**Autspaces: Content Moderation Considerations**

**Drivers of Harm**

**What are the risks drivers generated by Autpspace’s strategy?**

|  |  |  |  |
| --- | --- | --- | --- |
| **Platform Type** | **Platform Strategy** | **FOE/Protection** | **Users** |
| Platforms primary functionality/affordance is in relation to the opportunity for users to upload their sensory experiences and for these to be made public.  This means that the platform **will** create user generated content (UGC) but is unlikely to facilitate significant interactions between users (UI) as the platform is not designed to facilitate conversations/interactions. | * The platform’s purpose is to collect real-life information about autistic people’s sensory experiences. As a result, there is a need for uploaded information to be an authentic representation of people’s experiences and as truthful as possible. It is clear that this content by its nature could be considered upsetting or offensive to others. * Whilst authenticity and freedom of expression are important for the platform; these will need to be balanced against the need to create platform that is viewed as a safe space by potential users (particularly considering the vulnerability of many of the users). * There are also reputational considerations for the platform; particularly considering the negative attention that an unsafe platform could bring to the community. | * As set-out in the platform strategy section, a balance will need to be found between ensuring freedom of expression and maintaining the platform a safe space. | Autistic people   * Higher to lower functioning * Differing experiences * Differing levels of vulnerability ranging from low to high. * Barriers to entry may be high based on previous negative experiences in relation to research/social platform. * Extremely high negative on the individual resulting from negative experiences on the platform.     Parents of autistic  People   * Parents may prevent their children from using the platform if they do not have sufficient confidence that is a safe space. * Possible translation of pre-conceptions and history (for example differences over techniques/practices) onto the platform.   Others   * Some third-party users may approach the platform with a counter-productive/negative agenda (for example, attempting to abuse/harass autistic people). * Platform may come under significant (potentially negative) media scrutiny as a de-facto representative of the community.   Community Members   * Potentially negative reaction from sections of the autistic community could be directed at and manifest itself on the platform. * Counterproductive arguments and differences within the autistic community could manifest themselves on the platform. |

**Design/Affordances**

*This is an initial stab, I’ll need to see the latest designs to be able to build this section out more, lets catch-up on that!*

**To what extent does Autspace’s platform design impact risk?**

|  |  |
| --- | --- |
| **Opportunities for user interaction/engagement** | **Design of the Autspaces as an exacerbator/facilitator of harm** |
| The austspace’s platform will allow users to input their sensory personal experiences which they can choose to be posted in the public domain, alternatively users can choose that their comments are not made public and are only used for research processes.   * Risks are minimal in the event of users choosing not to make their comments public, the only risk that is reasonably possible is if user experiences identify a real-life risk to harm for the individual or others. Processes will therefore need to be in place to vet non-published comments in a reasonable timeframe. * Risks of harm are increased in the event that comments are posted as users will be able to discover content that they may view as upsetting or offensive in some way. This risk is heightened due to the vulnerability of the Autspace’s user base. * Risks of harm being directed/inflicted at a particular user appear to be low, this is because users will not have the ability to interact/respond to each other’s comments, nor will they have the ability to private message [confirm both points]. * Uploaded comments should be considered as user-generated content and thus their will be a regulated duty of care under upcoming online harms legislation meaning that Autpsace’s will need to ensure effective systems and processes are in place to manage the risk associated with harmful content, ensure protection for children and ensure that illegal content (unlikely) is removed from the platform. | The design of the autspaces platforms and the affordance that it provides users to interact and discover content is both a key driver and also mitigator of user harm. We will therefore need to think very carefully about how the design of the platform can amend the quantum of risk. Without having seen the latest design of the platform, I am unable to offer specific comments; however key design considerations could include   * Does the design of the platform potentially facilitate discussion between users? Comments appearing chronologically on-top of each other could create a de-facto forum environment in which users can effectively respond to previous comments. This could present risks in-terms of abuse, harassment and virtual ‘pile-ons’. * How discoverable would harmful/upsetting content be for users? * Does the platform provide the possibility of filtering out potentially harmful/upsetting content for the most vulnerable users or otherwise warning users of potentially harmful content? * Are there any other opportunities users interaction provided by Autpsace’s (presumably no private messaging?) * Does the platform facilitate interactions that could occur off the platform, either on another site or in real life?   **Next Steps:** *Review of platform design to understand affordances and interactions in greater depth* |

**Operations**

**To what extend does the governance/compliance model either drive or mitigate the risks posed?**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Resourcing** | **User Pressures** | **Moderation approach/philosophy** | **Systems and processes** | **Enforcement** | **Lines of defence/resilience** |
| -The level of resourcing available to Autspace’s will directly impact on the scope, pace and effectiveness of the moderation approach.  -The level of resourcing we will need to be commensurate to the number of users of the platform and the wider moderation challenge (this is an unknown!)  **Questions to consider**  -What level of resourcing is available for moderation, how can this be tailored to demand?  -Who will be the moderators, where will they be sourced from?  -Will there be a technological component to moderation?  -We will need to anticipate the likely level of demand to determine commensurate resourcing and process requirements. We will need to ensure resilience and flex is built-in to this process to enable | -As mentioned, the number of users/comments will determine the risk and moderation challenge with higher traffic naturally leading to greater pressures.  -The exact pressures and the ways in which they will manifest into behaviours and moderation strategies is an unknown.  **Action-** We should attempt to map user numbers to moderation requirements; ie. X number of users = x number of hours/moderators etc. | As previously discussed, the overarching strategic approach to moderation particularly in relation to freedom of expression will impact on moderation requirements, resourcing and risk management procedures.  -If we assume that all comments are pre-moderated rather than immediately uploaded (and reactively responded to) (is this correct?), then we will need to ensure there is commensurate moderator capacity to meet this challenge.  **Action**  -To define our moderation philosophy/tolerance for risk.  -To confirm if we will pre-moderating experiences or taking a reactive moderation approach. | [Under online harms legislation](https://www.gov.uk/government/consultations/online-harms-white-paper/outcome/online-harms-white-paper-full-government-response), proportionate and systems and processes will need to be in place to manage the threat of harm. The two key areas of harm relevant for Autspace’s are:  -Illegal content  -Content that is legal but harmful to children.  Austpace’s will not be required to take steps against legal but harmful content for adults; In practice the biggest risks for this platform are probably most relevant for the platform so we should also ensure systems and processes take into account this harm also.  **Action:** Consider proportionate systems and processes that Autspace’s can put in place to manage the stated harms. | -T&C’s will need to be enforced in a clear consistent way to enable clear standards of behaviour to be understood as well as the consequences of misbehaviour.  -Because of the participatory nature of the platform, users must be clearly informed if action has been taken against content and why.  -There must also be the opportunity to challenge content takedown decisions (user redress)  -Steps taken in relation to users who are perpetual T&C violators should be a): sensitive to their needs, perspectives and experiences, b): be proportionate to the violation and c): should escalate over time under the assumption that persistent violators are unwilling or unable to engage.  -To ensure representation, perhaps there should be a moderation appeals panel comprised with people from the community?  **Action-** Consider how the enforcement and redress regime should be structure to be as clear, robust and effective as possible. | -We will need to ensure that effective lines of defence are in place to ensure resilience in the event of prior steps and processes failing allowing harmful content to be published, steps could include:  -A second opinion or clearance the chief moderator before a platform goes online  -Moderation panel (incl members of the community)?  -Technological solutions? |

**Proposed Next Steps [Very Preliminary and indicative- just a starter for 10]**

|  |
| --- |
| *1: Literature review: Moderation and T&C’s precedent and best practice [focussing on participatory platforms]*  *2: In-depth risk/review mapping of risks associated with Autspace’s*  *3: Development of T&C’s off the back of groups session and the T&C’s presented [quali analysis]. Users will need to be considered during this process, if time we could considered interviews or focus groups (time permitting)*  *4: Once T&C’s have been considered they will need to be tested with the community and stress tested against scenarios.*  *5: The overarching moderation approach will need to be operationalised.* |