Machine Ethics, Part One: An Introduction and a Case Study

David Gray Grant

The past few years have made abundantly clear that the artificially intelligent systems that organizations increasingly rely on to make important decisions can exhibit morally problematic behavior if not properly designed. Facebook, for instance, uses artificial intelligence to screen targeted advertisements for violations of applicable laws or its community standards. While offloading the sales process to automated systems allows Facebook to cut costs dramatically, design flaws in these systems have facilitated the spread of political misinformation, malware, hate speech, and discriminatory housing and employment ads. How can the designers of artificially intelligent systems ensure that they behave in ways that are morally acceptable–ways that show appropriate respect for the rights and interests of the humans they interact with?

The nascent field of machine ethics seeks to answer this question by conducting interdisciplinary research at the intersection of ethics and artificial intelligence. This series of posts will provide a gentle introduction to this new field, beginning with an illustrative case study taken from research I conducted last year at the Center for Artificial Intelligence in Society (CAIS). CAIS is a joint effort between the Suzanne Dworak-Peck School of Social Work and the Viterbi School of Engineering at the University of Southern California, and is devoted to "conducting research in Artificial Intelligence to help solve the most difficult social problems facing our world." This makes the center's efforts part of a broader movement in applied artificial intelligence commonly known as "AI for Social Good," the goal of which is to address pressing and hitherto intractable social problems through the application of cutting-edge techniques from the field of artificial intelligence.

One major part of CAIS' work is devoted to the development of public health social work interventions targeting the rapidly growing population of homeless youth in Los Angeles, where CAIS is based. In May of last year, the Los Angeles Homeless Services Authority reported that there were an estimated 6,000 homeless youth living in Los Angeles County in January 2017, a shocking 61% increase relative to their 2016 estimates. For the past several years, CAIS personnel have been working with local collaborators and domain experts at other institutions to develop a wide variety of interventions to address the needs of this highly vulnerable population, including social work public health interventions aimed at HIV prevention for homeless youth and predictive-modeling-based triage tools intended to help organizations supplying housing and services to homeless youth allocate scarce resources to those who need them the most.

During my time at CAIS, center researchers were in the early stages of adapting a substance abuse prevention program called TND (Towards No Drug Abuse) for use in residential shelters for homeless youth. Homeless youth have been found to use and abuse drugs—including alcohol, methamphetamine, heroin, and cocaine—to a much greater extent than their housed counterparts. Moreover, developing effective interventions to reduce drug use in adolescents has proved to be extremely difficult. Part of the problem has been that traditional, classroom-based substance abuse reduction programs have been led by authority figures that tend to have far less social influence on young people than their peers. To address this, TND makes use of highly-interactive, peer-led small group classroom sessions: the thought being that at-risk youth will be more likely to adopt healthy attitudes towards drug use when those attitudes are encouraged by their social peers.

The underlying theory here appears to be substantively correct: TND has yielded impressive results at reducing drug use among at-risk youth in school contexts. However, there is an important caveat. In a recent large-scale trial of TND, Valente et al. (2007) found that particularly at-risk participants not only failed to benefit from the program but seemed to be placed at an even higher risk of future drug abuse as a result of their participation. Valente et. al. hypothesized that this increase in risk was the result of deviancy training, a well-documented phenomenon that occurs when individuals are encouraged by their peers to adopt harmful or antisocial behaviors. When such high-risk youth are placed together in an intervention group, Valente et. al. concluded, they are likely to influence each other to use substances more, not less.

This poses a significant challenge for any effort to adapt TND for use with homeless youth, given their higher rates of substance abuse and more permissive attitudes towards it. However, the program's potential payoff in terms of meaningfully reducing the harms associated with serious drug abuse made this a challenge well worth taking on.

When I arrived at CAIS, center researchers were in the midst of developing an artificially intelligent software agent intended to improve the way the intervention allocates participants to small groups by taking deviancy training effects into account. The agent was designed to accomplish this through the use of (1) a predictive model of how attitudes towards drug use spread through homeless youths' social networks and (2) an AI planning algorithm that attempts to identify the best way of allocating intervention participants to peer-led groups. What makes a grouping "best" is of course subject to interpretation: here the best grouping was understood to be the one predicted by the agent's predictive model to result in the greatest aggregate reduction in future drug use by all participants.

Achieving this second goal requires solving what is known in mathematics and computer science as an "optimization problem." An optimization problem is the problem of finding the element from a set that is best, or "optimal," according to a given criterion (the "optimality criterion"). In this case, the set of possible solutions consists of all possible ways of allocating intervention participants to small groups. The relevant optimality criterion says that a grouping

is optimal if and only if the model predicts that it will lead to less aggregate future drug use by intervention participants than any other possible grouping. This means that only the *total* reduction in predicted drug use for the whole population of participants counts: how that drug use is *distributed* across individual members of that population is not relevant to whether a given solution is optimal. The motivating idea behind this way of formulating the problem is that the set of participant groups that is optimal according to this criterion represents researchers' best educated guess about which grouping will make the intervention most effective overall.

At this point you may be wondering why we need artificial intelligence to solve this problem. Without going into the technical details, the reason is that the problem is mathematically complex enough to render it computationally intractable to solve in the field using simpler, more traditional approaches to optimization. However, cutting edge AI-based techniques make it possible to approximate the optimal solution, and to do so quickly enough for the results to inform the team conducting the intervention's decisions about how to group participants.

This brings us to the central moral dilemma in our case study. Early on in my time at CAIS, center researchers were in the process of testing their newly developed agent on previously collected data about the members of a particular population of homeless youth in Los Angeles. To do this, they fed their sample data into the agent's predictive model, and then instructed the agent to recommend intervention groups for a hypothetical set of participants in that population. The agent made a surprising recommendation: it recommended that the intervention team put the youth currently at highest risk of drug abuse into one group, and divide the youth at lower risk among the remaining groups. The agent's model, it turned out, had predicted that grouping the hypothetical participants in this way would result in a substantial reduction in future drug abuse for the lower-risk youth—at the cost of a substantial predicted *increase* in future drug abuse for the higher-risk youth. The expected decrease in risk for the lower-risk youth was great enough that, despite the increased risk for the higher-risk youth, the resulting grouping was optimal from the point of view of minimizing predicted drug use at the population level. In effect, the agent's proposed plan would allow a minority of participants to "crash and burn" in order to maximize benefits for the group considered as a whole.

For obvious reasons, this made the research team deeply uncomfortable, and left them uncertain as to how to proceed. On the one hand, the hypothetical grouping of participants in question appeared to be the one best suited to achieve the intervention's primary goal: maximizing population-level benefits. On the other hand, achieving those benefits appeared to require—at least in this hypothetical yet realistic scenario—putting a minority of participants in harm's way. Was this tradeoff morally acceptable, they wondered? And if not, how should the agent be redesigned to address the issue?

The team eventually concluded, in part based on our discussions of the moral issues involved, that modifications to the agent were indeed necessary. While some tradeoffs of the kind just described might be morally acceptable, others would not. This in turn meant that additional safeguards were needed to ensure that such tradeoffs would be negotiated in a responsible way when the intervention was eventually deployed. In the next post, I'll explain why—and use that as a launching off point for our broader discussion of central problems in machine ethics and promising strategies for addressing them.