Should algorithms do what people want or what is good for them?: Exploring ethical dilemmas in online health and privacy

Steve Whittaker, Victoria Hollis, Artie Konrad & Jeff Warshaw, UCSC

We review our own studies of successful health interventions and privacy profiling. These studies show significant barriers to educating people about the implications of their online actions. We describe dilemmas that arise when what users *think* are beneficial personal behaviors are inconsistent with behaviors induced by an algorithm that do actually help them.

The problem: not knowing what's good for you

One successful line of work in our lab has been online health interventions. We have been designing reflective apps that help people manage problematic aspects of their everyday lives. These apps involve people actively reflecting about past personal experiences. For example, we have developed an app that allows people to record personal experiences. A simple algorithm presents these recordings back to participants for later reflection after some time has elapsed. Our work shows that such reflection improves psychological well-being (Isaacs et al., 2003). A different app helps people moderate problematic health related behaviors (aka 'bad habits'), such as overeating, smoking and drinking. This second app reduces the frequency of bad habits by encouraging people to reflect on past incidents involving their habits and to focus on the future emotional outcomes of indulging in these behaviors (Hollis et al., in press).

So far so good, you might think.

However here is the problem. In both cases, we interviewed participants who used the apps in a one month intervention. We explored participant intuitions about how the app is helping them. In both cases participants have flawed intuitions about how the app helps. Our objective data show that behavior change is helped by presenting participants with 'failure' cases, e.g. where they submitted to temptation and indulged their habit. Seeing these failures leads to more behavior change than presenting participants with 'success' cases where they successfully resisted their habit. However when asked about what past events they *wanted* to see, our participants said that they would prefer to see more positive (but objectively less helpful) events. Participants felt it was depressing and demotivating to focus on past failures. These discrepant observations are supported by related work showing that although participants show multiple long term health benefits from repeatedly reflecting on past traumas (Pennebaker & Chung, 2011), they nevertheless report this has short term negative consequences for mood (Sloan and Marx, 2004).

Here then is the ethical dilemma: Should algorithms do what people *want* or what is *good for them*?

This is clearly a complex issue, and we have yet to ask our participants to evaluate the trade-offs between short-term pain versus long term gain. Whatever our participants answer, this brings up an important ethical question about how future interventions should be conducted. What if participants say that they prefer to accentuate the positive when we *know* that this is less likely to help them? Should we (in the spirit of user-centric design) honor their wishes and show them past events that are less likely to improve their health?

Of course this is not an isolated case. The firestorm around the Facebook social contagion study (Kramer et al., 2014) reveals many of the same issues. The social contagion intervention showed that manipulating people's Facebook feeds to make their contents more positive led participants to make more positive and fewer negative posts. The opposite was also true; that seeing more negative posts in one's feed reduced positive and promoted negative posts.

Now that Facebook knows how to make people happier, what should Facebook do with this information? Should it interfere with people's feeds to make them more positive or leave well alone? And this isn't isolated to social contagion. Facebook could also do other large scale social manipulation about people's relationships and well-being. We know that certain patterns of focused Facebook communication benefit well-being and social ties. Directly messaging another user or commenting on their post improves well-being and deepens social ties (Burke et al., 2014, Burke et al., 2010). This contrasts with passive feed consumption that depresses well-being, and leaves social relations unchanged. So, should Facebook manipulate their algorithms and UI to actively encourage their users to engage in beneficial rather than damaging behaviors? If a participant never messages or comments on posts in their feed, should Facebook shut that user down?

A (failed) intervention: seeing very private information that can be inferred about them does not modify users' behavior

We believe that this is an emerging and very general problem where users must compare what they want versus what might be good for them. But what can we do about this? In other privacy work we have looked at trying to coerce people into behaving in their own best interests.

It is well known that people are ignorant about what can be inferred about them from their online actions. We explored this by developing an algorithm that can accurately infer personality profile and social motivations from a relatively small number of tweets or Facebook posts (Warshaw et al., in press). Interviews show that participants are surprised by the accuracy of our algorithm. However being shocked does not necessarily lead users to change their behaviors. When we asked users for their reactions to these profiles, rather than deciding to modify their social media behaviors, or refusing to share this information, they showed passivity and helplessness arguing that whatever actions they might personally take, these profiles would inevitably emerge. Despite the algorithm being unfinished and occasionally inferring incorrect personality traits, users treated it as an expert process that they were hesitant to contradict. This is surprising, given that people tend to see themselves as the ultimate authority on their own personality, pointing to the importance of understanding people's trust in algorithms as a concept, and of finding ways to increase distrust for imperfect systems.

Although this is a single intervention in a single domain, it speaks to the difficulty of increasing users' privacy even when they disagree with an algorithm.

Three workshop questions

1. How can we better communicate to users the algorithmic implications of their online behaviors? Algorithms are complex and users are often unaware of what can be inferred about them, or how using a app changes their behavior. How can we show users the relations between behaviors and outcomes, whether these outcomes are for health or privacy?
2. How can we better educate users about cost/benefit trade-offs for online behaviors? Many people cannot envisage a world without Facebook because it brings so many social benefits. So even showing what private information is known about them may not cause them to stop using Facebook, but could we better educate them about the implications of particular Facebook behaviors?
3. What can be done about the fact that algorithms are a moving target that's always improving? Explaining to users what is known about them now, does not tell them what *might* be known about them in the future.

References

Burke, M., Marlow, C., & Lento, T. (2010). Social network activity and social well-being. ACM CHI 2010: Conference on Human Factors in Computing Systems, 1909-1912.

Burke, M. and Kraut, R. (2014). Growing closer on Facebook: Changes in tie strength through site use. ACM CHI 2014: Conference on Human Factors in Computing Systems.

Hollis, V., Konrad, A., Whittaker, S. (2015). Change of Heart: Emotion Tracking to Promote Behavior Change, Proceedings of the 2015 Conference on Computer-Human Interaction (CHI '15). ACM, New York, NY, USA.

Isaacs, E., Konrad, A., Walendowski, A., Lennig, T., Hollis, V., & Whittaker, S. (2013). Echoes from the past: How technology mediated reflection improves well-being. Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems

Pennebaker, J. W., & Chung, C. K. (2011). Expressive writing: Connections to physical and mental health. Oxford handbook of health psychology, 417-437.

Sloan, D. M., & Marx, B. P. (2004). A closer examination of the structured written disclosure procedure. Journal of consulting and clinical psychology, 72(2), 165.

Warshaw, J., Matthews, T., & Whittaker, S., Kau, C., Bengualid, M., & Smith, B. (2015). Can an Algorithm Know the "Real You"?: Understanding People's Reactions to Hyper-personal Analytics Systems. Proceedings of the 2015 Conference on Computer-Human Interaction (CHI '15). ACM, New York, NY, USA.