Ethics and Artificial Intelligence in Public Health Social Work

David Gray Grant
Harvard University
dggrant@fas.harvard.edu

1 Introduction

Autonomous software agents based on artificial intelligence have many potential applications at the intersection of public health and social work, known as public health social work. These applications are enormously promising, but often pose novel ethical problems for researchers that can be difficult to assess, much less to resolve. The aim of this chapter is to show how analytical tools from moral philosophy and theoretical computer science can be combined to better understand these problems and to develop strategies for addressing them in practice.

I will focus on a specific set of problems that arise in the development of public health social work interventions based on artificial intelligence. Specifically, I will focus on a class of problems that I will call "beneficence problems" (the rationale behind the name will become clear below). Beneficence problems occur in the context of public health social work interventions that are partially planned, in the field, with the help of an artificially intelligent autonomous software agent (call these *AI planning interventions*).

Consider cases like the following. Suppose that researchers are developing a public health social work intervention to be conducted by some intervention team with some target population. The goal of the proposed intervention is to provide some specific set of benefits to the target population, such as (for example) reducing the incidence of HIV in the population. In the field, a software agent will assist the intervention team by recommending an intervention plan, consisting of some sequence of actions to be performed by the intervention team. The agent's goal is to identify and recommend the intervention plan that, if carried out, would benefit the target population in the intended way to the greatest extent

possible. The intervention team believes that the intervention plans that will be recommended by the agent are likely to benefit the target population, considered as a whole, to a greater extent than alternative interventions they could perform. However, they also believe that the software agent may under some foreseeable conditions recommend a plan that, while well-suited to maximizing benefits to the target population, will also pose a significant risk of harm to some of that population's members. Moreover, they have the capability to predict the individual- and population-level expected benefits and harms of particular intervention plans the agent considers. What should the researchers do?

The answer is far from obvious. Cases like this one pose a dilemma between two moral duties whose importance is widely recognized in the fields of both public health and social work. The first is the duty not to conduct interventions that are expected to harm others, otherwise known as the *duty of non-maleficence*. The second is the duty intervention teams have to conduct interventions that they expect will provide the greatest possible benefits to the populations they work with, otherwise known as the *duty of benefit maximization* The challenge posed for the design of the autonomous agents involved in such interventions is how to ensure that the intervention plan that is eventually conducted by the intervention team will demonstrate appropriate respect for both duties. Call the problem of meeting this challenge in a particular AI planning intervention a beneficence problem, and call an AI planning intervention that poses a beneficence problem an AI intervention*.

As we shall see, it can be difficult in practice to modify the software agent used in interventions of this kind so that there are reasonable guarantees that both duties just mentioned will be respected. That is, it can be difficult to ensure that the intervention plans the agent recommends will both (1) minimize the expected risks of the intervention to a degree compatible with the duty of non-maleficence and (2) maintain the expected benefits of the intervention at a level consistent with the duty of benefit maximization. On the one hand, if the balance researchers strike between minimizing expected harms and maximizing expected benefits weights expected benefits too heavily, favoring the duty of benefit maximization, then the risks posed to particular individuals may be too significant to be morally justifiable. On the other hand, if the balance struck weights expected harms too heavily, favoring the duty of non-maleficence, then the intervention may be

rendered ineffective enough to obligate researchers to abandon the proposed intervention in favor of a more beneficial alternative. Either kind of failure would render the proposed intervention morally impermissible to conduct, in the absence of further modifications.

My ultimate objective in what follows will be to provide practical guidance about how to approach designing the software agents used in AI interventions* so that both extremes are avoided. The plan for the chapter is as follows. To make the discussion more concrete, I begin in section 2 by setting out a real-world beneficence problem currently being faced by researchers at the Center for Artificial Intelligence in Society at the University of Southern California. In section 3, I provide a more formal definition of beneficence problems and set out some necessary technical background. Since implementing appropriate safety constraints requires a working understanding of the moral duties involved, section 4 discusses the duties of non-maleficence and benefit maximization in more detail. Section 5 proposes a framework, offered in a tentative spirit, for addressing conflicts between the two duties in public health social work interventions. Section 6 sketches some potential strategies for operationalizing this framework in the context of AI planning interventions. In section 7, I close by drawing some conclusions about the concrete beneficence problem described in section 2.

2 Case study: adapting TND Network for homeless youth

Researchers at the Center for Artificial Intelligence in Society (CAIS) are in the process of adapting a drug abuse prevention program called TND (Towards No Drug Abuse) Network for use in residential shelters for homeless youth. The program divides participants into small groups of approximately five members. The goal of the program, which relies heavily on peer interaction, is to use the power of peer influence to help participants develop the attitudes, skills, and confidence required to resist drug abuse. This program has been demonstrated to be effective in schools with at-risk youth—an important result, given how difficult drug abuse has proven to address—but current versions have a significant downside. In a recent large-scale trial of TND Network, Valente et al. (2007) found that particularly at-risk participants failed to benefit from the program. In fact, the program seemed to place them at even higher risk of future drug abuse. Valente et. al. hypothesized that this increased risk was the result of *deviancy*

training, a well-documented phenomenon[1] 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.

Researchers at CAIS are currently attempting to modify the program in two ways.[2] First, to make it more suitable for homeless youth at high risk of serious drug abuse living temporarily in residential homeless shelters. Second, to address its deviancy training effects. To achieve the second goal, CAIS developed an autonomous software agent to help improve how the program divides intervention participants (the current residents of a medium-term residential shelter for homeless youth) into groups. The agent's goal is to maximize the intervention's positive effects of peer influence on participants' future drug use, and minimize its negative (deviancy training) effects.

To do this, the agent uses a peer influence model (and more specifically, a linear threshold model; see Kempe, Kleinberg, and Tardos (2003)) to predict how attitudes about drug abuse will spread from person to person as a result of the intervention. (The model first predicts changes in participants' social networks that will occur as a result of participation, based on who is assigned to their participation group.) The intervention team conducts interviews with each participant in order to acquire information about their past drug use, who is in their personal (or "egocentric") social network, how strongly they are related to those individuals, and to what extent those individuals are known by the participant to have used drugs, currently or in the past. The agent's influence model is then updated based on this information, allowing it to make predictions about how various choices of intervention groups will affect participants' future drug use behavior. The agent then attempts to identify and recommend the "optimal" set of intervention groups—the set of intervention groups predicted by its model to result in the greatest aggregate (or population-level) reduction in future drug abuse by all participants.

Researchers hope that this modification to TND Network, which we can call *TND Network-SA* ("SA" for "Software Agent"), will render the program even more effective at reducing drug abuse by minimizing potential

[1] Dishion, McCord, and Poulin (1999).

[2] See Center for Artificial Intelligence in Society (2017) for more information.

deviancy training effects. This is particularly important given that their target population, homeless youth, has an unusually high proportion of individuals who currently abuse or have abused drugs in the past, which increases the risk that the intervention will lead to deviancy training by strengthening participants' social ties to other youth who view drug abuse favorably.

On early tests on sample data, however, researchers discovered a potential beneficence problem. Researchers fed previously collected data about the social networks of members of a particular population of homeless youth in Los Angeles, California (where the center is based) into the agent's influence 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 aggregate 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 risk at the population level.

CAIS researchers judged that the model's predictions here were reliable enough for the risk of significant harm here to be genuine if the hypothetical intervention were actually conducted. As a result, the intervention poses a beneficence problem, as defined above.

3 Beneficence problems

In this section, I provide a more formal definition of beneficence problems in general. In the field of artificial intelligence, a *planning problem* is the problem of identifying the sequence of actions that is best suited to achieving some goal. A particular planning problem can be defined by specifying the following:

1. An *agent* of some kind (e.g., a human, a software agent, a robot, or a team consisting of some combination thereof);

2. A *goal* to be achieved by that agent;
3. A set of possible *states* of the world—particular ways for the world to be;
4. A set of possible *actions* that the agent could perform in order to achieve that goal;
5. The agent's *beliefs* about the starting state of the world, e.g., in the form of a probability distribution over the states referenced above;
6. An optional set of *constraints* on how the goal can be achieved, used to rule out as possible solutions to the problem actions or sequences of actions known to be infeasible or otherwise undesirable for the agent to perform.[3]

In a *beneficence problem*, the agent in question includes both (a) a human team that is planning a public health social work intervention, and (b) a software agent that will assist that team in planning the intervention. The goal to be achieved is the goal of maximizing benefits of a particular kind for some target population. The states are various possible features of the intervention context that are relevant to predicting the intervention's effects. Both the intervention team and the agent have relevant beliefs about the starting state of the world prior to the intervention. The intervention team has information about the starting state, and supplies that information to the software agent in the form of input. This input is used to generate a formal representation of the starting state in terms that the agent is capable of understanding. There may or may not be any preexisting constraints on how the team achieves the relevant goal. The software agent and the intervention team will plan different aspects of the intervention, depending on what planning tasks are being offloaded to the software agent.

We need to specify two further features of the planning problem to complete our definition of beneficence problems. The first is that it has an explicitly moral constraint on how the goal of maximizing benefits for the target population can be achieved: that it be reasonable to expect that the intervention plan that the intervention team conducts based on the agent's recommendations will appropriately respect the duties of benefit maximization and non-maleficence. Call this the *beneficence constraint*. The second is that the agent is capable, or could be made capable, of providing useful predictions about the expected benefits and harms of the intervention

[3] Russell and Norvig (2010).

for particular members of the target population.

This last feature is what distinguishes AI planning interventions that pose beneficence problems (AI planning interventions*) from other public health social work interventions that generate similar conflicts between the duties of benefit maximization and non-maleficence. With the greater ability to assess individual-level effects that AI interventions* offer comes a greater responsibility to minimize potentially harmful effects. This leads to difficult questions about how to resolve the associated dilemma between beneficence and non-maleficence.

4 Moral duties and beneficence problems

A *moral duty*, as I will use the term here, is a relatively general moral obligation that applies across a range of possible contexts. As I have mentioned, beneficence problems pose a dilemma between two different and sometimes conflicting moral duties that public health social work professional have: maximizing the benefits of interventions they conduct and avoiding causing harm to particular individuals. These two duties are combined in a foundational principle of human subjects research ethics, the principle of beneficence. The Belmont Report states the principle as follows:

> "Persons are treated in an ethical manner not only by respecting their decisions and protecting them from harm, but also by making efforts to secure their well-being. Such treatment falls under the principle of beneficence. The term 'beneficence' is often understood to cover acts of kindness or charity that go beyond strict obligation. In this document, beneficence is understood in a stronger sense, as an obligation. Two general rules have been formulated as complementary expressions of beneficent actions in this sense: (1) do not harm and (2) maximize possible benefits and minimize possible harms."[4]

As the last sentence brings out, there are really two principles here. The first, "do not harm," is ancient in its origins,[5] and is commonly referred to as

[4] National Institutes of Health (1979).

[5] The Hippocratic Oath, for instance, was written as early as the fifth century BC, and requires physicians to swear as follows: "I will utterly reject harm and mischief" Wikipedia (2017).

the principle of *non-maleficence*. The second, "maximize possible benefits and minimize possible harms," is a version of the utilitarian injunction to bring about "the greatest good for the greatest number." I will refer to it as the principle of *benefit maximization*. The duties of non-maleficence and benefit maximization are widely accepted in public health and social work, whether or not human subjects research is involved, though there is significant controversy over how to address conflicts between the two duties.[6]

It's worth emphasizing that the benefits the duty of benefit maximization refers to are expected benefits, net benefits, and aggregate benefits. They are *expected benefits* because there is uncertainty involved. The intervention could have various possible effects, and researchers are really only in a position to estimate how beneficial or harmful those effects might be under various possible conditions, and assess the expected value or disvalue of the intervention across the full range of possible ways things could turn out. They are *net benefits*, because benefits are allowed to compensate for harms in determining the total benefit provided, at least for some types of harms and benefits. (The harms might be greater than the benefits, in which case the net benefit would be negative, a net harm.) And they are *aggregate benefits*, because they are calculated by summing the expected benefits for each individual to produce an aggregate score representing the total expected net aggregate benefit for the population as a whole. We can also call them *population-level* benefits for this reason, as opposed to *individual-level* benefits.

The duty of non-maleficence refers to expected harms in the same sense that the duty of benefit maximization refers to expected benefits. Since an intervention's effects cannot be predicted with certainty, public health social work professionals are enjoined to avoid risks that they are in a position to foresee, and weight those risks both by their severity and by how likely they

---

The more familiar phrasing, "first, do no harm," appears to have been coined by the English surgeon Thomas Inman in the 19th century Sokol (2013).

[6] On the public health side, see for instance Mann (1997) and Childress et al. (2002). On the social work side, the NASW's Code of Ethics explicitly endorses a version of benefit maximization ("The primary mission of the social work profession is to enhance human well-being and help meet the basic human needs of all people, with particular attention to the needs and empowerment of people who are vulnerable, oppressed, and living in poverty") and repeatedly enjoins social workers to avoid activities that would harm clients. See National Association of Social Workers (2008).

are to occur. The level of expected risk an intervention poses is therefore both a function of the likelihood and severity of the relevant risks. However, the relevant harms are neither net nor aggregate: the duty directs public health social work professionals to avoid expected harms to particular individuals, regardless of whether they are compensated for by benefits to those same individuals or to others.

It will simplify the discussion in what follows to confine our attention to cases where the expected harms and benefits of an intervention are different sides of one and the same coin: the benefits alleviate some potentially harmful condition a person can be in (e.g., being disposed to abuse drugs), whereas the harms make that condition worse. This is true in TND Network-SA: the relevant type of harm is being influenced to increase future drug abuse behaviors; the relevant type of benefit is to decrease future drug abuse behaviors. Restricting our attention in this way will let us set aside issues of intrapersonal compensation, such as determining whether a person is harmed on balance when they are made more likely to abuse drugs but less likely to contract HIV.

Some further distinctions between the two duties will also be helpful for our purposes. One such distinction that is worth making explicit (even though it's obvious) is that the duty of non-maleficence requires that specific individuals be treated in certain ways, whereas the duty of benefit-maximization requires that larger populations be treated in certain ways. Call moral duties of the former kind *individual-oriented duties* and moral duties of the second kind *population-oriented duties*. One thing to notice right away is that the agent used in an AI intervention* will need the capability to predict both individual- *and* population-level effects in order to help the intervention team make determinations about whether the duties of maleficence and benefit maximization are satisfied.

Another crucial distinction between the two duties for our purposes is that the duty of non-maleficence is what is sometimes known in moral theory as a "perfect duty," whereas the duty of aggregate benefit maximization is an "imperfect duty." On at least one way of understanding the distinction, *perfect duties* specify types of actions that are morally required or morally prohibited. The duty of non-maleficence, then, is a perfect duty because it prohibits us from performing any action that would harm another person. *Imperfect duties*, by contrast, require us to pursue certain goals. The duty to maximize benefits is most naturally understood as an imperfect duty,

as it requires those designing and conducting public health social work interventions make reasonable efforts to maximize the expected aggregate benefits of those interventions. However, the duty does not specify what kinds of means should be used to achieve this goal: it does not require or prohibit actions of any specific kind, except to say that the interventions in question should be well-suited to promoting the goal of maximizing population-level benefits.

There are also (a) imperfect and individual-oriented duties and (b) perfect and population-oriented duties, some of which are highly relevant for thinking about researchers' obligations with respect to beneficence problems. Space constraints prohibit more than a cursory discussion of these additional duties and others, but it's worth mentioning a few in particular. First, social workers are generally understood, as acknowledged by the National Association of Social Work's code of ethics, to have a general duty to promote the interests of their clients, at least in certain respects.[7] This includes ensuring that services provided to clients—such as enrolling them in a public health intervention—are reasonably expected to be of benefit to them, except under special circumstances. This duty is both imperfect and individual-oriented.

Second, public health social work professionals are generally understood to have a moral duty to ensure that the burdens and benefits of the interventions they conduct are fairly distributed among various subpopulations.[8] This is a perfect and population-oriented duty. Researchers should take special care, then, when designing AI interventions* that may impose unfair burdens on some subpopulation. In AI interventions* as I have described them, the information about social status necessary to determine whether this duty is satisfied by particular intervention plans may well not be represented in the formal model the agent uses to plan the intervention. Building this information into the model can help the researchers developing the intervention to determine whether the agent is like to recommend plans that may distribute benefits unfairly, and could potentially be used by the agent itself as it assists attempts to identify

[7] National Association of Social Workers (2008).

[8] See Faden and Shebaya (2016) regarding the field of public health. The NASW code of ethics makes clear that social workers are obligated to promote social justice, including the fair distribution of resources of various kinds, in their work National Association of Social Workers (2008).

and recommend potentially desirable plans.

I will set aside these further duties in what follows, and confine my attention to the duties of benefit maximization and non-maleficence. However, these other moral obligations are worth keeping in mind, and it is important to note that public health social workers have other moral duties that will eventually need to be factored in as well when beneficence problems are addressed.

One reason the distinction between imperfect and perfect duties is useful for our purposes is that it maps roughly onto the distinction between goals and constraints in the literature on AI planning. Imperfect duties specify moral goals that human agents should pursue; perfect duties specify moral constraints that human agents need to observe as they pursue these goals and others. The similarity here is not just superficial: it mirrors the way that perfect and imperfect duties typically interact with one another. Insofar as the two conflict, perfect duties are generally understood to have priority over imperfect duties. The constraints on action imposed by perfect duties, that is, restrict the range of actions an agent is morally permitted to perform as she pursues the goals imposed on her by imperfect duties. Ordinarily at least, an agent is not permitted to violate a perfect duty for the sake of better promoting the goal an imperfect duty obligates her to pursue. We will consider some potential exceptions to this general rule below.

This suggests a natural strategy for accommodating perfect and imperfect duties in the design of AI interventions*: accommodate imperfect duties through planning goals; accommodate perfect duties through planning constraints. In an AI intervention*, the software agent assists the intervention team by planning certain aspects of the intervention on their behalf. In order to help the intervention team avoid violating imperfect duties, appropriate changes can be made to the goals of the planning problem the agent seeks to solve on their behalf. In order to help them avoid violating perfect duties, changes can be made to the planning problem's constraints.

I'll talk more about how this might work in section 6, but first we should consider what obligations public health social work professionals have when the duties of non-maleficence and benefit maximization conflict.

5 A framework for resolving conflicts

Ethicists working in a variety of fields have long recognized that the duties of beneficence and non-maleficence can come into conflict in ways that generate difficult ethical dilemmas. The Belmont Report acknowledges this possibility early on:

> "[The] role of the principle of beneficence is not always so unambiguous. A difficult ethical problem remains, for example, about research that presents more than minimal risk without immediate prospect of direct benefit to the children involved. Some have argued that such research is inadmissible, while others have pointed out that this limit would rule out much research promising great benefit to children in the future. Here again, as with all hard cases, the different claims covered by the principle of beneficence may come into conflict and force difficult choices."[9]

A beneficence problem presents researchers with a dilemma with exactly this structure: in some foreseeable intervention contexts, the software agent may recommend an intervention plan that maximizes expected aggregate benefits for the target population by generating expected benefits for some of its members and expected harms for others. The duties of non-maleficence and benefit maximization appear to conflict in these cases, posing the question of how to resolve the conflict.

I mentioned above that perfect duties typically place stringent constraints on how imperfect duties may be pursued. Given this, a natural first thought is that these conflicts are easy to resolve: researchers are simply prohibited from imposing expected harms on particular individuals for the sake of producing greater aggregate benefits for the larger population. However, there are good reasons to think that there will be many exceptions to this general rule in the context of public health in general and social work in particular.

There are various ways to defend the claim that public health social work interventions may sometimes violate the duty of non-maleficence, consistent with being morally justifiable on balance, but one standard (if

[9] National Institutes of Health (1979).

not uncontroversial) argument is as follows.[10] The moral justification for particular public health interventions is often understood to derive from the justification of the public health system as a whole. What justifies the sum total of interventions conducted by a public health system, in spite of the fact that particular interventions impose net costs (including harms and other kinds of costs) on some individuals, is that each individual has a strong reason to prefer living under a public health regime that allows individual interventions to trade off costs for some individuals to achieve comparatively greater benefits for others. That reason is that a system prohibited from intervening to improve the health of a population at a cost to some of its members would be far less effective overall. Provided that the system is designed to ensure that the burdens an individual is subjected to from any given intervention are compensated for by benefits from other interventions, the net result should be that everyone experiences greater benefits on balance from systems that allow such tradeoffs to be made than those that do not.

Excise taxes on cigarettes, for example, improve aggregate wellbeing in a population by reducing the incidence of various diseases, but accomplish this at a cost to those smokers who are not persuaded to reduce their intake. Similarly, access restrictions on performance-enhancing drugs such as Ritalin benefit populations by reducing their abuse, but leave some individuals who would benefit from them worse off. What makes these tradeoffs morally acceptable? According to the foregoing argument, these tradeoffs are justifiable because those individuals who are negatively affected nonetheless have good reason to prefer living under a public health regime that allows them to be made—provided that there are restrictions on how significant the expected burdens can be in particular cases. Call this argument the *higher order justification argument*.

This argument is at least initially plausible as it applies to public health social work interventions in particular. The individual members of the populations served by public health social workers have a reason to prefer being served a system that allows public health social work interventions to make the kinds of tradeoffs just described, provided that safeguards are in place to ensure that the end result is that each individual expects to benefit more if those tradeoffs are allowed in some cases than if they are not. This justifies relaxing the requirements of the duty of non-maleficence as it applies

[10] My presentation of the argument below owes much to Faden and Shebaya (2016), section 2.1.

to individual interventions to at least some degree, allowing for net expected harms to individuals in some cases.

Suppose, then, that this argument, or another one with the same conclusion, is correct. If so, then some exceptions to the principle of non-maleficence as it applies to public health social work interventions can be justified. It remains to be seen how public health social work professionals should go about determining whether the expected harms imposed by a specific public health intervention are morally justifiable. This is of course a complex question requiring extensive treatment, but a few preliminary points can be made. I will briefly sketch a framework, based on existing work in public health ethics,[11] that offers a useful and at least initially plausible proposal about how public health social work professionals should go about negotiating conflicts between the duties of non-maleficence and benefit maximization.

The framework posits three moral duties that it takes to apply to public health social work interventions where the duties of non-maleficence and benefit maximization conflict. The first two duties apply to interventions considered in isolation; the third applies to interventions considered in light of available alternative interventions.

First, the duty of necessity. Suppose an intervention is believed to offer certain magnitude of expected benefit to the target population and a certain magnitude of expected harm to some of its individual members. The duty of necessity requires that the expected harms of the intervention be necessary to achieve its expected benefits. If the intervention can be modified in ways that reduce those expected harms while maintaining a comparable level of expected benefit, then the intervention must be modified accordingly. Doing so reduces the severity of the intervention's violation of the duty of non-maleficence. It also increases the degree to which the duty of benefit maximization is satisfied by the intervention, since expected harms for individual members of the target population entail reduced aggregate expected benefits for the population considered as a whole.

Second, the *duty of proportionality* requires that the expected benefits of an intervention be great enough to justify the expected harms involved—that the two be suitably "proportional" to one another. The idea here is that sufficiently great benefits can outweigh comparatively minor harms. We can

[11] The framework I develop is based in large part on the framework proposed in Childress et al. (2002), though it differs in important respects.

see this logic at work in an argument mentioned in the second quote from the Belmont report above (my emphasis added):

> "A difficult ethical problem remains, for example, about research that presents more than minimal risk without immediate prospect of direct benefit to the children involved. Some have argued that such research is inadmissible, while *others have pointed out that this limit would rule out much research promising great benefit to children in the future*."

The implicit argument here is that the expected benefit of research of this kind for future children is so great that it justifies the uncompensated risks posed to the children that will be enrolled as research subjects, because the relevant risks and benefits are "proportional" given the relative importance of those two moral considerations.

Why should we accept the duty of proportionality as a necessary condition on permissible interventions? Here is one argument. That an intervention is expected to generate substantial benefits for the target population counts in its favor, from a moral point of view, relative to alternative options. Other things being equal, it is better to perform interventions that produce greater aggregate benefits, as recognized by the duty of benefit maximization. Further, that an intervention poses a significant expected risk of harm to some of the target population's members counts against it, again from a moral perspective. But the two kinds of moral considerations just mentioned are not equally important. Expected risks count more against an intervention than expected benefits of a comparable magnitude. This way of thinking about the comparative moral importance of expected risks and expected benefits recognizes the fact that the duty of non-maleficence imposes a meaningful constraint on the way expected benefits can be pursued by public health professionals. If the two were not weighted differently, in this way, then the duty of non-maleficence would not impose a meaningful constraint.

Third, the *duty of tradeoff optimization*. Unlike the first two duties, this duty requires that the expected benefits and harms of an intervention be compared to those of viable alternative interventions. Suppose an intervention team is faced with a choice between possible interventions that all satisfy the duties of necessity and proportionality—various versions of their own intervention, perhaps, and other alternative interventions

they could perform. How should they choose among these possibilities? According to the duty of tradeoff optimization, researchers are obligated to choose the intervention from the set that strikes the best balance between expected benefits and expected harms—the expected benefit/expected harm profile that best satisfies the duty of proportionality.[12] The thought here is that the elements of a set of interventions can each satisfy the duty of proportionality, consistent with it being true that the expected benefits and harms are more "proportional" for some of the interventions in the set than for others. More proportional interventions strike a more morally desirable balance between the conflicting demands of the duty of non-maleficence and benefit maximization, and so should be preferred.

The tentative proposal I am offering here is that AI interventions* that satisfy all three of these duties—of necessity, of proportionality, and of tradeoff optimization—are morally justifiable insofar as the duties of non-maleficence and benefit maximization is concerned. This is just a first step and requires much further refinement and assessment in light of other ways of approaching conflicts between the two duties. Moreover, other moral duties, such as those mentioned above, are not yet taken into account by the framework. However, I think the framework is at least initially plausible as far as it goes, and—despite its partial nature and the complexity of the issues involved—we can use it to draw a number of tentative conclusions about how AI interventions* should be designed.

We need to introduce one final element into the mix before we get down to brass tacks and consider how to operationalize the framework just described. As I mention above, autonomous agents such as those considered here make it possible to make more accurate individual-level predictions about an intervention's benefits and risks than was previously possible. One question researchers need to consider as they attempt to satisfy the foregoing three moral duties is to what extent they are obligated to use these tools to improve their estimates of those individual-level risks. This question is important because the goal of achieving a more accurate and comprehensive estimate of the expected individual-level harms of an AI intervention* is to some extent in tension with the goal of maximizing expected benefits for the population as a whole. One reason for this is

[12] How to understand "best" in the preceding sentence is a further question that I will set aside here, but it will almost certainly depend on the nature of the argument that is given for allowing exceptions to the duty of non-maleficence.

that there is only so much research time that can be spent improving an intervention in the variegated respects that are morally desirable. Another reason is that many ways of improving a software agent's estimates of expected individual-level effects for particular intervention plans introduce additional computational complexity, which can in itself reduce the agent's ability to optimize population-level benefits.

6 Operationalizing the framework

We can now return to the question of how researchers should proceed when they believe an AI planning intervention they are developing has a beneficence problem. Assume researchers are designing the software agent associated with such an intervention.

Let's begin by revising our original definition of beneficence in light of the framework I have proposed for resolving conflicts between the duties of benefit maximization and non-maleficence. Specifically, we can remove the beneficence constraint and replace it with a new constraint: that it be reasonable to expect that the intervention plan that the intervention team actually conducts, based on recommendations from the agent, will satisfy all three of the moral duties just described—the duty of necessity, the duty of proportionality, and the duty of optimal tradeoffs. This in itself is progress, as the redefined problem gives us a better idea of what is required to respect the duties of benefit maximization and non-maleficence when they come into conflict.

As is probably clear by now, it is extremely unlikely that this constraint can be fully operationalized in the design of the agent, so that it never recommends plans inconsistent with it. Determining whether the three moral duties involved are satisfied by a particular intervention requires time, experience, consideration of the merits of possible alternative interventions, and open discussion with other professionals and stakeholders (including members of the target population) with a diverse enough set of perspectives to minimize the chances that something important will be overlooked.[13] Formulating general and operationalizable rules that will provide reasonable assurances that they are never violated by the intervention plans that the

[13] There are also independent reasons to consult stakeholders in particular. See Childress et al. (2002)'s discussion of a principle they call "public justification."

software agent in an AI intervention* recommends is almost certainly impossible.

What this means is that human judgment will always be required—the intervention team will need to assess the intervention plans recommended by the agent to determine whether it satisfies the duties of necessity, proportionality, and optimal tradeoffs. However, the software agent might nonetheless be designed in ways that make it less likely than it would otherwise be that the duties will be violated in the field. I will sketch out three possible strategies.

First, the agent could be designed so that, when it recommends a particular intervention plan, it provides the intervention team with supplemental information that is relevant to determining whether the recommended plan satisfies the constraints of necessity, proportionality, and optimal tradeoffs. Call this the *interpretability approach*.

How might the interpretability approach look in practice? Researchers at CAIS are currently redesigning the agent to be used in TND Network-SA to supply the intervention team with more information about the expected individual- and population-level harms and benefits associated with the plans it recommends. With this information in hand, researchers will be in a much better position to assess whether the moral duties associated with beneficence problems are satisfied. An obvious first step here is for the agent to be programmed to calculate the change in expected value predicted to be achieved by conducting a particular intervention plan, relative to conducting no intervention, for each individual member of the target population, as well as for the target population considered as a whole. This would require, inter alia, that the agent's model include what we can call an "individual-level value function." Let an *individual-level value function* be a function from pairs of states and individual members of a target population to real numbers on the interval between 0 and 1. The value assigned to a pair is the value for the individual of being in that state.

If the change in expected value for an individual (or population) associated with a particular intervention plan is positive, then the intervention is expected to benefit the individual (population) relative to doing nothing; if it is negative, then it is expected to harm the individual (population) Providing this information to the researchers lets them know how conducting the recommended intervention plan is expected to positively and negatively affect both particular individuals and the population

considered as a whole. This would at least put the researchers in a better position to assess whether the proportionality and necessity constraints are satisfied, in conjunction with the information already available to them.

Second, the agent could be redesigned so that it automatically rules out plans with features that can be determined in advance to violate one of the three duties in the framework. This could be accomplished using what artificial intelligence researchers call "safety constraints." Call this the *safety constraint approach*. How might this work? One possibility is to model the planning problem solved by the agent as a constrained partially observable Markov decision process (constrained POMDP). A *partially observable Markov decision process* (POMDP) is a mathematical tool for modeling planning problems where there is uncertainty about the initial state and how it will evolve over time based on actions performed by the agent.[14] This makes POMDPs useful for modeling beneficence problems, where both kinds of uncertainty are present. In a constrained POMDP, the agent's goal is to maximize the value of one function while simultaneously bounding the values of other functions within a specified range.[15] Researchers could determine a level of expected harm that they judge would be impermissible to impose on an individual in the preponderance of foreseeable starting intervention states, and direct the agent to bound the expected change in individual-level value functions associated with each member of the target population accordingly. An advantage of this strategy is that the intervention team would have to reject recommended strategies less often, which could dramatically reduce the amount of time it takes to identify a morally acceptable intervention plan.

Another variant on the safety constraint approach would be to generate recommended intervention plans using a two-step procedure. In the first step, the agent would eliminate all possible end states that exceed a level of expected harm for individuals that they have decided is unlikely to be acceptable. It would then work backwards by considering the states that are possible in the preceding step, and eliminating any that have a high probability of resulting in one of those end states. This could be repeated until the step immediately following the initial state is reached. Pynadath and Tambe (2001) show how this can be done in planning problems that can be

[14] See Russell and Norvig (2010) chapter 17, section 4.

[15] See Isom, Meyn, and Braatz (2008).

modeled as vanilla Markov decision processes. Whether the strategy could be extended in a computationally tractable way to beneficence POMDPs is beyond the scope of this chapter, but is worth exploring.

Third, researchers could implement what Pynadath and Tambe (2001) call safety conditions. *Safety conditions* specify conditions under which the software agent assisting a human team should transfer control to its human operators in order to request more information or a decision of some kind. The general idea behind safety conditions is that the human operators of a software agent designed to solve a planning problem often have important information about the problem that the agent does not. In many cases, this information cannot be reliably captured or processed by the agent as it decides how to act, regardless of how well it is designed.[16] In principle, such conditions could be used to direct the agent in an AI intervention* to present the intervention team with a selection of possible intervention plans, each with a different risk/benefit profile. If feasible, this would put the intervention team in an even better position to determine how to satisfy our three moral constraints, since they will have more information about the effects of different possible intervention plans. Note that, for this to work, the agent will need to choose alternative plans in an intelligence way, by recommending plans that make the tradeoff between expected individual-level harms and population-level benefits in meaningfully different ways. Call this the *safety condition approach*.

These three approaches could of course be combined, if computational limitations allow. Various other strategies are also of course possible. In any case, strategies like these have considerable potential to help the intervention teams in AI interventions* select an intervention plan that promises morally desirable expected population-level benefits without imposing morally unjustifiable expected individual-level harms.

7 Conclusion

By way of closing, I want to return to the case study we started with, and evaluate it in light of the ensuing discussion. As that discussion suggests, beneficence problems can be and often are difficult to resolve. However, what to say about our case study in particular is, I think, relatively clear.

[16] Pynadath and Tambe (2001).

First, the researchers in our case study knew one thing with confidence from previous research, and that was that putting all of the highest risk participants into a single group would expose them to serious additional risk. The knew that building or strengthening social ties among individuals with a history of serious drug abuse makes them significantly more likely to abuse substances in the future, or to abuse them to a greater extent. They also knew that subjecting the high-risk participants to this increased risk of future substance abuse would in turn increase their risk of a variety of negative outcomes, including death. Importantly, they knew these things without needing to consult the predictions made by the agent's predictive model. By contrast, the potential benefits of the proposed intervention for the other participants were more speculative. While TND Network had delivered promising results in previous trials, their modified version had yet to be tested, and targeted a different population in a different context. The duty of proportionality requires that the anticipated benefits of an intervention at the population be sufficient to justify any anticipated harms at the individual level. In the case at hand, it seems like a stretch to suggest that the well-known and quite serious risks that the high-risk group would be subjected to by the intervention plan could be justified by the comparatively more speculative benefits anticipated for the larger participant group.

Second, the duty of necessity requires that the anticipated harms of an intervention be necessary to secure its anticipated benefits. But in this case, there was a clear alternative to enrolling the highest risk youth in question and placing them all in the same intervention group: simply declining to enroll them in the first place. After all, the reason that confining the highest-risk individuals to a single participation group was predicted to maximize the population-level benefits of the intervention was that doing so would prevent them from having a negative influence on the lower-risk participants. Simply removing the highest-risk individuals from the intervention, and finding an alternative way to include them in the activities of the community, would allow the anticipated benefits for the lower-risk participants to be realized without putting the higher-risk group in harm's way.

Third, as mentioned above, social workers have special duties to their clients, including a duty to ensure that services provided to them can be reasonably expected to benefit them. It is hard to see how a social worker could, consistent with this duty, conduct such an intervention knowing that it was more likely to harm the higher-risk participants than to help them.

I conclude that the specific hypothetical intervention plan that began our inquiry would be morally wrong to conduct. However, I hope that it is by now clear that the impermissibility of that specific intervention plan does not cast significant doubt on the viability of TND Network-SA in general. By modifying the artificial agent used by the intervention in the ways I have suggested, researchers can help ensure that the interventions they eventually do conduct will meet a genuinely desperate need—helping homeless youth to avoid substance abuse—while at the same time showing appropriate respect for the interests of those homeless youth who are most at risk.

…

References

Center for Artificial Intelligence in Society (2017). *Social Network-Based Substance Abuse Prevention for Homeless Youth*. https://www.cais.usc.edu/projects/social-network-based-substance-abuse-prevention-for-homeless-youth/. Accessed: 2017-08-15.

Childress, James F. et al. (2002). "Public health ethics: mapping the terrain". In: *The Journal of Law, Medicine & Ethics* 30.2, pp. 170–178.

Dishion, Thomas J., Joan McCord, and Francois Poulin (1999). "When Interventions Harm: Peer Groups and Problem Behavior". In: *American Psychologist* 54.9, pp. 755–764.

Faden, Ruth and Sirine Shebaya (2016). "Public Health Ethics". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2016. Metaphysics Research Lab, Stanford University.

Isom, Joshua D., Sean P. Meyn, and Richard D. Braatz (2008). "Piecewise Linear Dynamic Programming for Constrained POMDPs". In: *Aaai-2008* 1, pp. 291–296.

Kempe, David, Jon Kleinberg, and Éva Tardos (2003). "Maximizing the spread of influence through a social network". In: *Kdd*, p. 137. arXiv: 0806.2034v2.

Mann, Jonathan M. (1997). "Medicine and Public Health, Ethics and Human Rights". In: *Hastings Center Report* 27.3, pp. 6–13.

National Association of Social Workers (2008). "NASW Code of Ethics (Guide to the Everyday Professional Conduct of Social Workers)". In:

National Institutes of Health (1979). "The Belmont Report". In: *The Belmont Report Ethical Principles and Guidelines for the Protection*

of Human Subjects of Research February 1976. arXiv: arXiv:1011.1669v3.

Pynadath, David V. and Milind Tambe (2001). "Revisiting Asimov's First Law: A Response to the Call to Arms". In: *8th International Workshop on Intelligent Agents VIII*, pp. 307–320.

Russell, Stuart and Peter Norvig (2010). *Artificial Intelligence: A Modern Approach*. Upper Saddle River, New Jersey: Pearson Education, p. 1151. ISBN: 9780136042594.

Sokol, Daniel K (2013). ""First Do No Harm" Revisited". In: *BMJ* 347.October, f6426.

Valente, Thomas W. et al. (2007). "Peer Acceleration: Effects of a Social Network Tailored Substance Abuse Prevention Program Among High-Risk Adolescents". In: *Addiction* 102.11, pp. 1804–1815.

Wikipedia (2017). *Hippocratic Oath*. https://en.wikipedia.org/wiki/Hippocratic_Oath. Accessed: 2017-08-15.