# Considerations of Moral Ontology and Reasoning in Human-Robot Interaction

**Lundy Lewis[1] and Dorothy Minior[2]**

[1]Department of Information Technology
Southern New Hampshire University
Manchester, NH USA
l.lewis@snhu.edu

[2]Lundy Lewis Associates, LLC
Mason, NH USA
dminior@verizon.net

**Abstract**

In this paper we discuss moral ontology and reasoning as it relates to HRI. We review several theories of *ought(A)*, where we examine the meaning of "ought" and the ontological status of *A*, e.g. as abstract properties of acts, individual act propositions, states-of-affairs, and practitions. We discuss several forms of moral deliberation based on those theories. We argue that a discussion of moral ontology and reasoning will play a role in understanding the ethical and social implications of HRI, and we offer three conjectures towards that end.

## Introduction

The questions of concern in this workshop are as follows:

1. How can notions of human identity be affected in the context of HRI?
2. How can understandings of human consciousness be affected?
3. How can concepts of human freedom be affected?
4. How can human social behavior be affected?
5. How can ideas of human moral status and moral responsibility be affected?
6. How can presumptions of human uniqueness be affected?

In this paper we argue that the exploration of these questions would benefit from the consideration of ontologies and reasoning mechanisms that may or may not be shared between humans and robots. In particular, we are concerned with the moral status and responsibility of both humans and robots, and thus our paper is concerned primarily with question #5.

The paper explores historical work on moral ontology and reasoning in the context of deontic logic – the study of the logical structure of discourse and reasoning about obligation, prohibition, and permission – and concludes with a discussion in the context of current HRI. The overarching questions that guide our discussion are these:

(i) When a robot reasons "I ought to do A" (OA) and then proceeds to actually do A, what exactly is *A* from an ontological point of view, and what is the inference procedure that leads to the doing of A?
(ii) Should a robot's usage of A and OA and the forms of deliberation thereof correspond to a human's usage?
(iii) How do the answers to questions (i) and (ii) influence our ideas about the Workshop's questions 1 – 6 above, in particular #5?

The paper is organized as follows. Section 2 provides some historical theories concerning that to which the notions of obligation, permission, and forbiddance apply. Section 3 discusses several forms of moral deliberation based on those theories. Section 4 concludes the paper with a discussion in the context of HRI.

## Theories about the *A* in Ought(A)

The main theories about the A in ought(A) are that A is (i) a property of an individual act, e.g. beneficence, (ii) a property attributed to some individual act, e.g. Tom's giving X dollars to charity Y at time t is beneficent, (iii) a generic state-of-affairs, (iv) a predicate attributed to some agent, and (v) a practition, e.g. Tom to do A. Below we discuss these theories in turn. We note that deontic logicians are often motivated by paradoxes in existing systems and devise new systems accordingly. In this paper, however, we aren't so concerned with logical paradoxes, but rather the implications of the underlying ontology of the systems.

### *A* is an Act-Property

In the seminal 1951 paper "Deontic Logic", von Wright argues that the kinds of thing to which the deontic words "obligatory," forbidden," and "permitted" apply are act-properties, e.g. beneficence is obligatory, and marriage is permitted (von Wright 1951). When deontic words are attached to act-properties, the result is a proposition. In the formula OA, A is an act-property and OA is a proposition, e.g. "Marriage" is an act-property and

"Marriage is permitted" is a proposition. The logic of act-properties is thus analogous to the logic of propositions in classical propositional logic.

### *A* is an Act-Property Applied to an Act-Individual

An alternative theory by Hintikka holds that deontic words apply to act-individual/act-property constructs (Hintikka 1981). If lower-case letters represent act-individuals and upper-case letters represent act-properties, then La might mean that Allen saying that he has a thousand dollars in his pocket has the property of being a lie. In this way, Hintikka brings in the machinery of quantification logic into moral deliberation and inference, e.g. (x)O~Lx might mean that all acts ought to be of kind not-lying.

### *A* is a State-of-Affairs

In a third theory, von Wright argues that deontic operators are about states-of-affairs (SoA) propositions (von Wright 1981). A *generic* SoA proposition is a timeless, Platonic entity that can be talked about abstractly, but can be used descriptively or prescriptively also. When a generic SoA proposition is used descriptively to say something about the world, it becomes a truth-value bearer -- it truly or falsely describes some partial state of the world. When used prescriptively, a generic SoA encourages action to see to it that the proposition becomes true. If "A" represents the generic proposition that a window is open, "OA" says that one ought to see to it that the window is open.

Also in this new system, an operator "/" is introduced to represent conditionality of a generic SoA coming into being. O(A/B) can be read: One ought to see to it that A when B. If, for example, A represents the SoA that the window is closed and B represents the contingency that it starts raining, then "O(A/B)" says that one ought to see to it that the window is closed should it start raining. The description to the left of "/"tells us how the world ought to be, when it is as the description to the right says that it is.

### The Foregoing is Muddled and Confused: Consider the Ought-to-Do

Peter Geach, in a fourth theory, questions theories that focus on the ought-to-be-ness of propositions and SoAs (Geach 1982). Geach argues that such theories are muddled and confused. He argues that 'ought' is an adverbial qualifier that modifies predicates and the resulting predicate applies to one or more agents. Consider that Allen ought to take Lucy home. This is an agent/predicate construct, and the predicate is an adverb/predicate construct: The adverb "ought" modifies the predicate "to take Lucy home", and the resulting predicate "ought to take Lucy home" is attributed to Allen.

Geach argues that the attraction of the analogy between moral deliberation and formal logic, and consequent emphasis on the ought-to-be-ness of propositions, was a fatal step of earlier theories. The movement obliterated what is really the essence of moral language, viz. that obligations have to do with people and not facts or states-of-affairs, i.e. the focus should be on ought-to-do-ness rather than ought-to-be-ness.

### A New Concept: Practition

In a fifth theory, Castaneda argues that moral statements divide into (i) those that involve agents and actions and support imperatives (the ought-to-do) and (ii) those that involve states of affairs and are agentless and have by themselves nothing to with imperatives (the ought-to-be). On Castaneda's view, the arguments for moral operators of the ought-to-do type are *practitions* (Castaneda 81). Prescriptions and intentions together constitute the class of practitions. Consider the following tenors of prescriptions:

| | |
|---|---|
| (1) Order: | Allen, take Lucy home. |
| (2) Request: | Allen, please take Lucy home. |
| (3) Advice: | Allen, you'd better take Lucy home. |
| (4) Entreaty: | Allen, I beg you, take Lucy home. |
| (5) Obligation: | Allen, you should take Lucy home. |

(1) - (5) all have the same core, viz. Allen to take Lucy home. This core is a prescriptive to-do expression bereft of any intentional clothing. The intentional clothing of a to-do expression shows itself via the circumstances in which the expression is uttered, e.g. the intonation of the utterance or with words like "please, "you'd better", "I beg you" as in (1) - (5) above.

An "intention" has the very same character as a prescriptive to-do utterance except that the agent in question is oneself. Intentions are in first-person, while prescriptions are second and third-person. An example of an intention is my emphatically thinking to myself "I shall take Lucy home!" The pure intention here is "I to take Lucy home" and the mandate clothing is expressed in writing by the exclamation mark.

### Discussion

In the theories above, we can see a shift in thinking about the ought-to-be-ness of facts and states-of-affairs and the ought-to-do-ness of acts by moral agents. We argue here that moral thinking culminates in action, assuming that the action is possible. Practitions, in particular first-person intentions of the form "I to do A", are the last conceptual

elements of the causal chain that begins with deliberation and ends in action. Supposing that one sincerely thinks "I to do A", one immediately finds oneself A'ing if it is indeed possible to do A. On the other hand, we argue that practitions do not always instigate their corresponding actions even if the action can be performed. They do not entail action when they are intermediate steps in moral deliberation. I might conclude "I to do A" and "I not to do A", in which case further deliberation is required to solve my dilemma and thus act accordingly (Lewis 1986). In the next section, we review various forms of moral deliberation that take this point into account.

## Forms of Moral Deliberation

The construction of a automated moral deliberator has at least three advantages: (i) it forces one to be exact about models of human moral deliberation, (ii) it provides a tool with which to test the consequences of various moral systems for both humans and robots, and (iii) it fosters a sort of shared ontology, which we have argued would be beneficial for exploring the social, ethical, and religious implications of HRI.

Moral deliberation culminates in action. The end of the chain of moral deliberation just before action will be a thought of the form "I to do A." However, a model of moral deliberation whose last element just before action is an act-property or an act/property construct is incomplete. There is a conspicuous hiatus between, e.g., "Beneficence is obligatory" and one really performing acts of beneficence, and this problem needs to be rectified. It might well turn out than ought-modified act-properties will play some part as a type of premise involved in moral deliberation, but as forms of moral conclusions, they will not do.

The most reasonable candidate for the sort of conclusions reached in successful moral deliberation is Castaneda's ought-modified practition. Practitions are noemata, or thoughts, which are genuine intentions to act in certain ways. The statement "I ought to do A" is a pure practition "I to do A" saturated with oughtness. At this juncture, let us consider the premises and forms of moral deliberation on which such conclusions are based.

Consider two possibilities. On Castaneda's view, moral conclusions are grounded in the following rule (Castaneda 1974):

> It is obligatory$_e$ that X do A, if and only if there is a natural number h and some normative system n such that both it is obligatory$_{n(h)}$ that X do A, and there is no normative system m and no natural number k less than or equal to h such that it is forbidden$_{m(k)}$ that X do A.

On this view, an agent will conclude "I ought$_e$ to do A" (recall that ought$_e$ entails the pure practition "I to do A" which precedes action) just in case the ought$_e$ statement carries more weight than any other competing ought-statement, where the weight of an ought-statement is inherited from the ranking of the normative system to which it belongs. The salient points of this view are that ought-statements are defeasible, and that it is possible to calculate just when one ought-statement is defeated by another.

In another view by Loewer and Belzer, deontic conclusions are grounded in inference rules of the following form (Loewer and Belzer 1983):

(1) One ought to do A when B is the case
(2) Premise 1 is not defeated
(3) One ought to do B *or* B is already settled
(4) A is possible
---
Thus: (5) One ought to do A

An example is this: As a general rule, one ought to keep one's promise (A) when he makes a promise (B). However, one ought not keep one's promise if the keeping of the promise will endanger somebody's life (thereby defeating premise (1)). If it is settled that one has made a promise and that the keeping of the promise will endanger somebody's life (3), and it is possible not to keep the promise (4), then one ought not keep the promise (5). In this example, the general rule "One ought to keep the promises one makes" is defeated.

## Discussion in the Context of HRI

Let us imagine an ordinary interaction between a human and a robot, and let us assume that the robot is equipped with some form of moral deliberation similar to those discussed in the preceding section. We'll have to think in the future, for I believe that today's robots don't include such mechanisms for moral deliberation; see (Anderson and Anderson 2007) for a good discussion. What is required for one to accept the robot as a morally responsible agent, i.e. to be an ethically acceptable agent?

**Conjecture 1**: We conjecture that the robot will have to offer a cogent explanation of its ethical decisions and actions. If the robot performs action A, and we ask the robot why it performed A, then we would expect a rationalization along the lines of the forms of moral deliberation in Section 3. Of course the robot's rationalization will involve content in addition to form, but nonetheless there is a common ground for discussion and argumentation between it and the human. The robot and the human might well find that differences hinge

upon respective normative systems having different weights, similar Castaneda's rule of moral inference.

**Conjecture 2**: A longstanding controversy in the acceptance of robots is the issue of intensionality. How can a robot, which is presumably a physical thing without a mind, entertain intensions and mental goings-on in the way that we humans presumably do? The argument goes that it can't, and so robots will never be ethically or socially responsible. However, let us recall the popular theory of epiphenomenalism – that a human's mental activity is a by-product or side effect of a human's physical brain activity. We can observe brain activity directly, but not so with mental activity, we must infer mental activity base on physical actions. We argue that we can side step the intensionality controversy in that a robot's activity is just a side effect or by-product of physically activity as well, e.g. silicon and software. We have seen in Section 2 that at a syntactic level, words that express intensions may be couched in the form of a practition, i.e "I to do A" saturated by oughtness, and that a practition is the last in line in a stream of moral deliberation, where the next in line is some action grounded in the practition. In fact, we may argue that we have no direct evidence of another person's mental goings-on save in virtue of actions, including actions of explanations.

**Conjecture 3**: Finally, we note that in the preceding we have used "it" to refer to robots. We believe that if a robot can provide cogent explanations of action that form a basis for robot/human discussion and argumentation (Conjecture 1), and if we can side step the issue of intensionality in robots (Conjecture 2), and everything else being equal, then a natural consequence will be that we can in good conscious refer to a robot as a *he* or *she*. That is, we will ascribe moral responsibility and consciousness to robots in the same way that we ascribe them to ourselves. If this indeed becomes the trend in the future, then our thinking about the Workshop's questions 1 – 6 will be affected in a profound way.

## References


Anderson, M. and Anderson, S. 2007 "Machine Ethics: Creating and Ethical Intelligent Agent," *AI Magazine*, Vol. 28, No. 4, Winter 2007.

Castaneda, H. 1974. *The Structure of Morality*, Charles C. Thoman: Springfield, Illinois.

Castaneda, H. 1981. "The Paradoxes of Deontic Logic: The Simplist Solution to All of Them in One Fell Swoop: New Studies in Deontic Logic (editor: Risto Hilpinin); Reidel: Dordrecht, Holland; 1981; 37-85.

Geach, P. 1982. "Whatever Happened to Deontic Logic"; *Philosophia* (Israel) 11 (1982); 1-12.

Hintikka, J. 1981. "Some Main Problems of Deontic Logic", *Deontic Logic: Introductory and Systematic Readings* (editor: R. Hilpinin); Reidel: Dordrecht, Holland, 1971 (1981); 59-104.

Lewis, L. 1986. The Ontology, Syntax, and Computability of Deontic Logic. Ph.D. diss., Dept. of Philosophy, Univ. of Georgia, Athens, GA.

Loewer, B. and Belzer, M. 1983. "Dyadic Deontic Detachment", *Synthese* 54 (1983); Reidel: Dordrecht,Holland; 295-318.

von Wright, G. 1951. "Deontic Logic"; *Mind* 60; reprinted in *Contemporary Readings in Logical Theory* (editors: I. M. Copi and J. A. Gould); Macmillan Co.: London; 1967; 303-304.

von Wright, G. 1981. "A New System of Deontic Logic"; *Deontic Logic: Introductory and Systematic Readings* (editor: Risto Hilpinin); Reidel: Dordrecht, Holland; 1981 (1971); 105-120.