# 1: Ethical Business Plan

## 1.A. Company Name

The name of our company is TrueScan AI. The reason we chose this as our company name is because our goal is to use AI as a fact-checking tool and scan news articles to “find the truth” hence the name TrueScan AI.

## 1.B Long-Term Vision Statement

### 1.B.1 Goals

Our main goal for TrueScan AI is to provide a resource for users to input an article where the AI scans it and lets the user know which parts of the article are factually correct and which parts are inaccurate or considered “fake news”. TrueScan will also provide advice and information to the user on what to look for when reading articles to determine whether they’re accurate or not.

### 1.B.2 Idea Origination

Our group initially thought up a list of technologic ethical issues that modern society faces today. The issue of misinformation when it comes to the spread of news through social media is something that we decided to focus on. We then brainstormed together what kinds of solutions could be made to deal with misinformation where we finalized on AI that is specifically curated to read through an online news article and fact check it.

### 1.B.3 Purposes/Values/Mission

The purpose of TrueScan AI is to provide users with a reliable and trustworthy AI tool that highlights the factual accuracy of the news media they read on social media platforms. The main values that we follow at this company are integrity, transparency, and accountability. These are the three main pillars that our product, TrueScan AI, stands on so that users can trust our results without worrying about whether the information the AI states also has to be fact checked.

### 1.B.4 Key Questions

The main question that everyone would most likely ask about our startup, is why do we need to exist as a company? And the answer to that is simple: We live in a time where misinformation has been at an all time high with no safe guards that let people know how accurate the information they're reading from is. TruescanAI provides a solution to combat misinformation and to educate people on what to look for when reading news articles.

## 1.C Strategy with Ethical Impacts and Ethical Safeguards

### 1.C.1 OKR 1

#### 1.C.1.1 Objective and Key Result

**Objective**

An objective that my team’s company is trying to achieve is to establish a cooperative integration of TrueScanAI with major social media platforms to embed its fact-checking features into their platforms.

## Key Result

We are planning that by the end of the first fiscal year, we want to secure and implement integration of TrueScanAI into at least two major social media platforms (Reddit, Facebook, Twitter/X). Having this integration will allow users of these platforms to have direct access to our fact-checking tools whenever they come across an online news article they find suspicious. We are rolling out integration onto a platform once every five to six months.

We chose this objective because it’s been shown that the majority of US Americans get their news from social media as their primary source. According to research done by Pew Research Center, [1] about 53% of US Adults state that they get their news from social media. Popular social media websites such as Twitter/X, Facebook, Reddit, and Instagram are places where users not only receive their news but also where these users discuss with one another about news, making these sites the perfect places for misinformation to spread like wildfire. By integrating TrueScanAI into these sites, we can reach users at the exact moments they come across news articles. This OKR is very important to achieve for us as a company because it allows us to distribute our product, TrueScanAI, to where it’s most needed, where instead of relying on users to go to our website and input a news article, we can bring TrueScanAI’s features and capabilities into social apps that most people already use.

## Stakeholders

There are three stakeholders for our business, there our customers, the end-users, who are the ones that directly utilize our product and its related services. The social media companies are our second stakeholders, who we will negotiate with in successfully integrating TrueScanAI into the platforms. And the third stakeholder is our TrueScanAI company. We will provide the AI tool and its related features to both the customers and to the social media companies.

Our customers are those of the general public who use platforms like Twitter/X, Reddit, Facebook, and Instagram to get their news and see the happenings of the world. They are the ones who will be benefited by the integration. The demographics of customers are going to be very wide, it will include consumers of teenagers and adults of all ages who rely on social media platforms to get their news and interests. They'll people with varied income and education levels, but those with less formal education or low income might have fewer alternative news sources than others, making TrueScanAI to be extremely useful for them. Our target users span all genders and racial backgrounds since misinformation is something that targets people of all backgrounds and not just specific ones. Because of this, TrueScanAI can help those of minority and majority communities alike. The relationship between our company and our users is a relationship of trust. When our users use our integration in their social apps, there must be some level of trust that when TrueScanAI states whether something is accurate or not, they can believe what it says. Otherwise, if our tool shows any form of bias or it fails in its fact-checking, that could alienate users from using TrueScanAI.

The social media companies like Facebook, Twitter/X, etc. are to be our most crucial stakeholders since we need their cooperation for the integration of TrueScanAI. Without their agreement, embedding our AI onto the platforms would be very difficult or impossible to implement. It's in the interest of these social media companies of this OKR because of external pressure and internal goals. For example, the social media platform [2]X is considered to be one of the biggest sources of fake news and disinformation according to the European Union. So these platforms have come under scrutiny for false news to spread, integration of TrueScanAI can help these platforms show their commitment to accuracy and user safety. However, these social media platforms may have their interests conflicted, the reason being that these social networks rely on user engagement for their platforms to thrive and generate income, and since fake news drive clicks, the TrueScanAI integration may drive down user engagement. The relationship we have with these companies is one of cooperation, where we imagine negotiating pilot programs with willing companies. Perhaps Twitter/X might be our first integration (for instance, Twitter already has a Community Notes feature where they might integrate our AI to enhance it).

As the startup behind TrueScanAI, our company is directly invested in this OKR. Having TrueScanAI integration in social media platforms will give us wider user adoption and will benefit our revenue model. The TrueScanAI team will be responsible for building integration interfaces to handle potentially thousands of requests from platform users. We are also responsible for maintaining the accuracy of the AI. If our tool makes a major mistake on a big platform, it could become a public relations issue for both us and the social network. Thus, this OKR forces our company to meet high standards. We have to manage stakeholder relationships carefully, keeping the platforms happy by aligning with their policies and up time needs, while also keeping end-users satisfied by providing correct and unbiased results. The relationship between our company and the social media companies will likely be formal partnerships, with users, it’s more of a service-provider relationship.

### 1.C.1.2 Metric(s) with Experimentation

Success for this OKR will be measured by achieving at least two live platform integrations within the first year at a rate of one partnership with a social media platform every five to six months, alongside having strong user engagement and satisfaction. A simple metric where we will keep count of the number of integrations that we have achieved. As an experiment, we might approach a spectrum of platforms. One large (e.g. Twitter) and one medium sized to compare negotiation processes. By mid year, if progress is slow with major players, we will pivot to integrating with a smaller platform or even a browser extension for users as an interim step. This can be seen as an experimentation in strategy. For example, we plan a pilot integration with a volunteer user group on Reddit via a browser plugin. We'd recruit, say, 100 Reddit users to use a TrueScanAI plugin that auto-checks any news links on their feed. We'd observe how that pilot goes. If successful, that data will be used to persuade Reddit's management for an official integration.

Another metric we plan to measure is the active user engagement. The number of users that are using TrueScanAI via the integrated platforms is something we will measure the number of active users engaging the TrueScanAI by implementing analytics tracking within each integration. For example, once TrueScanAI is live on platform X, we will track events such as "user clicked TrueScanAI on article" or "TrueScanAI result viewed." This data will feed into dashboards of daily and monthly active usage.

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

Pursuing this OKR brings several ethical implications and potential issues. When dealing with user data, automated judgments on online content, and influential social media platforms we have to consider the possible harm that could happen.

One of the main concerns is Privacy. Integration means data will be exchanged between social media platforms and TrueScanAI. For TrueScanAI to fact-check content, the platform might send us the text of posts or articles a user is viewing. Even if it's just public article URLs, we might also receive metadata like user ID or context. There's a risk that this data could be mishandled or over-collected. A real world case underlining this risk is the [3]Facebook-Cambridge Analytica scandal. In that incident, a third-party app harvested data from millions of Facebook users without proper consent and used it for political targeting. Users might also feel uncomfortable knowing that an AI is "watching" what they read on social media. Some users might ask, "Is TrueScanAI recording everything I click on? Could that information be used to profile me or shared with others?" If we cannot adequately address these questions, the ethical impact is a loss of user trust and a genuine invasion of privacy.

The second main concern is the actual accuracy and handling of misinformation that TrueScanAI has to do. With the way that TrueScanAI operates, it will be making judgments about what is "factually correct" or not. If our system is flawed. For instance, if it labels something false that is actually true, or vice versa we could end up spreading misinformation under the guise of fighting it. This is ethically troubling: users might be misled by incorrect flags, and content creators might be wrongfully discredited. A real world case of something like happening occurred when [4]Facebook's own fact-checking partner wrongly labeled a BMJ investigative article as "Missing context" and "partly false", which ended up causing the article to be censored on the platform. This example shows how a poorly implemented fact-check system can violate the rights of content creators and the rights of readers to access information.

| **Stakeholders** | **Financial Risk** | **Privacy Risk** | **Conflicting Interest Risk** | **Violation of Rights Risk** |
|---|---|---|---|---|
| Social Media User (Customers) | Low | Mid | Mid | Mid |
| TrueScanAI Company | High | Mid | High | Low |
| Social Media Platforms | Mid | Low | High | Mid |

From the Ethical Impact Risk Table shown above, we can easily show what the stakeholders' risks are for our business idea. For Social Media user, our customers, we assess financial risk as low. Using TrueScanAI is expected to be free for end-users (the service would be provided by the platform or supported by ads or other revenue streams, not by charging users directly). There's no direct cost to being fact-checked. Privacy risk for users is mid (medium) because there is some personal data exposure. While we try to minimize data collected, the very act of integrating means some user actions such as reading/sharing articles could be logged. If not handled properly, this could lead to personal data being inferred. Conflicting interest risk for users is marked as mid. By this we mean the conflict between what the user wants and what other stakeholders might do. Users want an accurate, unbiased tool that respects their autonomy. However, the platform or even TrueScanAI might have other interests that conflict with delivering the pure truth. This could manifest in users not getting the full benefit of the service. Violation of rights risk for users is mid. The main right at stake is freedom of expression and access to information. If TrueScanAI incorrectly flags content important to a user, that user's ability to share or see information is impeded. For example, a user might want to share a controversial opinion article, if TrueScanAI slaps a warning on it, fewer people engage or the user feels discouraged from speaking. That can be seen as a mild form of their speech being censored.

For the company, TrueScanAI, financial risk is high. Achieving this OKR is financially make-or-break. We'll likely invest significant resources into integrations (engineers, partnership deals) hoping for long-term returns. If something goes wrong ethically like a privacy scandal or a big mistake, we could lose those partnerships or face lawsuits, which would be financially devastating. Privacy risk for the company is mid. Here we interpret it as the company's risk exposure related to privacy issues. If we mishandle user data, the company faces legal penalties and damage to our reputation. Conflicting interest risk for TrueScanAI is high. We as a company sit at the intersection of multiple interests: truth vs. profit, platform demands vs. user trust, growth vs. ethics. This is arguably one of the toughest risks to manage. For example, if a major social network offers us a big contract but asks that we customize our fact-check criteria to be more lenient on certain content (to avoid controversy), we face a conflict, either take the money or stick to principles. Violation of rights risk for the company is low. Unlike individuals, the company itself doesn't have human rights like privacy or free speech in the same way. The company's "rights" could be interpreted as our right to operate freely in the market, which could be affected if, say, a government bans our service somewhere due to political reasons but that's not a direct rights violation but more of a business limitation.

Now for our social media platforms like Facebook and Twitter/X, financial risk is mid. Integrating TrueScanAI could have financial benefits like having increased user trust, avoiding fines for misinformation, better user retention among those who value credible content. But it also has downsides. If a chunk of users dislike fact-checks, they might leave the platform or reduce engagement, which hurts ad revenue. Privacy risk for the platform is low. The platforms already collect huge amounts of user data, integrating TrueScanAI doesn't drastically increase their exposure beyond what they already handle. In fact, they might share some data with us, but likely under strict contracts. The platform will want to minimize what data we get to decrease the chances of any data leaking out. Conflicting interest risk is high for platforms. This ties closely to earlier points with platforms having to balance the integrity of information with user engagement and profit. If our AI flags something that is generating a lot of ad revenue and clicks for the platform, that can be a conflict of interest with the social media network since they now have to balance between getting money from misinformation or getting rid of it. Violation of rights risk for platforms is mid. Platforms themselves don't have human rights, but here we frame it as the risk that through the platform's actions, rights might be violated, leading to backlash or legal troubles. For example, if through our integration the platform ends up effectively censoring certain content, they might be accused of violating user's free expression rights.

### 1.C.1.4 Ethical Safeguards

To safeguard user privacy, TrueScanAI will adopt a privacy by design philosophy from the outset. This means we only collect and process the minimum data necessary for fact-checking, and we do so with user consent and transparency. When integrating with a social platform, we will structure the API such that we receive only the content to be checked, and nothing identifiable about the user viewing it. Moreover, we will not share or sell any usage data to third parties. These decisions will be formalized in our data policy. To design these safeguards correctly, we will involve privacy experts. Consultation with a data protection officer or an expert in privacy law.

To prevent feelings of manipulation or censorship and to respect user autonomy, we will ensure TrueScanAI's integration is transparent in its function and give users some control. Whenever TrueScanAI flags or analyzes content, it will be accompanied by a clear indication and explanation. For example, if we label a post as "Contains false information", there will be an obvious icon or label (perhaps a small "TrueScanAI" check mark or warning symbol). Importantly, users can click on this label to see a detailed explanation of why it's flagged.

## OKR 2

### 1.C.2.1 Objective and Key Result

***TrueScan AI will provide accurate and reliable information to users when analyzing articles.***

Our milestone is to launch a public beta version of TrueScan AI that achieves at least 85% factual accuracy in detecting misinformation across 1,000 user-submitted articles. Key stakeholders would include the users submitting the articles, the company, and social media platforms that we partner with. Additionally, the people publishing the articles that we fact check may be impacted and as such should be considered stakeholders.

### 1.C.2.2 Metric(s) with Experiment(s)

In order to measure our goal of 85% factual accuracy across 1,000 user-submitted articles, we will conduct the following experiment:

1. Select a dataset of 1,000 articles from various topics.
2. Have professional fact-checkers label each claim in those articles as True, False, Partially True or Unverified.
3. Run the same dataset through TrueScan AI's model.
4. Compare the AI output to that of the professional fact-checkers.

### 1.C.2.3 Ethical Impact(s)/Issue(s)

One ethical impact is the issue of how we gather the data we use to train our AI model. As seen in the case of Bartz v. Anthropic PBC [5] it is important to make sure that any data used to train our model is ethically sourced and full credit is given for any sources we use.

**Expected Ethical Impact Risk Table**

| **Stakeholder** | **Financial Risk** | **Privacy Risk** | **Conflicting Interest Risk** |
|---|---|---|---|
| User | low | low | mid |
| Company | high | low | low |
| Social media platforms | mid | low | mid |

| | | | |
|---|---|---|---|
| Publications | high | low | high |

- **User Stakeholder:** The financial and privacy risks are all low for the user because the use of this tool should not impact them financially or have any effect on their privacy. There may be a conflicting interest risk if the information given to the user conflicts with their world view.
- **Company stakeholder:** The financial risk to the company is considered high because we could be sued by publications if we claim they are spreading misinformation.
- **Social media platforms stakeholder:** The financial risk to the social media platforms is mid because they would be spending money to use our tool. The conflicting interest risk is high because many social media platforms are used as a tool for misinformation.
- **Publications stakeholder:** The financial risk to the publications whose articles our tool analyzes is high because if they are found to publish inaccurate articles often, they may see a drop in revenue. Similarly, the conflicting interest risk is very high because they obviously will not want our tool to call them out when they are spreading misinformation.

### 1.C.2.4 Ethical Safeguards

The most important ethical safeguard will be to make sure that our data is properly sourced and all claims that our tool makes are fully cited. To implement the safeguards, we will make sure our AI model only makes a statement about the veracity of information in an article if it can provide a source for that statement. If the tool states that a claim made by the article is correct, it must provide a source that proves that claim. If the tool states that a claim made by the article is inaccurate, it must provide a source refuting that claim. If the tool is unable to verify whether a claim made by the article is true or false, it must simply say that it cannot find any sources related to that claim. In this instance, we must be sure the model does not state that the information is true or false, simply that it cannot find enough information. By providing these sources we not only make sure that we are properly crediting the data that our model is trained on, but we also improve the users' trust in our tool and give them insight into how they can detect misinformation on their own. The use of citations as an ethical safeguard is supported by Miles Brundage, who states that "In order for AI developers to earn trust from system users, customers, civil society, governments, and other stakeholders that they are building AI responsibly, they will need to make verifiable claims to which they can be held accountable." [6]

## OKR 3

### 1.C.3.1 Objective and Key Result

***Drive measurable user adoption of TrueScan AI.*** Our milestone is to increase monthly active users to at least 5,000 verified unique users within the first evaluation period, with at least 40% returning for two or more scans per month. Key stakeholders for this OKR include the end users, the company itself and the publications whose articles will be scanned by our tool. Another important stakeholder would be educators, librarians and academic programs. These groups may be interested in using our tool for educational purposes and could help to increase our adoption.

### 1.C.3.2 Metric(s) with Experiment(s)

In order to measure our goal of at least 5,000 verified unique users within the first evaluation period, with at least 40% returning for two or more scans per month, we will conduct an A/B experiment. Condition A will be the control group, which will be given the standard TrueScan AI experience. Condition B will be the treatment group, which will see an adoption-optimized experience. New visitors will be randomly shown either Control or Treatment. Then we will track whether the visitor completes at least one scan. Over the next 30 days, we will track whether that user returns to perform two or more scans. Finally, we will count the number of monthly active users and return rate for each user, comparing these numbers between each group.

### 1.C.3.3 Ethical Impact(s)/Issue(s)

One ethical impact of this OKR is the possibility of prioritizing growth and adoption over accuracy, transparency and bias mitigation. This was an issue that Facebook faced when it introduced a metric called "Meaningful Social Interactions" (MSI) in 2018 [7]. The goal of this metric was to increase interactions that seemed positive, such as comments, replies and conversations. However, this ended up with an increase in divisive content because that content drove more "meaningful interactions."

**Expected Ethical Impact Risk Table**

| **Stakeholder** | **Financial Risk** | **Privacy Risk** | **Conflicting Interest Risk** |
|---|---|---|---|
| User | low | high | mid |
| Company | mid | mid | high |

| | | | |
|---|---|---|---|
| Educational groups | low | mid | low |
| Publications | high | low | high |

- **User Stakeholder:** The financial risk is low for the user because the use of this tool should not impact them financially. The privacy risk is high because tracking repeat-use for this OKR increases data retention. There may be a conflicting interest risk if the information given to the user conflicts with their world view.
- **Company stakeholder:** The conflicting interest risk to the company is considered high because pressure to meet these metrics could conflict with ethical accuracy and transparency obligations.
- **Educational groups stakeholder:** The financial risk to schools and universities is low because they would face little financial exposure. The privacy risk is mid because we may collect article URLs containing personal or academic information.
- **Publications stakeholder:** The financial risk to the publications whose articles our tool analyzes is high because if they are found to publish inaccurate articles often, they may see a drop in revenue. Similarly, the conflicting interest risk is very high because they obviously will not want our tool to call them out when they are spreading misinformation.

#### 1.C.3.4 Ethical Safeguards

The most important ethical safeguard will be to make sure that we do not prioritize meeting our growth metrics over accuracy or fairness. One way to do this is by using fairness audits based on the idea of information justice, as suggested by Neumann, De-Arteaga and Fazelpour [8].

According to this framework, "information justice" must be considered for multiple groups: the sources of information, the subjects of information, the seekers of information, and the sources of evidence used to fact-check the information. Some ways to measure the effectiveness of this safeguard could include measuring error rates across different subjects/communities, analyzing which groups' content gets flagged as misinformation disproportionately, and measuring how and where we source our evidence used to fact-check information.

# 2: Cultural Policy

## 2.A. Core Values

We at TrueScan AI want people to know us as a careful, honest, and accuracy-driven business company. Our core values are that of integrity, transparency, fairness, user empowerment, and accountability. Integrity is an important value to us because it means our fact checking is guided by evidence and not affected by outside influences such as advertisers, social media platforms, or political groups. Transparency is something we show to our users because it increases trust in us whenever something is flagged and the reasoning of why it is flagged is supported by detailed explanations and cited sources. Fairness is about how we actively test for and correct bias so that no viewpoint is treated systematically worse by our models. User empowerment is idea where our product is used to assist and bolster our user’s critical thinking skills by not acting as a replacement for their judgment, but as a tool that teaches users how to assess the news on their own. Accountability means we accept that our decisions may cause harm if we flag something incorrectly, so we keep logs, publish summaries of our impact, and correct any errors publicly.

## 2.B Motivation

We are motivated by a love of informed, critical communities and a fear of unaccountable AI misinforming people all over the world. The people of TrueScan AI enjoy working with computing, journalism, and we value teams where engineers, designers, and social scientists can work together. Inside of the company, we have a focus of going to the safe and steady route, raising ethical concerns, and documenting trade-offs over certain decisions. This can lead to form habits in the company of having clear documentation, an open discussion of risks, and collaborations with external watchdogs. We want our users to view TrueScan AI as a place where technical accuracy and ethical concern are at the same standard of quality.

## 2.C Summary

1. Evidence-driven
2. Transparent
3. Fair
4. User First

# 3: Ethics Policy

## 3.A Core Items

Here at TrueScan AI, our ethics policy focuses on five items. The pursuit of truth and evidence, the focus on privacy and minimization of data, fairness and bias mitigation, transparency and contestability, and finally accountability and governance. We believe these five items are not only what defines us as a company but what will be the way our business thrives.

### Truth & Evidence

All flags done to news articles that have been analyzed must be traced back to verifiable sources. Should any new evidence appears for ongoing news, we will publish corrections of any kind to related news articles we've analyzed.

### Privacy & Data Minimization

When it comes to the use of data, we will only process the contents of a given news article or URLs. This strips any kind of identifiable information of the user as much as possible. Encryption of data in transit, avoidance of long data retention, and never selling any kind of behavioral data profiles.

### Fairness & Bias Mitigation

We use a diverse array of training data and use that data to test performance across different viewpoints and communities to see if there is a skew in preference in our data, and treat user bias reports as first-class signals.

### Transparency & Contestability

Every label links to an explanation that is supported by reliable sources, and users and publishers can appeal the decisions we've made through an in-house support service where human reviewers can review any support tickets submitted.

### Accountability & Governance

We maintain our own internal review process, logging incidents, publish summaries of flags and appeals, and we reserve the right to refuse any deployments that conflict with this policy.

## 3.B Board

Our first person to be part of the board is Angie Drobnic Holan who is the Editor-in-Chief of PolitiFact. She is a strong fit to work for our board because she's is nationally recognized as a leader when it comes to fact-checking, as being an editor for PolitiFact, one of the most influential organizations when it comes fact-checking. She has overseen hundreds of investigations, U.S. election cycles, public health misinformation, and international claims. She's been part of committees with the International-Fact-Checking Network (IFCN), assisting in the development of fact-checking standards globally. My second person to be part of the board is Dr. Renee Hobbs. She is one of the most prominent media literacy scholars in the world and is known for her work on news interpretation, evidence evaluation, and how to respond to misinformation. She's a strong fit for TrueScan AI because how her past work focuses on teaching practical critical-reading strategies, which is exactly the kind of educational information that TrueScan AI aims to provide. She's also published extensively on misinformation, propaganda, source evaluation, and how students learn to interpret online information. And for my third person to be part of the board is Kate Crawford. She's an Australian researcher and the co-founder of the AI Now Institute whose work focuses on the social and political impacts that AI can cause. Having her part of the board due to how her expertise in how data systems can affect rights, power, and social media. All of which maps directly into AI fact-checking.

## 4: YouTube Presentation

https://youtu.be/asztZ-Jq27g

## 5: References

1. Anon. 2025. Social Media and News Fact sheet. (September 2025). Retrieved October 13, 2025 from https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/#:~:text=Digital%20sources%20have%20become%20an,over%20the%20last%20few%20years

2. Peter Suciu. 2023. X is the biggest source of fake news and disinformation, Eu warns. (October 2023). Retrieved October 13, 2025 from https://www.forbes.com/sites/petersuciu/2023/09/26/x-is-the-biggest-source-of-fake-news-and-disinformation-eu-warns/

3. Rebecca Coombes and Madlen Davies. 2022. Facebook versus the BMJ: When fact checking goes wrong. *BMJ* (January 2022). DOI:http://dx.doi.org/10.1136/bmj.o95

4. Rosalie Chan. The Cambridge analytica whistleblower explains how the firm used Facebook data to sway elections. Retrieved October 13, 2025 from https://www.businessinsider.com/cambridge-analytica-whistleblower-christopher-wylie-facebook-data-2019-10

5. Anon. 2024. Bartz v. anthropic PBC, 3:24-cv-05417 – courtlistener.com. (August 2024). Retrieved September 23, 2025 from https://www.courtlistener.com/docket/69058235/bartz-v-anthropic-pbc/

6. Brundage, M. (2020) *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims*. doi:https://doi.org/10.48550/arXiv.2004.07213.

7. Anon. 2021. Facebook revelations: What is in cache of internal documents? (October 2021). Retrieved November 29, 2025 from https://www.theguardian.com/technology/2021/oct/25/facebook-revelations-from-misinformation-to-mental-health

https://www.courtlistener.com/docket/69058235/bartz-v-anthropic-pbc/

8. Terrence Neumann, Maria De-Arteaga, and Sina Fazelpour. 2022. Justice in misinformation detection systems: An analysis of algorithms, stakeholders, and potential harms. *2022 ACM Conference on Fairness Accountability and Transparency* (June 2022), 1504–1515. DOI:http://dx.doi.org/10.1145/3531146.3533205