Table of contents

Abstract.................................................................................................................. 3
Introduction........................................................................................................... 3-4
Research in Spain.................................................................................................. 4-6
Research in Stockholm......................................................................................... 6-7
Umeå in February.................................................................................................. 8-9
Research in Bucharest......................................................................................... 10-15
Umeå in March.................................................................................................... 15-16
Umeå in April...................................................................................................... 16-21
Umeå in May....................................................................................................... 22-23
Result................................................................................................................... 34-35
Reflection............................................................................................................. 35-36
Abstract

Social media content moderators perform a critical function of modern civic society: they act as first responders on social media platforms, getting confronted with harmful digital content on a daily basis. Without them, the Internet would be at the mercy of the unpredictable, global social climate, likely making the Internet a less safe, less peaceful place to be. Nevertheless, disturbing content screening cannot easily be addressed with the future’s planned full integration of artificial intelligence. Content moderation is not a technical challenge that can be solved by employing artificial intelligence as the solution to the content moderation related issues. Automated tools, for example, which are built to recognise creative permutations of language, fail their job. Cultural nuances and linguistic specificities are challenging to algorithms. For this reason, it is important to keep content moderation as a human process, because algorithms are not infallible in the content screening workflow and despite all the struggles content moderators face with screening content. The design concept explores how the emergence of deep fakes can be tackled by social media content moderators without relying too much on algorithms and computer power, but on human skills.

Introduction

Without human-powered content moderation, disturbing, harmful, hateful, and misleading content – crafted by dark impulses of human nature – can quickly surface the Internet without hindrance. Content moderators have this powerful editorial role of deciding what can stay and should be removed from their responsible sites and platforms. Particularly social media platforms are highly reliant on those critical, yet invisible, and undervalued gatekeepers of social media content production chains. However, social media content screening happens at a high cost. The nature of the material, content moderators are exposed to every day, might have deleterious effects on one’s psyche. Recently filed lawsuits proceeded by former content moderators against social media companies reveal how psychologically taxing this work is: What do you do when work is entering into your psyche and interpersonal relationships? How can you feel intimate with your partner the same way again after overseeing such content? How would you deal with anxiety issues, sleep loss, loneliness, and dissociation in and outside of such work environments?

This project “Is seeing still believing”, has explored in great length, whether resilience can sustain content moderators’ mental health. Research field trips to Spain, Romania, and Stockholm have resulted in a shift of the design approach. The initial aim of designing for resilience concentrated on a service design approach, attempting to answer the questions mentioned above. The conducted research, however, concluded that the tackled problem was too large in scope. Consequently, the concept has been narrow-scoped to an interaction design approach, which focuses on a deep fake
analyzing tool. This tool manifests itself in the form of an interface design for social media content moderators.

Research in Spain
My objective from the brief was to understand the content moderator’s work better, and help them to increase their resiliency, so they could cope better with potential mental health issues caused by the nature of their work. Therefore, I decided to follow content moderators in their daily work routines in Spain for a short period of time.

I arrived two weeks earlier in Spain to take the time to mentally prepare myself for the mandatory job training. I took the time to read myself through the content moderation literature, but also made myself clear about what I wanted to achieve during two weeks of training. I wanted to connect with my new colleagues and hear what they have to say about their profession. Taking active part in this training was important to me as well. I wanted to experience what content moderation is all about in spite of the recent, critical news coverage. During my first day of work training, I realised that new joiners had to train on another floor and thus were separated from the actual content moderation team. This changed my plan. Instead of talking to experienced content moderators, my research had to rely on new colleagues’ work experiences only. I assumed that experienced content moderators have more to share about their work life than new joiners. I was able to adapt quickly to the new, friendly work environment. Despite the overwhelming amount of policies a content moderator has to learn and implement, the training as a whole was easy to follow and to pass for most of the new joiners. We were reminded frequently that to be able to do this job, policies should be executed and not be questioned. For this reason, the trainer paradoxically challenged us to ask anything during the training that might be doubtful to us in terms of policies and the work in general. In consequence, there were surprisingly frequent discussions going on between the trainer and the trainees during policy training. Discussions were raised for example by ambiguous policy definitions or questionable policy implementation rules on specific situations. The trainer acknowledged that “our world is anything but black and white”. But for this job, it was necessary to differentiate between good and bad things in accordance with the client’s policies. Those discussion rounds allowed me to bring in my actual research in parallel. Whenever the chance was there, I asked critical questions about this nature of work. The trainer was surprisingly open and patient in regards to our questions. I asked things such as, “What does the future hold for this profession while machines are trained to do our job?”, or “What kind of work-related health issues might content moderators get affected by in the long run?”. What was more challenging to me was to connect with my colleagues in a short period of time. This is why I found it necessary, not only to spend time with them during breaks, but also after work. I felt that doing an interview in my circumstances, took at least twice the time as if a designer would do a regular interview setup with an interviewee. The way I approached the conversations with my colleagues was
to mix up semi-private related topics with work-related topics. I would start the conversation about video games for example, and would then steer that topic slowly back to content moderation. One the one hand, I found this type of immersive research rewarding, as I got to know people from a personal level as well. On the other hand, the immersive research did not leave me much space to reflect on what I have been doing after a work day. In the beginning, I started out to take 1-minute video diaries about my research, but unfortunately stopped. I relied mainly on my notes I took during work hours and the impressions I got from after-work hangouts. At the end of my training, I believed to have a good picture of the average content moderator starting this job. I got a good impression of why they started this work in the first place, what expectations they had, and what fears had accumulated since they started. One large piece that was still missing in my research was the resiliency aspect in context of content moderation. Are content moderators trained to be resilient against the content they have to review everyday? If yes, how is resiliency being trained? Is resilience actually something someone can learn? These questions pushed me to seek psychological help that was available on site. I had the opportunity to schedule three sessions with three different psychologists sharing with me different perspectives on what resiliency is and how to implement it into one’s work and private life. Interestingly, two of them were former content moderators themselves before they started working as consultants. Talking to them was the most stressful moment in my entire training. While I was asking questions about content moderation in general and how this type of work might affect one’s mental health, they also took their turn and asked questions about my intentions and background. I was wondering how long I could keep up pretending to be someone else, as I felt I had to answer their questions with sincerity to some degree. In exchange for information I had to give personal information back. One interesting thing I learned during those conversations is that one psychologist wished that she could comment on the policies whenever she felt that the policy-in-use needed corrections or that her case was an edge case and thus no policies were applicable to her problem. In her view, making comments would have helped her to produce more qualitative results in case solving. Moreover, the psychologists helped me to understand that resiliency is something that cannot be used as a remedy towards work related mental health issues, as I imagined. Designing for resilience, it seemed, was something that requires a good amount of time and professional attention to be able to see results. I concluded that my research approach was leaning more towards a service-related design project rather than an interaction design project.

In retrospect, the content moderation work in Spain was not as disturbing as the news media depicted, partly because the social media platform serves for a young, still relatively manageable target audience. Partly, because the research trip was limited to training only. Access to the actual content moderation production site was reserved to project-related employees only. This research would have been more richer, if conversations were planned more in advance, particularly with experienced
content moderators, who were difficult to reach with. Catching up with experienced colleagues on Slack would have been an alternative way to interview them about their motives of why they are doing this type of work for a longer period of time. Furthermore, it would have been beneficial to continue to document the research process with the video diaries that were implemented in the beginning, as they are a more accessible, enriched medium compared to handwritten notes. Overall, the research approach could have been more open from the start and less guided by the initial brief. The research direction was directed too much by the brief’s objective, without giving free space to explore other trajectories that might have led to other interesting insights. Nevertheless, this work experience still gave a rich complementary view on content moderation, alongside the secondary research that has been done simultaneously. Both primary and secondary research were mostly aligned in regards to the way content moderator’s daily work routine is understood.

Research in Stockholm

The results from my research in Spain showed that improving a content moderator’s resilience would benefit more from a service approach rather than, for example, through an interface application. The involvement of professional support for psychological matters would require people, not pixels. Back in Sweden, I started to look for professionals with expertise in resilience training and practice at work. One promising touch point I found was this psychiatrist working for the Military Academy Karlberg in Stockholm. The psychiatrist gave me a good overview of her role and function inside the Military Academy’s strategic management. Among other things, the management planned out a programme that strengthens the soldiers’ mental and physical well-being before and after deployment in conflict areas. In response to my question, whether resilience can be learnt or not, the psychiatrist answered that resilience training as such was not offered in the programme. However, the psychiatrist pointed out that all homecoming soldiers were required to participate in follow-ups organised by the strategic management. Those follow-ups took shape in the form of conversational meetings and therapy sessions, and should be honored as a closure from “battle mind” to “home mind”, before they are going home to their families.

In addition to that, the psychiatrist introduced a simple tool that measured stress based on a colour-system. This tool “Combat Operational Stress Control” (COSC) was used by soldiers deployed abroad to evaluate themselves but also fellow soldiers who might have suffered from combat stress problems, such as post-traumatic stress-disorder (PTSD). I was able to take away two practical applications from this conversation. One is about creating closure and the other one is about self-monitoring and monitoring stress.
Furthermore, I contacted a cancer patient from Hamburg via phone call. I wanted to gain a better understanding of what resilience means. From the patient’s point of view, resilience is a mindset; it is something that takes time to learn and is reciprocal to adverse events he experienced in his lifetime – marked by loss, addiction, and cancer. Although he knew that his cancer is incurable, he was still able to make jokes and stayed positive throughout the conversations, despite the constant pain he had caused by the chemotherapy. The patient refused to be entirely defined by the cancer. He rather chose to fight it off by continuing his life as usual, as much as he could. In times of weakness and desperation, his wife’s company and support was his anchor point. I conclude that not only the mindset, but also one’s social supportive circle is a significant factor in being able to grow stronger resilience. People need human care and support in order to be able to cope and outgrow their crises.

This phase showed that the research turned away from the first opportunity described in the initial brief and instead emphasized more on opportunity two and three, which are both more service design - oriented. This directional change became evident when the research was widened from the overarching topic content moderation to the theme resilience.

Looking for examples of practical applications in resilience training took priority and were found in the military training. Finding a deeper comprehension about the meaning of resilience was clarified and sharpened through conversations with a cancer patient. Throughout this phase, it became apparent that designing for resilience meant to design reflective and caring dialogues, as illustrated with the military and the cancer patient. In the context of content moderation, some production sites already offer psychological consulting hours on a voluntary basis. The question is whether such consulting hours are effective enough to support content moderators processing disturbing content and whether they are actually aiming to improve one’s resilience in the long run. Designing around this space of mental health would be less about self-help initiatives, but rather more about how to connect content moderators with, for example psychologists or close family members, to have those reflective and caring dialogues concerning the things they have seen. This phase was predominantly occupied with mental health in general, PTSD, and resilience. It could have been helpful to do secondary research on existing research around similar professions suffering from PTSD beforehand, such as drone combat fighters. This might have sped up the conversations with the military and the cancer patient more and led to results with closer relation to content moderation.
Umeå in February

One week before the research examination, I received input from different people I got in touch with a while ago.

A PHD student and former police officer from Umeå University gave me input about his PHD work. We discussed the police officers’ mental health and how the police training might have an effect on it. He claims that the seeds for mental health related issues are sown already during their training to be officers. This conversation sparked the idea of whether resilience training was actually necessary for content moderators in the first place. What if it’s possible to filter out sensitive, vulnerable applicants for content moderation jobs already during the interview screening. The screening would then need to be redesigned in a way so that only “resilient” applicants would make it through the job interview process. In his opinion, a refined intelligence test could be used as an indicator to find out whether someone is resilient or not.

I talked to an old friend, who currently lives in Berlin. I asked him to join a content moderation job interview and report me back once he completed it. He went through a 2.5 hour job assessment center in Berlin, which included a general presentation about the company, a test, and a personal interview with each candidate. His experience was valuable to me, as I was able to compare it more objectively with my content moderation job interview experiences.

An Interaction designer from Microsoft made me realise how large my chosen thesis topic is, just by reading my brief. She encouraged me to scale my research down to something more tangible. We discussed possibilities of improving the content moderator’s workflow through the support of technology. Artificial Intelligence was one suggestion to combat harmful, disturbing content while reviewing.

I spent the rest of my time preparing the research presentation. My primary goal was to show what content moderators actually do day by day. The work experience in Spain helped me to illustrate that. Based on my main insights I got from the research I also wanted to convey the message how easy it actually was to become a content moderator. I believe that anyone can become one, unless the job applicant cannot do shift work and cannot read English. The input from my old friend confirmed this claim.

Lastly, I wanted to show some sort of design directions through simple early concepts, which were not actual concepts but more like wiggly compasses I could use to navigate with up until the next milestone. The early concepts were informed by previous conversations I had with, including the
psychologists from Spain, the psychiatrist from the military academy, and the Interaction designer from Microsoft.

The “panic button” concept was inspired by the psychiatrist’s input about creating closure from “battle mind” to “home mind”. I have found that creating a temporal healthy distance to bad content could be one approach to mitigate the possible mental health issues caused by reviewing content. The “panic button” can be seen as an actual button placed on the work desk. Once that “panic button” is pressed, the content moderator can leave the work desk for a break(out), without directly affecting his/her performance metrics, usually determined by algorithmic management.

The algorithmic filter concept was based on the conversation I had with the Interaction designer. Filters, like the ones found in photo apps and in other social media platforms, could be integrated into the content moderator’s workflow. Whenever a content moderator has to watch a disturbing video clip, the A.I. could signal the content moderator ahead via “trigger warnings” that some scenes in the video the moderator is about to get exposed may upset or disgust him/her. The content moderator could either decide to distort those scenes with a filter, so he/she does not get exposed by it fully or simply skip it.

The third concept revolves around placing comments on policies. Content moderators must use those policies to complete user reports and decide what can stay or need to be removed from the social media platform. But correcting or commenting on those policies is usually not part of the standard workflow. Probably because comments would just slow down the handling time for solving cases. This idea was coming from a psychologist, who shared with me her past content moderation job experience.

This phase marks the converging process of the accumulated research done in the past two months. The research results were shown during the research examination gateway, with emphasis on the findings from Spain and Stockholm. The feedback received from professionals helped to scope down the project ambitions from the initial brief to a more specific problem. From this point, resiliency is not the main focus in this research anymore, but the daily work life of the content moderator. Instead of trying to cater users’ psychological needs with a service, the project is moving more towards something that is directly related to the content moderator’s context, for example making their working tools better. In response to that, three early concepts had been presented as approaches for designing new supportive “tools” in the field of content moderation.
Research in Bucharest

After the research examination I decided to leave Umeå again for another content moderation job opportunity in Romania I applied successfully months ago. Three factors made me go. Firstly, I felt that I needed to validate my early concepts in a real-world work context. Secondly, I haven’t talked to that many experienced content moderators, yet. And thirdly, the previous job opportunity in Portugal did not expose me to any real harmful, disturbing content, which made me curious about other social media platforms’ content moderation production sites.

Bucharest job interview

The job interview for the content moderation position in Romania was one of the most transparent content moderation job interviews I ever had. The interviewers spoke openly about what to expect from this job, by for example showing samples of disturbing, harmful content as well as testing my judgement on case examples. As in all other content moderation job interviews, I pretended to be a student who was looking for a new orientation in my life. It was challenging to convince them why I wanted to move to a country I have never been to before. I made clear that change was necessary in my life and therefore any country would have done the job.

Bucharest arrival

Once upon arrival in Bucharest, I went to my Airbnb apartment, which was just 70 meters away from the office. This close distance to work allowed me to go back anytime during lunch time to write down notes or catch up with my tutor on Zoom. Since I already went through a similar onboarding procedure in Portugal, I did not feel the need to prepare myself for the first week training in Hungary. The first 1.5 weeks of training was instructor-led in a dedicated training space. Once the trainee passed the self-paced assignments, trainees would continue the training at the actual work space. Each trainee would sit next to an assigned colleague, who supervised the trainee. This nesting period was the most valuable time as a design researcher, since it gave flexible time to shadow and talk to experienced content moderators. I planned to use that flexible time to co-create with content moderators. I imagined developing thoughts on the future of content moderation by sketching with content moderators together.

After three weeks of training, the trainee was supposed to work on it’s own.

Week One
The first work day started as an official onboarding day, through a formal introduction to the new work space, colleagues and company culture. Before the new joiners could see the work space, all of them had to sign a work contract. Additionally to that, content moderators also had to sign a piece of paper stating their awareness of the fact that they might be reviewing harmful, disturbing content at work.

The first week of training was similarly structured compared to the training in Portugal. Learning policies by going through presentation slides together with the trainer was the main focus in the training. In a content moderator’s job, the hardest part period is surprisingly the training, since new joiners have to learn so much within weeks. “So much information, but you cannot use [apply] it somewhere else.” as stated by an employed content moderator.

Moreover, the trainer introduced best practices on how to solve cases most efficiently. She referred us to an ever growing living online document filled with a great amount of constantly updated information pages and rulebooks we should always keep an eye on while solving cases. She recommended us to always cross-check with this living document. The living document should always be displayed on our second monitor screen:

It is good to have the information page open, this is where you have all the information, all the materials for the training decks. It is handy to have this open. Yes, I am doing my best, but I can’t remember everything, and there are updates I should check.

In regards to my question about what role a content moderator is taking on a social media platform, she replied: “I don’t think we are censoring or anything, I think we are educating them. [...] This process lets people know that you need to learn how to address certain people correctly.”.

**Week Two**

In this research field trip, I planned to engage people outside of my training class as well. I was particularly interested in experienced content moderators as I had not many opportunities to talk with them on the last research field trip. Furthermore, I also wanted to engage with people who are connected with content moderators in general, such as trainers, qualitative analysts, and team leaders.

One of the new joiners started out as a content moderator a year ago in another country and got transferred to my location. We got to know each other quickly and spent time after work whenever possible. Her revelations about the content moderator’s job was disillusioning. Content moderation job is “like any other job”. Contrary to my expectations, she actually enjoys being a content moderator. For one reason, because she gets excited by watching “gory” content. For another reason, she is from time to time helping people by removing content such as revenge pornography. In her opinion, this is one of the most meaningful and gratifying things that can happen to her at work. However, she also states that “I want to do something else in my life. I am just looking for something that gives me a real
professional career.". She thinks that the content moderation career path is a dead end with no real future perspective. Her concerns were hardened by a team leader’s comment on the future of content moderation. He works as a team leader, who is partly in charge of the data labeling department: "We [Human labellers] are feeding the machine with data. Super valuable for [the client], because it improves the platform, it improves the user experience, it grows their market as well. I think it will be the future.". The team leader goes on with: I think it will take many, many years. But until then, we are improving [the client’s] algorithms, to be able to do the work full-time."

Another qualitative analyst I talked to describes the future of content moderation as follows: "How can we use A.I. to help us solve cases. What is the minimum input from the agent we need?". The qualitative analyst took part in a creative workshop with several other stakeholders trying to come up with alternative futures of content moderation. One vision he shared with me regarded the content moderator, "[...] who shouldn’t be the experts in the policies, because they are complex. So let’s make a fast system that asks agents a couple of questions. Let the agent answer the questions. [...] And let the system come up with the solution.".

One of the core systematic problems that many content moderation production sites hold is that policies are created as reactions to the ever-changing global social climate on social media. The client assigned those policies from top to down to the subcontracted content moderators, without having a good understanding and oversight of how those policies should be practically used by moderators: "Most of the time, [the client] didn’t really know what their policies were. They create the policies, but they don’t know how to implement it." – stated by a qualitative analyst.

**Week Three**

After two weeks of training, I started to realise that the work-related mental health issues depicted by the news media, is a present problem but not as bad as covered by the news. This is one reason why I moved more and more away from trying to tackle problems in the field of mental health. I got more interested in the question of how to make the work process of content moderators more efficient, overall better.

During my nesting period, I was able to engage with a few experienced content moderators. One of them was my mentor, who also worked on child sexual exploitation (CSE) cases. I was able to shadow her while she was going through child porn reported from around the globe. Seeing such content made me think what her emotional journey would be like in a long time span. I felt I was not in the position to ask her directly, how she feels when working on CSE. But I was able to pick up some bits and pieces that reflected her emotional state towards CSE cases throughout the conversations we had together. Overall, she made a professional impression handling those cases. She stated that she must see and jump through all the CSE videos in order to identify the age and other factors, but once she completed her CSE target, she stopped thinking about it, she said.
I chose not to stick only with my mentor, but also talked to a few other moderators at the work space. I learned different kinds of workflows and “hacks”, depending on who I was sitting next to. I observed common struggles among content moderators when solving cases. Particularly edge cases were interesting: “The fucked up stuff is so easy to detect, but then, some of the cases are so difficult to detect because you need to understand the context quite well.” Edge cases are challenging cases content moderators sometimes have to review on. For example, one content moderator received a user report stating that she wants a revenge porn video to be removed from the platform. In order to remove something, users usually have to prove their identity connected to the reported video. Unfortunately, her face on the video was not identifiable. The only thing that would identify that woman was her tattoo that was visible on the video. However, according to the policies, the moderator can still not help her, as the moderator must be able to identify the reporter clearly on the reported video. “Let’s say, you have a sex video, but you cannot see the face, but you can see the tattoo right. All of such subtle details make it not that easy to solve the case.”. In such a moment, it becomes challenging for the moderator, when one is expected to solve a case, but has no appropriate policy at hand that can guide one’s decision-making. Being under time pressure, the moderator cannot spend too much time on one single case. The interaction between user and moderator for example, is usually limited to three responses. Which sometimes gives the reporter the impression that he/she is talking to an automated system. The moderator needs to move on in order to keep up with one’s daily minimum target of case solving. As the moderator solves cases, qualitative analysts from other geographic work locations evaluate the moderator’s performance metrics. The qualitative analyst does spot checks on completed cases. Just forgetting to label a case’s type of violation already counts as a mistake. In consequence, content moderators are not only evaluated by how well they solve cases, but also their productivity is of importance for the client as well: “A good moderator makes no mistakes. And know how to solve cases. That’s a good moderator. A lot of cases per day. Productivity and quality altogether. And then, you become good.” – stated by an employed content moderator.

The Co-Creation Attempt

The planned co-creation workshop at the work space turned out to be one of the biggest challenges in my research so far. When I looked for workshop participants, I realised that participants in any kind of workshop setting need to be “on the same page” with the facilitator in order to be able to co-create together. This means that the facilitator needs to clarify the intention of the facilitator’s workshop. If the participant does not fully understand why the facilitator wants to be creative with someone, especially when the participant has never done co-creation workshops before, the willingness to corporate will likely tend to zero. At this point in time, I was not willing to reveal to them my intentions about my design research. I also did not dare to pretend to be just doing creative exercises with moderators. The more I left my content moderator role, the more suspicious people became. This was
clear to me, when I did my questioning rounds with a couple of moderators. In regards to my questions, I got responses from content moderators like: “If you want to improve policies, you need to talk to a team leader or a qualitative analyst”., or, “Our tools are pretty much fixed. Like, you have the tools, you have the preferences. It is coming as it is. So you cannot really modify it.“. On that account, I decided to ideate through questions, based on “what if’s”. The what-if questions led me to one interesting idea about involving user communities with the content moderation work. At first, the content moderator started to oppose my questions with logic: “It can be easily taken into the wrong direction. Someone is voting “this is right!”, that climate change isn’t real. And a bunch of morons vote like that, too. I don’t trust people to be true about certain things.”. But once I explained to her just to imagine how things could be different, if she would control the platform, she began to become creative: “Maybe the community has decided, but this is not the final answer. A public vote that I [the content moderator] can take into consideration.”. Her idea came close to the early concept “commenting tool”, although from the general public’s perspective. It was even more difficult to find a group of people to do this exercise. I changed my plan and singled out one content moderator I could ideate with together after work. But even outside of the work environment, it was still difficult for me to animate people to be creative, without telling them anything about my research project. In those conversations I was still able to validate my early concepts from the last research presentation. It turned out that neither the “panic button” concept nor the “filtering tool” were useful to content moderators. Content moderators usually take breaks when they have encountered something bad in front of the screen. Furthermore, a filter would be considered as “waste of time”, as moderators need to see the content without distortion, in order to be able to make a clear judgement on that case.

Reflection

It was insightful to see and compare two different companies operating for two different social media platforms. One similarity both companies were sharing was the work space setup. Each work desk was equipped with two monitors. Details like these helped to create a believable storyboard for the design concept later in the design process. One dissimilarity was that each company for example, had its own up and running “back-end” interface designed to investigate users and their violations. A much needed redesign on those interfaces would likely increase the content moderator’s productivity. However, the aim of this research was not to inform the design with anything that supports the logic of platform capitalism. But rather to produce insights for a more humanized, sustainable way of doing social media content moderation.

Content moderators were invited by the design researcher to participate in a creative dialogue about the future of content moderation. Although content moderators were willing to take part, it was difficult for them to show their creative spirit, partly due to lack of transparency communicated by the facilitator. The research circumstances did not allow the facilitator to be transparent about the
workshop's intention and thus it was not possible to design with content moderators together in an open way, as it's normally the case with creative workshops. Furthermore, the workplace did not seem to be the right, open environment to do “creative exercises” either. Thus, it was a challenge to materialise any creative output with them together. Still, the dialogues were helpful insofar as some “what-if” based discussions led to interesting thoughts. Moreover, content moderators were helpful in validating the early concepts presented in the last research presentation. None of the early concepts were shown to them directly, but rather explained to them through “what if” questions. In consequence, participants could have been briefed better about the workshop intentions. A quiet workshop-friendly place could also have been organised beforehand. Moreover, more courage could have been shown during the dialogues – with playfulness and use of materiality such as paper and pen.

Umeå in March

I had to consolidate the findings of three weeks of intensive research in Bucharest for the upcoming Half-way presentation. The objective of the presentation was to show more in depth the job of a content moderator, to show the validations I gained from the research field trip, and how I am going to use my insights in my next steps. I received feedback on my presentation draft from my tutor and from a service designer based in Oslo. The service designer encouraged me to leave behind my opportunities described in my brief and be more open for other growing opportunities. My tutor encouraged me to think about how my way of conducting research might influence my design practice. I have acknowledged the feedback but couldn’t implement all of it due to the lack of time.

Future casting

After having created the body of research originated from Spain and Romania, I felt prepared enough to transform the research into design. At this point in time, I imagined the design outcome to be something physical that would express an inspiring, and yet believable narrative about the near-future content moderation work place. Therefore, I wanted to work with future casting, a workshop toolkit that helps painting bold provocations about the future, together with other people. The workshop would have been initiated by confronting people with “fake news” articles based on my research findings. Those articles would then inspire workshop participants to rethink about today’s content moderation practices and in a next step create three potential futures of content moderation together with others. The potential futures would be used to create a contextual setting for the design to live in. The design outcome itself is still not defined yet, but will be, once the “habitat” of the design is decided.

Feedback after Half-way presentation
The extensive feedback from the examiner and external tutor confirmed previous comments made by the service designer and tutor. The presentation showed three large possible design directions (ethics, policy, and technology) I could take. During the presentation it was difficult to see which direction I wanted to go, making it also hard for the examiners to decide which direction they should pay attention to for giving feedback. Another feedback was that it was sometimes not clear from which source my insights came from, as I was just reading them out, but not mentioning who was speaking. All the insights were reflections from interviewees. Another issue that was flagged was the ethical part of my research. I have mentioned the research ethics briefly during the exam, but it was not enough to make the examiners understand how my conducted field research influenced my design practice and how I generally would cope with this situation in the future. Those points needed to be fixed before the next milestone.

Umeå in April

Looking back at the brief again, the project slowly shifted the focus away from designing a service, to designing interactions in existence with an interface concept. The research findings have revealed a small fragment of the complexity of what lies behind the problematic employee’s work experience and wellbeing in the context of content moderation. Solving a service design problem in this scope would require a more open communication between different stakeholders, which was not given from the start due to the research circumstances. As a result, the research efforts have become more sympathetic to a more narrow-scoped interaction design project that still aims to design for the content moderator.

The panic button, the filtering tool, and the commenting tool

From the three design directions presented in the last Half-way presentation (panic button, filtering tool, commenting tool), I decided to advance with the “commenting tool” concept. For the reason that the “commenting tool” concept seemed to empower content moderators more in their work life than the other two concepts. In hindsight, the two concepts could be seen more as preventative measurements against disturbing, harmful content. And yet, they would not bring any significant change to the content moderation process themselves.

The commenting tool

The “commenting tool” concept, however, would make the content moderation workflow more transparent for everyone at work. Everyone could see how policies have been juggled in the past and present, through the use of the commenting tool. Comments on policy implementations on cases adds value to the content moderator, because they can add more insightful information or suggestions for
improvements for usually not-up-to-date policies. One of the big challenges content moderation production sites face in fact is to keep up the pace with the global social climate circulating impactfully around the world. The ever changing impulsive, fluctuating global social climate is one of the main reasons why it is difficult to keep the hundreds of policies up to date within a user-unfriendly designed conglomerate of webpages. The commenting tool could relieve the pressure of constant updates and follow-ups on policies and instead propose a more transparent and visible interface design that empowers content moderators to tackle even difficult edge cases without worrying about getting lost in it. Because comments offer orientation on what has been done before and what needs to be done in the future in the context of case solving.

**Theses**

One of the feedback I received from the last Half-way examination concerned my “design principles”, which were not actual design principles, but in hindsight were more like a collection of key words describing key insights. I reworked my “design principles” into theses. These theses serve as a frame for envisioning alternative futures of social media content moderation. Each thesis is derived from past readings in previous design studies seminars, including “Do Artifacts Have Politics” by Langdon Winner, “Changing Things” by Johan Redström and Heather Wiltse, and “Platform Capitalism” by Nick Srnicek.

The theses inspired me to rethink my chosen “commenting tool” concept, as the current idea did not prove to be consistent with the well positioned theses. The commenting tool would for example not necessarily increase transparency for platform users in the way content moderation production sites currently operate. Furthermore, this concept would still follow the logic of platform capitalism; It will increase the quality of content moderation work overall, but it will not systematically shift any values beyond efficiency and productivity.

**Digital juries**

The paper “Digital Juries: A Civics-Oriented Approach to Platform Governance” I found during my desk research challenged my current commenting tool concept, as it presented a way of democratising social media content moderation. In a conducted social experiment, the authors Fan and Zhang invited the general public of Internet users to be involved on a panel of digital jurors, who would take over similar content moderation tasks as content moderators complete today. The new premise was that inviting the general public to work on the content moderators’ cases is more inclusive and thus less “platform logical” than designing a commenting tool exclusively for content moderators inside the production site.

**Throwback**
When looking back at my notes, I noticed that this new concept direction had been already in discussion, even before the “Digital juries” paper was published. During research in Bucharest, a content moderator had this idea of a public vote:

“Maybe the community has decided, but this is not the final answer. A public vote that I [the content moderator] can take into consideration.”

The paper, however, suggested two distinct ways of processing flagged content with the integration of a digital juries online system in mind. Practically, the jury, formed by selected online users from the general public of Internet users, would need to form a majority through voting for a verdict to happen for any open case. The content moderator is not included in this process. The other approach is to accept the verdict decided by a majority in the digital juries as a recommendation only, so that the final decision still has to be made by the content moderator. Based on the research, the recommendation version would be preferred by the content moderator: “ [...] because they [user community] don’t know what to moderate or how.” – stated by an experienced content moderator. It was important to me to include the content moderator’s voice in this new concept, which made me select the recommendation verdict over the final enforcement verdict. Implementing the digital juries as a recommendation system for content moderators would require more information on how the interactions between the general public of Internet users and the content moderator would look like in detail.

Overall, the content moderator’s voice validating the current concept idea, gave me confidence that this direction can still be considered useful for content moderators, although it seems slightly far-fetched from the feasibility point of view. The question now was how to package all this in a believable story.

**Future casting**

At the end of my previous Half-way examination, I explained my next steps in the design process. The steps included a future casting workshop with fellow students that should help me to situate my design concept adequately in the right time, in the right place, with the right value propositions I am standing for as an interaction designer.

Due to the current Covid-19 pandemic, live group workshop sessions had to be canceled. Thus, I chose to do the future casting on my own and still came to satisfying results. While preparing the future casting, I oriented myself on the “Futures Toolkit” published by the English Government Office of Science. This toolkit provided me with useful tools for “gathering intelligence about the future”, and “exploring the dynamics of change”, “describing what the future might look like”.

Not exactly knowing where I was heading to, I started out with intensive desk research to “gather intelligence about the future”. The “horizon scanning” method helped me to look for early warning signs of emerging trends and developments that could have an impact on the social media content
moderation production sites in the future. I scanned through the google news search engine and used for search keywords such as “content moderation”, “community manager”, “content screening”, “human reviewer”, “automated content moderation”, “facebook”, “230a section”, etc. I also looked into future trends published by the Pew Research Center, for example “Many Tech Experts Say Digital Disruption Will Hurt Democracy” or “The Future of Free Speech, Trolls, Anonymity and Fake News Online”. Another source I used were “Tech trends” reports regarding the future of work, the tech industry, and the news media.

The next step involved explorations on how these trends and developments might combine and what impact they might have. For this, I made use of two methods described in the tool for “exploring the dynamics of change”.

The “Driver Mapping” method was used to identify drivers of economic, environmental, legislative political, technological, and societal nature. Once drivers were generated, evaluated by their importance and certainty, drivers would then be used on the “Axes of Uncertainty” as “uncertainties”.

The “Axes of Uncertainty” method is used to describe multiple, alternative futures for the context of social media content moderation. The axes would be fed with uncertainties generated by the driver mapping method and should result in short, detailed scenarios, describing plausible outcomes and impacts.

A completed Axis results in a 2 x 2 matrix, representing four plausible futures based on what I observed from references of the past and present. In order to populate those axes, I wrote for each uncertainty an opposite state to compare with (show example image).

I generated as many axes as I could across all six drivers. I then chose two uncertainties that originated from different drivers (economic, environmental, legislative, etc.) and placed them on opposing axes. In order to generate unimaginable plausible futures, I tried to connect two uncertainties that contrasted each other as much as possible.

Adding a descriptive headline to each quadrant’s plausible future would further help to identify risks and opportunities (near-term opportunity, long-term opportunity, near-term risk, long-term risk, etc.). Usually, the created axes of uncertainty would be discussed together in a group setting. But I chose to explore the quadrants of interest in more depth on my own. I decided to go for those quadrants that usually have not been seen in strong association content moderation before. Cross-pollinations between content moderation and with “subscription economy” or “synthetic media” for example, produced thought-provoking materials for the next activity in line; scenario thinking.

The aim of scenario thinking is to build stories that describe alternative ways of the future of social media content moderation. Each scenario explores how different conditions might support or constrain social media content moderation. Through this, different behaviours of key actors, such as the government, citizens or businesses, might surface under those different conditions. I did not
intend to use scenarios as predictions, but rather use them as challenging, stretching or controversial pictures of the future to challenge and develop my conceptual design work further. Due to the lack of time, I was not able to implement the produced materials for the “scenario thinking” activity. Instead, I made use of already existing future scenarios. The journalist Sam Guzik, for example, created futures about social media platforms which I used to continue on, where I left off with my previous work.

This future casting has taken me to many places, from readings about artificial intelligence, media manipulation and disinformation online, the future of news media, the future of digital life, the future of work, social media, to technology ethics. The results of future casting helped me to identify “weak signals” coming from the branch “synthetic media”. In particular “deep fakes” generated by artificial intelligence, are of interest. This finding was an outcome of the axes of uncertainty exercise. I have found it compelling to imagine how deep fakes will become a problem for society, as such technology will likely accelerate the decrease of trust in the media. In the report “deepfakes and cheapfakes – the manipulation of audio and visual evidence”;

Social media platforms are one of the main responsible nurturings and spreading media manipulation and disinformation online around the Internet. The least they can do is to enable social media content moderators with appropriate tools to fight against disinformation, such as deep fakes, which have the potential to distort reality, making it difficult for people to distinguish what is fake and what is not. How then, can content moderators process deep fake cases in their every day work routine?

**Putting all things together, the final concept**

The commenting tool for content moderators has changed to a commenting tool for the general public.

Involving the general public in content moderation work can work under two conditions. It is important to take into consideration what kind of Internet audience shall be attracted to content moderation work, how Internet users will be attracted to do this type of work, and what type of content will the selected jury be reviewing on. To be able to answer these questions, the fictional company “Good Loop” was invented. Good Loop will used to create a believable, coherent narrative around the concept:

Good Loop is an intermediary partner between social media platforms and online news paper outlets. Good Loop makes profits by solving difficult edge cases for social media platforms. Those platforms pay Good Loop to complete difficult cases they have trouble solving with. Good Loop does not accept all kinds of cases. Cases that are appropriate to review by the general public will be accepted, for example deep fakes. Disturbing cases, such as child sexual exploitation cases will not be accepted. In order to find suitable people to solve the selected cases, Good Loop approaches online newspapers that provide a relatively educated readership and also have an installed paywall system in place. Good Loop will offer online newspaper outlets to integrate Good Loop’s paywall option into their existing ones. In return, Good Loop pays these newspaper outlets for being part of their subscription options
and further helps the outlets to position themselves as active misinformation fighters. From the perspective of online newspaper readers, they have the option to either pay for the content the newspaper offers or work for it as “temporal content moderators”.

Online newspaper readers would solve only appropriate cases provided and selected by Good Loop. Solving cases will be restricted to a certain amount of time, depending on how many articles the reader wants to consume. To put this ecosystem more in a near-future time frame, readers would particularly solve deep fakes in the light of the US election 2020. Therefore, also a work interface has to be designed for temporal content moderators.

This concept was mainly driven by the question of how to attract the general public of Internet users doing content moderation work. But why should people do content moderation work for free? This question led me to the human verification system “reCAPTCHA” provided by the tech company Google. “reCAPTCHA” blocks spam bots while allowing human users. It has become an omnipresent (inter)face on the web, which made me think to look for similar kinds of interfaces for this concept as well. I was searching for a touch point where people can be crowdsourced for content moderation work. In the light of the emerging “subscription economy”, I found digital paywalls a suitable interface to begin with. The idea around “Good Loop” was predominantly framed by the theses created from the research. The theses steered the initial commenting tool concept to strive more an “illogical platform capitalism behaviour”. This behaviour would result in a more inclusive, more transparent way of operating content moderation production sites. The general public for example, gets included in the content moderation process. People can participate as digital juries on the work interface provided by Good Loop. People would discuss and express their opinions about how to solve the case at hand, and on that account inform the content moderator with an agreement on a verdict recommendation that can become crucial for the final decision-making by the social media content moderator.
Reality-check
The delivered concept at the process gateway presentation needed a reevaluation. I came to realize that the concept tackled too many things at the same time. The concept attempted to involve the general public in social media content moderation work. It wanted to cooperate with online newspaper outlets. And it wanted to provide an interface that connected the social media platform, the online newspapers and the general public for the purpose to solve cases to support social media content moderators. Involving the general public has its limitations. For example, how can Good Loop be sure that a random collection of individuals seeking out the content moderation duty can be trusted? They can manipulate the results and bend results to suit specific political agenda. Concerning the online newspapers, paywalls are not particularly the most welcomed interfaces by Internet users. It is difficult to estimate whether people would actually be willing to "work" for the content of interest, or rather just pay or go to other outlets. Trying to put everything together in one interface seemed to be the most challenging, as multiple parties’ needs had to be considered in the design.

Benchmarking
I looked into various design projects found on the Internet and tried to understand what it was that made their projects good design outcomes. I summarised each design project with three questions. What was the problem the project tackled, what was its proposed solution, and where did the weak signal or inspiration come from. Successful design projects, that are successful in being awarded with recognition, usually answered the questions mentioned above in a short, concise and comprehensible manner. At this point in time, the concept could not be summarised in such a way. I decided to rethink the concept once again.

Looking back at my thesis goal
I have written down my personal goal for this project in the beginning of the thesis period. I revisited this goal to make myself clear what I wanted to achieve as an Interaction designer. A reworked on the concept again having this goal in mind: For my thesis, I want to design something like a fortune cookie, that surprises my audience with an insightful and foresightful message. This message, packaged in a moving, visionary narrative, seeks and communicates a question that presents a new and critical perspective on the daily life of tomorrow's content moderator. The designed message resonates with the audience isofar that it provides insights in my strategic and visionary thinking as well as the fact that this fortune cookie is also practical/useful – in the sense that you can eat it with relish.

The Future Mundane
The design approach "The Future Mundane" by Nick Foster, has heavily influenced the new concept direction. The Future Mundane helped to push the level of concept design that does not feel like science fiction, but focuses more on the relationship between people and technology.
The weak signal

“A weak signal is an indicator of a potentially emerging issue that may become significant in the future. Weak signals supplement trend analysis and they can be used to expand on alternate futures.” – Mikko Dufva.

The weak signal relevant for my concept concerns the emergence of synthetic media. Synthetic media technologies will become commercialised and widely available. Without the infrastructure in place to help people distinguish between synths and humans, the likelihood of misinformation campaigns will rise. Social media platforms will be facing huge problems in contending with an increasing volume of misinformation – in the form of deep fakes particularly. Deep fakes have the ability to undermine public trust in the media and above all democracy. Because of the Deep fakes’ misleading nature, they could be treated the same as all other Internet content, showing up in search results, broadcasted through the Internet of things, and throughout social media. They can bring significant harm to one’s safety and reputation, not only to celebrities and political figures, but also to civilians, who can be targeted with character assassination and cyberbullying.

The content moderator and synthetic media

This concept is aiming to build tools for social media content moderators working on detecting, taking down, and labelling deep fakes and other synthetic media. There is no sustainable way to detect synthetic media in real time and at scale any time soon; detecting deep fakes and other synthetic media will always be a “cat and mouse game”. This is why social media platforms place their bets in automated content moderation instead of manual investigation by humans, as technological progress may soon make manual investigation impossible. Despite all that, this concept believes to keep content moderation as a human process, by empowering content moderators with tools machines cannot and will not be able to use and operate like humans can in near-future time.

Final Concept
The final concept's proposal suggests to expand and crowdsource the current content moderation work force with support of the general public. Collective intelligence will work to identify synthesised and manipulated media, highlight verified facts and block unreliable sources.

Making detection tools available for users to flag manipulated media on their own and taking advantage of the general public’s cognitive and sense-making ability, will widen the angle of a moderator’s perspective on individual cases.

The general public is generally defined as visitors and subscribers of online newspapers such as "The Washington Post" or "Die Zeit". It is assumed that online newspaper readers are more receptive to world affairs compared to crowdsourcing platform workers on Amazon Mechanical Turk.

Most of today's content moderation production chains are heavily relying on outsourced companies moderating user-generated content. From the industrial-scale platform’s point of view, scaling up content moderation is needed in order to be able to respond to the growing number of users and user-flagged reports in a timely manner.
The concept has been developed to a crowdsourced variant of content moderation. This direction has been framed by the previous formulated 6 theses during the research phase.

### Theses

**T1**  Content moderation is **fluid** in its actions

**T2**  Content moderation has **politics**

**T3**  Content moderation becomes **transparent**

**T4**  Content moderation is acknowledged as **critical** in the workflow of social media

**T5**  Content moderation continues to rely on **human labour**

**T6**  Content moderation de-prioritises the **logic** of platform capitalism

One of the theses expresses transparency as an important element that needs to be increased in the content moderation workflow. A more transparent work environment can make the work conditions better for social content moderators, which the current system does not allow. Crowdsourcing
content moderation invites the general public, particularly online newspaper readers, to participate in the content moderator’s work routine to act as a supportive role in evaluating cases.

This role takes the similar shape of a qualitative analyst, who usually checks the work done by content moderators. The general public is understood as a form of collective intelligence and further is used to empower the content moderator making better judgements on case work. Particularly the contextual and sense-making abilities of a global crowd can be useful to for example find missing information and detect irregularities in challenging cases. As an example, a crowd cross-checks manipulated media’s assertions in partnership with professional fact-checking organisations. This partnership will be discovering false sources and correcting them accordingly, and furthermore ensure to change in size and scale depending on the social media platform’s content moderation needs. It is assumed that the crowd has the advantage over professional fact checkers in terms of scalability. The content moderation practice in its current form requires a large number of human moderators to be able to act and react to the ever changing global social climate.

The provider of human verification systems “reCaptcha” inspired the crowdsourced content moderation to be as much accessible as possible to a large number of people to ensure scalability. Another influencing paper that informed greatly the concept is “Digital Juries: A Civics-Oriented Approach to Platform Governance”. This social experiment provided a good overview on what is possible and not in terms crowdsourced content moderation with Amazon Mechanical Turk workers.
Figure 2. After completing onboarding, participants see information about one of three possible cases. Below information about the case, they see instructions and space for providing input and interaction, depending on the condition. After all inputs are in and a minimum time has passed, the final decision and actions are posted on the right pane. In the jury conditions, the right pane updates in real-time if any juror chooses to change their vote.
This paper helped to position the concept in a more believable narrative as the paper showed evidence on how practical and useful crowdsourced content moderation was. Other influencing elements building the narrative in a near-future world were the design approach “The Future Mundane” and the film “Her” by Spike Jonze. They helped to find time-relevant matters that were used to make the concept feel to be relevant today and in the near-future, such as the Covid-19 pandemic and China-Us political affairs.

There are three ways to make crowdsourced content moderation attractive or present to the general public.

1. Crowdsourced content moderation is marketed as a civic responsibility job for the good and thus is on a voluntary basis, comparable to donating blood to people.
2. Crowdsourced content moderation is offered as an optional choice between other choices to be made, similarly to common interfaces such as privacy and pay walls.
3. Crowdsourced content moderation is mandatory to all platform users, comparable with Youtube users forced to watch five seconds to two minute long commercials before being able to consume content.
PAYWALL SYSTEM – WORK INSTEAD OF PAY FOR YOUR CONTENT
The second option has been selected based on desk research findings about the “Future of Truth and Misinformation Online” stated by the Pew Research Centre:

The information environment will not improve: The problem is human nature. “A small segment of society will find, use and perhaps pay a premium for information from reliable sources. Outside from this group “chaos will reign” and a worsening digital divide will develop”

Moreover, other desk research findings validated the above mentioned scenario: PEW research papers state the following:
“Privacy will be largely a luxury of the rich, who will pay extra for internet service providers, services and perhaps separate networks that protect privacy and security.”, or “The greatest challenge facing society is determining how much privacy and autonomy we are willing to cede in exchange for convenience and features...”. Regarding the keyword “reliable sources”, sources such as the Future Today Institute for example speaks of a “Truth decay in an era of synthetic media.” Synthetic media is nothing far-fetched, rather it has started to manifest in today’s world already. Twitter has updated its platform policies with an additional policy rule called “Synthetic and manipulated media policy”, which comes into effect this year 2020. Scholars raise concerns about the impact of fake news becoming just as real as real news. This paints the picture of a future, where the only credible source of information might be direct personal contact. But is that desirable for society?
How can we turn this dystopian scenario into a more positive, compelling, and actionable alternative?
Three factors have been found to make this future scenario actionable, one focusing more on the human side than the other, and another one advocating a combination of both human and machine factor:

Human-driven scenario:

1. “The information environment will improve, because people will adjust and make things better: crowdsourcing will work to highlight verified facts and block those who propagate lies and propaganda.” – Pew Research Centre

Machine-driven scenario:

1. “The information environment will improve because technology will help label, filter or ban misinformation and thus upgrade the public’s ability to judge the quality and veracity of content.” – Pew Research Centre

Designing technical and human systems:

The rise of "fake news" and the proliferation of doctored narratives that are spread by humans and bots online are challenging publishers and platforms. Those trying to stop the spread of false information are working to design technical and human systems that can weed it out and minimize the ways in which bots and other schemes spread lies and misinformation.

The design concept nurtures from all three factors. Crowdsourced content moderation is understood as a human-centered workflow, and yet, is supported by machines. A mutual effort of humans and machines can enhance the social media content moderation production chain. Currently, human moderators do mainly the content moderation. However, social media platforms have announced to invest their resources into automating content moderation. The concept attempts to keep the human in the loop, by promoting the human contextual and sense-making abilities that are required to successfully do this type of work and furthermore cannot be done by machines, at least not in the near-future. This is particularly true for language related cases, such as hate-speech cases. Whereas cases showing nudity and other types of sexual content can be reliably detected by algorithms through pattern-recognition. Cases relating to manipulated media, for example deepfakes, is still an emerging threat landscape that still needs to be evaluated.
The concept presents deepfakes as a major problem in the near future and tries to hinder the publication and sharing of deepfakes with the fictional service company “Good Loop”. Good Loop operates as an intermediary between social media platforms and newspaper outlets. Good Loop reallocates cases to the right time and right place.
Good Loop presents itself as a plugin for content moderators and as a service for the general public.
Future work

Conversations with people made clear what risks and opportunities were connected to the concept.

Why would social media platforms agree to work together with amateurs / non-specialist instead of hiring professionals such as content moderators and fact-checkers to do content moderation? This concept is not arguing to replace content moderators by amateurs, but see the general public, the collective intelligence as a valuable source of truth based on majority and a complementary type of content moderation platforms can resort to. Another advantage is that crowdsourced content moderation is relatively easier to scale than the traditional content moderation practices. Which becomes important in a constantly growing social media market.

Why is it that the concept puts the focus more on human moderation rather than automated moderation?

Based on the conducted field research it has been observed that automating moderation is still far from being able to take the content moderation job over. Understanding the context of any reported case is still key to solve it. Algorithms are trained for a specific context and need to be re-trained again to be able to understand something beyond the trained context.

How can good work ethics be ensured while people are doing content moderation work in exchange for news content?

The concept speculates that the information environment will not improve in the near-future. Reliable information sources will become expensive and scarce as misinformation and disinformation will pollute all imaginable means of communication. This will result in a more concerning attitude towards information hygiene sought by the general public. Therefore, the general public is willing to do the content moderation work, not to only get the valuable, reliable news content, but also to help mitigate the spread of falsehood.

What countermeasures are provided to fight against misuse of this concept?

Bad actors could intentionally manipulate the service for their own political and personal gain. One way to mitigate this is to implement
a group chat discussion-based content moderation practice, where people discuss together on specific case work. Yet, group chats stretch its limits as it is still challenging to gather and synchronise participants to be in the same chat room.

**What can be done with high error rates and thus more production of fake data?**
The participation of a large number of people is key to see clear patterns in case-related verdicts.

**How to avoid bias from participants and newspapers?**
The service cannot avoid subjectivity on certain cases and themes. But what it can do is to aggregate diverse newspaper outlets together into one overall set of recommendations to acknowledge and see multiple perspectives.

**How will large amounts of recommendations made by the general public be evaluated?**
The recommendations need to be dissected into machine-readable data so that bits and pieces of a recommendation can be categorised for the evaluator / content moderator.

**Who is designing those cases for the general public?**
The content moderator, who wants to be evaluated or needs to be evaluated builds the case for the general public and makes sure that the case is anonymised so that user and privacy rights are not violated. Templates will be offered to speed up the case building process. It is assumed that the content moderator’s work improves when considering the general public’s voice will be taken into account.

**Reflection**

The final design concept is a result of a successful scoping down from a service design project to a Interaction design project. The complexity of the initial brief led to a more humble approach returning back to the emphasis on tools rather than social and mental products. As an Interaction designer I learned to make things instead of thinking and overthinking is also a valid way of moving forward with your design. It just takes a little bit of courage. I also learned that an Interaction designer should not take for granted that people understand your process, instead always start from scratch when explaining your design. Whoever is planning to do a service design project should be conscious about the fact that the design methods used for service design requires the designer / facilitator to be transparent and open with the participants and other stakeholders. Otherwise, the information exchange and communication in general will not run smoothly between designer and potential user
and stakeholders. This project contributes to the interaction design community insofar that it shows that design can be done even in the most isolated places. The concept is a classical interaction design project dealing with screen-based interactions since the theme deals mainly with visuals.