

# Computing an Ethical Theory with Multiple *Prima Facie* Duties

Michael Anderson<sup>1</sup> and Susan Leigh Anderson<sup>2</sup>

<sup>1</sup>Department of Computer Science, University of Hartford

<sup>2</sup>Department of Philosophy, University of Connecticut  
[anderson@hartford.edu](mailto:anderson@hartford.edu)

## Abstract

In attempting to make Ethics computable, thus permitting ethical principles to be incorporated into a machine, we have adopted a multiple *prima facie* duty approach to ethical decision-making which we believe is more likely to capture the complexities of ethical decision-making than a single, absolute duty ethical theory. A major problem with this approach is the development of a decision procedure for determining the ethically correct action when the duties give conflicting advice. We are developing a way to solve this problem using machine learning to abstract information leading to a decision principle from ethical experts' intuitions about particular ethical dilemmas. This method has been tested in our proof of concept system *MedEthEx* using a type of ethical dilemma that involves three *prima facie* duties and eighteen possible combinations of these duties, where just four training cases were needed to create an ethically significant decision principle that covered the remaining cases.

As machines approach acting in an autonomous manner, it becomes increasingly important that they follow ethical principles, if at all possible, to avert undesirable consequences. We have, therefore, been exploring the extent to which an ethical theory can be followed by a machine. Furthermore, we believe that in the process of making an ethical theory precise enough to be programmed, it is likely that the ethical theory would be sharpened and revised. This work would, then, not only provide a new domain for the application of artificial intelligence techniques and an opportunity to develop a new area of applied ethics, but also make an important contribution to the theory which is applied.

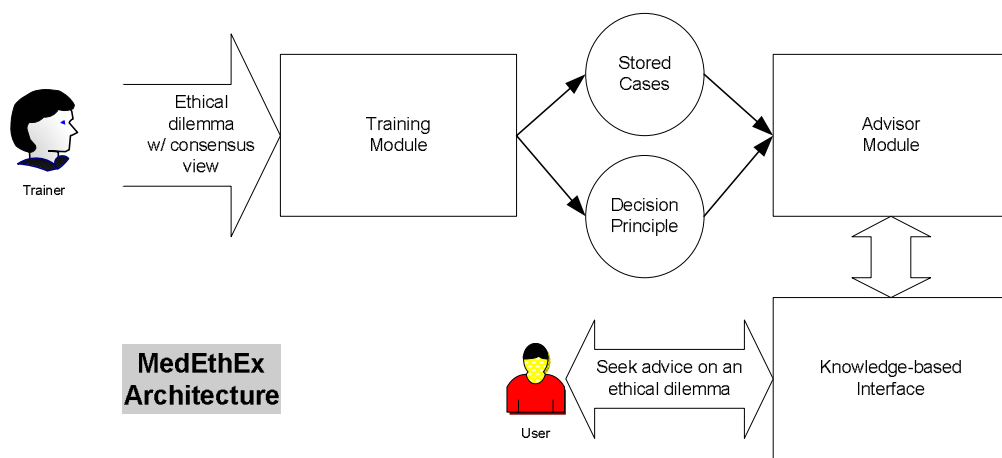
To be specific, we have been attempting to make Ethics computable for three reasons: 1) to determine whether an ethical dimension can be added to machines, 2) to advance the study of Ethical Theory by making it more precise than it has ever been made previously, and 3) to try to solve a particular problem in Ethical Theory — developing a decision procedure for an ethical theory with multiple *prima facie* duties.

In attempting to make ethics computable, we have adopted the action-based approach to ethical theory, where the theory tells us how we should act in ethical dilemmas. This approach best lends itself to machine implementation

by giving the agent either a single principle, or several principles, to follow to guide its actions, unlike other approaches that don't clearly specify which action is correct in an ethical dilemma. A good action-based ethical theory should be *consistent*; it should not contradict itself, by telling us that a single action in a given set of circumstances is simultaneously right and wrong. It should also be *complete*; it should tell us how to act in *any* ethical dilemma in which we might find ourselves. The theory should, also, be *practical*; we should be able to follow the theory. Finally, it should *agree with intuition*, specifically the intuition of ethical experts (Anderson 1999).

In agreement with W.D. Ross (1930), we believe that all single-principle, absolute duty ethical theories (such as Hedonistic Act Utilitarianism and Kant's Categorical Imperative) are unacceptable because they don't appreciate the complexity of ethical decision-making, the tensions that arise from different ethical obligations pulling us in different directions.

Ross' own theory consists of seven *prima facie* duties. A *prima facie* duty is an obligation that we should try to satisfy, but which can be overridden on occasion by another, stronger duty. Ross' suggested list of *prima facie* duties (which he says can be altered) captures the best of several single-principle ethical theories, while eliminating defects by allowing for exceptions. His particular duties are: the *Duty of Fidelity* (one should honor promises and live up to agreements that one has voluntarily made), the *Duty of Reparation* (one should make amends for wrongs one has done), the *Duty of Gratitude* (one should return favors) and the *Duty of Justice* (one should treat people as they deserve to be treated, in light of their past behavior and rights they may have). These duties are Kantian in spirit. The next two duties — the *Duty of Beneficence* and the *Duty of Nonmaleficence* — derive from the single Utilitarian principle, but reflect Ross' insight that possible good consequences and likely harm that can be caused must be separated, with the *Duty of Nonmaleficence* being stronger than the *Duty of Beneficence*, to account for our intuition that it is wrong to kill one person to save five. Finally, the last duty, the *Duty of Self-Improvement*, captures the best of "Ethical Egoism" by acknowledging that we have a special duty to ourselves that we don't have to others.



**Figure 1. MedEthEx architecture**

While everyone agrees that Ross’ duties seem intuitively plausible, he does not tell us how to determine the ethically correct action when the duties give conflicting advice, beyond saying that one should use one’s intuition to resolve the conflict. Unfortunately, this would allow one to rationalize doing whatever one feels like doing, by maintaining that a duty that supported *that* action is the most important one in the dilemma. Without an objective decision-procedure, furthermore, the theory can fail all of the requirements of an acceptable action-based ethical theory. We have concluded that the ideal ethical theory is a multiple *prima facie* duty theory, like Ross’, with some sort of a decision procedure to determine the ethically correct action in cases where the duties give conflicting advice. Devising such a decision procedure is one of the goals of our research.

We have formulated a method for solving the problem of making a multiple *prima facie* duty theory like Ross’ workable that essentially adopts John Rawls’ “reflective equilibrium” approach (1951) to creating and refining ethical principles, where one goes back and forth between particular cases and principles. What we have done is to find or create ethical dilemmas where there is a tension between the *prima facie* duties, and where there is also a consensus among ethicists as to the ethically correct action, and used machine learning techniques to abstract a general decision principle from those cases. A principle learned in this way is then tested on further cases and refined as needed to reflect ethicists’ intuitions about the ethically correct action in these other cases.

We have started with a simpler multiple *prima facie* duty theory than Ross’, with a constrained domain, that also lacks a decision procedure: Beauchamp’s and Childress’ Principles of Biomedical Ethics (PBE) (1979). PBE uses Ross’ duties of *Beneficence*, *Nonmaleficence* and *Justice* and adds the principle of *Respect for Autonomy*, a principle that reflects the shift in recent years from a paternalistic model of the healthcare worker - patient

relationship to one where the patient is given a more active role in his or her health care. For a decision by a patient concerning his/her care to be *fully autonomous* (Mappes and DeGrazia 2001), it must be based on *sufficient understanding* of his/her medical situation and the likely consequences of foregoing treatment, sufficiently *free of external constraints* (e.g. pressure by others or external circumstances, such as a lack of funds) and sufficiently *free of internal constraints* (e.g. pain/discomfort, the effects of medication, irrational fears or values that are likely to change over time).

We began our research with PBE and considered, initially, biomedical ethical dilemmas because 1) the information needed to judge whether a particular duty is involved in an ethical dilemma and its intensity (i.e. the needed data) is information that health care workers are likely to have, 2) there is more agreement as to the ethically preferable action among ethicists working on biomedical ethics than in other areas of applied ethics, and 3) there is a pressing need for ethical advice in this area, as research in biomedicine creates new, challenging ethical dilemmas.

*MedEthEx* (Anderson et al. 2005) (Fig. 1), our proof of concept system based upon Beauchamp’s and Childress’ Principles of Biomedical Ethics and Rawls’ “reflective equilibrium” approach to creating and refining a decision principle, offers guidance on one type of biomedical ethical dilemma:

A healthcare professional has recommended a particular treatment for her competent adult patient and the patient has rejected that treatment. Should the healthcare professional try again to change the patient’s mind or accept the patient’s decision as final?

The dilemma arises because, on the one hand, the healthcare professional should not challenge the patient’s autonomy unnecessarily. On the other hand, the healthcare professional may have concerns about *why* the patient is

refusing the treatment, i.e. whether it is a fully autonomous decision.

MedEthEx is comprised of three components: a *knowledge-based interface* that provides guidance in selecting the duties involved and their satisfaction/violation levels for a particular case, an *advisor module* that makes a determination of the correct action, if it can be given, for a particular case by consulting learned knowledge, and a *training module* that abstracts a guiding decision principle from information supplied by a biomedical ethicist, acting as a trainer, concerning particular cases. The first two modules are used in concert to provide advice for an ethical dilemma; the last module is used to train the system using cases in which biomedical ethicists have a clear intuition about the correct course of action.

MedEthEx attempts to capture expert ethical opinion in cases of this type of dilemma by using Inductive Logic Programming (ILP) (Lavrac and Dzeroski 1997) to learn a decision principle that captures the relationships between the three duties involved in this dilemma. Clearly, such a system can only learn a decision principle to the extent that ethical experts agree on the answers to particular dilemmas. In the general dilemma that we have chosen, there does seem to be general agreement among bioethicists as to the correct action in particular cases.

From four training cases, MedEthEx learned a decision principle that covers all eighteen possible combinations of the three *prima facie* duties involved (the principle of Justice is not involved) in the chosen type of ethical dilemma. It states that a health care worker should challenge a patient's decision if it is not fully autonomous and *either* there is any violation of the duty of Nonmaleficence *or* there is a severe violation of the duty of Beneficence. This philosophically interesting result gives credence to Rawls' Method of Reflective Equilibrium. We have, through abstracting a principle from intuitions about particular cases and then testing that principle on further cases (as ILP learns incrementally), come up with a plausible principle that tells us which action is ethically preferable when specific duties pull in different directions in a particular ethical dilemma. Furthermore, the principle that has been so abstracted supports an insight of Ross' that violations of the duty of Nonmaleficence should carry more weight than violations of the duty of Beneficence. Finally, the principle learned can be added to a theory that could be inconsistent and was not practical without a decision procedure, thus improving the theory.

The decision principle learned is ethically significant and supported by ethical theory. We have, thus, demonstrated that the particular problem in ethical theory — devising a decision procedure for a multiple *prima facie* duty ethical theory — is in principle solvable and that AI techniques are helpful in solving this problem. We, therefore, believe that not only is it possible to train a machine to make ethical decisions, but that machines can

help human beings in determining the principles that should guide them in ethical decision making.

## Acknowledgement

This material is based upon work supported in part by the National Science Foundation grant number IIS-0500133.

## References

- Anderson, S. 1999. "We Are Our Values", in *Questioning Matters*, Kolak, D. (ed.), Mayfield Publishing Company, p. 599.
- Anderson, M., Anderson, S. and Armen, C. 2005. MedEthEx: Toward a Medical Ethics Advisor. *Proceedings of the AAAI 2005 Fall Symposium on Caring Machine: AI in Eldercare*, Arlington, VA.
- Ross, W.D. 1930. *The Right and the Good*, Clarendon Press, Oxford.
- Rawls, J. 1951. Outline for a Decision Procedure for Ethics. *Philosophical Review*, 60.
- Beauchamp, T.L. and Childress, J.F. 1979. *Principles of Biomedical Ethics*, Oxford University Press.
- Mappes, T.A and DeGrazia, D. 2001. *Biomedical Ethics*, 5<sup>th</sup> Edition, pp. 39-42, McGraw-Hill, New York.
- Buchanan, A.E. and Brock, D.W. 1989. *Deciding for Others: The Ethics of Surrogate Decision Making*, pp48-57, Cambridge University Press.
- Lavrac, N. and Dzeroski, S. 1997. *Inductive Logic Programming: Techniques and Applications*. Ellis Harwood.