Summary of WSJ Article

In 2018, Facebook made a major change to their news feed algorithm that sparked a lot of controversy. Facebook had been steadily seeing a decrease in engagement on their platform and this caused worries from management that the platform would soon become obsolete. The change was intended to **shift their algorithm from promoting relevant content to promoting more meaningful interactions between friends and family**. It was framed as a sacrifice by the company, noting that the expected time users spent on the platform would go down, but the time spent would be more valuable to the user. However, due to the new emphasis on reshared content, the algorithm change **began promoting more provoking and hateful content**. Content that sparked controversy was rewarded, and this included fake news, violence, and generally toxic content. This impacted how news publishers and political parties chose to post on facebook. As a response to the change, **more negative and radical content started to be posted to retain engagement**. In order to reduce the negative effects this algorithm change had, data scientists on Facebook's integrity team proposed several solutions, **many of which Zuckerburg resisted to preserve user engagement.** Some categories of content implemented solutions that effectively reduced promoting misinformation, but was not adapted broadly for this reason.

Summary of Emotional Contagion Article

The emotional contagion article is focussed on a study that **exposes Facebook users to different amounts of positive and negative content on their news feed, and studies its emotional effect on the user** through the user's posts. The study found that when the amount of positive posts, defined as posts containing at least one positive word, were reduced on users' feeds, the number of positive words in their own posts decreased while the number of negative words increased. The opposite was true for when negative posts were decreased, resulting in a higher number of positive words in the user's posts and lower number of negative posts. In addition to the main finding, the study also found that users exposed to less emotional posts were also less likely to express themselves emotionally through their own posts.The **results support the hypothesis that emotional contagion through social media occurs**, even without the direct, in-person interaction that previous studies have indicated as a requirement for emotional contagion.

Ethical Concerns in the Articles

There are many ethical concerns regarding what was done by Facebook through the algorithm change. Perhaps the company's intentions were to do good initially, but it was their response after finding out that the change rewarded toxic and false posts that was unethical. Even after it was clear that the company was aware of the negative effects the change had, they still chose to keep it in place and **resisted proposed solutions that would impact the improved user engagement**. This is a clear violation of the justice principle of the Belmont report, which states that the subjects of the burden should be the ones who reap the reward. In this situation, it is the **Facebook users who are the subjects of the burden, while the company is the one who are benefitting from this change**. The change did have some positive effects for the users in that posts from closer connections were promoted, but it is vastly outweighed by the negative effects on the users. The emotional contagion study showed the effects that negative social media posts have on users and Facebook was actively promoting these kinds of posts as well as misinformation. A fundamental recurring principle of ethics is to do no harm, and Facebook has

clearly **harmed its users** in this case. Even when put against one of the simplest frameworks of ethics, where you describe the research to your grandma, it would make anyone **uncomfortable to describe promoting outrage and misinformation to improve user engagement.**

The ethical concerns with the study conducted in the emotional contagion article are more subtle compared to the Facebook files but still present. The article claimed to have obtained informed consent through the agreement of Facebook's Data Use Policy, but they did not. Participants were **not informed on the details of the study and more importantly about the risks associated**. Furthermore, they **did not give users the ability to opt out of the study**. The age requirement needed to sign up for a Facebook account is 13, meaning that there are minors on the platform. There was no action taken to account for this fact and **offer any sort of protection toward minors or other vulnerable groups**. The participants were pulled into the study without their knowledge and subject to the study's potential negative impacts – a **violation of the principle to do no harm**. Additionally, the justice principle of the Belmont report is also violated here as well. There is clearly **only one group that is bearing the burden of the study**, where each person should have been treated equally.

Questions Our Company Should Think About

Let's start at a high level. **Where in our operations could ethical concerns arise?** We have many different products and applications that we provide for our customers, but not all of them may need to be closely examined in regards to ethics. For example, the scheduling team may have a lot less to worry about in terms of ethics compared to say the analytics team. It is important to first identify potential areas so we can prioritize our resources adequately. Even if there aren't ethical dilemmas being faced on the team right now, we should identify where there is possibility for misuse. Like for the case of Facebook, the team in charge of the news feed and more specifically what kinds of content is shown to the user is an area where a lot of ethical issues may arise. We should identify areas or teams in our company that may also need to deal with these issues so we can implement specific processes around them proactively. Like Facebook, we are also a company that has access to a lot of data, but just because we can do something, it doesn't mean that we should. Additionally, we should think about areas **where patient risk, fairness of treatment, and provider usage are involved** as these are areas that have the highest potential and impact for harm.

Next, let's narrow the scope a bit to have more concrete questions to pursue and answer. One area I would like to further explore is fairness of treatment. Relating to the study from the emotional contagion article, where one group was unfairly negatively affected, we want to make sure that we are not subjecting any patients to a less quality treatment, even if it benefits other groups. One way this often happens, is when demographic data is involved. We should determine, **are we properly accounting for different demographic groups when building our applications?** There are many different scenarios where this can manifest. For example, a team may be considering switching to a completely paperless form of registration. We should consider if this would disproportionately affect certain demographics due to the lack of technology. Different cultures, income brackets, and age groups have differing preferences and ways of doing things, and we should make sure to account for these so that we are not unfairly benefitting one group at the cost of another.

Now that we have a more concrete question to guide us, let's explore specific ways that we can attempt to answer this for some of our applications. One of our teams that deals with demographic data is our predictive analytics team, as demographic data is often used in the models. **Are our models disproportionally benefitting a specific demographic group?** Depending on the model, this could happen in different ways. Generally speaking, we should make sure that if we release a model, it **performs well on all cohorts**. For example, we have a model that helps detect early signs of <anonymized for privacy> in patients. The model may have a high evaluation statistic overall, but if say it is predicting significantly worse on minority patients, then we are unfairly subjecting them to worse care. In the case of the sepsis model, the potential impact could be lives. The good news is that **we have all the data we need to answer this already**, since they are all inputs into the model. We should have processes in place to **perform cohort analysis** on each of our models to make sure no demographic groups are under performing, and continue to monitor the model's performance of the different cohorts even after they are initially released. Model performance as a whole can often start drifting, but there are cases where they start drifting for only a specific group, and we should be ready to detect and correct this.