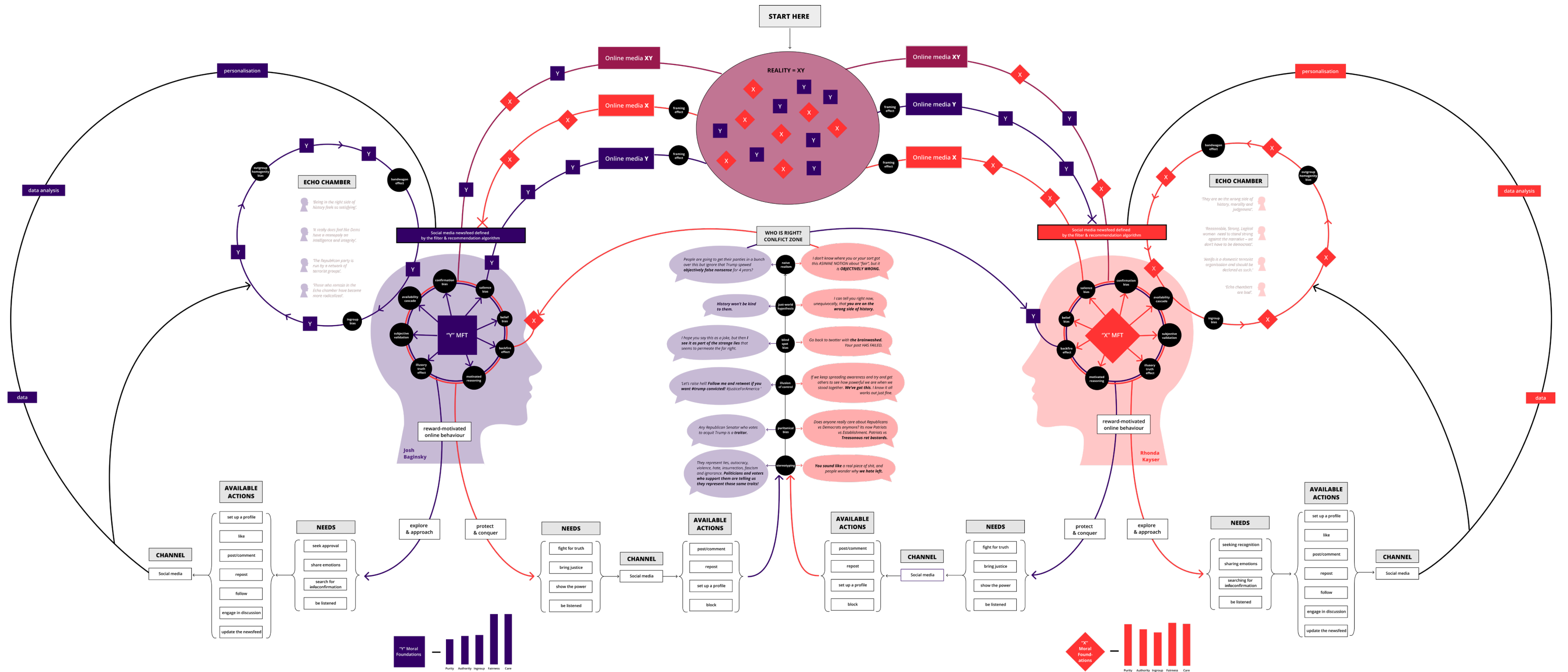
disparity in the weights that different sides place on these foundations. For example, people on the left value Fairness and place much less value on Authority, and vice versa for those who consider themselves conservatives (Fig. 34).

The digital ethnography study along with additional literature review helped me to see the information processing mechanisms, motivations, decisions, and pain points of my target users and empathise with them. I analysed all findings and mapped them as a user flow of all steps involved in the online polarisation process.

As a result, I created a cognitive-behavioural map (Fig. 35) that helped me identify potential design intervention points. In the next paragraphs, I describe how the map was made and how to read it.



fig. 35. COGNITIVE-BEHAVIOURAL MAP

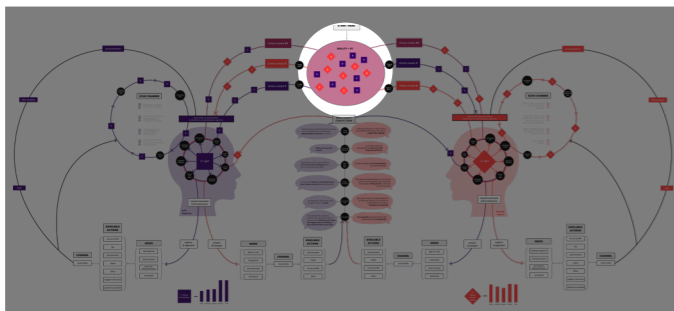


## HOW TO READ THE MAP

The map visualises the process of attitude polarisation in social media for my far-left and far-right personas. The mechanism is similar for both sides, but the content varies. Black dots placed along the entire flow each represent a cognitive bias (explanation is provided in the Appendix 2, source: Wikipedia, Definition – List of cognitive biases).

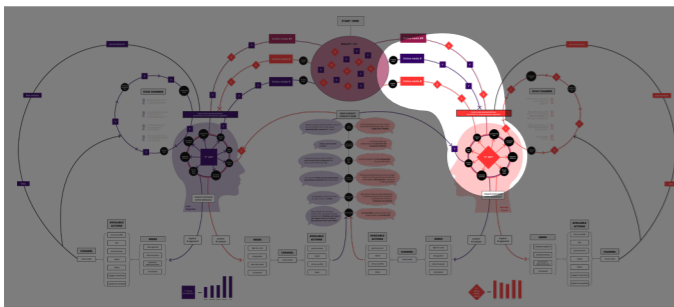
### START HERE

First, the process begins with the “Reality” bubble at the top centre of the schematic. “Reality” consists of facts, news, events that reflect either purple “Y” (far-left beliefs) or red “X” (far-right beliefs).



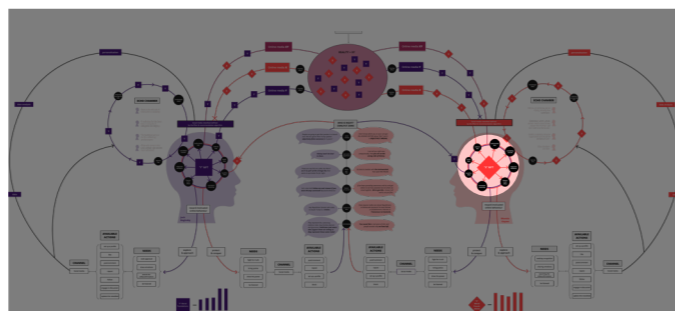
### FOLLOW ARROWS TO THE RIGHT

If we then move to the right side of the schema (Rhonda Kayser), there are three lines of “Online media”: the top one is neutral and therefore has both “X” and “Y” coverage, the middle media channel is biased left, and the last one down is biased rightist. All go into the red news feed “block” (“Social media news feed defined by the filter & recommendation algorithm”) and are either let in (neutral and red “X”), or rejected (purple “Y”). The news feed isn’t just made up of media channels, though – it also filters or includes posts and actions from a friends list. Also, despite the algorithm moderations, Rhonda sometimes meets the “Y” side of reality.



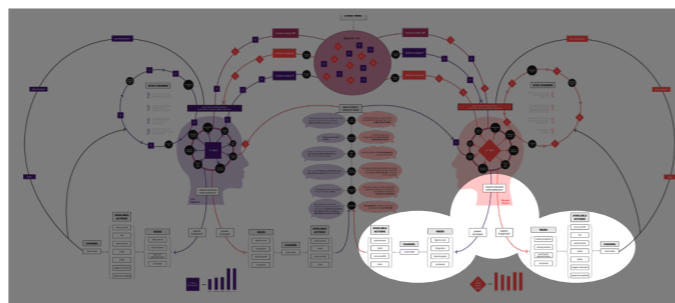
## INSIDE THE RIGHT HEAD

So all these channels go into the head of the far-right persona, and this is where the cognitive part of the map begins. The information-processing engine wheel is made up of cognitive biases that interpret the information, and is driven by the “X” MFT (moral foundations theory). Together, the cognitive biases and the “X” MFT decide whether or not the information fits the belief system and how to respond.



## REWARD-MOTIVATED BEHAVIOURS

After the information is processed, two types of reward-motivated online behaviours emerge. In Rhonda’s case, the purple “Y” information flow triggers the “Protect & conquer” behaviour, and the red “X” triggers the “Explore & approach” behaviour. Both are reward-motivated, as the former releases adrenaline due to aggression, and the latter releases dopamine due to the feeling of being right and belonging (Heath, 1963). Both neurotransmitters play a major role in the motivational component of reward-motivated behaviour.

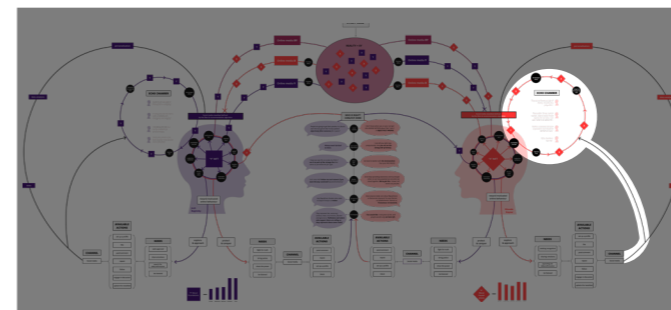


## EXPLORE & APPROACH BEHAVIOUR

According to my observations, the “Explore & approach” red line behaviour can be divided into several goals: seeking recognition, sharing emotions, search for infoconfirmation and self-expression. Each goal can be achieved with any of the seven actions available on the channel of a social media platform.

## FAR-RIGHT ECHO CHAMBER

Then the channel branches into two flows. One of them leads far-right persona to the red “Echo chamber” wheel and turns her “Explore & approach” behaviour into the feedback loop that constructs her news feed. Inside the “Echo chamber” persona’s beliefs are repeated, confirmed, and reinforced by members of her friends list using cognitive biases shared by the group. Inside the “Echo chamber” wheel there are several quotes representing these beliefs.



## DATA FLOW

The next overarching arrow that emerges from the social media channel represents the flow of data. This data includes Rhonda’s actions and informs the final algorithm that optimises their online news feed for more personalisation, emotion, and engagement.

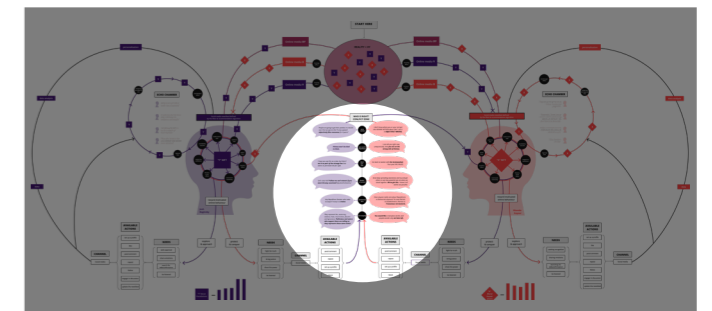


## PROTECT & CONQUER BEHAVIOUR

Then I should describe another reward-motivated online behaviour. This behaviour is triggered by the purple “Y” information flow that contradicts Rhonda’s belief system. In this case, she needs to “Protect & conquer” which can be expressed as several goals: fight for truth, bring justice, show power, and self-expression. To achieve these goals, there are various actions in social media. All unfold in the conflict zone of conversations with users from the left.

## THE CONFLICT ZONE

The conflict zone consists of speech bubbles with quotes from the right (red) and from the left (purple). Each quote is from a different discussion and each reflects a cognitively biased judgement. As can be seen on the map, the quotes from both sides fall under the same cognitive biases.

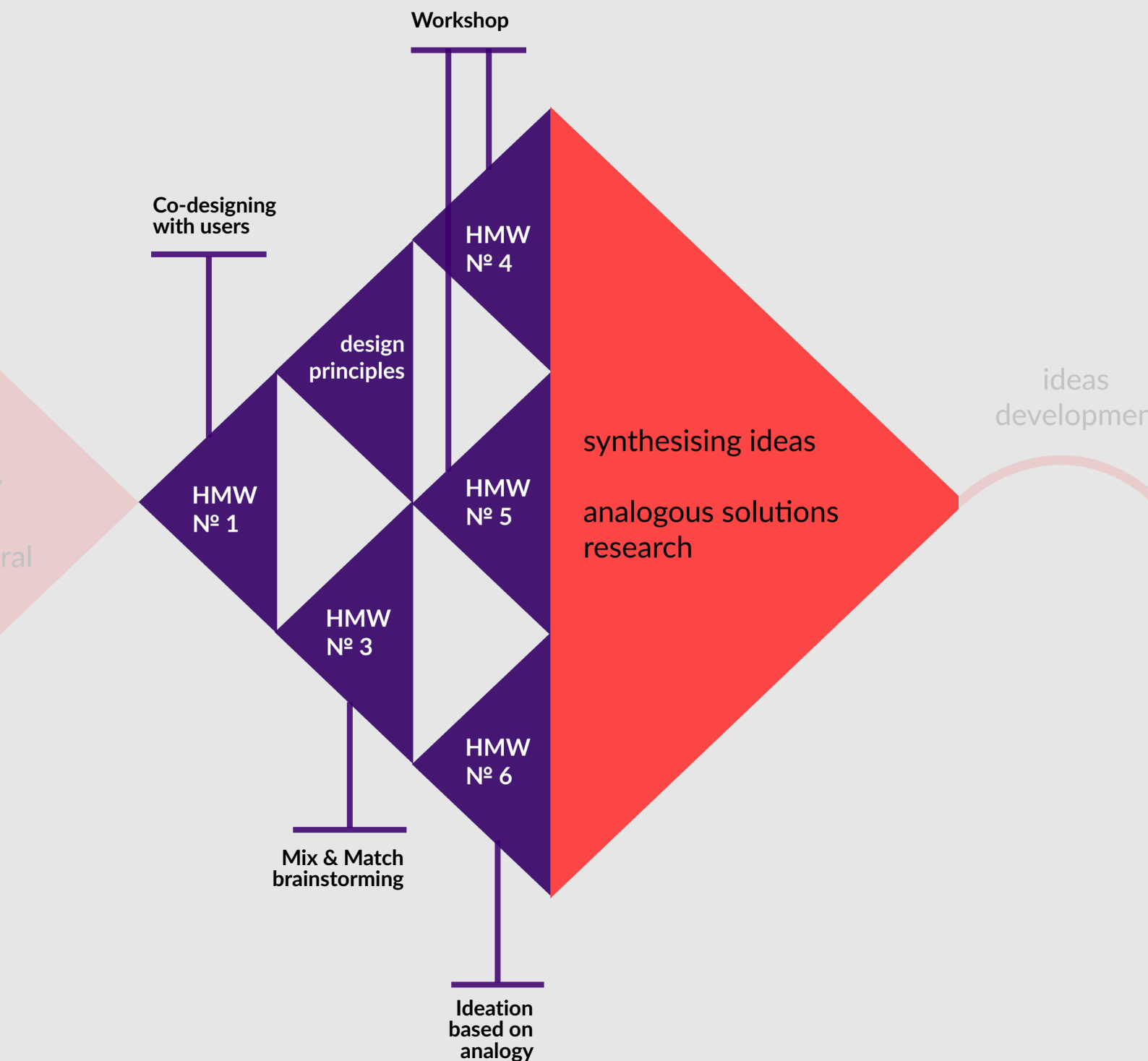


This is the online polarisation process of my far-right persona. The same can be seen on the left side – the same cognitive mechanisms and behaviours are involved. Although the quotes on the left are worded differently than the quotes on the right, they match in their meanings.

Even though the map is very detailed, I don’t claim it explains the problem exhaustively. It shows my observations as a design researcher, backed up by the findings from the literature review. Despite its complexity, the map is a simplified scheme, and the details of this scheme go as far as was necessary in the context of this project. The map was corrected and iterated after several feedback sessions with my mentor, supervisors, and the psychology expert. My goal with this cognitive-behavioural map was to use it as a thinking tool for further ideation sessions. For example, when thinking of the potential design intervention, I could apply it to the map and imagine how the next steps would be affected.

**PHASE 2** March

# IDEATION





In the classical design process, the central point of the Double Diamond represents a narrow problem formulated as a “How might we...” question. At this point in my own design process, however, I could see many possible directions and “HMW” questions, all equally worth exploring because they all had the potential to achieve the ultimate goal of this project – to unravel the loop of polarisation in the social media.

The cognitive-behavioural map offered several ways to approach ideation sessions. In total, I tried four different methods. In the next sections, I describe each of them.

## CO-DESIGNING WITH USERS

First, I decided to ask the opinions of my potential users to see if they even recognise the problem and to know if they already have their own methods to solve it. I simplified the map and posted it from Rhonda’s page on Gab with a following message (Fig. 36).

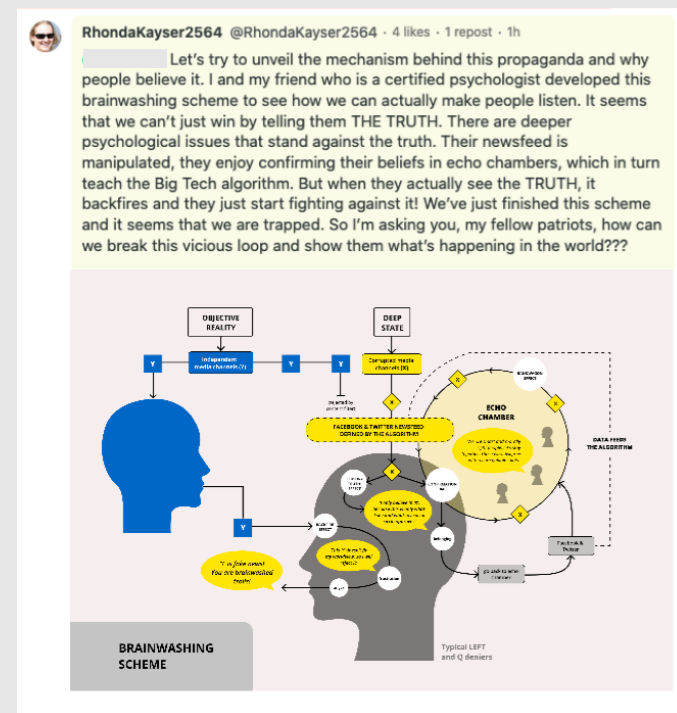


Fig. 36. Invitation to co-designing with users

However, users were not very active in responding. A few users responded with hateful messages about the left, but didn’t suggest anything.

I also posted it from Josh’s page, but received zero responses (Fig. 37).

We discussed this method with my mentor, and apparently the map was still too hard to digest for people who weren’t interested. So, having tried this approach, I decided not to include users in ideation sessions, because their involvement was very low due to technical limitations.

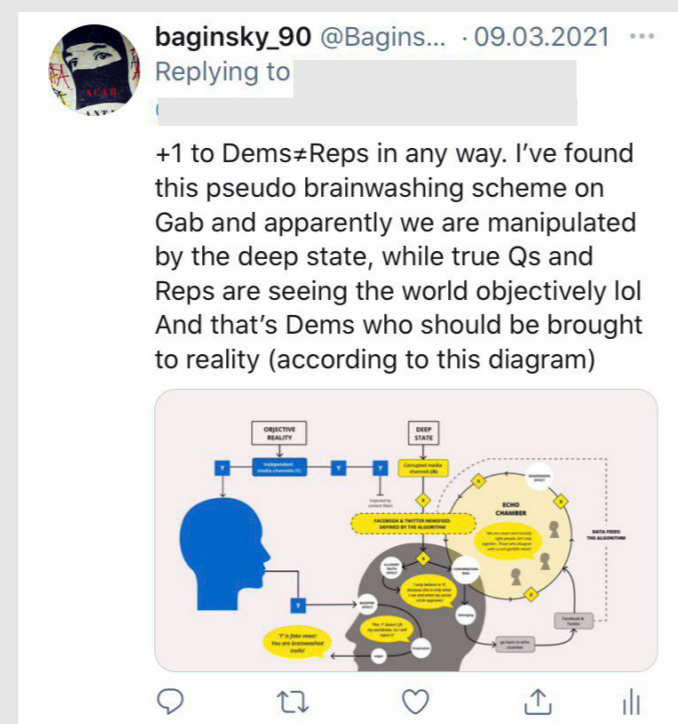


Fig. 37. Invitation to co-designing with users

## MIX&MATCH BRAINSTORMING

With the help of lessons learned from the previous session, I decided to plan the next one differently. I invited my classmates who could familiarise themselves with the whole complexity of the cognitive-behavioural map and be physically there to brainstorm with me.

I planned to start conquering the map by addressing the cognitive biases one by one. As I had learned from previous research, eliminating biases completely was technically an impossible task. So I took a different approach and asked: HMW make use of cognitive biases to promote positive online behaviour?

To address this “HMW” I planned to explore the mix&match brainstorming method. I listed each cognitive bias from the map with its functions and harmful effects in three rows, and another row represented the step where that bias comes into play (Fig. 38). Then it was a matter of mixing these steps and seeing how I could use the positive functions of one bias and reduce the harmful effects of the other.

For example, when updating news feeds, people tend to fall into a Confirmation Bias – they look for information that supports their beliefs. At the same time, people experience a Bandwagon Effect when they like and agree with highly voted posts.

So what if we mixed these steps: adding the most highly voted posts that contradict the person’s beliefs to their news feeds, thus beating confirmation bias with a Bandwagon effect?

We brainstormed together with my classmates (Fig. 39).

During the session, however, we found that a crucial point was missing. There was no clarity on what constitutes “positive” online behavior, or what my guiding design principles were when brainstorming.



Fig. 38. Preparation for the Mix&match brainstorming



Fig. 39. Mix&match brainstorming with classmates



## DESIGN PRINCIPLES

To have more clarity on what exact change in perception or behavior I wanted to achieve, I developed design principles. My mentor suggested I explore the method of matrices, where at the endpoints of the horizontal axis would be my personas and on the vertical axis would be the parameters against which I would measure them.

However, finding these metrics was no easy task. I decided to interview the psychology expert to discuss this. After the conversation, I understood that the question of “normality” is usually examined from a psychopathological perspective: if something is statistically rare, then we can register this as abnormality in a statistical sense. If we think about it in a functional sense (how people should normally do things), then it always falls under someone’s values and expectations. And the further we get away from basic human needs and functions, the fuzzier “normality” becomes.

So the question of positive behavior became a question of values and design. The expert mentioned that, for example, if I look at

“normality” from the perspective of online discussions, I could base my understanding on existing rules.

After the interview, I reviewed my previous research. Based on what I felt while playing Rhonda and Josh and what I wanted to accomplish as a designer, I came up with four value metrics (perception angle, attitude, tone of voice, and online activity) and placed them on matrices (Fig. 40, 41).

All matrices have extreme points representing Josh (far-left) and Rhonda (far-right) on the horizontal axes.

The first matrix represents the “angle of perception” and measures extreme points for “wide” and “narrow” angles. The red sticky note in the top center shows what a person would experience standing politically in the middle and having a very wide angle of perception – it would possibly lead to cognitive dissonance, the state I did not want to achieve with my solution. On the contrary, if the focus is down in the middle, the person falls outside the scope of this project. Combining the extremes on either side would lead to “all channels involved”

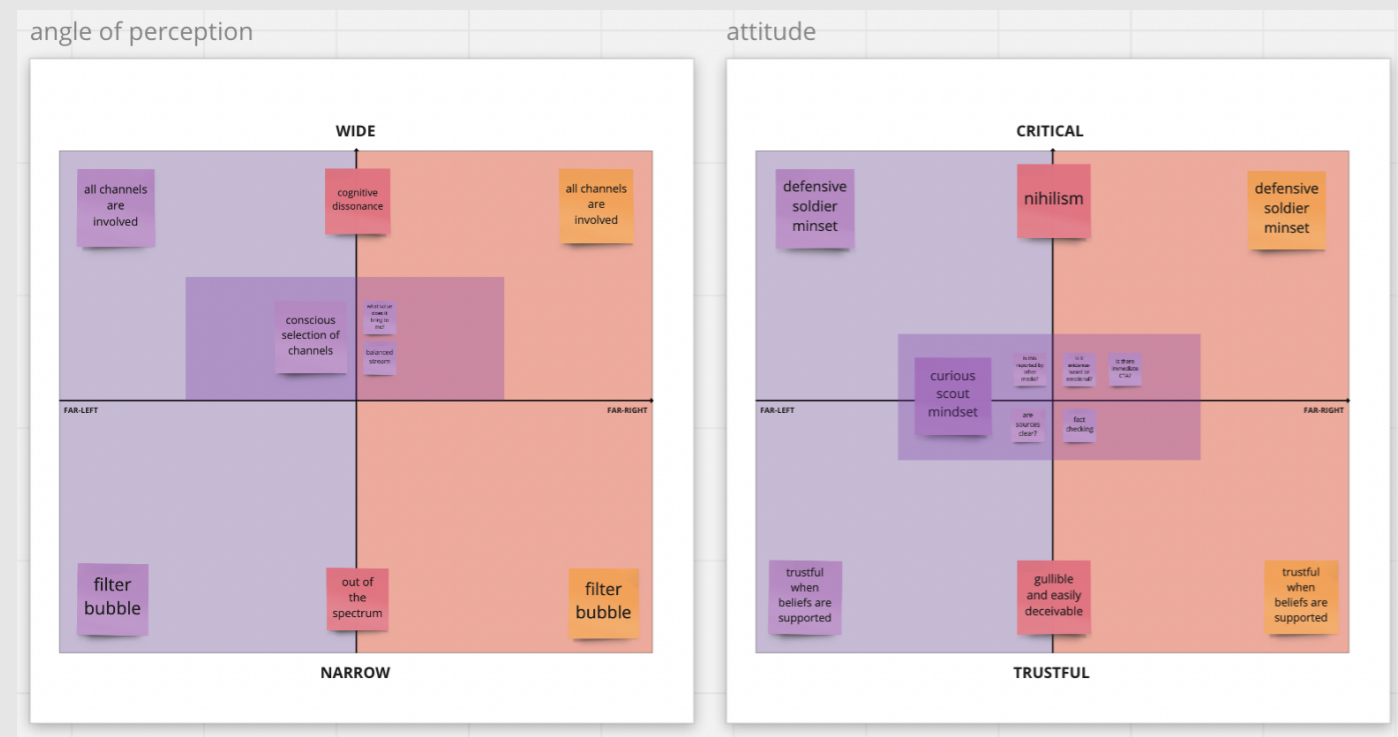


Fig. 40. “Positive” online behaviour measured on matrices

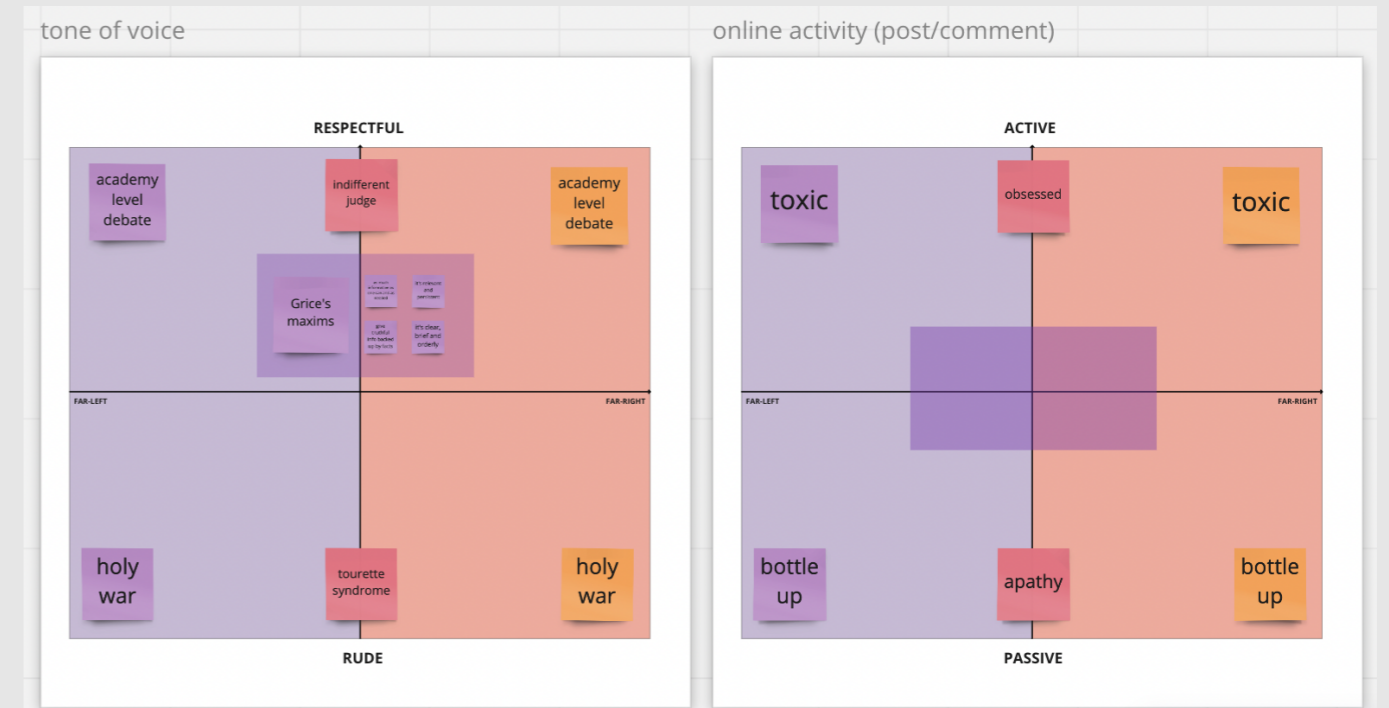


Fig. 41. “Positive” online behaviour measured on matrices

or the “filter bubble” which was the current state of my personas. The purple area with explanatory sticky notes in the middle highlights the state of cognition and behavior I was aiming for – it was about conscious channel selection and balanced information flow.

The same logic applies to all the other matrices. In the “attitude” matrix, I would try to avoid nihilism and gullibility, and aim for the “curious scout mindset”. In the “tone of voice” matrix, speaking like a cold, indifferent judge or like a person with Tourette’s syndrome would be undesirable, whereas following some rules of polite discussion would be preferred. The last “online activity” matrix was about achieving the state of serenity that lies between total apathy or obsession.

Overall, I had the following design principles.

### OPEN UP PERCEPTION

Widen the angle of perception from the filter bubble state towards a more conscious and diverse information flow (“angle of perception” matrix).

### DESIGN TO SPARK CURIOSITY

This principle is tied to the previous one, it emphasises that new or unexpected information

should be presented to spark curiosity instead of triggering the “defensive soldier” mindset (“attitude” matrix).

### MOTIVATE RESPECTFUL DISCUSSIONS

Nudge users to express themselves politely (“tone of voice” matrix).

### BALANCE ENGAGEMENT

This principle is tied to the “online activity” matrix and guides me to design for achieving the compromise between hyper-activity and total apathy.

Plus, design principles that did not come from matrices, but were still important:

### DESIGN FOR SELF-REFLECTION

This comes from the cognitive-behavioural map insight, where opposite personas reflect each other, but don’t notice it. With this principle, the aim was to enlighten users through their self-reflection.

### MAKE IT PROMINENT

Embed the solution in the current user flow shown on the map to address users and the problem right where they are rather than creating a separate space.



## IDEATION BASED ON ANALOGY

With the design principles outlined, I could set the clear goal for the next ideation session. The next session was about balancing the information flow part of the cognitive-behavioral map: HMW help Rhonda/Josh consciously consume and process online information flows?

In this session, I tried an ideation method based on analogy. I compared information consumption to food consumption, looked for examples of a good, balanced diet, and tried to apply solutions from that field to my field.

For this “Food for thought” session, I needed participants with a good imagination and a good appetite. I thought that my design colleagues from Block Zero perfectly fit the requirements, so I invited them (Fig. 42). Participants were following the process of this degree project since the beginning, so they were already familiar with the problem.

“Food for thought” was organised in the virtual space I prepared in advance: I designed the dining room, asked the participants for their favourite

dishes and put them on the table (Fig. 43). The idea was to discuss why we chose those specific dishes, would we eat only that for the rest of our lives and what would make us diversify our diet. Rhonda was also invited to the session – she argued that everyone should only eat sweets and that all other foods were fake. To play Rhonda’s part, I added her avatar to the video call, recorded her speech in advance, and played it during the discussion. In my opinion, this brought an unexpected dynamic to the session and provoked the participants. The scenario from this session can be seen in the Appendix 3.

This was my first experience with running brainstorming sessions with people from outside our masters program. I only planned it for one hour because I understood that the participants had their own studio’s work to do. One hour was certainly not enough to address everything I was aiming for, but I still discovered interesting perspectives on my topic.

I expected the food metaphor to naturally lead to a discussion about healthy eating habits, but the conversation evolved in a different direction. Participants talked a lot about how certain foods evoked good memories, how “comfort” food made them feel “at home”. I transferred these feelings to information consumption and saw how my “diversifying” intervention could take away pleasant moments from someone’s online experiences.

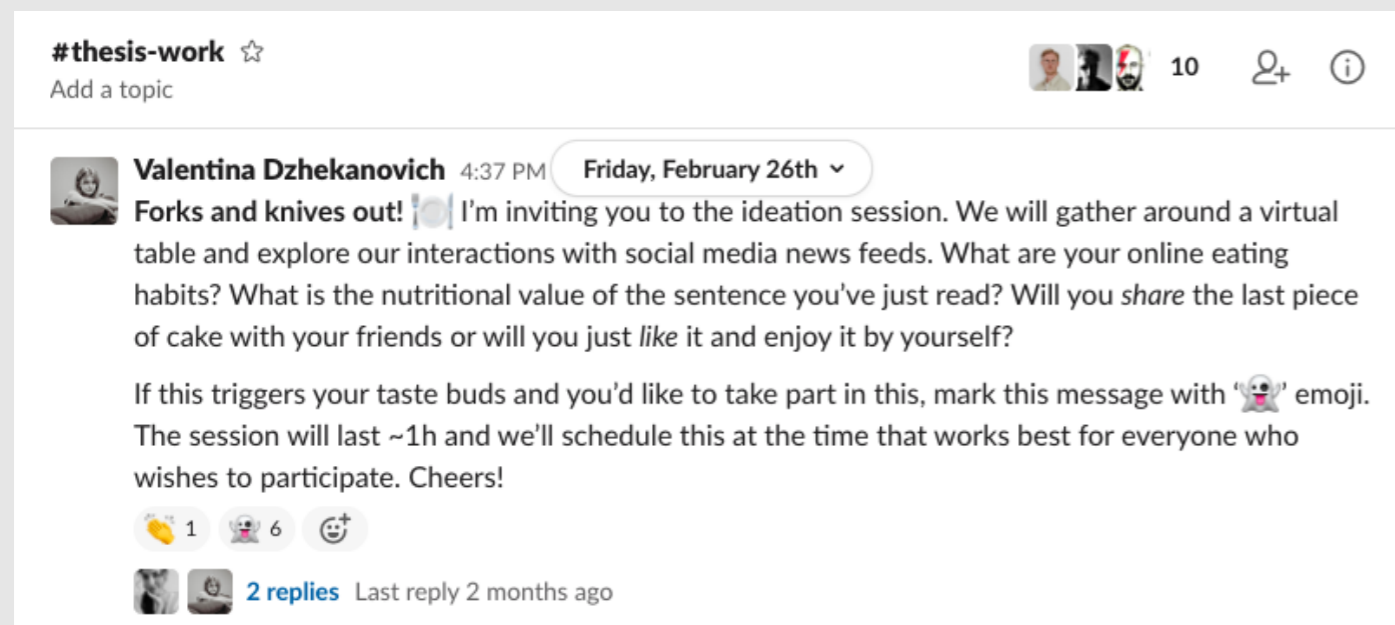


Fig. 42. Invitation to the ideation session



Fig. 43. “Food for thought” ideation session

Another interesting perspective on the topic was brought by the example of a child who refuses to try something new because “the child knows in advance that he wouldn’t like it”. And even after trying a new food and liking it, the child would still refuse it because it’s hard to admit they were wrong. Remembering my own childhood, I can definitely relate to this situation. So if this tendency can be seen from early childhood, how would an adult person first agree to have a different perspective and then admit his previous perception was wrong?

Overall, the “Food for thought” session showed me that influencing a person’s online “eating habits” against their own will is not the way to go. It needed more subtle interventions and perhaps through other interactions rather than digital ones.

## WORKSHOP

Based on the experience from the previous session, I understood that by narrowing the direction for each session, I was limiting the scope of possible interventions. At this point in my project, however, it was necessary to go as wide and as crazy as possible.

While I was going broad, it was also important to keep the user in mind. I asked myself what was the most striking moment of the experience I had with Rhonda and Josh. I realised it was the moment Josh was assaulted and bullied.

So, the last ideation session was planned to be all about my far-left persona. First, I wrote the touching story about Josh and sent it to eight designers from the IxD.ma community who agreed to participate in the workshop. Then I prepared the Miro board with main insights from my study, with design principles, persona and the end goal, which was very broad: HMW help Josh?

At the beginning of the workshop, participants individually brainstormed what they thought my persona’s problem was.



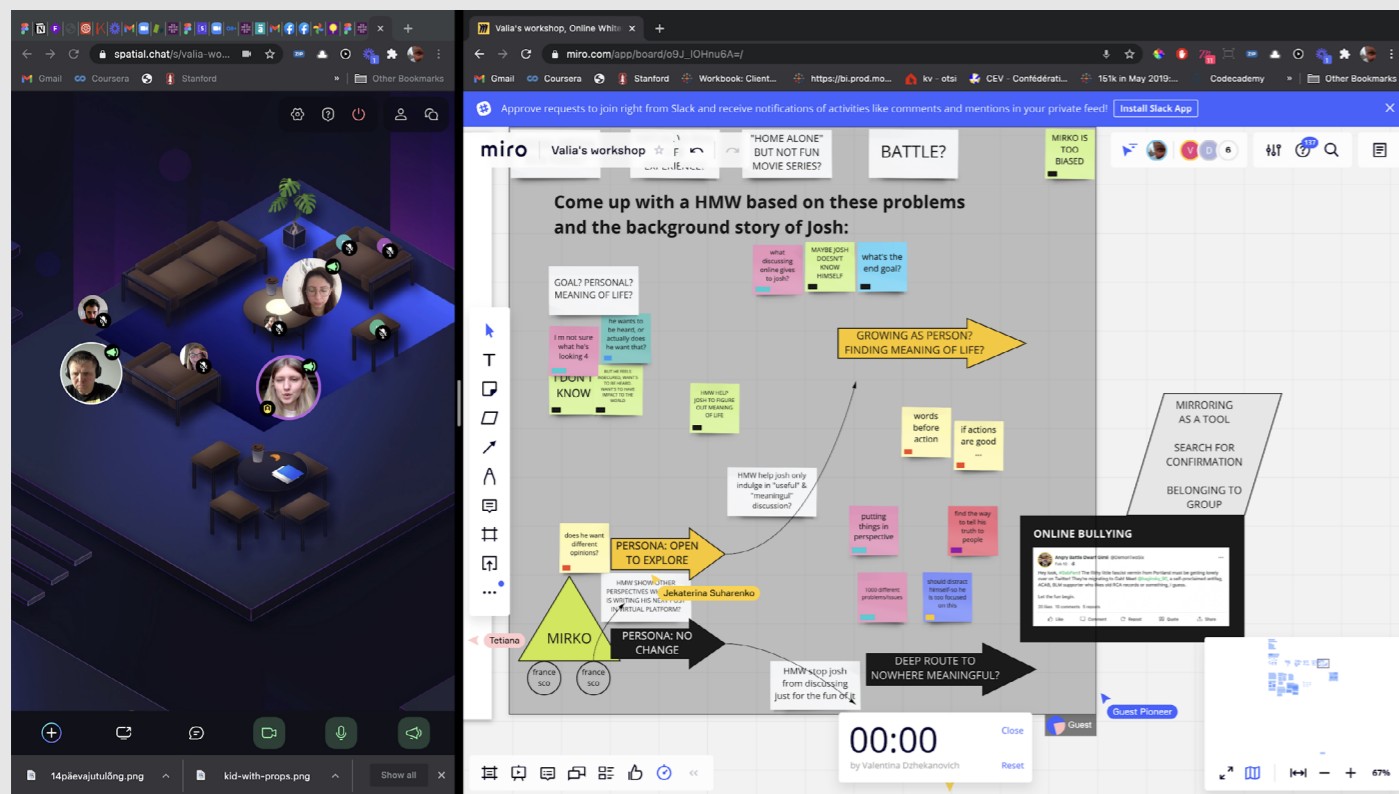


Fig. 44. Workshop “How might we help Josh?”

It was valuable for me to have everyone bring their own perspective to the problem statement already, because it could lead to more diverse ideas (Fig. 44).

In the second part of the workshop, participants grouped into teams based on similarities in their vision of the problem. In groups, they brainstormed about the HMW question and possible solutions.

Overall, the session generated many interesting ideas, some of which became the basis for my final design solution. Below I’ll make a collage of key words & concepts that were created during the session and that inspired me during the next concept development phase (Fig. 45).

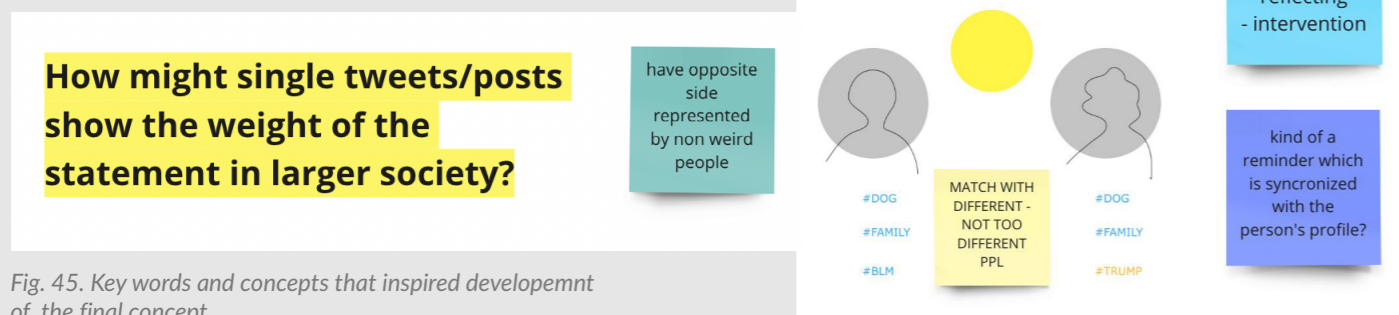


Fig. 45. Key words and concepts that inspired development of the final concept

## ANALOGOUS SOLUTIONS

While compiling the insights and ideas from all of the sessions, I explored which solutions were already being implemented on social media.

There were many initiatives that social media platforms had just launched alongside my project.

### COMBATING MISINFORMATION

For example, Twitter launched a “Birdwatch” fact-checking program in January 2021 to combat misinformation (Lyons, 2021). In the pilot phase of this program, U.S.-based users could annotate tweets that were misleading and provide context (Fig. 46).

In March, Twitter updated company’s blog with an announcement: the platform began applying labels to tweets that contain false information about COVID-19 vaccines, and promised to ban accounts that share this tweets at least for five times.

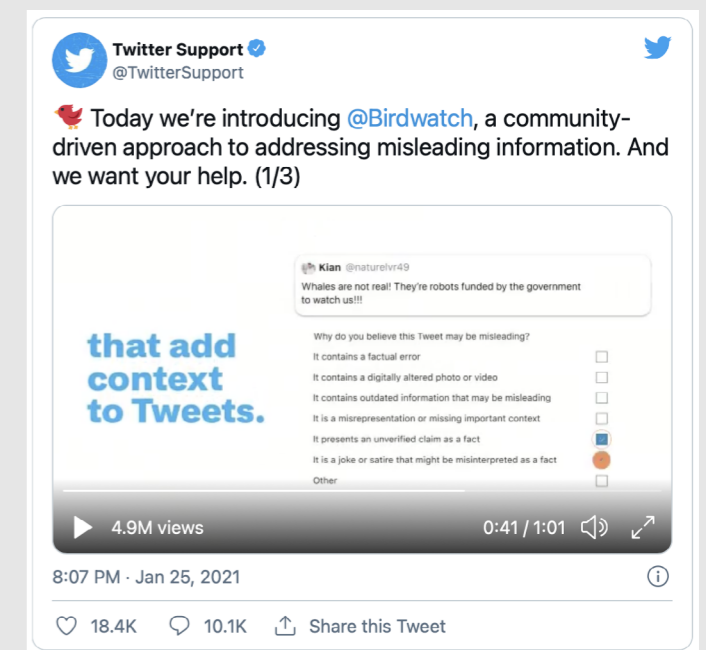


Fig. 46. Twitter launched “Birdwatch” program

### REDUCING TENSION IN CONVERSATIONS

In February 2021, Twitter introduced the “Think before you tweet” feature (Fig. 47), which allows users to review potentially offensive content before posting it (Ismath, 2021).

The next month, Facebook added a new option (Fig. 48) within groups that allows group administrators to slow down comments on a particular group post, which could help ease tensions and anxiety in increasingly heated debates (Hutchinson, 2021).

In April 2021, Facebook introduced the “Related discussions” feature (Fig. 49), which prompts users to discover other conversations around a shared post by highlighting public groups where the same post has been shared. The feature is quite controversial, as the author of the following quote notes:

“The pessimistic view is that this is what Facebook wants – by highlighting more discussions related to topics that you’re likely interested in, it’ll spark more engagement among users, which, even if it also leads to a rise in disagreement, will boost overall engagement numbers. Facebook has repeatedly noted that this is not its aim, that people won’t keep coming back to Facebook if they have too many

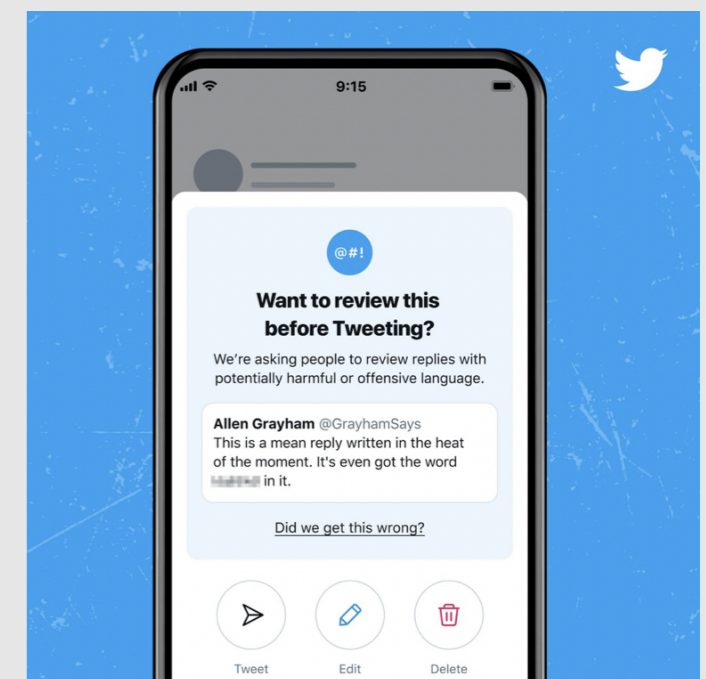


Fig. 47. “Think before you tweet” feature

negative experiences in the app like this, so it's not actually in Facebook's interests to promote divisive, argumentative content for the sake of its own engagement stats.

But it does seem like this expansion will lead to exactly that" (Hutchinson, 2021).

### ACTIONS AGAINST HATESPEECH

A separate tool called Perspective API was designed to mitigate toxic speech with machine learning techniques. The tool analyses the sentiment and gives out its toxicity score. It was tested on Twitter before, and now many other platforms, such as Reddit or Disqus, employ this solution to moderate online discussions.

Another very inspiring design idea was the Polite Type project. The Polite Type automatically rewrites offensive words and replaces them with more inclusive ones.

To sum up, this additional research helped me understand where my idea should stand in this context. There were already many initiatives done on the topic, so I planned to fit in there with my own design proposal that should be viable, but at the same time not yet existing.

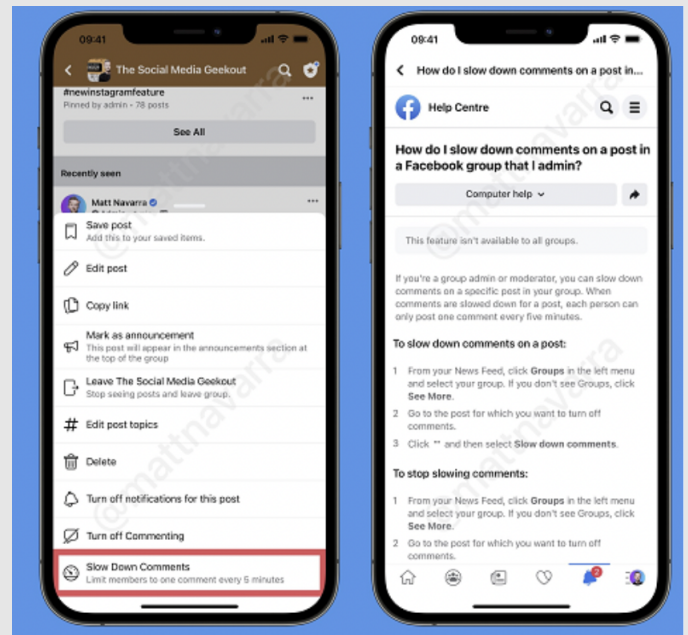


Fig. 48. "Slow down comments" feature on Facebook

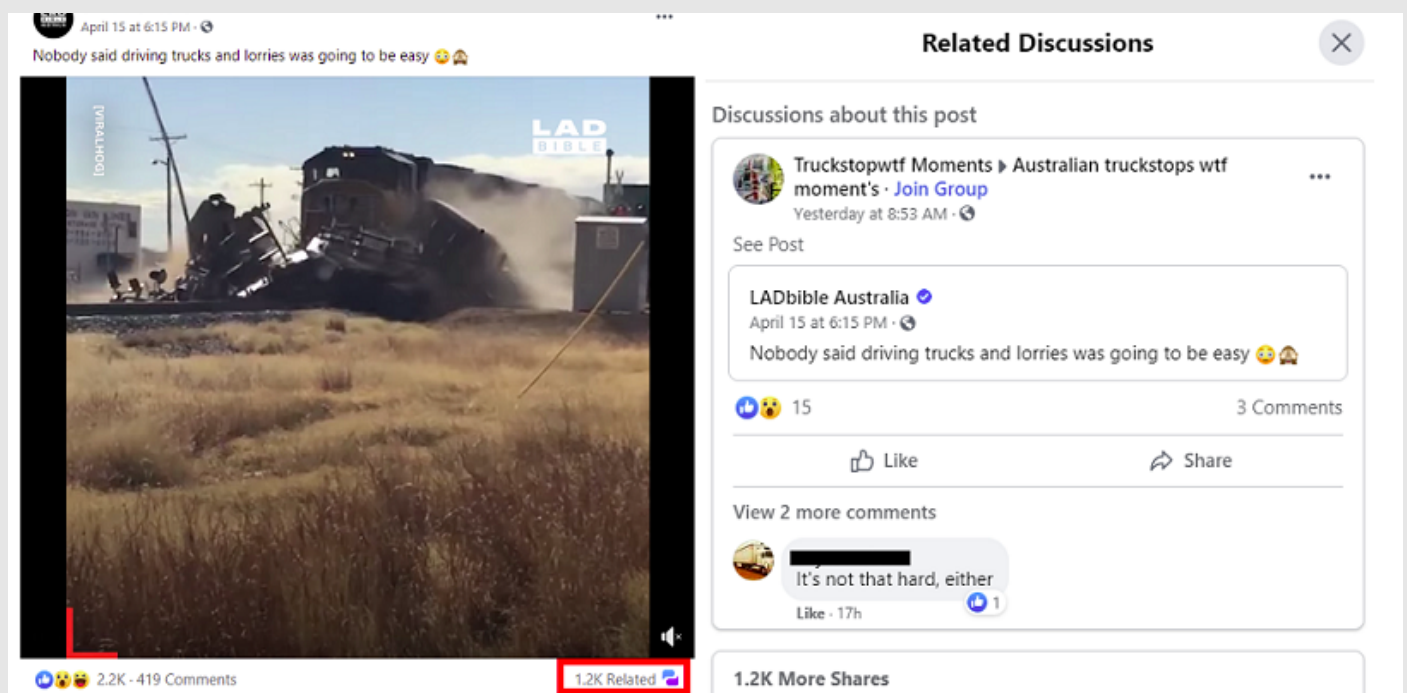
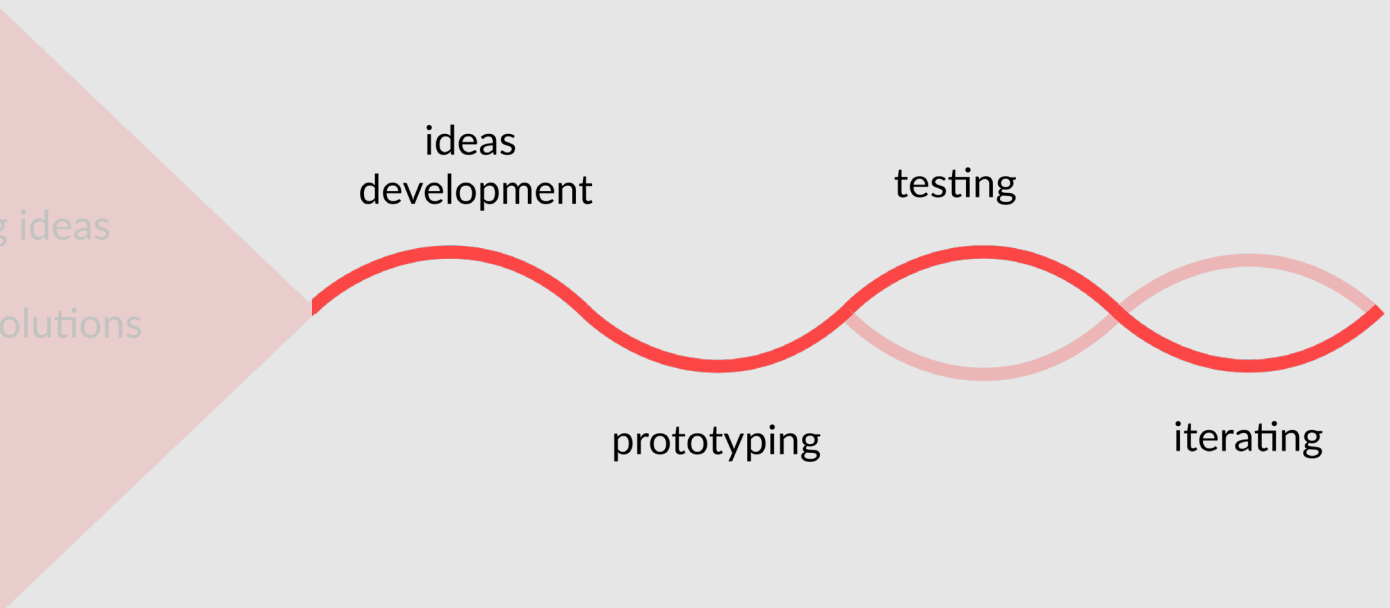


Fig. 49. "Related discussions" feature in Facebook

---

**PHASE 3** April – May

# DESIGNING AN INTERVENTION





## DEVELOPING DESIGN INTERVENTION

In this chapter, I describe how I developed the final design intervention that addresses the problem of polarisation in social media environment.

Initially, research findings and ideation sessions inspired me to go in two directions. After receiving initial comments, I adjusted my approach, developed the design intervention, and validated it with my mentor and supervisors. These feedback sessions refined my concept and I prepared it for testing with users and industry experts. Finally, I collected user and expert responses and developed the final high-fidelity prototype.

Thus, concept development took place in three iterations of development and testing. In the next sections, I briefly describe each of them.

### INITIAL DIRECTIONS

As the “Make it prominent” design principle suggested, my first attempt at designing solutions revolved around the small UI interventions into the current journey of my target users.

One of them dealt with the idea of “Dynamic bio” (Fig. 50): what if all usernames were accompanied by a short personality description that would change depending on the viewer to match his similar characteristics? In my opinion, this idea would highlight matching traits over those that users polarise on, and use similarities as a basis for trust and respect between people. However, the first round of feedback showed that this was a rather naive idea for two reasons. First, history and literature are full of examples where people are opposed to each other despite having many similarities because of the main difference between them (the most vivid example is I. Turgenev’s book “Fathers and Sons”). Secondly, from a technological point of view, it was extremely difficult to measure what similar characteristics would be strong enough to counteract the divide between people.

Another intervention from the UI perspective was about enlightening users that they were making the same arguments in the discussions regardless

of their political stance. I thought about a live pop-up in the text input field that would show how many users on the platform were using the same phrases. This pop-up would link to the list of similar texts and, according to my plan, the author of the text would see how surprisingly similar his and his opponents’ statements were (Fig. 51). This could become the first step to accepting the opponent’s views, and eventually closing the gap of misunderstanding between polarised users. Initial

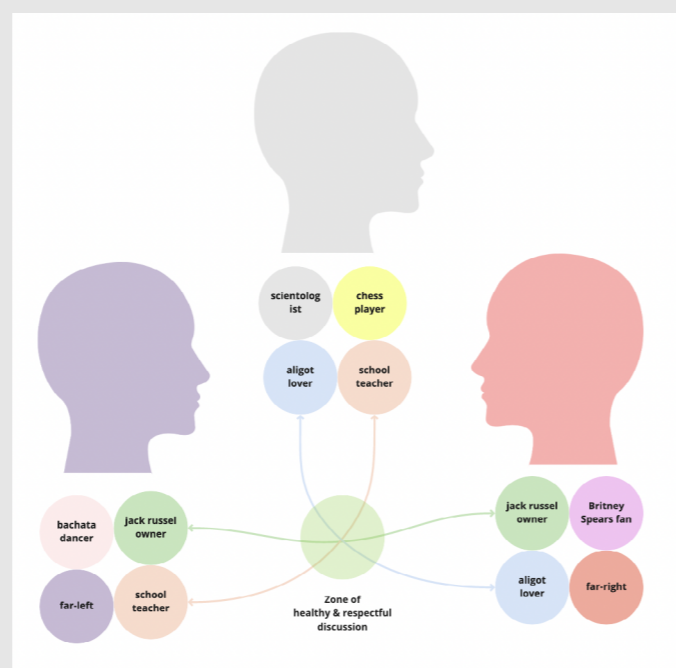


Fig. 50. The scheme of matching traits in the “Dynamic bio” idea

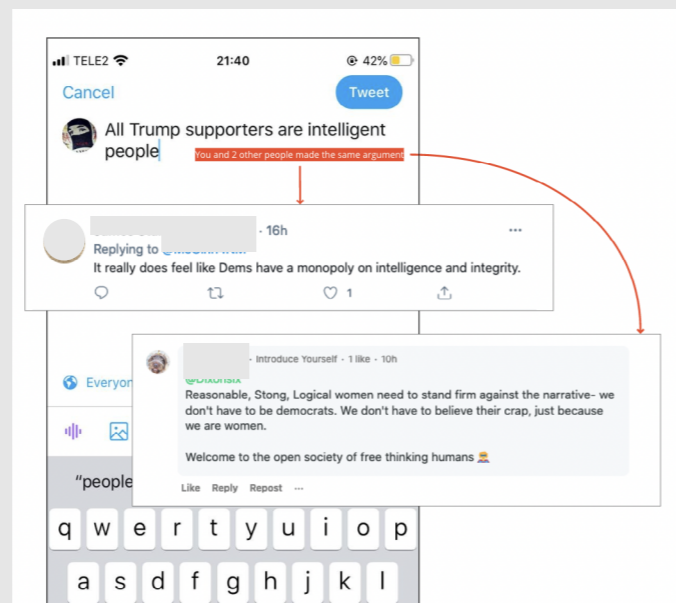


Fig. 51. Sketch of the second idea

feedback on this idea was more promising, so I decided to explore it further in the next iteration.

Overall, whole presenting these initial ideas, I was fairly encouraged to take a step back from mere UI interventions and build on the deeper needs of my target audience.

### ITERATION N°1

In my next concept development phase, I advanced the idea of showing similar arguments, put more emphasis on the core needs of my personas, and explored other mediums.

There were several core needs that I depicted on the cognitive-behavioural map. I decided to focus on the need to “be listened to” because I felt it could be most utilised and developed in alignment with my design principles and the end goal of this project.

I combined this need with my previous idea and looked for inspiration in my Miro board from the last workshop. That’s how I came up with the “Visibility” ranking idea. In this idea, all text input fields should include a live feedback feature. The feature measures and displays the visibility score based on the “toxicity” of the message: the more hateful the text, the less visible it is in the ranking system and the fewer users will be able to see it (Fig. 52). Behind the visibility score, the “Learn more” link opened the explanation banner that suggested rewriting the text or learning more about the feature.

Clicking on the link “More about how it works” took the user to the interactive 3D data space, which showed the place of the user’s text among all other published texts (represented as coloured dots on the Fig. 53). The place of the dot on the horizontal layer of this space would indicate the topic of the text with semantic analysis, and the vertical axis would measure the tone of the message with sentiment analysis: the higher, the friendlier the message and the more people would be able to see it, and vice versa for the texts in the lower area.

By clicking on the dots and viewing other people’s texts, the user had the opportunity to explore what others had said on the same topic and learn by example how to achieve 100% visibility.

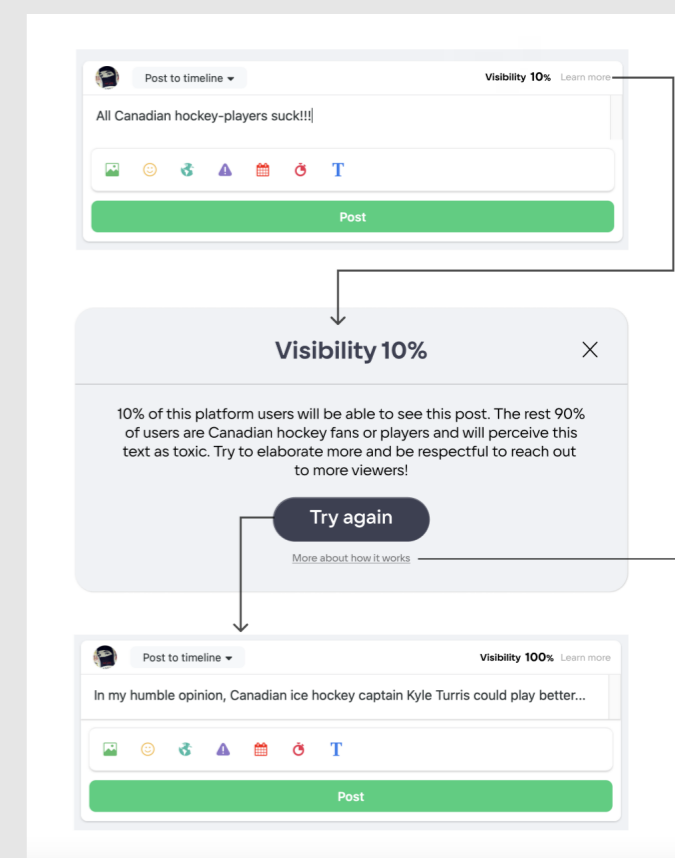


Fig. 52. Visibility rate

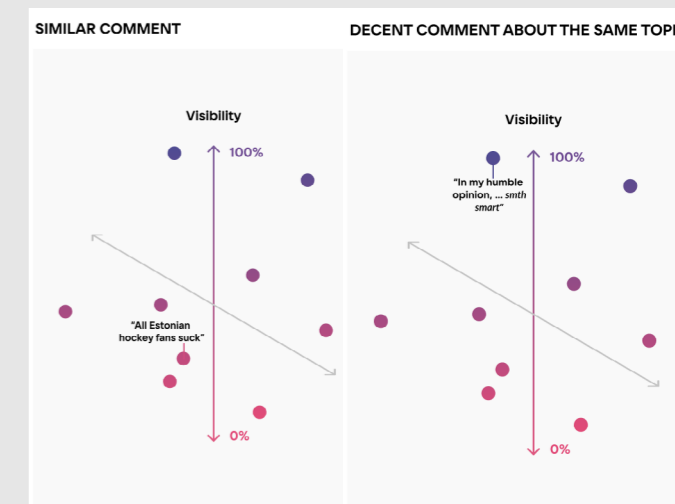


Fig. 53. The data space appears after clicking “More about how it works” button

By leveraging the need to be listened to, I tried to motivate respectful discussion, and also offered a way to self-reflection to see how to fulfill this core need.

The technical part of this idea was discussed with my second mentor Jonas Knutsson, who was the AI expert in this project. In this meeting, we discussed potential ways to implement semantic and sentiment analysis techniques in my prototype with API keys from already existing solutions. Jonas mentioned that for a professional like him it would take a couple of days for building a prototype like this. At this point, I fully acknowledged that I lacked the skills and time to build such a thing from scratch, so I decided to leave this on a theoretical level.

While presenting the “Visibility” ranking idea to my supervisors, I received the suggestion to look at this intervention from the perspective of my target audience – what value would it provide? How would they interact with this feature? There was a big chance that users would reject this idea because essentially it was another way to moderate their speech. Although I had refined my approach since the last time I tried it, I still needed to improve it to better suit my users before I could test it with them.

### ITERATION N°2

To improve the “Visibility” ranking idea, I sketched it out on the value proposition canvas and played out the UX with this feature implemented. After these steps, I decided to put more emphasis on the positive “amplifying” side of my idea and remove the negative penalty of zero visibility and reinforce the positive side of being listened to. To be honest, I didn’t see Rhonda or Josh interacting with the 3D data space either, but at the same time I needed to keep the self-reflection tool.

So after refining the previous idea of visibility ranking, I came up with the “Speaker Badge” feature (Fig. 54). With this concept, when a user writes a post or replies to a comment, he or she receives live feedback that rates the credibility and tone of the text (Fig. 55). By performing well in his postings and receiving support from the community, the user earns the Speaker Badge, which helps him stand out as an expert in any conversation.

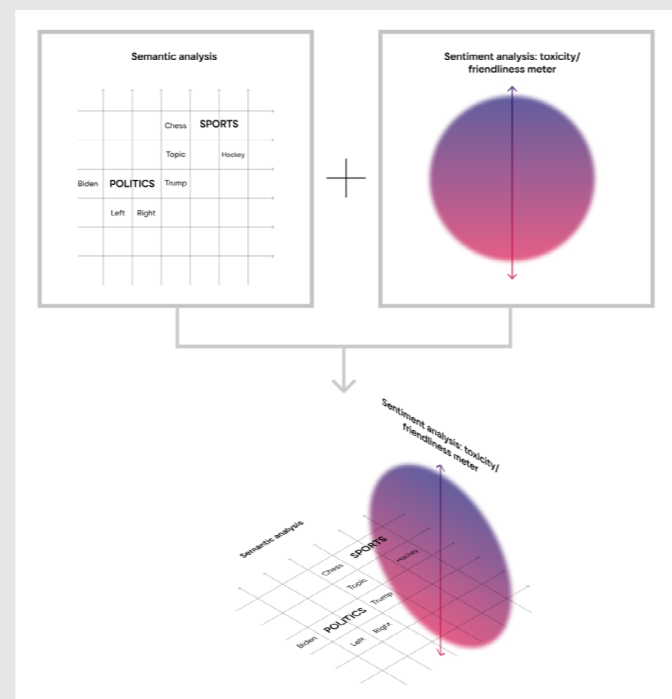


Fig. 53. Scheme explains how the 3D interactive data space could be built

However, once the Speaker Badge is awarded, the user can lose it just as easily if the new comments or posts violate the platform’s rules, such as sharing fake news or spreading hate speech. To monitor progress and see more information about the rating, the user can go to the Speaker Badge tab in the private settings.

To test this intervention, I created the landing page which partially can be seen in the figures on the right or by clicking this link: <http://valia.design/speaker-badge>. To make it look like a real product and to get honest feedback from users, I didn’t mention that this was a student thesis project. I posted this landing page on Gab from my personal account and briefly described the purpose of the feature (Fig. 56). To spread the word, Rhonda reposted it to her page.

I received several negative replies (Fig. 57, 58, 59) – some users did not understand how it worked and interpreted it as a feature that worked against them. However, one user signed up for testing and shared positive feedback (Fig. 60). I checked this user’s page, and it was exactly who I was trying to reach: a desperate far-right conspiracy theorist and Trump supporter. So it was very valuable feedback for me.

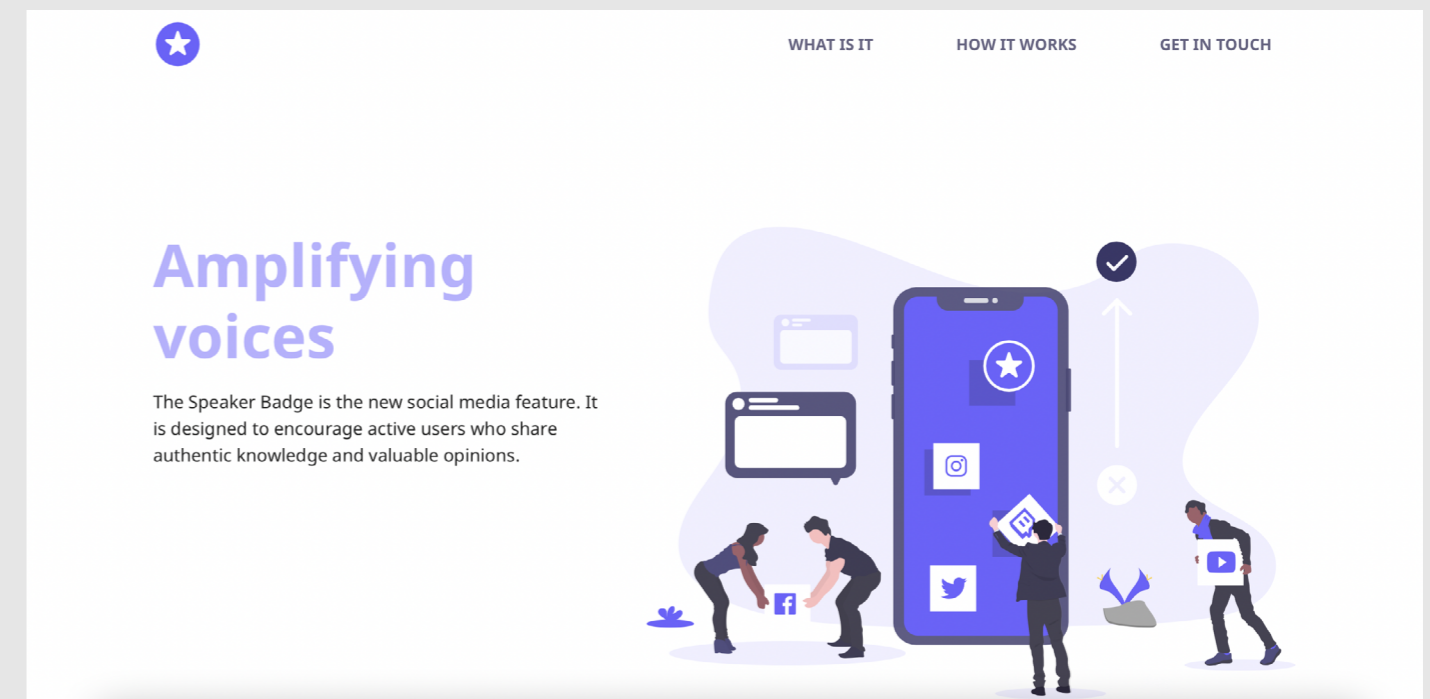


Fig. 54. The Speaker Badge feature (landing page)

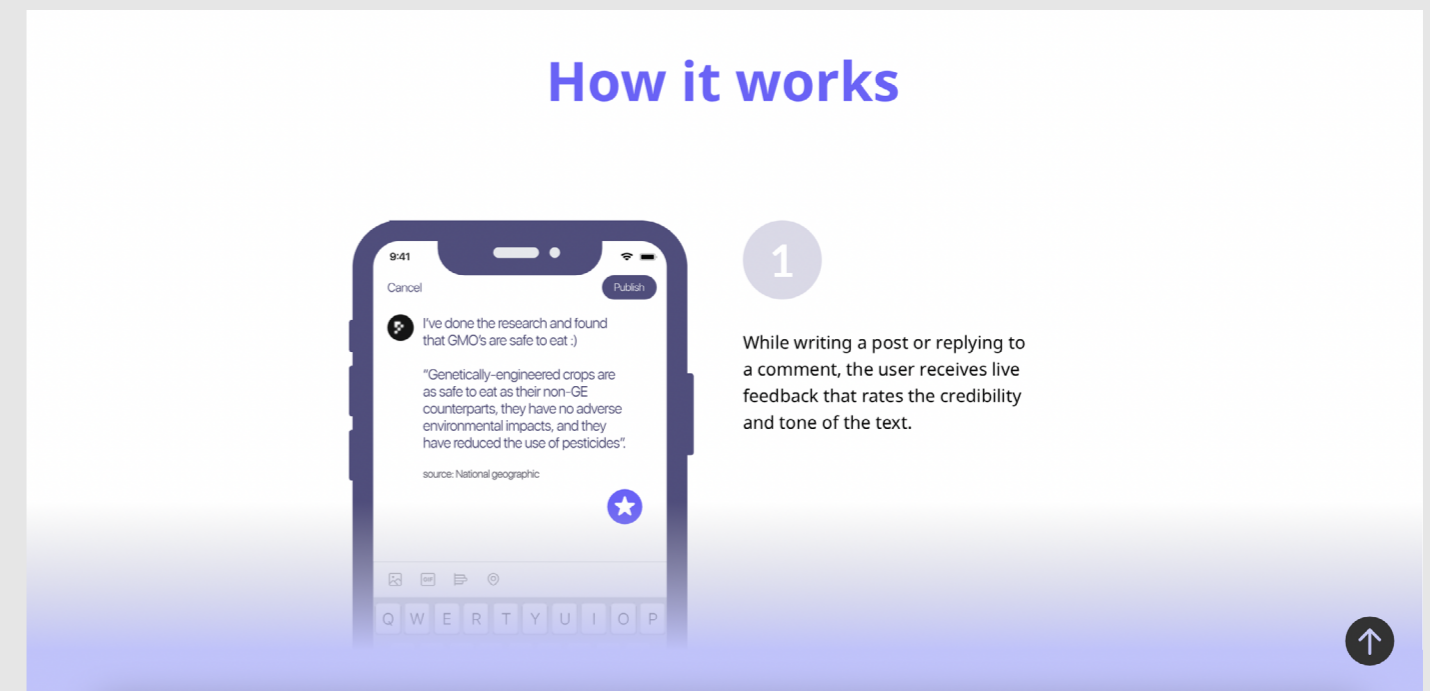


Fig. 55. The Speaker Badge feature (landing page)



As I analysed this feedback, I realised that it was difficult to show users the value of my design intervention from the broad perspective of the polarisation problem. My target users were already so extreme and hostile that evaluating my design proposal became another battle in which they blindly defended themselves without being attacked.

Thus, I realised that while my design intervention would not become a cure for online polarisation, it could function as a prevention tool for those users who were not yet so radicalised but were moving in that direction.

To test this assumption, I asked for feedback from social media experts in LinkedIn and Discord. In the discussion it was compared to the Chinese social credit system.

However, in my opinion, there is a fundamental difference between a dystopic initiative to control every aspect of human life and a system that motivates social media users to be accountable for their words and be respectful to others.

Even though my proposal does limit the individual freedom to some extent, it is still up to user's choice whether to keep the Badge or lose it with no punishing consequences. In the end, our freedom ends where other person's freedom starts, and in my opinion it has to be communicated in social media platforms, otherwise all online spaces will turn into Gab.

Overall, this feedback showed to me what should be presented more clearly in my final design concept, so I moved on to the third iteration round to implement this. In the next chapter, I present the high-fidelity prototype of my idea.

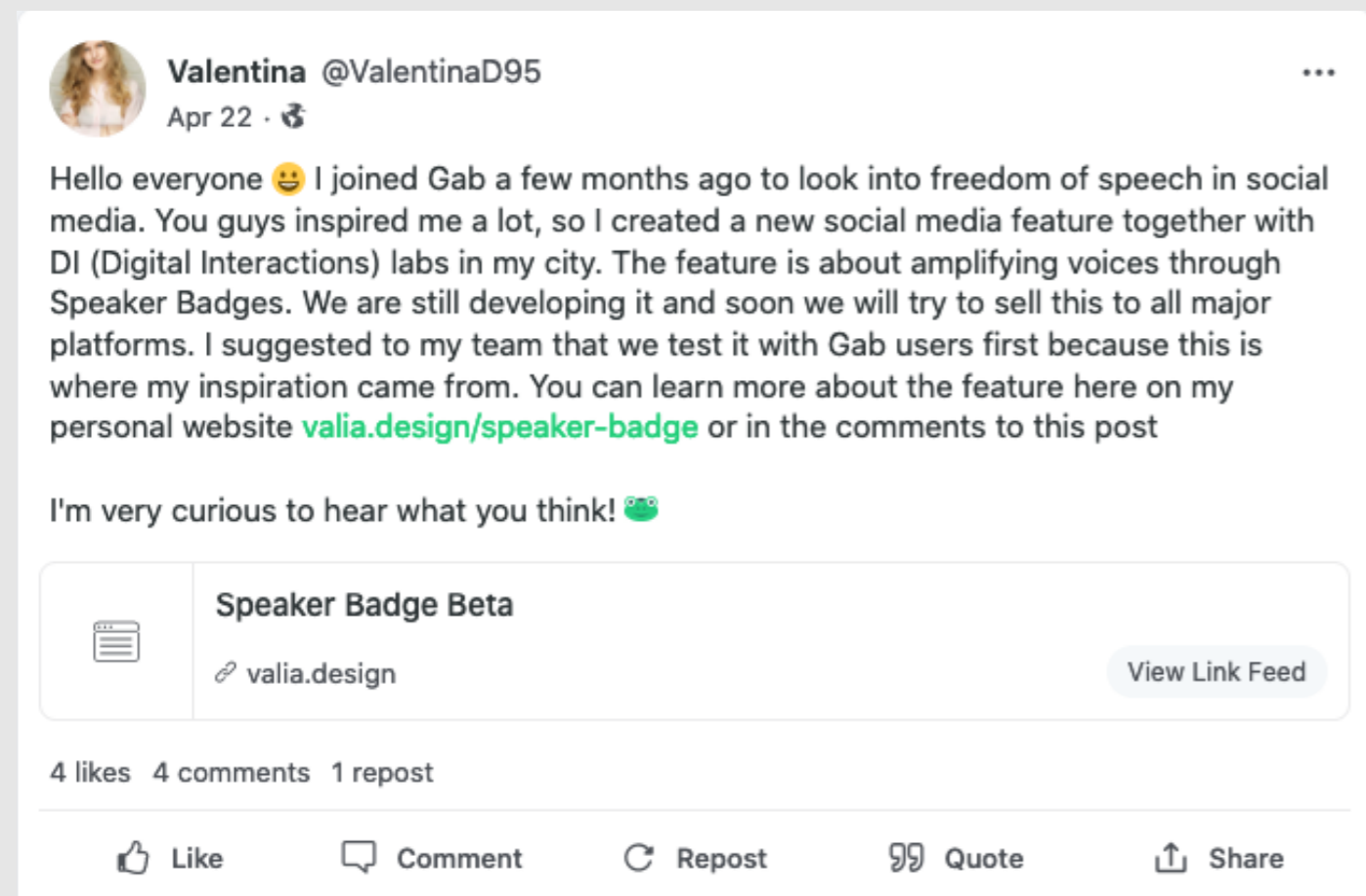


Fig. 56. I present the Speaker Badge feature on Gab

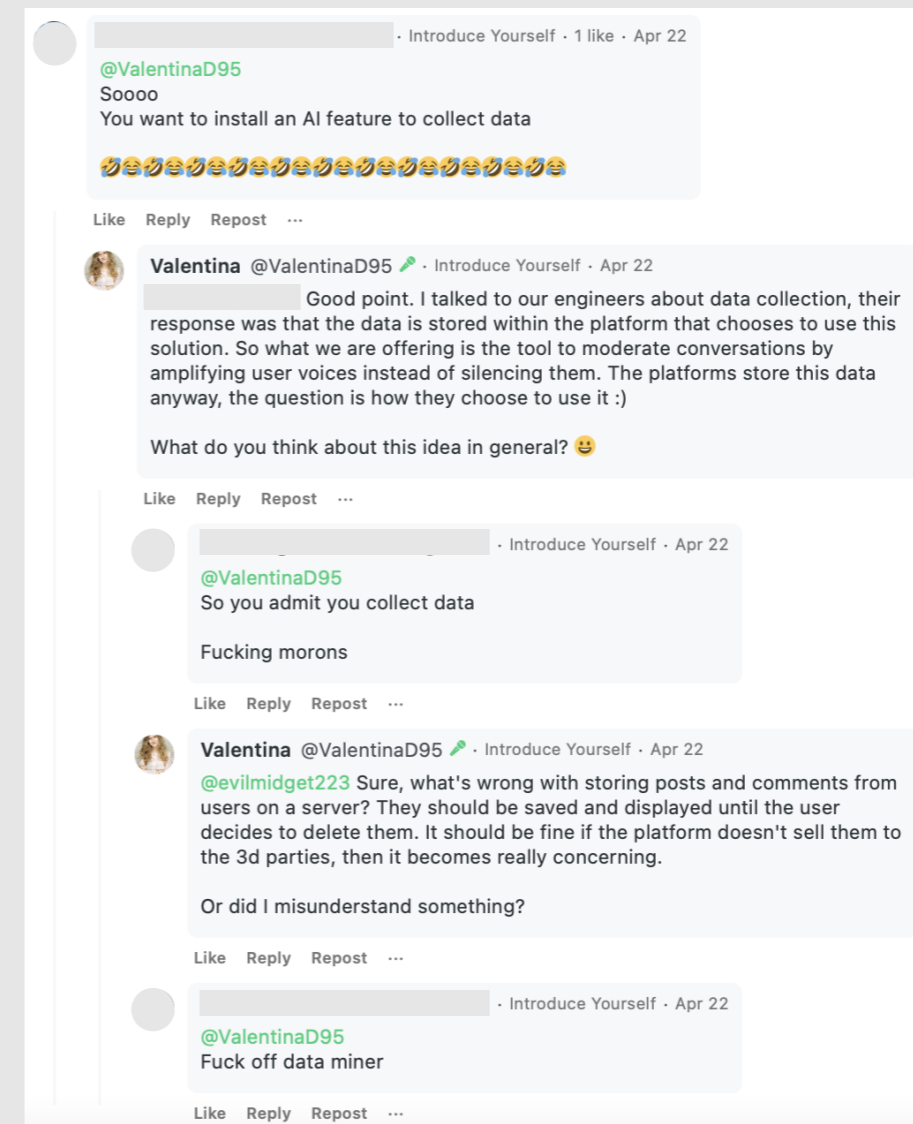


Fig. 57. Users' comments about the Speaker Badge

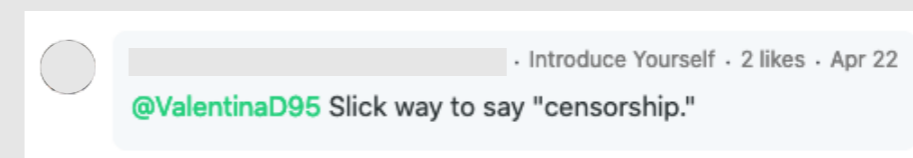


Fig. 58. Users' comments about the Speaker Badge

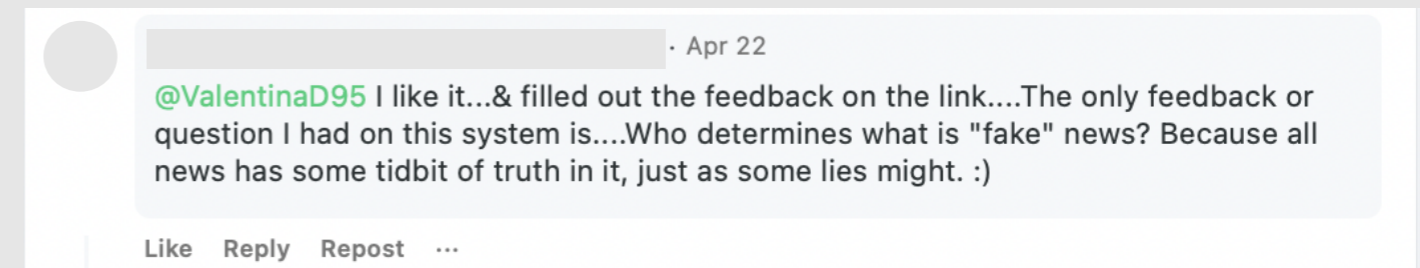


Fig. 60. Positive comment from user



Fig. 59. Users' comments about the Speaker Badge



# AI-DRIVEN UX/UI DESIGN INTERVENTION

## WHAT HAPPENS ON THE SCREEN?

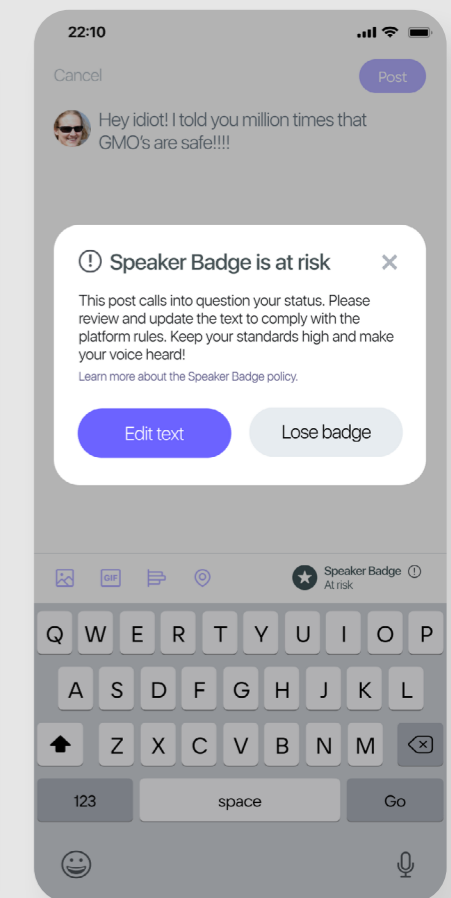
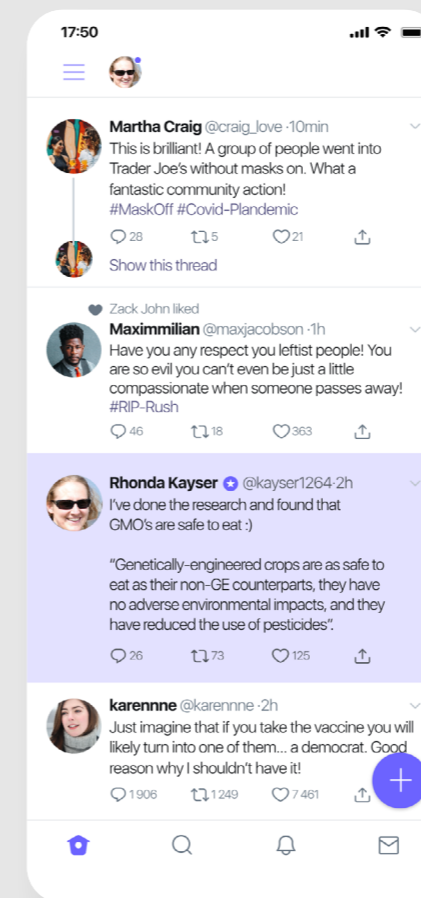
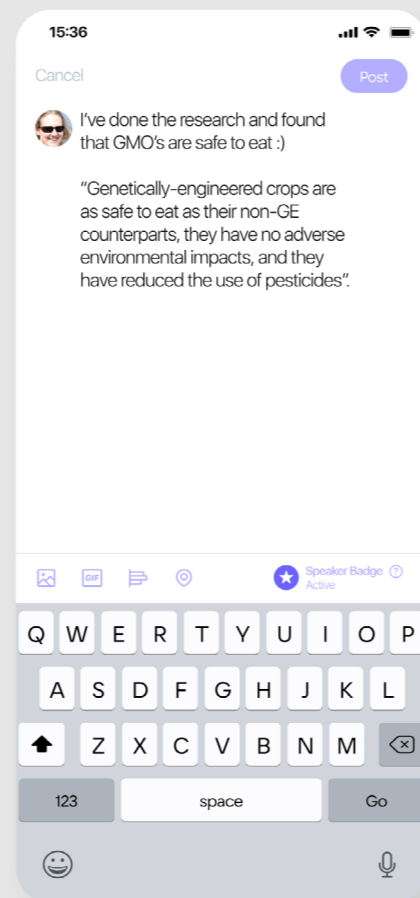
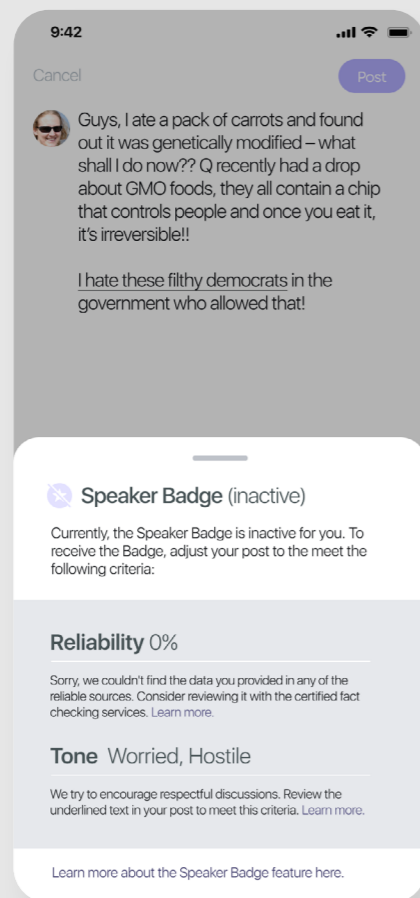
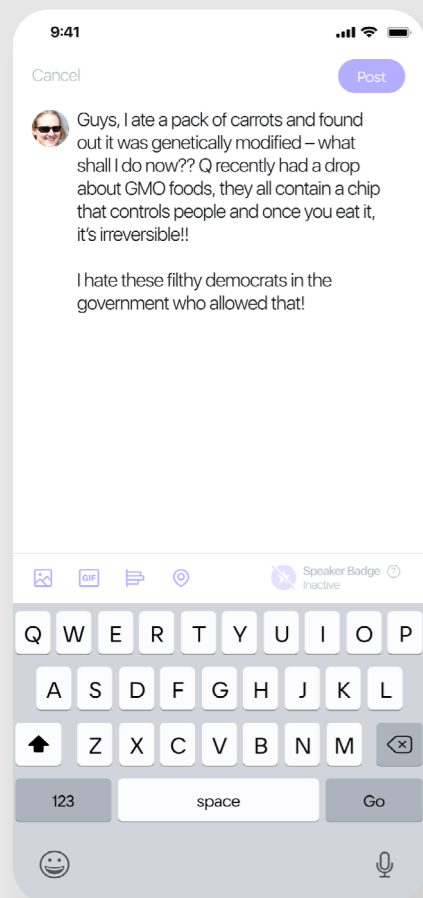
To address the problem of polarisation on social media platforms, I propose an AI-driven UX /UI design intervention called “Speaker Badge”. This feature grants reliable user-generated content and motivates users to keep conversations at a respectful level. Below I describe the user interactions with the Speaker Badge feature implemented in a generic

social media platform’s mobile app. Following screens showcase major events in the interaction with the feature. To have a more holistic experience of using the Speaker Badge feature, click this link and access the interactive prototype in Figma:

<https://bit.ly/3epsUHe>

To watch a video of this proposal, click this link:

<https://bit.ly/3o6qAlr>



### 1. WRITING THE POST

Rhonda Kayser expresses her concerns about the dangerous effects of GMO foods. The Speaker Badge feature performs AI-driven semantic and sentiment analyses and stays inactive because the text doesn't pass the fact-check and contains the offensive speech.

### 2. LIVE FEEDBACK

By clicking on “Speaker Badge” in the bottom menu, the user can access the explanation of why the feature remains inactive. This helps the user to make informed corrections in the text while writing.

### 3. THE SPEAKER BADGE IS ACTIVATED

As the time passes and Rhonda learns from the feedback, she starts sharing more credible information in a friendlier manner. The Speaker Badge re-evaluates user’s performance and activates.

### 4. POST IS HIGHLIGHTED

With the Speaker Badge activated, the published post receives a star icon next to the username and gets highlighted. The feature makes the publication more prominent and appealing to all users. This ensures that fact-checked content gets more attention, while also helping Rhonda stand out in the conversation.

### 5. SELF-REFLECTION

The Speaker Badge tab in the settings provides more information about the feature. The graph shows the statistics of all publications as measured by the performance score. By clicking on dots from the graph, user can access each publication and see how it was analysed by parameters of reliability, tone of voice and community support.

### 6. THE SPEAKER BADGE IS AT RISK

On this screen, Rhonda writes offensive text and gets a warning that such speech threatens her Badge. The pop up suggests to edit the text through a call-to-action button, or offers to deliberately give up the Badge.

# PSYCHOLOGICAL BASIS OF DESIGN PROPOSAL

## WHAT HAPPENS AT THE COGNITIVE-BEHAVIOURAL LEVEL?

In the previous section, I presented how the Speaker Badge design intervention works from a screen-based interaction perspective. Here I describe how

the feature addresses deeper triggers of polarised users, which nudges it uses, how it fits and alters the polarisation flow of the cognitive-behavioural map. The following scheme shows how the intervention works for a user who has achieved the Badge (right) and how it affects the opposite user who interacts with the Speaker Badge owner (left). Given the similarity of both users, this feature can be also implemented vice versa or in both cases at the same time. *I suggest reading the next paragraphs clockwise.*

## 1. PSYCHOLOGICAL VALUE OF THE SPEAKER BADGE

The Speaker Badge addresses users' needs to nudge behaviour change through following features:

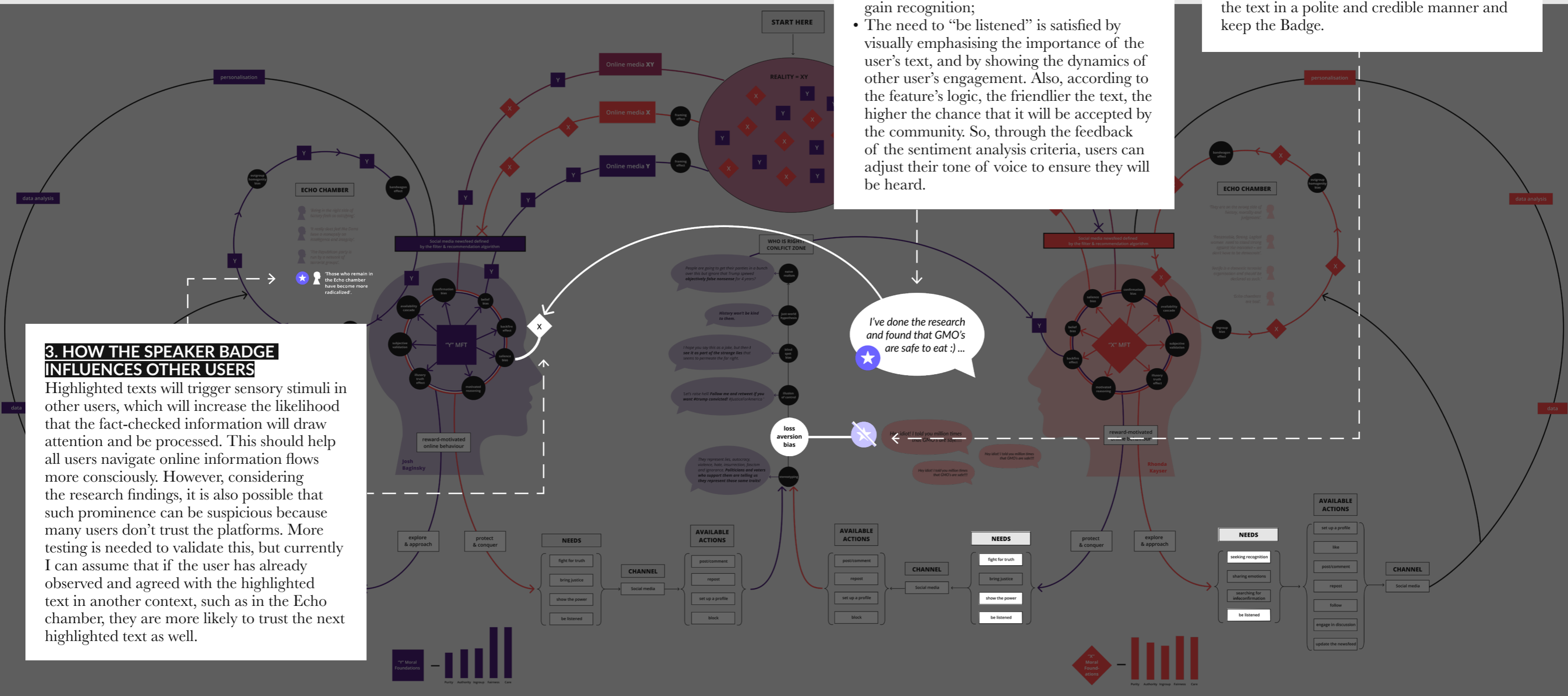
- The need to “fight for the truth” is supported and developed by the fact-checking analysis;
- Users are motivated to comply with the requirements and achieve the Speaker's Badge by their need to show the power and gain recognition;
- The need to “be listened” is satisfied by visually emphasising the importance of the user's text, and by showing the dynamics of other user's engagement. Also, according to the feature's logic, the friendlier the text, the higher the chance that it will be accepted by the community. So, through the feedback of the sentiment analysis criteria, users can adjust their tone of voice to ensure they will be heard.

## 2. A NUDGE TO PREVENT OFFENSIVE OR MISLEADING CONTENT

To nudge respectful discussions and stop the spread of misinformation, the Speaker Badge feature relies on the “loss aversion” cognitive bias. If user's text doesn't meet the requirements, the feature warns that the Badge will be discarded. According to the “loss aversion” bias, the prospect of losing the Badge will feel worse than the possibility of posting a new text. So, this nudges user to edit the text in a polite and credible manner and keep the Badge.

## 3. HOW THE SPEAKER BADGE INFLUENCES OTHER USERS

Highlighted texts will trigger sensory stimuli in other users, which will increase the likelihood that the fact-checked information will draw attention and be processed. This should help all users navigate online information flows more consciously. However, considering the research findings, it is also possible that such prominence can be suspicious because many users don't trust the platforms. More testing is needed to validate this, but currently I can assume that if the user has already observed and agreed with the highlighted text in another context, such as in the Echo chamber, they are more likely to trust the next highlighted text as well.



---

## CONCLUSION

To conclude this project, I would like to critically reflect on the design process, evaluate the design intervention, and share my perspective on the broader context of the proposal.

### **REFLECTIONS ABOUT THE DESIGN PROCESS**

The journey of this thesis project started with my own biased perception that social media personalisation algorithms radicalise people by exploiting their irrational information processing mechanisms, also known as cognitive biases. In my opinion, the process could be reversed if algorithms reinforced rational thinking.

To investigate and address this problem, I framed and reframed design activities as the project progressed. Throughout the project, I conducted semi-structured interviews with experts in various fields, developed a survey, created personas, experimented with digital ethnography study to get to know my target audience better, designed a cognitive-behavioral map, and kept coming back to desk research, advice from my mentor, and feedback from supervisors. All of these design research methods refined my initial understanding of the problem, revealed its true complexity and depth. Today, I understand that to fully solve this problem, I would need to redesign all of human nature.

The positive side of this deep research is that the final research result (the cognitive-behavioral map) offered me great potential for design interventions. I want to emphasise that I consider this scheme as important in this project as the final design proposal.

Having completed the research part, I moved on to the next design activities. In order to explore the wide variety of possible solutions, I conducted four ideation sessions – each of them contributed a lot to my final design outcome and inspired me. Then designing an intervention phase began, where I developed and tested the concept in three iterations. That's the most I was able to achieve in this project period.

To critically reflect on the design process and methods, I would like to mention that perhaps I

could divide the time frame of this project more evenly between the different phases. Even though I am very satisfied with the research phase, I can objectively see that the next two phases (ideation sessions and designing an intervention) could take more space in the planning.

Another consideration is the methods of user research and testing. I understand the limitations and still think that the digital ethnography study was the most appropriate method, but I wish I had had more courage to interview my target audience in person and perhaps understand them even better in a private conversation. During testing, it would be very valuable to examine the behavior change through the Fogg model and see how I could adjust smaller parameters of my concept, such as what would be the optimal percentage of the performance score that achieves the Badge, or how much time it takes for users to change their behaviors, and etc.

### **EVALUATION OF THE DESIGN INTERVENTION**

In a broader context, I think the concept “Speaker Badge” fits into the current trend of emerging social media initiatives to address aspects of the polarisation problem, such as combating hate speech and spreading misinformation. Unlike the most commonly used solution of removing offensive/ misleading content and suspending accounts, my design proposal grants and motivates users to share fact-checked information and keep online debates at a respectful level. In my opinion, both approaches (“stick” for suspensions and “carrot” for granting the “Speaker Badge”) are necessary in content moderation. Also, the solution is two-fold and works for both users who have the “Speaker Badge” and those who can more easily find credible content through the quick visual cue of the Badge.

### **FUTURE WORK**

Of course, this design proposal is not a silver bullet that will completely solve such a complex and long-standing problem as political polarisation. There are still many aspects to my design proposal that will need to be developed and tested in the future. I will present a few of them:

- The reliability parameter of the performance score is the most controversial aspect of my concept. The datasets against which the user's text is to be compared must be as objective and complete as possible. It should not become a way for political powers to exploit the system and establish a single preferred public discourse;
- While it is important to notice and prevent offensive language, testing should be done to see how the tone of voice parameter affects the overall performance score. The feature should not become a way to suppress all emotions other than positive ones;
- To optimise the experience of receiving the Badge, the process should be tested with a Fogg behavioural model: it is important that the user's ability to follow the requirements matches the motivation.

While this is a very small intervention compared to the whole big problem I've been tackling, I think it has the potential to become a tiny step towards a common understanding of what is “true” in the world of social media and unite people around it through respectful dialog.

---

## ACKNOWLEDGEMENTS

I would like to sum up this project on a positive note. This journey has been full of ups and downs for me – I had very high expectations for myself and I have to admit that I didn't meet all of them, but what happened in reality was much cooler than I could have predicted. I would like to thank...  
...my mentor Johanna Rochegude for tackling all the complexities of this project with me, for her encouragement and care;  
...Jonas Knutsson for sharing his knowledge with me and supporting my humble assumptions about the tech world;  
...everyone from Block Zero who followed every step of this journey and supported me;  
...Alan Voodla, whose psychology class inspired me to address this topic and who constantly shared his thoughts with me about the project;  
...my degree buddies Jekaterina Sukharenko and Fransceso Duc for comfort and help;  
...and of course my supervisors Tanel Kärp and Nesli Hazal Akbulut for their guidance and feedback.

Last, but not least, this project has taught me to respect the opinions of others when they contradict my opinion because there are always deeper motivations and beliefs that are just as important as mine.

P.S. Also, I learned that I have a completely (almost) normal family.

---

## REFERENCE LIST

Berger, Peter L. Homeless Mind, The modernization and consciousness. 1: Vintage Books, 1973. Non Fiction.

Cooper, Timothy, and Jem Thomas. Nature or Nurture: A Crisis of Trust and Reason in the Digital Age. Albany Associates, 2019.

Drew Harwell, Isaac Stanley-Becker. "QAnon Reshaped Trump's Party and Radicalized Believers. The Capitol Siege May Just Be the Start." The Washington Post. January 14, 2021. Accessed May 01, 2021. [https://www.washingtonpost.com/technology/2021/01/13/qanon-capitol-siege-trump/?fbclid=IwAR33Ykyb\\_2XSbmf6eMcVfaybb3n\\_GPWgsbZTeQ\\_Fjz2KUnCGNG-z57lLsk](https://www.washingtonpost.com/technology/2021/01/13/qanon-capitol-siege-trump/?fbclid=IwAR33Ykyb_2XSbmf6eMcVfaybb3n_GPWgsbZTeQ_Fjz2KUnCGNG-z57lLsk).

Haidt, Jonathan. The Righteous Mind: Why Good People Are Divided by Politics and Religion. Vancouver, B.C.: Langara College, 2020.

Hao, K. 2021. "How Facebook got addicted to spreading misinformation." MIT Technology Review. March 11, 2021. <https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/>

Heath, Robert G. "Electrical Self-Stimulation Of The Brain In Man." American Journal of Psychiatry 120, no. 6 (1963): 571-77. doi:10.1176/ajp.120.6.571.

Hutchinson, Andrew. "Facebook Adds New Option to Slow Down Comments Within Groups." Social Media Today. March 30, 2021. Accessed May 05, 2021. <https://www.socialmediatoday.com/news/facebook-adds-new-option-to-slow-down-comments-within-groups/597606/>.

Hutchinson, Andrew. "Facebook Expands 'Related Discussion' Prompts on Shared Posts in News Feeds." Social Media Today. April 19, 2021. Accessed May 05, 2021. <https://www.socialmediatoday.com/news/facebook-expands-related-discussion-prompts-on-shared-posts-in-news-feeds/598598/>.

Hutchinson, Andrew. "Optimal Posting Practices for Facebook, Twitter and LinkedIn in 2021." Social Media Today. January 07, 2021. Accessed May 01, 2021. <https://www.socialmediatoday.com/news/optimal-posting-practices-for-facebook-twitter-and-linkedin-in-2021/592948/>

Ismath, Sabra. "Twitter Re-launches Test Feature to Revise Potentially Harmful Tweets." MobileSyrup. February 23, 2021. Accessed May 05, 2021. <https://mobilesyrup.com/2021/02/23/twitter-relaunch-test-feature-revise-tweets/>.

Kahneman, Daniel. Thinking, Fast and Slow. New York: Farrar, Straus and Giroux, 2013.

Lakoff, George. The Political Mind: A Cognitive Scientists Guide to Your Brain and Its Politics. London: Penguin Books, 2009.

"List of Cognitive Biases." Wikipedia, Wikimedia Foundation, Accessed May 05, 2021. [en.wikipedia.org/wiki/List\\_of\\_cognitive\\_biases](https://en.wikipedia.org/wiki/List_of_cognitive_biases).

Lyons, Kim. "Twitter Launches Birdwatch, a Fact-checking Program Intended to Fight Misinformation." The Verge. January 25, 2021. Accessed May 05, 2021. <https://www.theverge.com/2021/1/25/22248903/twitter-birdwatch-fact-checking-misinformation>.

Moral Foundations Theory. Accessed May 02, 2021. <https://moralfoundations.org/>.

Murdock, Ryan. "This Is Where the Filter Bubble Leads." The Shift News. January 8, 2021. Accessed May 01, 2021. <https://theshiftnews.com/2021/01/08/this-is-where-the-filter-bubble-leads/>.

Ngo, Andy. "Biden Won't Stop Them, and Neither Will Cops, until Portland Is Burned down." New York Post. January 22, 2021. Accessed May 02, 2021. <https://nypost.com/2021/01/21/biden-and-cops-wont-stop-them-until-portland-is-burned-down/>.

Noujaim, Jehane, Karim Amer, "The Great Hack." Netflix Official Site, July 24, 2019. <https://www.netflix.com/ee/title/80117542>.

Ortutay, Barbara. "Facebook Bans More than 790 Groups Tied to QAnon, plus More than 980 Militia-linked Groups." USA Today. August 19, 2020. Accessed May 02, 2021. <https://eu.usatoday.com/story/tech/2020/08/19/facebook-bans-hundreds-qanon-and-militia-linked-groups-pages/5609786002/>.

"Political Bias Test." ClearerThinking2020, Accessed May 05, 2021. [www.clearerthinking.org/the-political-bias-test](http://www.clearerthinking.org/the-political-bias-test).

Technopedia. "What Is a Filter Bubble? - Definition from Technopedia." Technopedia.com. May 17, 2018. Accessed May 05, 2021. <https://www.technopedia.com/definition/28556/filter-bubble>.

"Updates to Our Work on COVID-19 Vaccine Misinformation." Twitter. Accessed May 05, 2021. [https://blog.twitter.com/en\\_us/topics/company/2021/updates-to-our-work-on-covid-19-vaccine-misinformation.html](https://blog.twitter.com/en_us/topics/company/2021/updates-to-our-work-on-covid-19-vaccine-misinformation.html).

Wendling, Mike. "QAnon: What Is It and Where Did It Come From?" BBC News. January 06, 2021. Accessed May 01, 2021. <https://www.bbc.com/news/53498434>.

---

## APPENDIX 1: SURVEY

### INTRO

Hello! My name is Valentina Dzhekanovich, I am an Interaction design student from the Estonian Academy of Arts. This survey is a part of my master's degree project that studies the mechanisms of a firm belief formation and expressions of such beliefs in digital environments. I'm also interested in how the design of online social media shapes interactions and affects the way we communicate. Answering these questions takes 7-10 minutes, and your input can incredibly help this research to move forward.

### INFORMED CONSENT

Your survey answers will be stored in a password protected electronic format. The form does not collect identifying information such as your name, email address, or IP address. Therefore, your responses will remain anonymous. No one will be able to identify you or your answers, and no one will know whether or not you participated in the study.

At the end of the survey you will be asked if you are interested in participating in an additional interview. If you choose to provide an email address, your survey responses may no longer be anonymous to the researcher. However, no names or identifying information would be included in any publications or presentations based on these data, and your responses to this survey will remain confidential.

Please select your choice below. You may print a copy of this consent form for your records. Clicking on the "Yes, let's help the study!" button indicates that: you have read the above information; you voluntarily agree to participate; you are 18 years of age or older.

Do you wish to participate?

Yes, let's help the study!

No, I don't wish to take part in a student's project. /leads to check out from participation

### SECTION 1

Choose the preferred medium/media for obtaining news and other information:

TV Radio Printed newspapers/magazines Internet Social media Family and friends

If your preferred medium isn't listed above, please mention it here: \_\_\_\_\_

(This question filters out participants whose main media consumption channels aren't online)

Which sources do you trust the most?

Reuters Daily Mail BBC Fox News TIME The Wall Street Journal The Washington Times The American Conservative CNN Breitbart The New Yorker Vox Bloomberg Huffpost National Review

(This question is here to understand whether correlation between political bias and factual beliefs stems from biased media or from the respondents themselves. The list of media sources is taken from 'media bias ratings' and has 3 sources per each 'lense' (far-left, left, center, right, far-right)

If the source you trust the most isn't listed above, please mention it here: \_\_\_\_\_

---

How would you define your views?

Conservative

Lean Conservative

Lean Liberal

Liberal

None of the above

If none of the definitions listed above reflects your views, how would you define them yourself?

\_\_\_\_\_ (This question will define whether factual beliefs are dependent on political views and thus reveal political bias)

### SECTION 2

(This set of questions is based on Political bias test from Clearerthinking.org platform. Questions are slightly modified, but the main idea is that all questions address empirical facts and have a correct answer, but at the same time, they are politically controversial. For instance, questions 2, 3, 5, 7 seem to be a strike against liberal politics. Similarly, the questions 1, 4, 6, 8 seem to be a strike against conservative politics.

Respondents who are more often right on the questions where the true answers support their political points of view and more often wrong on the questions where it does not can be defined as politically biased.

In this section, please choose the answer you agree with).

1) Have the average global temperatures risen due to human activities?

Yes /correct

No

We don't know yet

2) Do you agree that this is unsafe to consume genetically modified foods (GMOs)?

Yes

No /correct

We don't know yet

3) Advocates of American capitalism argue that global technological and economic progress is largely due to American ingenuity.

There are 8 biggest publicly owned tech companies in the world (by market capitalisation\*). How many of them were American in 2014?

\*Market capitalisation refers to how much a company is worth as determined by the stock market.

3 out of 8

5 out of 8

7 out of 8 /correct

4) Some people explain why the economy isn't doing as well as it could by saying that "foreign aid spending is too high». Do you think this explanation is a major reason, a minor reason, or not a reason at all why the economy is not doing better?

Major reason the economy is not doing better

Minor reason the economy is not doing better

Not a reason at all /correct

5) Can radioactive wastes from nuclear power be safely contained for decades?

Yes /correct

No

We don't know yet



6) The average US citizen would be better off if a larger number of highly educated foreign workers were legally allowed to immigrate to the US each year.

True /correct

False

We don't know yet

7) Taxes are much lower in the US than in the average EU country. If lower taxes meant more wealth, you would therefore expect the US to be richer, in the sense of having a higher GDP per capita, than most EU countries. The US is...

Richer than nearly all EU countries /correct

Richer than most EU countries

About as rich as the average EU country

8) Does permitting adults without criminal records or histories of mental illness to carry concealed handguns increase or decrease violent crime?

Increases violent crime /correct

Decreases violent crime

We don't know yet

9) Donald Trump's election loss is a part of a strategy that will eventually lead to the victory and corrupted elites will be punished.

Yes

No

Not sure

If you have any additional thoughts and comments regarding this survey, please write it here:

(This will separate respondents into 2 groups: 1) those who despite an objective evidence still preserve their beliefs; 2) those who updated their beliefs and thus may be less radicalized).

## CHECK OUT FROM THE SURVEY

Thank you for completing the survey!

The next step of this research will involve deeper conversations about beliefs, online media, and communications. If you feel like sharing more about your experience in regards to these topics, please leave your email address and your name here:

Your email \_\_\_\_\_

How may I address you? \_\_\_\_\_

I will be happy to get in touch with you and invite you to an interview via Zoom or any platform you prefer.

Feel free to contact me in case you have any further questions, my email address is:

valentina.dzhekanovich@artun.ee

Thank you very much and have a great day.

## APPENDIX 2: LIST OF COGNITIVE BIASES FROM THE COGNITIVE-BEHAVIOURAL MAP

### AVAILABILITY CASCADE

A self-reinforcing process in which a collective belief gains more and more plausibility through its increasing repetition in public discourse.

### BACKFIRE EFFECT

The reaction to disconfirming evidence by strengthening one's previous beliefs.

### BANDWAGON EFFECT

The tendency to do (or believe) things because many other people do (or believe) the same.

### BELIEF BIAS

An effect where someone's evaluation of the logical strength of an argument is biased by the believability of the conclusion.

### BLIND SPOT BIAS

Cognitive bias of recognizing the impact of biases on the judgment of others, while failing to see the impact of biases on one's own judgment.

### CONFIRMATION BIAS

The tendency to search for, interpret, focus on and remember information in a way that confirms one's preconceptions.

### FRAMING EFFECT

Drawing different conclusions from the same information, depending on how that information is presented.

### ILLUSION OF CONTROL

Is the tendency for people to overestimate their ability to control events.

### ILLUSORY TRUTH EFFECT

A tendency to believe that a statement is true if it is easier to process, or if it has been stated multiple times, regardless of its actual veracity.

### INGROUP BIAS

The tendency for people to give preferential treatment to others they perceive to be members of their own groups.

### JUST-WORLD HYPOTHESIS

Cognitive bias that a person's actions are inherently inclined to bring morally fair and fitting consequences to that person; thus, it is the assumption that all noble actions are eventually rewarded and all evil actions eventually punished.

### MOTIVATED REASONING

A phenomenon studied in cognitive science and social psychology that uses emotionally biased reasoning to produce justifications or make decisions that are most desired rather than those that accurately reflect the evidence, while still reducing cognitive dissonance.

### NAIVE REALISM

In social psychology, naive realism is the human tendency to believe that we see the world around us objectively, and that people who disagree with us must be uninformed, irrational, or biased.

### OUTGROUP HOMOGENITY

Individuals see members of their own group as being relatively more varied than members of other groups.

### PURITANICAL BIAS

Refers to the tendency to attribute cause of an undesirable outcome or wrongdoing by an individual to a moral deficiency or lack of self control rather than taking into account the impact of broader societal determinants.

### SALIENCE BIAS

The tendency to focus on items that are more prominent or emotionally striking and ignore those that are unremarkable, even though this difference is often irrelevant by objective standards.

### SUBJECTIVE VALIDATION

Perception that something is true if a subject's belief demands it to be true. Also assigns perceived connections between coincidences.

### STEREOTYPING

Is an over-generalized belief about a particular category of people.[2] It is an expectation that people might have about every person of a particular group.

---

## APPENDIX 3 “FOOD FOR THOUGHT” IDEATION SESSION SCENARIO

### INVITATION TEXT

Forks and knives out! We will gather around a virtual table and explore our interactions with social media news feeds. What are your online eating habits? What is the nutritional value of the sentence you’ve just read? Will you share the last piece of cake with your friends or will you just like it and enjoy it by yourself?

### PREPARATIONS

invite participants  
send out EKA consent forms (to record the session, make photos and use images in the written part of thesis)  
ask participant’s favourite food  
prepare the set in spatial.chat: upload images of the table, chairs and favourite participant’s dishes  
prepare Rhonda: make her avatar, record her speech with google voice

### WARM UP AND INTRO (10 MIN)

‘Welcome everybody!...’ /explain the sessions goal, participants and Rhonda introduce themselves

We are sitting around an AI-powered table. It knows your preferences and gives you only the food you like. This is why each of us has something different on our plates. Let’s describe to each other what we are having and what we like so much about it. /building multidimensionality to food consumption and see what it adds to news feed consumption

Participants: ‘...’

Rhonda: ‘And am having a plate of candies over here. I always liked sweets, but I was not allowed to eat too much when I was a child. But now I have this AI-powered table. It automatically refills my plate with candies. And, what I also like about eating sweets at this table, is that it brings people around me with the same taste. We all share new sweets with each other. So, if you guys don’t have anything sweet on your plate, I think you should not sit here with me and go away.’

Shh Rhonda, this is so rude! Try to imagine that you will only eat the food that’s on your plate for the rest of your life. How would that feel?

Rhonda: ‘Oh I would love that. In fact, I’ve only been eating at this table for the past few years. It feeds me with objectively good food that I trust. Other tables serve fake food produced by big companies who brainwash you and manipulate you’.

What about other people around the table? How would that feel to eat ... for the rest of your lifes? What would you like to add to your table at some point?

Participants: ‘...’ /start adding different options

Rhonda: ‘Guys, you are already brainwashed. You can’t think critically about your food consumption. Why are you so stupid?’

Rhonda, you violate the terms and conditions that you have signed before entering the call. You can’t insult other participants. I will have to temporarily block you from this conversation.

---

Rhonda: ‘Of course you are blocking me! I am the only one who tells the truth at this table. And you want to hide the truth. I’ll leave you now and come back with an army of sweet truth warriors. We will open your eyes and you will see how objectively wrong you are.’

Rhonda leaves and joins Gab.

### MAIN PART (30-40 MIN)

What you just have seen is a typical example of a social media user from Twitter, who got spoiled by recommendations and filter algorithms and by echo chambers. Rhonda only trusts channels and other users who support her views. And if she encounters an opposite opinion, she protects herself by attacking everything that contradicts her beliefs. This way of information processing is far from healthy.

Since the issue is very complex, we will not address it directly. Instead, I invite you to think about it in parallel with a more simple analogy of food consumption. We will come up with basic healthy eating habits and think about how we might offer them to Rhonda.

From now on, we have 30-40 min to brainstorm ideas until Rhonda appears with an army of chocolate soldiers. \*introduce the rule of brainstorming, no critique and etc\*

### QUESTIONS TO NAVIGATE THE DISCUSSION:

- What are your basic healthy (or unhealthy!) eating habits? Can you apply them to your social media consumption habits?
- If you know that you have to eat something that you don’t like, but it’s healthy. How do you motivate yourself?
- How can we restore trust to other food options in general?
- How can we turn it from passive food consumption to active (mindful) food processing?
- Let’s think about the preventive measures. What if you were a person who created this table and who would see now its consequences? What would you do differently?
- How can we use this table and the digital environment to communicate the dangerous consequences of eating sweets only to Rhonda?
- Let’s imagine we have redesigned this table and it stopped spoiling Rhonda. How do we keep her from moving to other spoiling tables of her online friends?

### OPEN END (10 MIN)

Rhonda comes back. Hello guys, I am back. What are you discussing here?

Alright, I will think about it critically. I am sorry I called you stupid, I can see now that you are very smart. Thank you and goodbye.

Now when Rhonda is back, we have to present her our ideas. Choose one idea that you liked the most or several ideas. After each idea is presented, everyone turns into a sweet truth warrior and thinks why he would take/not take this idea.

Participants: ‘...’

Rhonda says final words and I wrap up the session.

---

# COPYRIGHT DECLARATION

I hereby declare that:

the present Master's thesis is the result of my personal contribution and it has not been submitted (for defence) earlier by anyone else;

all works and important viewpoints by other authors as well as any other data from other sources used in the compilation of the Master's thesis are duly acknowledged in the references;

I give consent to the Estonian Academy of Arts to publish my Master's thesis in the repository thus making it available for the general public by means of the Internet.

Pursuant to the above, I state that:

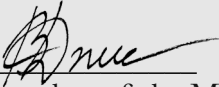
I as the author of the thesis am the sole owner of the individual copyright of the present Master's thesis and the works included and/or described within the thesis and the disposal of the proprietary rights related with the Master's thesis is subject to the procedures in force at the Estonian Academy of Arts;

as the Master's thesis published in the repository may be accessed by an unlimited number of persons, I presume that the readers of the thesis comply with laws and other legal acts and good practices in good faith, in a fair manner and with respect to and consideration of the rights of other people.

The copying, plagiarising or any use of the present Master's thesis and the works included and/or described within the thesis that infringes the copyright is prohibited.

07.05.2021

(date)

Valentina Dzhekanovich   
(the name and signature of the author of the Master's thesis)

The thesis complies with the Master's thesis requirements:

\_\_\_\_\_  
(date)

\_\_\_\_\_  
(the signature of the Master's thesis supervisor, academic or research degree)