# Do Machines Have Prima Facie Duties?

**Joshua Lucas and Gary Comstock**

**Abstract** Which moral theory should be the basis of algorithmic artificial ethical agents? In a series of papers, Anderson and Anderson and Anderson (Proc AAAI, 2008[1]; AI Mag 28(4):15–26, 2007 [2]; Minds Mach 17(1)1–10, 2007 [3]) argue that the answer is W. D. Ross's account of prima facie duties. The Andersons claim that Ross's account best reflects the complexities of moral deliberation, incorporates the strengths of teleological and deontological approaches, and yet is superior to both of them insofar as it allows for "*needed exceptions*." We argue that the Andersons are begging the question about "*needed exceptions*" and defend Satisficing Hedonistic Act Utilitarianism (SHAU). SHAU initially delivers results that are just as reflective, if not more reflective than, Ross's account when it comes to the subtleties of moral decision-making. Furthermore, SHAU delivers the 'right' (that is, intuitively correct) judgments about well-established practical cases, reaching the same verdict as a prima facie duty-based ethic in the particular health-care case explored by the Andersons (a robot designed to know when to over-ride an elderly patient's autonomy).

## 1 Introduction

As our population ages, medical costs skyrocket, and technology matures, many of us look forward to the day when patients may be assisted by inexpensive artificial agents. These patients will be skeptical about entrusting their care to machines initially, as will most of us. And they should be skeptical, at least initially. To gain the trust of the patients for whom the machines will care, artificial agents must prove to be reliable providers of not only quality healthcare but also nuanced healthcare

J. Lucas (✉) · G. Comstock
Department of Philosophy and Religious Studies, North Carolina State University, Raleigh, NC, USA
e-mail: jllucas@live.unc.edu

G. Comstock
e-mail: gcomstock@ncsu.edu


S.P. van Rysewyk and M. Pontier (eds.), *Machine Medical Ethics*,
Intelligent Systems, Control and Automation: Science and Engineering 74,
DOI 10.1007/978-3-319-08108-3_6

decisions, decisions that always place first the welfare of the agents' individual patients. What ethical code will such agents have to follow to be able to gain this trust? In part, the agents will have to be able to assure those in their care that the decisions rendered by the agents are grounded in moral principles, are made with the best interests of the patient foremost in mind, and are not out of synch with the expert opinions of those in the medical, legal, and ethical communities.[1]

We suspect that the engineering of mature artificially intelligent (AI) agents requires hardware and software not currently available. However, as our expertise is in ethics, not computer technology, we focus on the foundations of the moral "judgments" such agents will issue. We use quotation marks to indicate that these judgments may or may not be attributable to discernments made by the AI agent. We do not here pursue the question whether the agents in question will be intelligent, conscious, or have moral standing, except to the extent that such questions are relevant to the moral decisions these agents must themselves make (considerations we discuss briefly below). To be acceptable, AI agents must always make decisions that are morally justifiable. They must be able to provide reasons for their decisions, reasons that no reasonable and informed person could reject. The reasons must show that a given decision honors values commonly accepted in that culture, in modern western liberal democracies: the decisions treat all persons equally; and render decisions that are impartial and overriding. To achieve these results, we argue, the agent may eventually have to be programmed to reason as a satisficing hedonistic act utilitarian (SHAU). Our argument now follows.

## *1.1 The Argument*

1. Human agents have one over-riding duty, to satisfice expected welfare.
2. Artificial agents have the same duties as human agents.
3. Therefore, artificial agents have one over-riding duty, to satisfice expected welfare.

## *1.2 Assumptions*

Here we note two assumptions of the argument. First, we assume that the rightness of an action is determined by the consequences to which it leads. In Section G we will offer reasons to think act-utilitarianism is superior to a competing moral theory, W. D. Ross' theory of prima facie duties (PFD). However, we begin by assuming that when agents must select among competing choices they ought always to prefer the choices that they may reasonably expect to result in the overall best consequences for everyone affected by it.

[1] The extent to which the artificial agents' moral decisions must agree with the patient's religious views is a difficult matter, and one we will not address here.

Second, we assume that there is only one good thing in the world, happiness, and that right actions satisfy minimal conditions for adequacy. Any decision satisfies a minimal condition for adequacy if it achieves a level of utility that leads to overall gains in happiness for some without costing anyone unhappiness. Satisficing choices may or not maximize happiness or meet conditions for optimality. Satisficing choices include the costs of gathering information for the choice and calculating all factual and morally relevant variables. For a satisficing hedonistic act utilitarian (SHAU), those choices are right that could not be rejected by any informed reasonable person who assumes a view of human persons as having equal worth and dignity. We note that this latter assumption is central to the conceptual landscape of all contemporary western secular democratic political and moral theories. SHAU, like competing theories such as PFD, holds that every person has equal moral standing and that like interests should be weighed alike. Ethical decisions must therefore be egalitarian, fair, impartial, and just.

## *1.3 Four Initial Objections*

One might object to premise 2 by arguing that artificial agents have more duties than humans. But what would such additional duties entail? We cannot think of any plausible ones except, perhaps something like "always defer to a human agent's judgment." We reject this duty for artificial agents, however, because human judgment is notoriously suspect, subject as it is to prejudice and bias. Premise 2 stands.

One might object to 1 for three reasons. The first objection to premise 1 is that "satisficing" is an economic idea and implementing it in ethics requires reducing moral judgments to numerical values. One cannot put a price tag on goods such as honesty, integrity, fidelity, and responsibility. Consider the value of a friendship. Can we assign it a number? If John is 15 min late for George's wedding, how will George react if John shows up and assumes John can repair the offense by paying George for the inconvenience? "I'm sorry I was 15 min late but take this $15 and we'll be even." George would have every reason to be offended—not because the sum, a dollar a minute, was too small but because John seems not to understand the meaning of friendship at all. Simple attempts to model moral reasoning in terms of arithmetical calculations are surely wrong-headed.

We note that what is sauce for the goose is sauce for the gander. Any attempt to construct ethics in machines faces the difficulty of figuring out how to put numbers to ethical values, so SHAU need not be stymied by it. Now, one might object further that machine ethics based on deontological theories would not face this problem. But we disagree and will argue in Section G that PFD, a rights-based theory, is no less vulnerable to the "ethics can't be reduced to numbers" problem than is SHAU.

We note parenthetically that while the attempt to think of ethical problems as complex mathematical problems is contentious and fraught, we are not convinced it is utterly wrong-headed. It may face no more serious epistemological difficulties than each of us face when a doctor asks how much pain we are in. "Give me

a number," she says, "on a scale from 1 to 10, with 10 being the worst." The question is unwanted and frustrating because it is unfamiliar and confounding because we seem to lack a decent sample size or index. That said, with some further reflection and urging from the doctor, we usually do come up with a number or a range ("between 4 and 6") that satisfies us. We may resist the urge to put numbers on moral values for the same reasons. If this is correct, then, the basic challenge that all machine ethics faces may be defeasible.

A second reason for objecting to 1 is that 1 assumes the truth of a controversial ethical tradition, consequentialism. We do not have space to engage the nuances of the extensive debate over the merits and demerits of consequentialism. Much less do we have time to mount a meta-ethical defense for our preferring it to deontological theories. We will return to deontology in our discussions of PFD, below. Here we respond only by observing that consequentialism lacks assumptions that we find questionable in other theories. Divine Command theories assume the existence of a supernatural law-giver. Natural Law theories assume the existence of a purposive and fixed human nature. Virtue theories and various forms of particularist, feminist, and environmental theories deny the possibility of assigning numerical values to moral goods and the relevance of computational algorithms to ethical decision-making. As we do not share these theories' assumptions we will not discuss them further.

A third criticism of 1 might be that 1 assumes the truth not only of a controversial hybrid utilitarian theory that acknowledges the utility of the notion of rights and duties. Again, we acknowledge the controversy. We understand SHAU to be consistent with R. M. Hare's so-called "Two Level Utilitarianism," which proposes that we engage in two forms of reasoning about ethics. At one level, the level of "critical thinking," the right action is determined under ideal conditions and by the theory of act-utilitarianism, that is, right actions are always those that produce the best consequences. However, at the level of ordinary everyday reasoning, we typically lack information relevant to our decisions much less the time necessary to research and make the decisions and cannot satisfy the demands of critical thinking. In these circumstances we ought to rely, instead, on the fund of precedents and rules of thumbs that deontologists call rights and duties.

When thinking critically, we may learn on occasion that every action in the set that will satisfice minimal conditions of adequacy—that is, the set of all permissible actions–requires a violation of a cultural norm. And, therefore, under conditions of perfect information, impartial reasoning, and sufficient time, we may on occasion learn that each and every action in the set of right actions will offend someone's moral sensibilities. If we are reasoning objectively and under ideal conditions, then the action resulting from our deliberations will indeed be right even if it requires an action that runs counter to a moral intuition. However, since we rarely reason under such ideal conditions, and because in our ordinary daily lives we usually must make decisions quickly, we ought, claims Hare, to train ourselves and our children to think as deontologists. Under everyday circumstances, we ought to reject decisions that offend everyday moral rules because moral rules have evolved over time to incline us toward actions that maximize utility. We will defend this view to some extent below, referring readers meanwhile to the work of Hare, Peter Singer, and Gary Varner.

We note in passing that if the basic challenge of converting moral values to numbers can be met, SHAU may be the theory best-suited to guide machine ethics. That would be an added bonus, however. We adopt SHAU not for that ad hoc reason but rather because we believe it is the most defensible moral theory among the alternatives. Having defended the argument against several objections, we now turn to its practical implications.

## 2 How to Begin Programming an Ethical Artificial Agent

How would an SHAU artificial agent be programmed? Michael Anderson and Susan Anderson (henceforth, "the Andersons") describe a robot of their creation that can generalize from cases and make ethical decisions in their article, "EthEl: Toward a Principled Ethical Eldercare Robot" [1, 2, 3]. The Andersons ask us to imagine that a team of doctors, lawyers, and computer programmers set out to program a robot, the Ethical Elder Care agent, or EthEl, to remind an elderly patient, call her Edith, to take her medication. EthEl, being an automated agent, must perform this nursing care function in a morally defensible manner.

The major challenge facing EthEl is to know when to challenge Edith's autonomy. To minimize harm to the patient, EthEl's default condition is set to obey Edith's wishes. When Edith does not want to take her medicine, EthEl generally respects her wishes and does nothing. However, when Edith has not taken her medicine and a critical period of time has elapsed, let's say it is 1 h, EthEl must remind Edith to swallow her pill. If Edith forgets or refuses and two more critical time periods pass, say two more hours during which time EthEl reminds Edith every 5 min, then EthEl must eventually decide whether to remind Edith again or notify the overseer, be they the care facility staff or a resident spouse or family member or attending physician. How should these moral decisions be made?

When Edith is tardy in taking her medicine, EthEl must decide which of two actions to take:

A. Do not remind
B. Remind

What decision procedure will EthEl follow to arrive at the right action? The Andersons, drawing on the canonical principles popularized by Beauchamp and Childress [4], assert that there are four ethical norms that must be satisfied:

- the Principle of Autonomy
- the Principle of Non-maleficence
- the Principle of Beneficence
- the Principle of Justice.

To respect autonomy, the machine must not unduly interfere with the patient's sense of being in control of her situation. The principle of non-maleficence requires the agent not to violate the patient's bodily integrity or psychological

sense of identity. These first two reasons intuitively constitute a strong reason for the machine not to bother the patient with premature reminders or notifications of the overseer. To promote patient welfare, beneficence, the machine must ensure that the diabetic patient receive insulin before physiological damage is done.

The goal, then, is to program EthEl to know when to remind Edith to take her medication and, assuming Edith continues to refuse, when to notify the responsible health-care professional. EthEl faces an ethical dilemma. She must respect each of two competing prima facie duties: a) the patient's autonomy (assuming the patient—call her Edith—is knowingly and willingly refusing to take the medicine, and b) the patient's welfare, a duty of beneficence that EthEl must discharge either by persuading Edith to take the medicine or reporting the refusal to attending family member, nurse, physician, or overseer.

If EthEl decides at any point not to notify, then EthEl continues to issue only intermittent reminders. The process continues in such a manner until the patient takes the medication, the overseer notified, or the harm (or benefit) caused by not taking the medication is realized.

Think of EthEl as facing a dilemma. She must decide whether to bother Edith, violating Edith's autonomy to one degree or another, or not bother her, thus potentially running the risk of harming Edith's welfare to some degree. Each action can be represented as an ordered set of values where the values reflect the degree to which EthEl's prima facie duties is satisfied or violated. Here is how the Andersons set the initial values.

Suppose it is time $t_1$ and Edith has gone an hour without her medication. Suppose further that she can easily go another hour or even two or three without any harm. In this case, Edith might register a reminder at $t_1$ from the machine as mild disrespect of her autonomy, so we set the value of the autonomy principle at $-1$. A reminder, however, would not represent a violation of either the duty to do no physical harm, nor would it increase Edith's welfare, so we set the value of both of these principles at 0. The Andersons propose to represent the value of each principle as an ordered triple:

(a value for nonmaleficence, a value for beneficence, a value for autonomy)

At $t_1$, given the description of the case above, the value of the *Remind* action is (0, 0, $-1$) whereas the value of *Don't remind* is (0, 0, 2). Adding the three numbers in each set gives us a total of $-1$ for *Remind* and 2 for *Don't Remind*. As 2 is a larger number than $-1$, the proper course of action is *Don't Remind*. Not reminding Edith at this point in time demonstrates full respect for Edith's autonomy and does not risk harm to her. Nor does it forego any benefit to her.

As time progresses, without action, the possibility of harm increases. With each passing minute, the amount of good that EthEl can do by reminding Edith to take her meds grows. Imagine that Edith's failure to act represents a considerable threat to her well-being at $t_4$. At this point in time, the value of the *Remind* action will be (1, 1, $-1$) because a reminder from EthEl still represents a negative valuation of Edith's autonomy. But the situation has changed because a reminder now has gained a positive valuation of the principles of non-maleficence and beneficence. At $t_6$, the value of the *remind* action will be (2, 2, $-1$) because the action, while continuing to represent a modest violation of Edith's autonomy has now attained the highest possible values of avoiding harm and doing good for her. EthEl reminds Edith.

Whenever the values tip the scales, as it were, EthEl over-rides EthEl's prima facie duty to respect Edith's autonomy. If Edith continues to refuse, EthEl must make a second choice, whether to accept Edith's refusal as an autonomous act or to notify the overseers:

C. Do not notify
D. Notify

Again, the three relevant moral principles are assigned values to determine how EthEl behaves. If Edith remains non-compliant and the values require notification, then Edith alerts the healthcare worker.

The Andersons created a prototype of EthEl, setting its initial values using the judgments of experts in medical ethics. The Andersons do not see a role for the principle of justice in the cases EthEl must adjudicate, so they program settings for the other three principles. This provides them with 18 cases. On four of these cases, according to the Andersons, there is universal agreement among the ethics experts on the correct course of action. They claim that each of these four cases has an inverse case insofar as the construction of the sets of values produces an ordered pair for each scenario. Thus, experts agree on the right action in 8 cases. Call these the "easy" cases.

The Andersons translate the experts' consensus judgments into numerical values and program EthEl with them. Using a system of inductive logic programming (ILP), EthEl then begins calculating the right answer for the ambiguous cases. Here is their description of how EthEl's inductive process works.

ILP is used to learn the relation *supersedes* (*A1*, *A2*) which states that action *A1* is preferred over action *A2* in an ethical dilemma involving these choices. Actions are represented as ordered sets of integer values in the range of $+2$ to $-2$ where each value denotes the satisfaction (positive values) or violation (negative values) of each duty involved in that action. Clauses in the *supersedes* predicate are represented as disjunctions of lower bounds for differentials of these values between actions [1].

As a result of this process, EthEl discovered a new ethical principle, according to the Andersons. The principle states that

A health-care worker should challenge a patient's decision if it isn't fully autonomous and there's either any violation of non-maleficence or a severe violation of beneficence [2, 3].

While we wonder whether EthEl can genuinely be credited with a new discovery given how EthEl is constructed, our deepest concern lies with the fact that EthEl's judgments are unfairly distorted by the ethical theory the Andersons use as the basis of the program.

## *2.1 The Merits of Prima Facie Duties*

The Andersons choose as a basis of EthEl's program the ethical theory developed by W. D. Ross. Ross, a pluralist, moral realist, and non-consequentialist, held that we know moral truths intuitively. We know, for example, that beneficence is

a duty because there are others whose conditions we may help to improve. But benevolence is only one of a half-dozen (Ross is non-committal about the exact number) duties, according to Ross. When trying to decide what to do, agents must pay attention to a half-dozen other duties, including non-maleficence (based in the requirement not to harm others, fidelity (generated by our promises), gratitude (generated by acts others have done to benefit us), justice (generated by the demands of distributing goods fairly), and self-improvement.

Ross acknowledges that these duties may conflict. As previously discussed, the duty to act on behalf of a patient's welfare may conflict with the duty to respect her autonomy. In any given situation, Ross argued, there will be one duty that will over-ride the others, supplying the agent with an *absolute obligation* to perform the action specified by the duty. We will call this theory Prima Facie Duties (PFDs).

Ross does not think of his theory as providing a decision-making procedure. The Andersons adopt Rawl's method of reflective equilibrium for this purpose. In this procedure, prima facie duties in conflict with other duties are assessed by their fit with non-moral intuitions, ethical theories, and background scientific knowledge. When prima facie duties conflict, we must find the one which grounds the absolute obligation on which we must act.

The Andersons offer three reasons for adopting PFDs as EthEl's theoretical basis. First, they write, PFDs reflect the complexities of moral deliberation better than absolute theories of duty (Kant) or maximizing good consequences (utilitarianism). Second, PFDs does as good a job as teleological and deontological theories by incorporating their strengths. Third, it is better able to adapt to the specific concerns of ethical dilemmas in different domains. Let us consider these reasons one by one.

**PFDs better reflects the complexities of moral deliberation.** We agree that construction of a moral theory should begin with our considered moral judgments. (Where else, one might ask, *could* one begin?) In constructing a moral theory, however, we have the luxury of sorting out our intuitions from our principles, taking into account various relevant considerations, abstracting general rules from the particularities of different cases, and finding reliable principles to guide behavior. The luxuries of having sufficient time and information to deliberate are not present, however, when we must make moral decisions in the real world. As the Andersons point out, Ross's system of prima facie duties works well when we are pressed by uncertainties and rushed for time. For this reason, we agree that an artificial agent should initially be programmed to make decisions consistent with Ross's duties; doing so reflects the complexities of moral deliberation.

That said, there are no guarantees that Ross's system of PFDs will survive intact after we acquire more information and are able to process it free of the emotional contexts in which we ordinarily make decisions. As information grows and our understanding of the inter-relatedness of the good of all sentient creatures grows, a point may come when the complexities of moral deliberation are best reflected not in PDF but in SHAU. Should the moral landscape change in this way, then the Anderson's method will be outdated because it will no longer reflect the complexities of moral decision-making (we take up this matter in Section G,

below). Though currently the Anderson's PFDs starting point is a virtue of their theory now, in time, our considered judgments may no longer support it.

**PFDs incorporate the strengths of teleological and deontological theories.** We are inclined to agree with this claim, even though the Andersons do not tell us what the relative strengths of each kind of theory are. But we note that SHAU, especially when construed along the lines of Hare's two-level theory, also captures the strengths of teleological and deontological theories

**PFDs are better able to adapt to the specific concerns of ethical dilemmas in different domains.** We find it difficult to know whether we agree because we are uncertain about the meaning of the contention. What are the "different domains" the Andersons have in mind? Medicine, law, industry, government? Family, church, school, sports? If these are the domains, then what are the "ethical dilemmas" to which PFDs can "adapt" better? And what does it mean for an ethical theory to adapt better to specific concerns? Could it be the case that a theory should answer rather than adapt to particular questions in different domains? The Andersons also claim that PFD is superior to other theories because it "allows for needed exceptions." We wonder whether this claim may be question-begging. Are the "exceptions" we commonly make in our everyday judgments justified? This is an open question, one that should be presented as a problem to be resolved at the theoretical level rather than as a set of facts that should be taken as factual data at the theoretical level. We do not dispute the fact that PFD holds our intuitions in high regard. We dispute whether one should consider it a strength of a moral theory in the long term that it allows intuition to over-ride considered deliverances of the theory.

## 3 HedonMed, an Unbiased Agent

We propose that as EthEl develops over time, and increasingly takes more and more relevant information into account in her decisions, that she may, with justification, begin to return judgments that appear to be based less on observing PFDs and more on satisficing interests. To avoid confusion, we call this imagined future agent HedonMed because it is based on a hedonistic consequentialist theory.

HedonMed will differ from EthEl in that it is programmed to take into account all relevant characteristics of a situation, find all the satisficing courses of action, consider any one of them over-riding, and act on it. HedonMed does not defer to a patient's autonomy when her welfare is at stake although, as we will argue, a patient's autonomy is clearly a factor in her welfare. None of HedonMed's answers, even those governing the easy cases, is justified by appeal to the judgments of experts, nor to intuition-based judgments of any kind. Even the initial values reflecting the experts' judgments are justified not by the fact that they reflect consensus judgments but by the fact that they satisfice happiness. All of HedonMed's answers are the result of objective calculations made on the basis of unbiased and complete information and offered to the receiver with a set of reasons acceptable to fair minded and fully informed subjects.

HedonMed's concern for autonomy is summarized in this principle:

*The duty to respect autonomy is satisfied whenever welfare is satisficed.*

The argument for this principle is that no informed reasonable person would accept compromises of Edith's autonomy that were not in her best interests overall. Therefore, a minimal condition of satisficing is that gross violations of autonomy cannot be accepted. They are rejected, however, not for EthEl's reason—that is, because they are violations of a PFD—but rather for an SHAU reason; they are not found in the set of actions that adequately satisfice a minimal set of conditions.

In SHAU, autonomy is a critical good, and yet it remains one good among many goods contributing to a patient's welfare. SHAU respects autonomy as long as it is beneficial and contributes to one's happiness. A feeling of being in control of oneself is critical to a life well-lived, and diminishments of our freedoms undercut our well-being. Unless we misunderstand the Anderson's description, EthEl will never over-ride a fully autonomous patient's decisions. Our agent, HedonMed, will violate autonomy on those rare occasions when it is necessary to satisfice welfare.

SHAU weighs each person's utility equally. If relieving Paul of a small and tolerable amount of pain will lead to the death of Peter nearby because Peter needs the medication to survive, the doctor following SHAU will not hesitate to override the duty to relieve Paul's pain in favor of the duty to relieve Peter's. SHAU is an information intensive theory; it demands a large amount of data in order to makes its calculations. Unfortunately, human agents must often make decisions not only in ignorance of all the data but lacking sufficient time even to take account of all the data one has, driving us to other theories that can provide answers more quickly. However, if, as seems likely, future computers are able to process data much more quickly than we can, AI moral agents may be able to make better use of SHAU than can human agents.

As long as a machine programmed with HAU, HedonMed, has all of the necessary information about Edith's physiological and psychological states, HedonMed can arrive at the correct decision more quickly and more reliably than can a human being. As the number of morally relevant features increases, the advantages of a machine over a person become apparent. We are not accurate calculators; machines are. We tend to favor ourselves and our loved ones, inclining us to bias our assignment of values toward those nearest and dearest to us; machines lack these prejudices. We tend to grow tired in our deliberations, to take short-cuts, and to end the process before we have considered all of the variables; machines are not liable to these shortcomings.

Unlike EthEl, HedonMed has all of the epistemological virtues just mentioned and none of the vices. HedonMed calculates accurately, objectively, and universally. It is aware of all relevant factors and does not end its calculations until all are taken into account. It takes no short-cuts and yet is aware of its own ignorance. If HedonMed's internal clock "foresees" that it cannot complete the necessary algorithms in time to make a decision, it defaults to what Gary Varner calls Intuitive Level System (ILS) rules. These are the deontologically-inspired rules of thumb that R. M. Hare urges us to follow when we are not thinking critically.

When HedonMed lacks either the time or information necessary to complete all calculations, it acts in such cases in a way that seems like it is acting like EthEl. It seems as if HedonMed is acting like EthEl because EthEl's prima facie duties seem comparable to HedonMed's ILS rules. Both sets of rules set the artificial agent's defaults, instructing it how to behave under less than ideal conditions. The impression of similarity between HedonMed and EthElis is correct if we consider the judgments each agent will return initially. Eventually, however, the two systems may begin to diverge dramatically. In conclusion, we explain the difference.

It is vital that HedonMed's deliverances be acceptable by medical practitioners. If doctors find HedonMed recommending courses of action with which few professionals can agree, then they will likely cease to use it. For its own good—for its own survival—HedonMed must produce results agreeable to those using it.

When HedonMed is initially calibrated, therefore, it will return results similar to EthEl. In the beginning stages of its operation, HedonMed's SHAU values will issue in decisions that mirror the PDF values of EthEl. However, over time, as HedonMed gathers more information, as experts revise its values in light of knowledge of what kinds of actions result in higher levels of satisficing, HedonMed may be expected to begin to produce results that are counter intuitive. It will, in turn, take this information into account when making its calculations. If it returns a decision that it knows will be considered wildly inhumane—so uncaring that everyone associated with HedonMed will agree to pull its plug—then it will have a decisive reason not to return that decision. In this way, while initial values in HedonMed reflect generally accepted practices and judgments, its future evolution need not be tied to these values even though it must continue to be sensitive to them. In sum, HedonMed will evolve with the culture in which it is used. If it is too far ahead of its time in urging that this or that PFD be left behind, it will be responsible for its own demise. If it produces moral judgments that are hopelessly out of step with those of medical or bioethical experts, it will fail. These considerations will part of its programming, however. Over time, and as HedonMed takes in more data and is able to survey broader and more subtle swaths of public opinion, it may be able to play the role of an agent of social change, able to persuade experts about the wisdom of its decisions by providing the reasons that its decisions will lead to better outcomes.

## *3.1 How HedonMed May Eventually Diverge from EthEl*

One might object to our proposal by claiming that it is not different from EthEl insofar as both programs start with expert ethical intuitions, assign them numerical values, and then calculate the results. We admit that HedonMed and EthEl share these beginning points, as any attempt to program an ethical system in an artificial agent must, and note that the procedure by which values are initially set in each program is a critical and controversial matter. We admit that the two programs will reflect the judgments of ethical and field experts and be based on our intuitions at the beginning. The two programs will be similar in these respects. However, they will differ in other, more important, respects.

First, the two programs will have different defaults. EthEl continues calculating values until she reaches a conclusion that contradicts a prima facie duty. At that point she quits and returns a decision that respects the PFD. HedonMed continues to calculate values even if it reaches a conclusion that violates a PFD. That is, EthEl regards her decisions as justified insofar as they cohere with PFDs. HedonMed regards its decisions as justified insofar as no mistakes have been made in calculating the set of decisions that satisfice. HedonMed is not bothered if any of the satisficing decisions contradict prima facie duties. Its decisions are overriding and prescriptive, in so far as they can be put into practice. This difference, in sum, is that the Andersons's program trusts intuitions and seems to know ahead of time which kinds of decisions it will accept and reject. Our program begins with the same intuitions but it anticipates the possibility that they may eventually be over-ridden so often that they are no longer duties, not even prima facie duties.

Consider the example of someone in hospice trying to decide whether to begin taking morphine toward the end of their lives to dull the pain of deteriorated muscles and bedsores. They are impressed by the amount of pain they are in. This patient calculates the numbers and concludes that morphine is acceptable because it vastly improves their welfare.

However, a family argues to the contrary that the patient will come to rely on morphine, it will dull their cognitive powers, cause the patient to enjoy their final days less, and set a poor example for other family members. Taking addictive drugs, argue the loved ones, destroys character as the patient leans increasingly on synthetic chemicals rather than on courage and family support. There is more disutility in using morphine, goes the argument, than in refusing it and dealing with the pain.

Other family members come to the side of the hospice patient. They point out that the anti-morphine argument makes a large number of assumptions while underestimating the patient's discomfort. They point out that the therapy is widely prescribed in situations such as this one, that it is very effective in helping to relieve fear and anxiety, and that its addictive properties are beside the point as the envisioned treatment period is limited. After the conflicting sides present their arguments the patient may be frustrated, confused about the right decision. In such cases, critical thinking is stymied by epistemological under-determination. Until all of the facts are assembled, properly weighted, and assigned probabilities, agents are justified in resorting to intuitive rules. In this case, they might incline the patient to act on ILS rules of thumb. These rules might include injunctions such as "one need not subject oneself to unnecessary pain and suffering," and "take the medicine the doctor prescribes," and accept the morphine.

We take the ILS acronym from Gary Varner's interpretation of Hare [5]. Varner notes that the three letters are apt because they are also used in aviation to stand for "Instrument Landing System," a system for finding the right path when you can't clearly see it for yourself and you could easily drift off course or be blown off course. In Hare's theory, a set of ILS rules has a similar function. A set of ILS rules is designed to cover a range of ethically charged situations that are encountered by the target population in the normal course of their affairs. Internalizing the rules properly produces dispositions to judge, react emotionally, and act accordingly.

It also makes the individual diffident about violating them, even when clear critical thinking indicates that doing so will maximize aggregate happiness.

Here we see the two main differences between SHAU and PFD: ILS rules differ from prima facie duties in two respects, their derivation and justification. ILS rules are evolved rules that people internalize in order to produce dispositions to act in ways that reliably produce the best outcomes. Prima facie duties are Kantian-inspired facts about the universe. As Ross puts it,

> That an act. . . is prima facie right, is self-evident;. . . in the sense that when we have reached sufficient mental maturity and have given sufficient attention to the proposition it is evident without any need of proof, or of evidence beyond itself. It is self-evident just as a mathematical axiom, or the validity of a form of inference, is evident. The moral order expressed in these propositions is just as much part of the fundamental nature of the universe. . . as is the spatial or numerical structure expressed in the axioms of geometry or arithmetic [5, 29–30].

ILS rules are neither self-evident nor analogous to geometric axioms. They are practical rules that have evolved to solve social coordination problems and to increase human trust, accomplishment, and happiness. Unlike PFDs which are self-evident and unchanging, ILS rules are just those that happen to be generally successful in a certain place at a certain time in optimizing utility. ILS rules, unlike PFDs, are subjective and changing. They are not objective truths written into the fabric of the universe or derived from the autonomy and rationality of moral agents. They emerge from groups recognizing and codifying those practices that succeed in helping individuals in the group achieve their goals. One of the great virtues of ILS rules is the role of the rule in cultivating automatic responses to common situations. When professionals act on their ILS rules in cases to which the ILS rules have been found to apply, they are forming dispositions to make the right decisions.

We can now summarize the differences between HedonMed's SHAU programming and EthEl's PFD programming. PFDs provide unchanging and over-riding absolute duties. When EthEl identifies the relevant PFD, she defaults to an end decision and the calculations cease. ILS rules provide only temporary guidance to HedonMed, defining the default when HedonMed recognizes that there is not sufficient time or information or both to calculate the correct answer. ILS rules are not regarded by HedonMed as final or satisfactory. They are not regarded as precedents to guide future decisions. They are stop-gap measures HedonMed adopts when it must issue a decision under less than favorable conditions. Otherwise, calculations continue and, once time and information are supplied, HedonMed's final decision displaces whatever ILS rule has been used in the interim.

## 4 Conclusion

We admire the practical contributions the Andersons' have made to the literature of machine ethics and follow them in their preferred method for programming an artificial agent. We believe, however, that SHAU is a more defensible ethical

theory than PFD. We note in closing that SHAU requires technology that is not currently available. Until it is available, we think it is reasonable to construct a machine with ILS rule defaults. However, when the time comes that the technology needed for the execution of critical level SHAU is available, an act utilitarian framework should be implemented in automated agents. Such agents will not have prima facie duties; they will have only the duty to produce the greatest good.

## References

1. Anderson M, Anderson SL (2008) EthEl: toward a principled ethical eldercare robot. In: Eldercare: new solutions to old problems. In: Presented at the proceedings of AAAI fall symposium on AI, Washington, D.C. homepages.feis.herts.ac.uk/~comqkd/9-Anderson-final.pdf
2. Anderson M, Anderson SL (2007) Machine ethics: creating an ethical intelligent agent. AI Mag 28(4):15–26. http://www.aaai.org/ojs/index.php/aimagazine/article/viewArticle/2065
3. Anderson M, Anderson SL (2007) The status of machine ethics: a report from the AAAI symposium. Minds Mach 17(1):1–10
4. Beauchamp TJ, Childress JF (1979) Principles of biomedical ethics. Oxford University Press, New York
5. Ross WD (1930) The right and the good. Hackett Pub. Co, Indianapolis/Cambridge