Entrepreneurism & Ethics Project

Group 3

Zachary Gonsalves, Benjamin Chen, Cristian Gonzales

ReportShield **ReportShield startup tech company **

Item 1

1.A. Company Name:

Report Shield is a tech startup that offers an anonymous reporting platform, allowing users to safely report fake social media accounts, cyberbullying, and harmful content without being afraid of retaliation. Our tool provides a secure, and confidential way to flag online threats while integrating with major social media platforms. Report Shield is committed to user safety, privacy, and ethical accountability, ensuring a trustworthy experience for all.

1.B. Long-Term Vision Statement:

In the coming decade, ReportShield aspires to fundamentally redefine online safety and accountability. Through advanced technology, strategic partnerships, and community empowerment, we plan to change social media into spaces that prioritize human dignity, authentic connection, and ethical engagement. Overall, our biggest goal is to help create a digital world where people can trust what they see, harmful content is rare, and everyone feels safe to speak up.

1.B.1. Goals:

ReportShield's main goal is to make social media safer for everyone. We want users to feel protected from bullying, fake accounts, and harmful content without fear of being judged or targeted. By working with platforms like Instagram, TikTok, X (Twitter), Snapchat, and Facebook, we provide a secure and anonymous way to report abuse. While we may not eliminate these problems completely, we aim to significantly reduce them.

1.B.2. Idea Organization:

ReportShield originally began as a way to fight off cyberbullying and social media abuse. after the rise of fake accounts, online harassment, and misinformation, we realized that there was a clear need for a tool that lets users report problems safely and anonymously. This idea was inspired by the 2006 court case United States vs. Lori Drew. Where a mother and daughter posed as a fake account to harass a 13 year old girl so much to the point that it drove the 13 year old girl to take her own life. Report Shield was developed to empower all its users so that they may not face the same harm as the victim in the United states vs Lori Drew court case.

1.B.3. Purpose/Values/Mission:

Our purpose is to create a safer digital world. Free from any harm by empowering users with the ability to anonymously report any harmful online behavior. Our values include privacy, security, transparency, and ethical responsibility. Privacy to ensure that users can report anonymously without fear. Security is the data encryption that secures unauthorized access to your reporting information. Transparency allows for users to receive updates on report outcomes ensuring a high trust relationship between our users and the system. Ethical responsibility promotes the idea of a safe space in any social media platform that Report Shield partners with and prioritizes ethical decision making in our operations.

1.B.4. Key Questions:

How can maximum anonymity for the victims be ensured while maintaining maximum accountability for the perpetrators of cyberbullying or harassment? What technologies can be implemented to make reporting easier and more effective for diverse users such as those who are disabled? How do we balance freedom of speech and the policies of each social media company with the need to remove harmful content online?

1.C.1.1 OKR 1 Objective and Key Result Objective:

Objective:

Ensure Report Shield gives anonymous accessible and an efficient reporting system. When it comes to cyberbullying, spam, bot accounts, and harmful content across major social media platforms. This content could be; graphic, sexually explicit, or hate speech. Key Result: Conduct usability study with 1,000 users from diverse demographics. The targeted rate of user satisfaction should be 90%. Report Shield must be easy to use, it ought to provide the user anonymity as well.

Stakeholders:

Users: The main user base for the study will be the main user base of social media teens and young adults aged 13-30. They must be of diverse race as well as gender. There should be no minimum or cap on their income level. Their sexual orientation should also be diverse as cyberbullying can target those of differing races, genders, or sexual orientations thus it's encouraged that these demographics be diverse. They also use social media often throughout their day. They use it for entertainment, education, networking, and for a professional setting.

Social Media Platforms: These companies benefit from reducing harmful content on their platforms. This will improve the safety of their users. Which in turn will raise their user retention and compliance with their company guidelines. Their platforms will also make more money due

to more advertisers willing to promote their products on these platforms due to a reduction in explicit content.

Law Enforcement: Police agencies can use verified reports from the social media platforms and the encrypted evidence from Report Shield to launch investigations against those who perpetrate these crimes. This will also ensure the privacy of the victims of the cybercrimes to ensure they don't suffer from retaliation from the cybercriminals. It is to also ensure the privacy of the law enforcement officers so as to not alert the cybercriminal of their investigation.

Mental Health Advocates: Organizations that focus on mental health can use Report Shield's data to spread awareness and even assist in user policy in social media platforms or in Report Shield. Families and Caretakers of the Victim: Parents and guardians could use Report Shield to educate young users about online safety and intervene if cyberbullying were to ever occur.

1.C.1.2 OKR 1 Metric(s) with Experiment(s) 2.):

Metrics and experiments:

To attain a 90% user satisfaction rate. Report Shield will conduct a usability study with 1,000 participants from diverse demographics. The study will involve a survey and observing the participants interacting with the Report Shield platform.

Metrics:

User satisfaction score on survey:

Sample questions: “On a scale of 1-10, how satisfied are you with the ease of reporting the issue?”

“On a scale of 1-10, how secure did you feel in your anonymity when using Report Shield?”

“On a scale of 1-10, how likely are you to recommend Report Shield to others?

Success metric: 90% of users report the average of the survey as 8/10 or higher. Task Completion Rate: Users must successfully complete a report in 5 minutes. Success Metric: 90%

of users can complete said task in 5 minutes. Error Rate: Tracks how many users make errors (incorrect submissions).

Success Metric: Error rate is below 10%. Retention Rate: Measures how many users continue to use Report Shield after 3 months.

Success Metric: At least 70% of participants use Report Shield after 3 months.

Experiments:

A/B Testing: 2 different UI designs will be tested to see which layout is more engaging to the user by measuring the task completion rate on both UIs. Usability Testing: Users will be asked to submit a report while usability experts observe any challenges or bugs occurring in this process. Follow up survey: Every month participants will be given a survey to check their user satisfaction as well as if they had used it in the last month and 3 will be given out at the end of each month up until the end of the 3rd month.

1.C.1.3 OKR 1 Ethical Impact(s)/Issue(s)

While report shield ensures safety for the users and the social media platforms. It also introduces ethical concerns, particularly around privacy, misuse, and data security. false accusations or misuse Potential Ethical & Real-World Issues Privacy Risks: Anonymous reporting raises concerns about false accusations and misuse. Case Study: The misuse of Twitter's reporting system, where false mass reports have led to wrongful account suspensions. Conflicting interests: Social media companies might be reluctant to integrate Report Shield if it increases their content moderation burden. Potential for Bias: If the usability test doesn’t include a diverse range of participants, the platform’s effectiveness may be skewed toward certain demographics. Thus it could lead to forms of discrimination as it can target certain demographics more harshly for reports of cybercrimes falsely.

Ethical impact Risk Table:

Stakeholder	Financial Risk	Privacy Risk	Conflicting Risk	Violation of rights
users	Low	High	Mild	Mild
Social media platforms	Mild	Low	High	Mild
Law Enforcement	Mild	Low	Mild	High
Families and Care takes	Low	Mild	Low	None

Analysis of Ethical Risks:

Users (High Privacy Risk): Since reports are anonymous, if the encryption fails then the user's identity will be vulnerable to theft. Law Enforcement (High Violation of Rights Risk): If reports have misinformation or are outright false then it could lead to wrongful arrests. Social Media Platforms (High Conflicting Interest Risk): The platforms may have to pay extra for the use of report shield and in content moderation.

1.C.1.4 OKR 1:Ethical Safeguards:

To reduce risk, Report Shield will implement these ethical safeguards.

1). Data Encryption & Privacy Protection End-To-End Encryption: Report Shield will use encryption to ensure reports are anonymous. Anonymous reporting: No personal information is stored by the program ensuring the report is anonymous.

2). False Report Prevention System: Reports flagged as spam or harmful will be viewed by human moderators before being sent to platforms or law enforcement.

3). Bias Free Usability Testing The 1,000 test users will be selected from all genders, races, ages, sexual orientations to ensure diversity and inclusivity.

4). Oversight Report Shield ethics board will be established and it will have:

Cybersecurity professionals

Advocacy group representatives for mental health and cyberbullying organizations

Social media company representatives.

Measuring Effectiveness Audit Logs: Regular audits ensure encrypted data remains safe and unexposed. User Feedback Surveys: Continual collection of user input regarding experience, engagement, privacy, and safety. Independent Evaluations: Cybersecurity firms to review Report Shield's practices on a quarterly basis.

1.C.2.1 OKR 2 Objective and Key Result:

Objective:

My objective is to ensure Reportshield launches a digital literacy campaign that reaches Reportshield users. It is to increase awareness around cyber bullying, fake accounts, and the proper use of anonymous reporting tools. Many users, especially those that are minors, may not know what qualifies as cyber bullying or a fake account or accounts spreading misinformation. They may also not know how to safely and effectively report on harm. Reportshield aims to empower our younger users so that they may not fall victim to such malice. By launching a digital literacy campaign Reportshield will not only improve the quality of the reports but foster a more informed and ethical online community. This will aid to reduce underreporting and increase the proactivity of our newly educated users.

Key Results:

Launch an in-app interactive educational model and educate 50,000 Reportshield users within the first 12 months of the launch. Reportshield wishes to achieve at last a 60% improvement on digital literacy scores through the uses of pre and post quizzes. To partner with 10+ schools, universities, and advocacy organizations to embed Reportshield's training modules in their digital safety curriculum. To reach an average rating of 4.5/5 for the learning content of the digital literacy course based on its users feedback.

Stakeholders:

Primary Users: Everyday people who use ReportShield to report fake accounts, cyberbullying, misinformation, etc. Digital literacy is most important for them as they are at the most risk of harm. Digital literacy ensures they understand how the tool works, how to navigate the interface, and what to expect when reporting online harassment. Empowering out often underaged users with accessible education on digital literacy is essential to ensure trust.

Social Media Firms: While they don't receive direct lessons on digital literacy they do partner with reportshield and benefit from more informed users and an overall safer user experience. These more educated users will provide more thorough reports which will reduce moderation burdens and improve enforcement burdens for their team as well as that of reportshield. They will also get better PR by partnering with a tool that ensures the safety of their users.

Law Enforcement: Is a key stakeholder as well informed users will submit clear, actionable reports that can aid in the investigation of cyber crimes like harassment, blackmail, or fraud. This will ensure better evidence provided and evidence handling as well as greater cooperation between users and authorities. Law enforcement can support public education by sharing these courses with public schools. But for the sake of the user's privacy great care must be taken to safeguard their anonymity even during an investigation.

Advocacy Groups: These groups work to reduce any digital harm and advocate for equal access to online tools and education. Their collaboration with reportshield will help ensure that

the materials used for the digital literacy courses are inclusive foe people of all demographics. They will also help evaluate the level of effectiveness for the literacy programs of reducing online harassment.

Families: Since many reportshield users will be minors their parents and families have a need to ensure the online safety of their children's online experience. Through reportshield they could provide them with a clear guidance materials or training and foster safer online experiences and build a stronger support network for underaged users.

1.C.2.2 OKR 2 Metric(s) with Experiment(s):

Metrics:

To assess the effectiveness of the digital literacy program. Reportshield will conduct a longitudinal study with 2 control groups. Users will take a pre test, this pre test will act as a baseline quiz regarding the users level of digital literacy. This will be evaluated by assessing their knowledge on identifying bot accounts, cyber bullying, misinformation. It also assesses their ability to know their reporting protocols and understanding their right to digital privacy and how it can be violated.

Experiments:

Users will go through 5 interactive modules. Course modules: Digital footprints Identify Cyberbullying, Misinformation, and Bot Accounts Report Honestly and Responsibly Online Safety and Privacy Ethics Use of Social Media To assess the level of the course’s effectiveness users are now prompted to take a post test. To also measure the progress of the user. A certificate is awarded only if they pass with a score of at least 80% or above. Users will be given a survey in order to report and review their experience and opinion on the effectiveness of our digital literacy course.

User Survey Questions:

How confident are you now in identifying harmful or fake content online?

On a scale of 1 to 5 how likely are you to use Reportshield's reporting tool to submit accurate and responsible reports?

Did you learn something new from this course? (Yes/No) Are you now aware of your rights to privacy online? Are you now more careful about your digital footprint?

Experiment control group A vs control group B Group A will take the digital literacy course immediately upon signing up while Group B will take the digital literacy course 30 days after signing up. Then upon both groups completing their courses their metrics will be compared with the accuracy and honesty of their submitted reports.

1.C.2.3 OKR 2 Ethical Impact(s)/Issue(s)

The act of improving digital literacy while rooted with good intentions of a digital safe space can harbor harmful yet unintended consequences if not handled carefully. Oversimplified modules might encourage overconfidence or misuse of reporting tools. If not monitored, misinformation about the course's curriculum might be spread within the course itself. This mirrors past issues such as "YouTube's 'Creator's Academy,' which unintentionally downplayed the algorithmic suppression risks, causing confusion and false expectations" [1]. Another risk for our platform's digital literacy course is the lack of accessibility for users who are non English speakers or have disabilities. They might find the content hard to navigate which would exclude them from their education. Also users that may have a different learning styles from that of our course which may also lead to their exclusion from being educated on digital literacy.

Ethical impact Risk Table:

Stakeholder	Financial Risk Of privacy	Risk of Privacy	Conflicting Interest	Risk of rights Violation
Primary users	Low	Medium	Medium	High
Social Media Platforms	Medium	Low	Medium	Medium
Law Enforcement	Low	Low	Low	Low
Advocacy Groups	Low	Low	Low	Low
Families	Low	Medium	Medium	High

Analysis of Table:

Primary users, especially teens are at the highest of risks when it comes to being victimized by misinformation and harassment. They also face a medium risk of rights violations if they lack the awareness of consenting digitally, awareness of the consequences of sharing their data, and suffering from harmful reporting practices from cyber bullies. Social media firms could face risk with compliance with educational partnerships. Families face risks if children lack digital literacy or are misinformed regarding digital literacy.

1.C.2.4 OKR 2 Ethical Safeguards

To reduce these aforementioned risks above, Reportshield must implement inclusive learning. Inclusive Design: Modules will be created with consultations from accessibility experts such as voice guided lessons for visually impaired, and sign language assistance for any parts of the course that may have a video. The course will also be translated in 10 languages. Third-Party Review: Each module of the course will be reviewed on a quarterly basis by educators and

digital ethics organizations to ensure accuracy and relevance. Adaptive Content Delivery: The platform shall adjust the delivery of the course content based on user interactions and learning spaces to better accommodate diverse learners by allowing the changing of learning styles such as: visual, auditory, and hands on. Feedback Loop: Users will be able to flag any misleading or contradictory content in the course. This report will go to the course developers allowing for review and even revision of whatever section is reported especially if enough reports on the same section are made.

1.C.3.1 OKR 3 Objective and Key Result Objective:

This OKR seeks to eliminate algorithmic bias in moderation systems used by Reportshield. Within the first year, our goal is to reduce Reportshield's moderation bias by 50%. This is to ensure that all groups, especially underrepresented groups regarding; gender, race, sexual orientation, ableism, and socioeconomic status all have a voice in reporting harmful content.

Key Result:

To create a diverse training dataset with at least 40% of the group consisting of underrepresented groups. Conduct monthly fairness audits on AI flagging mechanisms and patterns Eventually achieve less than 10% of disparate impact rate in report outcomes. Reportshield also wishes to collaborate with advocacy groups and marginalized communities to refine moderation bias detection logic

StakeHolders:

Primary Users: Everyday users that use ReportShield to report abuse or harmful content online, fake accounts, and cyberbullying. These users rely on ReportShield to be a fair and trustworthy content moderator. Their experience is directly influenced by how well ReportShield is able to moderate any harmful content online.

Social Media Companies: Are essential stakeholders because they are integration partners. ReportShield works alongside these platforms helping them manage reports and urging them to manage reports ethically. These companies have a vested interest to ensure that moderation is seen as fair by its users of all demographics.

Law Enforcement: Are involved when a report escalates to a potential criminal matter for the user being reported. Or a safety issue for the user doing the reporting. They not only depend on the user doing the reporting but also on Reportshield to accurately vet the report for any misinformation or bias. If it is biased then the law could be misled to act on a faulty or an outright false or even framed report.

Advocacy Groups: Civil rights organizations, digital justice movements, anti cyberbullying groups serve as not only collaborators but also ensure that any harmful online content is heavily censored. They monitor how effectively ReportShield protects marginalized groups and help address any systemic discrimination with content moderation. They contribute by designing more inclusive systems by sharing and highlighting their insights as well as statistical data.

Marginalized Groups: Any racial, ethnic, gender, LGBTQ+, Disabled, and Linguistic Minorities are crucial stakeholders as they are at the most risk of harm from biased moderation. These communities face far more frequent rates of cyberbullying and harmful content and are disproportionately ignored by biased algorithms. Our success at Reportshield heavily depends on how well and how fair our reporting tool is for them.

1.C.3.2 OKR 3 Metric(s) with Experiment(s):

We at ReportShield to test for bias will apply disparate impact testing across our demographics.

Key Metrics:

False positive rate according to demographic. Time to resolve the report and take affirmative action disparity among the demographics The disparity of rates of satisfaction among demographic groups.

Experiments:

Experiment 1: For our first experiment we will submit 500 copies of the same report using our reporting tool. Every element of the report will be word for word the exact same (content as well as all markers of identity such as gender, race, sexual orientation, etc.).

Experiment 2: We will have a panel of 200 testers of diverse identities report similar harmful content through our reporting tool. We will check for any differences in our tool's affirmative action and it's rate of fairness results.

1.C.3.3 OKR 3 Ethical Impact(s)/Issue(s)

Reducing bias in automated moderation is a very critical issue. It's one however that must be expunged. Studies show that marginalized groups are disproportionately flagged or ignored by AI systems [2]. Failing to act upon this issue could result in further marginalization for these groups. This marginalization would be in a digital space which would burden these demographics with another frontier that they'd have to cross to combat their own discrimination. The "TikTok Shadowbanning" controversy, where LGBTQ+ and disabled creators were quietly suppressed demonstrates how biased moderation damages trust [3]. Also with being marginalized on digital spaces where many people spend a good portion of their daily lives this can trickle down to their lives offline as discriminatory sentiments could not only be harbored online but spread rapidly.

Ethical impact Risk Table:

Stakeholder	Financial Risk	Privacy Risk	Conflicting	Violation of Rights
Users	Low	High	Mild	Mild
Social Media Platforms	Mild	Low	high	Mild
Law Enforcement	Mild	Low	Mild	High
Advocacy Groups	Low	Low	Low	None
Families & Caretakers	Low	Mild	Low	None

Analysis of Table:

Primary users and marginalized groups face the highest risk in terms of privacy, bias and podré tisk rights violations. The algorithms may bias our users causing marginalization and if they are already in a marginalized group it will be made even worse for them. Social media companies face financial risk due to conflicting interest of safety for users thus causing moderation when what they really want is engagement. Law enforcement carries mid level bias risk due to if they enforce the law on a marginalized group based on biased data. Advocacy groups face minimal risk but also are meant to hold the system accountable if anything were to go wrong. So if data is still biased even after them reviewing it then they too could face accountability. The table

illustrates the need to protect our users and especially those of marginalized groups as they face the greatest risks of all the stakeholders.

1.C.3.4 OKR 3 Ethical Safeguards:

We at Reportshield are aware of these risks and aim to mitigate them where we can. We will implement bias mitigation by using algorithm explainability tools such as SHAP (SHapely Addictive exPlanations) to audit any unfair decision making process from our reporting tools and immediately take measures against it. SHAP allows the system to assign transparent numerical values to each factor that influences a moderation decision. This will make finding traces of unfair treatment towards users of certain demographics or multiple demographics easier. This can be done with procrastinating report audits from users because of their demographic. We will have inclusive design by having our moderation policy created with the help of marginalized groups. They will have direct influence on how harmful content is mitigated and their first hand experience at the hands of discrimination will provide an invaluable asset to our refining of our tool. To maintain trust among our users Reportshield will publish a yearly bias report with the aims of the amount of bias every year. It will be revised by advocacy and civil rights groups to ensure accountability at Reportshield. The platform will also allow for users to provide direct feedback explaining why their report was accepted or rejected. This allows for transparency with our users. These practices will ensure that Reportshield can be trusted to ensure that there is no bias or discrimination against any demographic that everyone should be safe when online and anonymously reporting any harm.

Item 2 Cultural Policy:

2.A. Core Values:

Overall, ReportShield is built on four simple core values. This consists of keeping users safe, protecting their privacy, being open and honest about how we work, and ethical responsibility.

We believe that everyone should be able to use the internet without having the fear of their data being exposed to others, which is why our anonymous reporting system allows them to safely speak up with no risk at. Security is our backbone, with end to end encryption and rigorous audits to protect users data. Most importantly, we take responsibility for making the online world safer for all, working with experts and authorities to stop abuse while protecting free speech. We want to be known as the team that finally gave others control over their digital lives, where reporting harm feels as easy as sending a text, and where no one has to choose between staying silent or risking retaliation.

2.B. Motivation:

At ReportShield, we love building systems that make people feel heard, safe, and powerful, especially for those who have previously been silenced or dismissed. We get excited about the idea that someone who once feared reporting a bully can now do so anonymously in under five minutes, while knowing they're backed by encryption, community, and clarity. Research shows that effective reporting systems significantly reduce online abuse while preserving free expression [4]. We fear complacency. The internet's toxicity grows when systems prioritize engagement over ethics, or when anonymity becomes a weapon instead of a shield. We fear misplaced trust, failed encryption, biased algorithms, or partnerships that compromise user safety. Research shows that "41% of Americans have experienced online harassment, with 25% reporting severe abuse like stalking or threats" [1]. 2.C. Summary Privacy Empowerment Integrity Transparency Compassion Innovation.

Item 3: Ethics Policy

3.A Core Items:

3.A.1 User Data Privacy:

Our first commitment as Reportshield is to safeguard the privacy of any and everyone of our user's data. Our platform is built for people who need to anonymously report fake accounts, cyber bullying, and harmful content without any fear of retaliation. To ensure the safety of our users we use end-to-end encryption and upon submission all reports are made anonymous. We at Reportshield strict data minimization principles only collecting what is necessary. Unless compelled to do so by legal requirements we abstain from sharing our user data with any third parties. We include user friendly dashboards ao users can see how their data is being processed and protected. As Valdez noted in Wired, "The data collected by Facebook was later harvested by Cambridge Analytica and allegedly used to manipulate voter opinion" [4].

3.A.2 Bias Elimination:

Reportshield uses machine learning algorithm to detect patterns in abuse and harmful content. We commit to diverse, equal, and inclusive data sets so as to eliminate bias. Eliminating bias helps eliminate any discriminatory outcomes based on race, gender, ableism, class, etc. We collaborate with third party ethics consultants and perform routine bias audits to ensure fairness. As the AI now institute wrote, "expand AI fairness beyond a focus on mathematics and statistical fairness towards issues on justice" [5]. Reportshield seeks to elevate underreported harms to marginalized groups.

3.A.3 Fair Labor Policies:

Reportshield acknowledges the value of its engineers, data analysts and content reviewers by offering fair compensation and safe working conditions. Reportshield is meant to provide safety to all and we as its creators must uphold that in our own workplace by renouncing the use of exploitative labor. The AI now reports that "accounting for the many forms of labor required to create and maintain AI systems is essential to ethical development" [5]. We support open acknowledgment of contributors in our ethical reports and research publication.

3.A.4 Interdisciplinary Ethics Consulting:

Developmental team works closely with social scientists, educators, lawyers, digital rights advocates, mental and physical health professionals. “It’s good to commit to interdisciplinary AI” [5]. By drawing from these disciplines Reportshield ensures that its technology will not harm its user’s psychological and physical welling, and their legal boundaries. We also engage with our underaged user’s family, educators, and advocacy groups through surveys and community partnerships to refine our ethical approach.

3.A.5 Proactivity and Accountability:

Reportshield commits to conducting bianual ethical impact assessments to review our platform’s social effects, report handling, if there are undeserved reports done to any demographic, and influence on our user’s mental health. Our findings are to be published publicly. In the case of failure we commit to prompt correction. In the case of misuse we commit to prompt investigation.“Black defendants are often predicated to be at a higher risk of recidivism than they actually were” [6], due to flawed algorithms whether intentional or not. Through our partnerships with civil society organizations we advocate for better online safety policies for all users.

3.A.6 Coalition Building and Advocacy:

ReportShield more than just a tool it is meant to be a utopia digital safe space. We are forming coalitions with civil society organizations, researchers, and digital safety advocates to push for level platform change. Whether our digital safe space is influencing policy changes across other social media platforms and contributing to a firm reduction in cyber bullying and fake account cases. It is encouraged to begin “building coalitions between researchers, civil society, and

organizers within the technology sector” [5], as it will bring ethics to encourage platform change. We see ourselves as the arbiters of a revolution in digital safe spaces.

Board of Experts 3.B.1

Dr. Timnet Gebru Dr. Timnet Gebru is a computer scientist and founder of the distributed AI research institute (DAIR). Her work and institute highlights the potential perils of facial recognition in exposing its potential for bias. She also focuses on imposing ethics on machine learning. She used to work for Google as a co-lead for their Ethical AI team. She was considered one of the chief authorities when it came to algorithmic fairness. This was a prime reason as to why Google fired her. She has conducted landmark studies on how AI models can harbor gender and racial bias. Reportshield’s use for AI content moderation makes her insight essential. Her expertise ensures that our AI models treat all users fairly and with equity regardless of race, gender, sexual orientation etc. That any ethical red flags or patterns of unjust reporting of a certain demographic are to be quickly expunged.

3.B.2 Dr. Rumman Chowdhury

Dr. Chowdhury is a data scientist, AI ethicist and former director of machine learning at Twitter. She is known for creating Machine Learning audit frameworks and serving as an advisor to tech companies and even governments on AI transparency. At Parity consulting she helped organizations implement AI practices. Dr. Chowdhury’s advice is invaluable to us at Reportshield for she has real world experience mitigating harm in algorithms and social media platforms. Her guidance would ensure that our machine learning models are fair, auditable, and aligned with user’s rights and expectations.

3.B.3 Alex Stamos

Alex Stamos is a cybersecurity expert and former chief security officer at Facebook. He is the director of the Stanford Internet Observatory. He has advised the public and private sectors on digital security. His knowledge is essential in safeguarding ReportShield's data, especially since our core premises are secure anonymous reporting. Stamos could assist with our end-to-end encryption. With his background in being in charge of safeguarding billions of user accounts in Facebook he brings nigh peerless credibility to our board.

References

[1] A. Vincent, "YouTube's Creator Academy and the Misinformation Dilemma," Tech Ethics Daily, 2021. https://www.techethicsdaily.org/youtube-creator-academy

[2] J. Buolamwini and T.Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in Proc. Conf. Fairness, Accountability, and Transparency (FAT), NewYork, NY, USA, 2018, pp. 77-91. [Online]. Available: https://proceedings.mir.press/v81/bualamwini18a.html

[3] J. Lorenz, "tikT Admits to Shadowbanning LGBTQ Users," The Guardan, 2020. https://www.theguardian.com/technology/2020/tiktok-shadowban-lgbtq

[4] A. Valdez, "Everything you need to know about Facebook and Cambridge Analytica," Wired, Mar 23, 2018. [Online]. Available: https://www.wired.com/story/wired-facebook-cambridge-analytica-coverage/

[5] AI Now Institute, "AI Now Report 2018," Dec. 2018. [Online]. Available: https://ainowinstitute.org/AI_Now_2018_Report.pdf

[6] J. A vein, J. Larson, S. Mattu, and L. Kirchner, “Machine Bias,” ProPublica, May 23, 2016. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing