Project Ethics

Normalizing obvious AI

I think it would be worth avoiding obfuscating how the system works in order to influence behavior. Being a ‘well-intentioned researcher’ isn’t a persuasive argument for hiding from a person that they are interacting with a bot and in an experiment. Maybe the tech should stand on its own and should work in spite of obviously being AI?

1. What if, in addition to de-escalating arguments, the bot aims to teach users *how* to engage with AI? This is vague, but people generally know they’re caught by “algorithms”. Even if they don’t like it they have no idea what to do.

2. Twitter wasn’t built for discussions, especially ones that require patience and context. It was built for off-the-cuff remarks, jokes, observations and breaking news. It simply is not made to be used to engage in serious discussions of politically charged ideas. It would be worthwhile to steer charged conversations into a light-hearted, joking realm or satire. Nothing surreptitious; a bot interjecting and reminding the people they’re online and in a venue that’s built to polarize.
    a. If the bot could figure out some kind of common ground the people arguing might have, and make a joke along those lines, that might be an interesting thing to try.
    b. Figuring out how well GPT-2 can handle satire would be interesting but probably not very fruitful.

Observing changes in user behavior

Would it be possible to gather data from interactions on our end only? Instead of tracking how a user changes over time (or tracking individual users at all), can we use engagement data from the bot’s tweets/follower count? Maybe also build in some other ways for people to voluntarily engage and track that? Maybe there are metrics that don’t involve tracking individual accounts that could still lead to valuable insights? I think giving people back their sovereignty and privacy online is important.

General thoughts

What depresses me the most about all of this is how the ideal fixes for these problems have been foreclosed on in this country. I think legislation and collective action are necessary but in large part because of the last ~40 years in this country’s political history, even incremental changes by these means is a long shot. And of course the companies won’t regulate themselves or fundamentally change their own business models. All of this to say, that’s why I

think building tools to mitigate the harm is important, and I think avoiding as many of the experimental design traps that the industry takes as necessary evils is also important.

Telling people to just stop using social media, a commonly prescribed solution, doesn't consider how atomized and isolated American society is. Social media has exacerbated this issue, but it is not the root cause. For better or worse much of social life is rooted online, and as long as there is an online there will continue to be a social dimension in some form.