Ethically Paternal: Value Directed Algorithms

Mark Rucker
MS Student
Systems Engineering
mr2an@virginia.edu

Sheung Tak
PhD Student
Corcoran Department of Philosophy
sl9jk@virginia.edu

March 19, 2017

Background

Algorithms have become an indispensable part of our decision making processes -- whether explicitly, like online dating services, or implicitly, like Facebook's news feed. They shape our understanding of the world. Such a reality is tacitly acknowledged in daily headlines concerning Google's and Facebook's responsibility for their content distribution.

Many of today's decision support algorithms seek to take a value neutral stance (Diakopoulos 2016). The belief often goes that by simply supporting a user's already held values algorithm designers themselves bear relatively little responsibility otherwise. In order to achieve this value neutral stance algorithms are usually designed to track what their users ***want*** based on their recent activities. Unfortunately, this design -- tracking what users want (explicitly or implicitly) instead of what is ***good*** -- has led to what people call an echo chamber, which has proved to cause undesirable polarizations in the society.

Having algorithms designed to track what is good for the users, however, leads to a different ethical challenge. There is an ethical concern about an algorithm's threat to our autonomy when it tries to nudge our decisions in a way that deviate from what we desire. The concern of paternalism is alarming especially when the working of many of the algorithms that affects our decision-making are usually not transparent to their users.

We believe that research in the ethics of paternalism and autonomy can help us find the much needed middle ground. Although paternalism doesn't have the best reputation, we will try to show that some versions of paternalism are actually ethically defensible and should be taken seriously. We believe that finding an ethically defensible notion of paternalism is the key to answering the programmer's dilemma. In particular, we will focus on algorithm for book recommendation. We hope this research is a first step to understanding how to design algorithms that help us make not just desirable, but good decisions.

Objectives

We aim to design an algorithm which can push back against an individual's immediate desires (i.e., act paternalistically) in an ethically defensible manner. To do this we will use statistical analysis and Inverse Reinforcement Learning to distinguish between a person's values and desires. We will then develop a book recommendation app which actively nudges the user towards what is good for him/her (their values) rather than simply suggesting what they desire (Hadfield-Menell et al. 2016). The hope is that by tailoring our book recommendations to balance a user's multiple values, the echo chamber effect can be mitigated. And by using the user's values instead of desires as the guideline for the algorithm, we hope that it can claim to be respectful to the users' autonomy (based on what Susan Wolf (1987) called the deep-self view of free will).

Hypothesis

1. Data features which capture a person's values have distinguishable statistical characteristics from features that capture a person's desires
2. Assistive cyber agents that interact with people and are built upon ethical frameworks from the ground up will be more trusted and effective than those which aren't

Methodology

In order to achieve the stated objectives of this project we will break the work into three phases:

Phase One. Philosophical work on paternalism has to be done for the idea to be represented through quantitative measures that are understandable to algorithms. To begin with, we will determine the source of ethical concerns about paternalism by reviewing recent philosophical work on paternalism (e.g., Coons & Weber (eds.) 2013). Then, borrowing philosophical resources from the Confucian ethics, which has a long tradition of trying to develop an ethically respectable idea of paternalism by distinguishing a person's everyday superficial desires and more primordial ethical sentiments/commitments (e.g., Flanagan & Williams 2010), and Gary Watson's (1975) influential theory of free will based on the distinction between a person's value and desire, we will defend a version of paternalism that is free from the ethical challenges. The basic idea is that paternalism is defensible as long as the act that goes against a person's desires is sensitive to the person's primordial ethical commitments or values. The conceptual distinction between value and desire is best illustrated in the case of an unwilling drug addict who *desires* to do drugs but *values* a life without drugs. The philosophical view we are trying to defend says that an act of paternalism that intervenes the person's crave for drugs to satisfy his desire is ethically benign as long as the act is sensitive to the person's deeper value to be healthy.

Phase Two. We will use the aforementioned philosophical account of an ethically benign version of paternalism to determine the key abstract components of a nudging algorithm for book recommendations. We want an algorithm that is capable of isolating a user's basic values from things that a user merely desires at the moment. This distinction could provide recommendations that nudge a user out of his/her comfort zone without being paternalistic in a problematic way that may threaten the person's autonomy.

In order to explore this question our plan is to reach out to Goodreads, an online social network centered on book reading, and ask if they would be willing to share anonymized user data sets with us. Former UVA students do work there, so we think it is a possibility. If this fails, our secondary plan is to set up our own mobile research study using an in-house app known as Sensus. Mark has experience deploying Sensus studies and so the risk of successfully carrying out such a study is relatively low. Finally, if the above two plans don't come about there is a well-known existing data set called StudentLife (Wang et al. 2014) that can be used to test our hypothesis concerning distinctions between human values and desires.

Recent work in Inverse Reinforcement Learning does support the idea that a user's values and desires can be effectively captured (Nouri & Traum 2012). However, relatively little work has been done with people to, one, identify the best data features from which to passively extract

these characteristics, and two, effectively distinguish between multiple kinds of goals (i.e., a person's desires versus their values).

Phase Three. Once a strong foundation has been built on philosophical and data analytic work, we will turn our attention to building a simple working application which provides book recommendations to users. This application will remain within the defensible ethical bounds. The reasons for building such an app as part of this project is to narrow the gap between practical application and theoretical research. Often times many unexpected hurdles have to be overcome in this process.

In building the app we will make use of Goodreads public book reviews. This dataset contains over 700,000 books and 10 million reviews. We will go through a similar process as we did with the StudentLife dataset in order classify books to match various kinds of basic values. This will likely involve considerable natural language processing; however, many other students in Mark's Predictive Technology Lab have extensive experience with this kind of work and will be available to give direction.

Combining our Inverse Reinforcement Learning work above with the book categorization we will be set to provide paternalistic book recommendations in an ethically defensible way (Abel et al. 2016). We will implement two nudging algorithms: one that strictly follows the philosophical ideas for ethical nudging and one that is a little looser. We will track user engagement, how many people thought the suggestions were helpful, and if the users felt the nudging overstepped its bounds. This final evaluation will test the second hypothesis, that in order for algorithms to work more closely with people a fundamental shift, placing ethics at the forefront, will need to happen for these tools to succeed.

Impact

- Substantial contribution to applied ethics by showing the plausibility of mild paternalism
- Interesting paths opened for work on algorithmic intervention research in other domains
- Potential future research between philosophy and systems in field of Machine Ethics

References

Abel, D., MacGlashan, J., & Littman, M. L. (2016, March). Reinforcement Learning As a Framework for Ethical Decision Making. In *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*.

Coons & Weber (eds.) (2013) *Paternalism: Theory and Practice.* Cambridge University Press.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM*, *59*(2), 56-62.

Flanagan, O & Williams, R. A. (2010). "What Does the Modularity of Morals Have to Do With Ethics? Four Moral Sprouts Plus or Minus a Few." *Topics in Cognitive Science* 2 (3):430-453

Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems* (pp. 3909-3917).

Nouri, E., Georgila, K., & Traum, D. R. (2012, August). A Cultural Decision-Making Model for Negotiation based on Inverse Reinforcement Learning. In *CogSci*.

Watson, Gary (1975). "Free agency". *Journal of Philosophy* 72: 205-20.

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., ... & Campbell, A. T. (2014, September). StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 3-14). ACM.

Wolf, Susan (1987). "Sanity and the Metaphysics of Responsibility". In Ferdinand David Schoeman (ed.), *Responsibility, Character, and the Emotions: New Essays in Moral Psychology*. Cambridge University Press: 46-62.