James Yang

DATA 512

11/5/22

Evaluating Future Impacts

Step 2 Notes (Scenario 5 – Improving Workplace Mental Health):

- One of the greatest worries of this scenario is pre-defining someone's future before it even happens.
- Firing someone for a behavior that they exhibited in the past should not represent who they are in the future.
- There might be bias on users who come from lower-income neighborhoods with higher drug rate or crime, and thus get laid off for something they had no control over.
- Actions that would be taken is a safeguard on these biases and additional layers ontop of the prediction model. These safeguards would include opinions from peers and families and actions exhibited externally.
- The situation in which a person has had their social media hacked can cause discrepancies if something were to go wrong.

Step 2 Summary: The scenario chosen is scenario 5, "Improving Workplace Mental Health". Having a smart employer service that evaluates a person's suitability for the workplace through social media, predictive modeling of mental illness, and other trends of postings about an individual is not only an invasion of privacy but can also be wildly misleading and ineffective. The issue starts with pre-defining someone's future before it even happens, creating a wave of misinformed precedent on people without them even acting upon the action. The same thought can be applied for criminal activity; it would be unfair to put someone in prison before they had done an illegal action because it was predicted that they would.

What would happen if someone's social media was hacked? Would their inputs be violently misconstrued and then eventually fired from every corporation known to earth? What if they accidentally posted a picture that didn't represent who they were? The scenarios of mistakes could go on.

Step 3 Notes:

- Risk Zone 1:

    - We would expect to collect measure and share very personal information (social media, likes, posts, news articles, etc.)
    - Hackers could take control of your account and create false posts, likes, etc. about you.
    - Someone could use this technology to undermine trust by creating false behavior. This includes not being yourself on social media, creating different posts that you normally wouldn't do, and just alternating your behavior in general.
- Risk Zone 2:
    - The business model benefits big corporations temporarily but it may hurt them as well.
    - It creates no benefit for either side because companies hire fake personas and people become these fake personas.
    - This technology would create very unhealthy engagement through not envisioning true self-persona.
    - To design a system that encourages moderate use of the idea, they would need to begin implementing some sort of data scalping about people to see if they truly line up with how their social media presence says they do.
- Risk Zone 3:
    - Bigger companies will have more access to this technology in comparison to smaller companies because of space, technology, expense, bandwidth, server, etc.
    - Larger companies has access to this asset.
    - In this case, we are not using machine learning to create wealth. It in fact is just altering human behavior.
- Risk Zone 4:
    - This technology certainly makes use of deep data sets and machine learning. There will definitely be gaps or historical biases just based on whether the user has social media or not. If the user stops using social media for a year, there will be a lot of gap in behavior regardless of whether that person has actually changed or not.

    - I have not personally seen instances of personal or individual bias enter into the products algorithms but they most certainly could based on a person's background and what they do.
    - The technology is definitely amplifying bias.
    - I think a push back against a blind preference for automation in this case will come in revenue from companies when they realize the workers who they decided to retain and hire are simply not cut out for the job.
    - The algorithms are certainly not transparent to the people impacted by them.
- Risk Zone 5:
    - A government body might use this algorithm to survey whether a military member is having wellness issues. However, they won't be in contact with social media at most times rendering that feature fairly useless.
    - Governments could create a hiring process that would render exactly the types of people in their society that they want to retain.
    - We are creating data that could follow users throughout their lifetimes based on their life choices. Any choice they make will be on permanent record with their hiring.
- Risk Zone 6:
    - Does not need to be collecting the data and if it is being sold it is a complete invasion of privacy.
    - Users do not have the ability to access the data that they are being collected on, but if they were I would wonder if that would impact how they would view themselves causing even more harm on mental health.
    - Bad actors could manipulate the data and cause harm to someone.
- Risk Zone 7:
    - The technology we are building has a clear code of rights for users. It is most likely not easy to read, access, and understand.
    - The technology could imply some underlying weights on a person if they were let go because of a previous algorithmic analysis on them.
    - All users are certainly not treated equally in this scenario.
- Risk Zone 8:

- People could use this technology to bully and stalk others simply by taking a deep dive analysis on their social media profiles.
- There are countless numbers of illegal activity that could arise from counterfeiting or impersonating others to manipulating data to ransomware on competitors.

Step 3: Risk Analysis

The three risk zones which were most relevant to scenario 5 were Hateful & Criminal Actors, Machine Ethics & Algorithmic Biases, and Surveillance State. For Machine Ethics & Algorithmic Biases, the technology makes use of deep data sets and machine learning by creating predictive models that allow for smart employment. The technology is most definitely amplifying existing bias because it is inherently learning off bias data. Social media posts are generating by humans and their likings. When a model learns off of these social media posts to try and translate whether the user is showing symptoms of health issues, there is reason to believe that there is going to be bias in determining that factor. Bigger corporations such as FAANG are more than likely responsible for developing the algorithm along with social media companies. There is a lack of diversity in the people responsible for designing the technology as well because it is emphasized on social media, but not outside data. What if the person never uses social media but instead spews racist comments in their household instead?

For Surveillance state, government body may utilize this technology to increase its capacity to infringe upon the rights of the citizens privacy. Training predictive data on people's lives is very unethical as nobody wants to have a model tell them whether they are fit for work. Governments could use this data to harness potential threats for school shooters, rapists, etc. However, the procedure of using this data would be highly unethical and could be deemed inaccurate as well. This model is creating data that could follow users throughout their lifetimes, affecting their reputations, and impacting their future opportunities. It is doing this by giving the user a track record of what they have done in the past, and reiterating it in different stages of their life by giving them a metric of "fitness to work". Ideally, we would not want anyone to use the data except the model, and even then it doesn't seem right that this data is being collected in the first place.

For Hateful & Criminal Actors, people could use the smart employer service to bully, stalk, and harass people based on their recommendations of who they are. Their likes, social media posts, and anything in between could be a source of harassment because of their "perfect"

persona that they are trying to create for employers. Illegal activity could certainly arise around the technology through racial, age, cultural, and many other biases occurring when deciding whether to retain or let go of an employer. The technology could be weaponized by manipulating people's self-image and who they truly believe that they are. It could create an influx in suicide, or even be detrimental to the working class as a whole.