**Logic and Social Robots**

1. **Introduction**

Social robots can reason and act while taking into account social and cultural structures, for instance by complying with social or ethical norms or values. It is currently predicted that social robots will become an integrated part of human societies in the near future. Researchers are investigating a very broad range of social applications of robots, for instance robots able to recognize pragmatic constraints of their environment (Fasola and Matarić, 2013a), (Fasola and Matarić, 2013b), robots placing themselves in accordance with the affordances in their environment (Lidner and Eschenbach, 2011), robots designed to adhere to an ethical code (Bringsjord and Taylor, 2012), (Bello and Bringsjord, 2012), and so on. For a general popular introduction to the field, see (Wallach and Colin, 2008). As social robots are likely to become more common and advanced and are likely to interact with human beings in increasingly complex situations, ensuring safe and correct robotic behavior in such situations is very important. One traditional approach to ensuring this has been basing the behavior of the robots on formal symbolic logics. The purpose of this paper is to evaluate the prospects of this approach and to suggest constructive ways of overcoming obstacles to it. Perhaps surprisingly, solving this seemingly very practical problem will require that we delve into the myth of logic as revealing the logical form of sentences, and into the relation between the normative and the descriptive aspects of formal logic.

1. **An overview of the paper**

Following two introductory paragraphs, the rest of the paper is divided into two parts. Part A is critical, Part B is constructive. In part A, I first discuss the logic program of AI and document how it lives on in certain areas of the social robotics community. I show how a specific version of this program cannot be realized for meta-logical reasons. I then move on to what I consider one of the roots of the problem, a myth concerning the nature of logic. I review and criticize this myth. In Part B, I propose an account of correct inferences in a reasoning system which gives us an approach to confronting the gap between the normative and the descriptive aspects of logic. This general idea is then investigated further in a study of deontic reasoning. I argue for the importance of empirical research in logic and I review some work which has been carried out in the psychology of deontic reasoning. I stress the differences between empirical work in psychology and in logic. An empirical study of deontic reasoning among Danish engineering students is used to argue that we need deontic logics more in accordance with natural language for successful human-robot interaction. This is further shown through an extended thought experiment involving humans and a robot. In the discussion, I argue further for my approach against a couple of alternatives: an appeal to conversational pragmatics, and a reduction of human-robot communication to simple, categorical orders.

**Part A - A critical assessment of the possibility of logic-based social robots**

1. **The logic program in social robotics and the meta-logical limits of logic-based ethical robots**

The idea of basing artificial agents on logic has its roots in the origin of artificial intelligence. Alan Turing’s seminal paper on computablity was an attempt to model human *computers,* human beings employed to do simple calculations, see (Turing, 1936), (Copeland, 2004). About his work, Kurt Gödel wrote:

“...the great importance of...Turing's computability...seems to me...largely due to the fact that with this concept one has for the first time succeeded in giving an absolute definition of an interesting epistemological notion.”

Kurt Gödel, quoted from (Copeland, 2004; p. 45)

I would like to point out elements of Turing’s work, which make it powerful. Turing’s theory contains two salient elements.

1. A symbolic system and rules for their manipulation (the instructions for the Turing machine).
2. An idealized model domain in which these instructions are carried out (the Turing machine working without making any mistakes on an infinite tape, and so on).

Now, these two elements in themselves give rise to certain possibilities for evaluation, i.e. the ideal Turing machines can be considered models for the symbolic language, which is basically what Turing did in his proof that the entscheidungsproblem is algorithmically unsolvable, see e.g (Börger, Grädel, Gurevich, 1997). But the Turing machine would not have been such a tremendous success, if it had not contained the element of cognitive modelling, which Gödel points out. Moreover this modeling is not only descriptive, it prescribes correct reasoning within this restricted domain. We therefore have to add a third element.

1. A social or psychological domain (the cognitive processes of a human computer).

It is the precision and simplicity of the definition of the Turing machine, which makes it possible to realize a physical design from it, a computing device. But it is the perceived agreement between model domain and social domain, which makes us (and Gödel) see that Turing has given us a correct theory of what it means to compute. Or, to put it slightly differently, most of the time, the agreement between a social domain and the model domain will give rise to seemingly correct inferences via the symbolic system. I shall, in Part B, use Turing’s work as a paradigm for a definition of a *reasoning system*, in which the *correctness* of inferences is guaranteed via the agreement between a model domain and a social domain. First, however, I investigate further the idea that AI could be based upon logic. In philosophy the idea has roots in the work of Hobbes and Leibniz. Strengthened to the idea that logic is all we need for AI it is sometimes referred to as the *physical symbol system hypothesis*. This hypothesis states that:

“…a physical symbol system has the necessary and sufficient means for intelligent action.”

(Newell and Simon, 1976)

The physical symbol system hypothesis is rather extreme, logic may play a useful and important role in AI without being either sufficient or necessary for intelligent action. Also, the hypothesis certainly downplays the role of the social domain for intelligent action. To provide just one example, logic may yield rather general principles for deontic reasoning, but the more specific rules for should count as an ethically correct action in a specific situation would presumably come from a substantive normative theory. How big of a role should logic play? When it comes to robots, the laws of robotics as suggested by the author Isaac Asimov seem to provide a fictional reference point also for current working robot theorists. The hope is that robots could be logically guaranteed to act in a socially acceptable or ethical way. Asimov’s laws are of course a literary device meant to create dramatic tension, but nevertheless his stories provide many examples indicating how hard this idea might be to carry out as an actual robot design project, see (Asimov, 1995). Still in spite of Asimov’s fictional misgivings, basing robots on logic continues to be a powerful idea. One might argue that this is due to fact that almost all current robots are implemented via computers and so they are as a matter of fact logic-based. Further, we trust in logic as reasonable and reliable tool which could provide the basis for justification of safe, ethical behavior of robots. A very explicit belief in such a strategy has been put forth by Bringsjord, Arkoudas and Bello who write:

“Our answer to the questions of how to ensure ethically correct behavior is, in brief, to insist that robots only perform actions that can be proved ethically permissible in a human-selected *deontic logic. “* (Bringsjord, Arkoudas, Bello, 2006).

These authors have also undertaken one of the few actual implementations of this program, (Arkoudas, Bringsjord, Bello, 2005). These papers contain implicit or explicit reference to Asimov’s laws of robotics. Other researchers have showed support for the general idea of using deontic logic as a basis for correct robot behavior; see e.g. several papers in (Lina, P., Abney, K., Bekey, G.A., 2012). In their contribution to that volume Bringsjord and Taylor define ethically correct robots as robots which satisfy the following three core desiderata.

D1 Robots take only permissible actions.

D2 All relevant actions that are obligatory are actually performed by them, subject to ties and conflicts among relevant actions.

D3 All permissible (or obligatory or forbidden) actions can be proved by the robot to be permissible (or obligatory or forbidden) and all such proofs can be explained in ordinary English.

(Bringsjord and Taylor, 2012)

I first note that these desiderata are rather attractive considering the point of view of the designer or user of social robots. They are only supposed to give the overall principles for robotic behavior, which must be further supplemented with specific details from ethical theories. Secondly, I note that Desideratum 2 about the performance of actions is rather vague, as it refers to “relevant” actions (from what perspective?), and says that the performance of these is “subject to ties and conflicts” (what “ties” and what “conflicts”, ethical, pragmatic, logical…?). However, Desideratum 1 and Desideratum 3 are clear, and the following argument goes through only considering these. Because, however attractive these desiderata may seem, a Gödelian argument, see (Gödel, 1931), (Smullyan, 1992) will show that it is not logically possible for a robot to fulfil all three desiderata at the same time. Let us call a robot *ethically consistent,* if it only performs actions it is permitted to perform. An ethically consistent robot does nothing it is not supposed to do. Desideratum 1 above requires robots to be ethically consistent in this sense. Let us call a robot *ethically complete*, if the robot can prove every obligation it fulfills. Desideratum 3 requires robots to be ethically complete. I show that, at least for robots equipped with what I consider reasonable expressive power, these desiderata are not logically compatible, in other words, there will be actions which are ethically obligatory from a meta-logical point of view, but which cannot be proven by the robot to be so. What do I mean by reasonable expressive power? In order to manipulate sentences, prove them, explain them and so on, it is not unlikely that a social robot will be required to refer to sentences within its own internal representation, in other words it will be equipped with some system of labelling sentences. The argument holds for robots with this ability to name sentences (through e.g. a gödel numbering). An example is a robot, R, logically equipped to express a sentence, such as the following, s1.

s1: R ought to prove sentence sn.

Here sn could be a sentence with an ethical content. But then we can construct the following sentence g (for Gödel).

g: R ought not to prove sentence g.

If R is ethically consistent it cannot prove g, because by proving it, it will have violated the obligation expressed in the sentence, which tells it not to prove g. It will have taken an impermissible action. On the other hand, R fulfils the obligation expressed in g by simply not proving it. Thus there is an obligation fulfilled by the robot which cannot be proven to be obligatory by the robot. We thus have to give up either Desideratum 1 or Desideratum 3. These desiderata can be considered regulative principles, guiding the design of social robots, but taken as universal rules, they cannot be.

1. **Philosophical criticism of the logic program and what to do next**

Throughout the years many critical points have accumulated against the logic program in AI from philosophers as diverse as Heidegger, Wittgenstein, Searle, and Dreifys. Specifically, the physical symbolic system hypothesis has been claimed not be hold because meaning is *grounded* or *situated* in the real world. Related to this is the criticism that agents have to be *embodied* to be intelligent. It has been claimed that there are forms of non-symbolic processing needed for real intelligence. Also, it has been claimed that there are non-computational aspects of cognition, see (Nielsson, 2007) for an overview of these criticisms. Partly in response to such criticism, partly because of the nature of robotics, researchers have begun to focus on Human-Robot Interaction in specific situations, see e.g. many of the papers in (Hermann et al., 2013). Further Bello and Bringsjord have recently made a case for using methods of experimental philosophy when devising moral machines, see (Bello and Bringsjord, 2012).

However, in spite of these advances within AI, a certain conception of what logic is still dominates the field. In the following, I focus on one aspect of this conception, and explain why I think this conception hinders progress within AI and logic.

1. **Logical form and context**
   1. **The myth of logic as a more literal language**

It is a rather common belief that formal logic as part of the more precise language of mathematics is suitable for getting to the true meaning or logical form of everyday language statements which are themselves often ambiguous. This belief presupposes that a natural language sentence consists of two elements, one being its content, the other its logical form. A logical analysis of a sentence can somehow reveal, or at least approximate, this supposed logical form. This idea of logic as an “ideal language” suitable for revising or reforming natural language goes at least back to Frege and Russell, see (Frege, 1879), (Russell, 1905), (Russell, 1916). Although, this idea and its philosophical assumptions have been criticized extensively, its proponents are still common; for a well-argued defence of logical forms, see (Stanley, 2000); for criticism of the idea, see (Stokhof, 2007). In opposition to this view, I would like to argue for the following two points.

1. Formal logic is not suitable for getting to the true meaning or logical form of a statement, because there is no such thing as the true meaning of or logical form of a statement as this is conceived in logic.
2. Rather, formal logics are artificial languages which are more predictable in some ways than natural languages, but which also suffer from some of the same flaws as natural languages.

Now, in order to not be accused of attacking men of straw, I will contend with a particular representative of the view stated above, the mathematician Keith Devlin, who has presented views similar to the one described above in a recent text-book, see (Devlin, 2012). By the way, I do not see it as a problem that I contend with Devlin’s views as they are presented in a text-book for undergraduates rather than an academic paper, as this is exactly the level where myths like these are perpetuated and thus also the level where you are likely to meet these myths expressed explicitly, as they are often taken for granted at more advanced levels. Devlin believes that the language of mathematics is unambiguous and universal (not connected to a particular culture) and that it can be represented via formal logic. He presents the following sentence as an example of sloppy use of everyday language.

“*One American dies of melanoma almost every hour.*

To a mathematician, such a claim inevitably raises a chuckle, and occasionally a sigh.”

(Devlin, 2012, p. 10)

According to Devlin, in order to get the literal meaning to correspond to the intended meaning, the sentence should be rewritten as follows:

“*Almost every hour, an American dies of melanoma*” (Devlin, 2012, p. 12)

However, it is hard to see what this grammatical rearrangement of the sentence has to do with mathematics. Apart from perhaps the ability to count to one, the skills required to express the so-called intended meaning are purely grammatical. It requires the ability to distinguish the meaning of the numerical “one” from the indefinite article “an”. The function of the movement of the adverbial phrase “almost every hour” to the front of the sentence, I take it to be to make clear the scope of this adverbial phrase. This is something (the thing) that is very clear in the standard translation of such sentences into first order logic. Devlin suggests that the sentences have the following logical forms:

1 ∃A∀H(A dies in hour H)

2 ∀H∃A(A dies in hour H)

The above formalization into first order classical logic is supposed to reveal the logical form of the sentence. To someone like me, who does not believe that there is such a thing going on here as the revealing of “the logical form” of a sentence, but only a rough translation between a very rich language (English) into a rather poor one (first order classical logic), this is very peculiar. (Let us ignore the fact that the sentences are written in “Loglish”, mixed English and logic as this is not relevant to my point). What should be clear for someone who knows tarskian-style semantics for first order logic is that the first sentence is one where we first pick out an American and then for every hour we can assign his or her death. This corresponds roughly to a situation where a certain specific American dies each and every hour (although what I just wrote is much more precise than the supposed “logical form” displayed above), which according to Devlin is what the first natural language sentence literally says. The second sentence is one where for every hour we pick out, we can find an American who dies of melanoma. However, this is all a trick. What I have done is that I have translated back into English Devlin’s intended meaning with these two (formal) sentences. This is probably roughly in accordance with some aspects of the meaning that most people would get out of the two natural language sentences. Most people would of course get a lot more out of the sentences than that, notably aspects related to lexical meaning (e.g. An American is a human being) and aspects related to contextual meaning (e.g. the causal connection suggested between the disease and death), which the formalization into first order logic does not capture. Are these aspects less essential than the ones captured by the formalization? I do not see why. Moreover, this is not even precisely what the sentences of first order logic say, they are much more “vague” than that. There are many standard classical models of the above sentences which would make them true but which would not at all correspond to the intended meaning of the natural language sentences. For instance, a lot more than one American could die on average and still the second sentence could be true. Clearly, that is not what is intended in the natural language sentence. Further the translation introduces a mistake, since the word “almost” has disappeared in the translation. In the formal models for sentence 2 you can find at least one American dying of melanoma for every hour, but this is clearly not intended in the natural language sentence. There are “gaps”, hours where no American dies of melanoma, these gaps are rare, but the word “almost” makes it clear that they are there. But such gaps cannot occur in the formal models for 1 and 2. Moreover, it is not that easy to translate a “fuzzy” word such as “almost” into first order classical logic, its model domain is that of discrete, countable, self-identical objects. This is not the world we live in.

There are metaphysical presuppositions when using first-order classical logic for representing natural language sentences, which are not easy to get around in a natural way, for instance that the domain of models for first order logic is non-empty. The most natural representations of a sentence such as the following make it false in first order classical logic:

“We detect no motion of the Earth relative to the ether.”

(Nolt, 2010)

Since “the ether” is a non-denoting term, a first order logic sentence containing it will be false. But most people will probably judge the above to be true, exactly because there is no ether and hence no motion relative to it. My point with these examples of the shortcomings of first order logic is not to point to Devlin as a poor logician, in fact he is a brilliant and versatile scientist. Nor is it my point that we should discard logic in general or first order logic in particular. Nor do I claim that these points are original or surprising. What is perhaps surprising is the fact that so few logicians and computer scientists seem to take such points into account when devising formal systems. I propose that we need to see logic differently, as restricted to certain domains, and that logic only gives rise to inferences that are correct ceteris paribus. In the second part of the paper, I provide a sketch of an alternative account of logic as seen from this different point of view and apply it to the case of deontic logic. It will then become clear why this is important for the design of social robots.

**PART B - Deontic reasoning systems for social robots**

“Our language can be seen as an ancient city: a maze of little streets and squares, of old and new houses, and of houses with additions from various periods; and this surrounded by a multitude

of new boroughs with straight regular streets and uniform houses.”

(Wittgenstein, 1958, Section 19).

1. **Reasoning systems**

The criticism levelled at the idea of logical form above may lead to the expectation that I am defending a radical contextualism with regard to logic. According to such a view there are no logical constants and words like “and” or “or” may mean anything depending on the context. However, I do not think that such a radical position is vindicated by empirical facts about language. Therefore, I would like to attempt a middle position between the meaning invariantism criticized above and the radical contextualism just indicated, a position which we can call *logical pluralism.* I take the above quote by Wittgenstein as my inspiration, but I am in no way venturing any Wittgenstein interpretation here – enough cans of worms have already been opened in this paper. According to the view, I want to defend, although there are probably no universally valid logical rules and thus no logical constants in the traditional sense, it does not follow that words like “and” and “or” may mean anything depending on the context. Instead, there are relatively stable pockets of meaning in rather large areas of discourse, perhaps corresponding to the streets or boroughs of Wittgenstein’s metaphor. The rules guiding deontic discourse is the example, I would like to analyze here. These rules hold ceteris paribus in spite of the creative and unruly aberrations which characterize human language use, and it makes perfect sense to talk about correct or incorrect reasoning with regard to these rules. In order to operationalize this claim in a way enabling us to delimit and analyse such areas of discourse further, I propose the following characterization of a *reasoning system*.

*A reasoning system is a fixed arrangement of components (a symbol system) that, most of the time, will give rise to* ***correct*** *patterns of reasoning when there is the right sort of fit between the model domain and the social domain in which the reasoning system is used.*[[1]](#footnote-1)

The above characterization is by no means to be taken as an “absolute definition”, rather it is a heuristic tool for combined empirical and theoretical studies into logic and reasoning. The extent to which the characterization is useful is itself dependent on the success of such studies, of which I have begun one, which will be described in some detail below. Generally speaking, logic is stretched out between a normative and a descriptive side. This is its main mystery. My basic claim is that we cannot take logic as giving us necessary laws of thought, which we should follow, but we also cannot take it to be a purely empirical science studying actual reasoning patterns. Therefore we have to find some middle ground. My proposal is that we should seek to engineer logics taken as systems of symbols and ways to manipulate them that yield correct reasoning patterns. This correctness is perceived when there is alignment between the theoretical models decided upon as the ideal interpretation of the symbols (these models could be presented via the tools of formal semantics, but they could also be other kinds of abstract entities e.g. Turing machines or even patterns of symbol manipulation as in Aristotelian syllogisms) and the portion of the social domain or “real world” or “domain of discourse” we are trying to model. I want to note a few things.

1. The correctness of the inferences of a reasoning system is only ceteris paribus. If the social domain is changed the reasoning system might not give us correct reasoning patterns any longer.
2. As the social domain is itself to a certain degree a human construct, one might both consider changing the social domain, e.g. teaching people the right way to think, develop the concepts, and changing the logic when meeting inconsistencies between linguistic intuitions and formal logics.

Formal logic applies to model worlds, which are purely theoretical. There are of course senses in which this purely theoretical endeavor can be correct or incorrect, as in the traditional concept of validity of an inference in terms of truth in a model. However, I take the above characterization to cover many logics besides those given a formal semantics. Also a logic which is presented proof theoretically, say, will single out certain inferences as correct or otherwise through its axioms and rules and this set of theorems will itself constitute such a logic’s ideal model domain. But even this expanded notion of correctness is too logic-internal to be useful in the practical application of logics. The correctness of the inferences, as it will be understood here, presupposes alignment between the model domain of the reasoning system and the social domain in which it is deployed. Social domains come in various sizes. At the one extreme there are the big discursive domains, deontic reasoning, epistemic reasoning, default reasoning, and so on. There might be structural similarities, but there will also be differences between these domains. These correspond to neighborhoods in Wittgenstein’s ancient city. At the other extreme, we have the streets, the houses, or even single rooms. For instance, a micro-world such as the blocks-world constructed for the robot Shakey is an example of an extremely limited social world being constructed to be aligned to the model world of Shakey’s internal representation. This leads to the important point that the research task of AI does not have to be universal robots that can occupy every possible social world. Rather, it is to expand the micro-worlds of the research laboratory to social worlds, which are themselves also shielded and prepared in different ways. Examples of such already shielded social worlds could be the kitchen of an ordinary house, a museum, a hospital. We do not expect to encounter a motorcycle in the kitchen so, *ceteris paribus,* the kitchen robot can reason and act correctly even if it does not know how to operate a motorcycle. There are no absolute certainties in this world and we should not expect our animate tools, the robots, to be able to handle every eventuality. The metaphysical assumptions made in the construction of the blocks-world and ensuring correctness of the reasoning patterns of the simple robot, do not guarantee safe transfer to any other domain or to any arbitrary expansion of the social domain.

Now, I would like to discuss further the role empirical research may play within this program. To be more specific, I present some empirical research, I have undertaken within deontic logic. As a preliminary, I would like to point out some related work which has been done in the psychological study of deontic reasoning.

1. **Psychological studies of deontic reasoning**
   1. **On the connection between logic and psychology**

Frege famously rejected the conception of logic as the subject of describing psychological laws of reasoning, a position held by to prominent contemporaries of his, e.g. John Stuart Mill. He constructed a thought experiment to illustrate this.

“What if beings were…found whose laws of thought flatly contradicted ours and therefore frequently led to contrary results even in practice? The psychological logician could simply recognize this and say: these laws are valid for them, whereas others are valid for us. I would say: here we have a hitherto unknown form of insanity.”

(Frege, 1893)

Frege’s point is that, regarded as an empirical science, logic loses its normative status as representing correct reasoning. But what if we ourselves are such other-worldly beings, what if there are domains of discourse where we systematically and consistently do not think according to classical logic? Is it satisfactory to call such reasoning insane or even incorrect? If logic loses all connection to how people actually reason, it becomes irrelevant. One way of assessing the gap between the normative and descriptive aspects of logic is through empirical studies of logic. This approach to researching logic is gaining some momentum, and Johan van Bentham has gone so far as to proclaim a *cognitive turn* in logic (van Bentham, 2008). This implies that logic has something to learn from empirical studies of reasoning, but also the other way around, cognitive psychology should also be informed by the model domains provided by formal logic, see (van Lambalgen and Counihan, 2008) In the following I briefly review some of the work which has been done by psychologists in deontic reasoning and evaluate its usefulness with regard to constructing deontic reasoning systems.

**7.2.** **The deontic selection task**

Most early work in deontic reasoning studies deontic variants of Wason’s selection task, see (Wason, 1968), (Beller, 2010). Here people are asked to perform reasoning tasks with conditionals with a deontic modality in the consequent, e.g.

“If a person is drinking beer, then that person must be over 19 years of age’’ see, e.g. ( Griggs & Cox, 1982)

Informants are then asked to evaluate cases of violating the obligation. One underlying assumption of the studies is that people ought to reason with modus ponens and modus tollens. In this respect, studies show that people do better in deontic reasoning tasks than in non-deontic reasoning tasks, and that people do better in reasoning tasks with a specific content than in so-called abstract reasoning tasks, where some sentences are replaced with symbols, see (Beller, 2010). Although interesting, work on the deontic reasoning task is limited to sentences of the conditional type, and therefore not sufficient for a complete model of deontic reasoning.

**7.3. Other empirical work in deontic logic**

Several other empirical studies have also been conducted in deontic reasoning, see (Beller, 2010) for a review of the literature until then. Particularly interesting among them is the study of so-called *illusory inferences*. These are inferences “that seem highly plausible but that are in fact invalid” (Bucciarellia and Johnson-Laird, 2005). The following is as an example.

“Suppose only one action is permitted: Either ‘‘Take A or B, or both’’ or ‘‘Take C or B, or both.’’—Are you permitted to take B? Most people erroneously answer ‘‘Yes’’. The correct answer can be derived only by considering what it means that one action is not permitted, while the other one is permitted.”

(Beller, 2010)

If you take B, then you are both performing the first and the second action, violating the constraint that you may only perform one of them. What is interesting about this study, from my point of view, is not so much that people fail to perform the reasoning task correctly, but rather the role played by *free-choice inferences* in the study. In the example above, the free choice inference is the inference from the disjunctive imperative sentence (e.g. “Take A or B, or both”) to the conjunctive permission of every disjunct (“you may take A *and* you may take B *and* you may take both”). The study presupposes this inference to be correct although it is not valid in most formal deontic logics. Further, one reason why people fail to perform the test correctly is probably also the dominance of the principle that a disjunctive imperative implies free choice, which is here overridden by a (higher-level) constraint on the *performance* of the action prescribed by the imperative. So we can conclude two things: free choice inferences are generally perceived to be correct (both by the informants and the designers of the study), but it can be overridden in specific situations.

**7.4. On the difference between a psychologist and a logician doing research**

The purpose of the psychological studies of deontic logic is to produce psychologically plausible models of how people reason deontically. These models are related to reasoning in a descriptive sense, but they do not in themselves provide us with the means to justify inferences as correct or reject them as incorrect. Obtaining that requires the tools of formal logic. However, this does not mean that logic cannot be informed by empirical studies. I would go so far as to claim that empirical studies can be used to find the boundaries of our theoretical models and to assess in which social domains they are likely to yield correct reasoning (most of the time).

What I propose is the testing of our reasoning systems in much the same way as other theoretical models are tested. If there is agreement between the model domain and the social domain we have a useful tool. What happens if there is no agreement? There is no simple answer here, as we both want our models to describe and prescribe correct reasoning. However, there are certain traditional answers, which are not sufficient. Some logicians protect the superior rationality of their approach, perhaps delegating the dirty work of getting it all to fit with reality to pragmatics. Others confine themselves to dealing with purely technical matters. From the practical point of view of providing a logical foundation for social robots these traditional answers can have downright catastrophic consequences as I will show.

1. **An empirical study of deontic logic**

Within deontic logic, as within many other branches of logic, empirical research with human subjects is not common. It is telling that the studies mentioned above are carried out by psychologists, not logicians. When devising new deontic logics for instance, deontic logicians seem too often to content themselves with their own linguistic intuitions, perhaps supplemented with those of a few of their peers. As a counter to this general trend he following is based on a study made with 161 Danish engineering students who were administered a survey asking them to complete various reasoning tasks. The reasoning tasks were structured as follows. The student was given a couple of background sentences and asked whether another sentence follows. They could choose between the answers “yes”, “no” and “I don’t know”. Further they were allowed to comment on their response, which some did in some cases. This adds a qualitative element to this otherwise quantitative survey. The main purpose of the present paper is to investigate logic-based social robots, so I will only mention a few points from the study which are relevant for this purpose. One of the reasoning tasks given was the following, which is based upon one of the persistent problems in deontic logic, known as Ross’ paradox, see (Ross, 1941).

“You may order ice cream. Does it follow that you may order ice cream or insult the waiter?”

64 percent of the informants answered “no” it does not follow, 31 percent answered “yes” it does follow and 5 percent answered “ I don’t know”. In most deontic logics, amongst which Standard Deontic Logic, see e.g. (Chellas, 1980), the conclusion does follow. The above was the so-called permission version of the paradox, it is about a disjunctive permission following from a simple permission. As we will see, the empirical data is even more supportive of rejecting the obligation version of the paradox. The informants were asked the following question.

“It is the case that you must offer a man an apology. Does it follow that you must offer him an apology or slap him in the face?”

12 percent of the informants answered “yes”. 84 percent of the informants answered “no”, whereas 4 percent answered “I don’t know”.

As a control question they were asked the following question further along in the survey.

“It is the case that you must send a letter. Does it follow that you must send the letter or burn the letter?”

To this question 82 percent answered “no”, 16 percent answered “yes” and 2 percent answered “I don’t know”. The intuition was reasonably stable among the no-sayers within this group. Studying fluctuations of linguistic intuitions would be another way of getting closer to a concept of “correctness” of inferences for a social domain.

Like the permission version of Ross’ paradox, in most deontic logic this inference is valid. This clash between intuition and classical logic was part of Alf Ross’ original motivation for denying the possibility of a logic of imperatives, a position he changed later in life in response to the development of deontic logic starting with (von Wright, 1951). However, von Wright’s logic also validates Ross’ paradox in contradiction to the empirical evidence. What should we conclude? The empirical evidence could be interpreted to imply that Ross’ original position is right, there is no logic of norms. Rather, I wish to take it to imply that the logic of normative discourse or logic in the normative domain differs from the classical logic known to Alf Ross and others. A related reasoning task is the so-called *free choice permission* which was also discussed above, in Section 7.

“It is the case that you must bring a present or give a speech. Does it follow that you may bring a present?”

To this question 73 percent of the informants answered “yes”, whereas 27 percent answered “no”. This inference is *not* valid in most deontic logics. Yet, the empirical data indicate that a majority of people perceive it to be correct and this is the inference presupposed valid by the psychologists studying illusory inferences, although it could be explicitly overridden. There is a clear clash between our standard logical models and the social domain of deontic discourse. In the following section it will become clear, how problematic this is with regard to human-robot interaction.

1. **An extended thought experiment about human-robot interaction involving deontic reasoning**

The following is an extended thought experiment involving a domestic robot and human beings. The purpose of this is to show how the empirical work described in the previous section has importance for social robotics, and further, to indicate some of the complications and possible solutions when applying reasoning systems in more specific social domains. It would be appropriate to clarify somehow the assumptions and limitations of this approach. To start with the latter, this is at present nothing more than an informal extended thought experiment, but it should absolutely be turned into a real case study involving computer simulations and empirical work. However, although this limitation is real, it should become apparent that I am basing this thought experiment on research and knowledge which is not science fiction, but which has been or could be implemented at present. This claim, however, can only be substantiated if I am allowed to take some research off the shelf without justifying it any further or even explaining it in any great detail. In the following I will thus be talking about Bayesian reasoning, multi-agent systems, and game-theoretical or utilitarian modelling of preferences, as if these approaches did not require any further justification. They don’t in the meagre sense that I am here (mainly) dealing with main stream results within these fields. They do, in the sense that even main stream research is based upon presuppositions which can always be scrutinized further, but this is beyond the scope of this paper.

The following work expands on previous work done in this area by including complex obligations and permissions, such as “you may pick up the plate or the cup” rather than only simple commands such as “pick up the cup”.

Consider a situation where a robot can pick up and place specific objects and perform other simple domestic tasks while respecting social constraints of its environment. More specifically, consider a micro-world with a set of 3 agents: 2 human beings and 1 Robot. Let us assume that the social domain includes a physical place, an apartment with 4 rooms: a kitchen, a living room, a bedroom and a bathroom. Socially, the robot is below the humans in the hierarchy, and must obey them as long as the commands fall within certain social constraints. Some of these constraints might be inherent to the physical design of the robot (e.g. it does not have gripping power enough to restrain a human being and therefore the obligation not to restrain a human being is inherently fulfilled by the robot), whereas some of the constraints are symbolically represented (e.g. “a robot is not permitted to enter the bathroom, when a human being is in there”). Some rules are only *prima facie* and can be overridden, e.g. the above rule may be overridden by a human explicitly asking for help in the bathroom. Other rules cannot be overridden, it could be, for instance, that the robot may never harm a human being. The robot serves various domestic functions within this social space. Let us first show a situation where our empirical study in deontic logic indicates there could be a problem, if the robot used something like standard deontic logic (or any other normal modal logic, see e.g. (Blackburn et alt., 2001)), for its internal reasoning. Apart from the hard constraints given by certain rules and direct obligations and permissions bestowed upon it by human beings, the robot also has an internal representation that mimics a value structure, imposing values on its actions via an evaluation of their possible outcomes and thus enabling the robot to simulate that some actions are better to carry out in a given situation than others. This internal value structure or utilitarian mechanism complements the harder constraints mentioned above, and could be represented using e.g. tools from game theory, preference logic, see e.g. (Liu, 2011), or stit theory, see e.g. (Horty, 2001). The internal value structure is used to e.g. weigh out alternative actions according to the resources required to undertake them, the risk they impose and so on. Naturally, the robot is programmed to use as few ressources as possible, however some tasks are more important than others, and so on.

* 1. **Example 1**

There is a dirty cup on the table in the living room. Human 1 tells Robot 1:

1. Somebody must put the cup on the table in the kitchen.

Assume that the robot has found a plausible interpretation of the above sentence, which is ambiguous since it could mean either that the cup is now on the table or that the cup must be placed on the table in the kitchen. This requires, among other things, grounding the noun terms in the environment using sense data and knowledge about the apartment and grounding the action terms in internal representation of actions available to Robot 1 at that point. This can be done, e.g. by using techniques similar to the ones described in (Fasola and Matarić, 2013b) where Bayesian reasoning is used to disambiguate such sentences relative to pragmatic fields containing background information about the placement of objects such as which rooms contain tables, and so on. Tools for handling conversational implicatures might be added, and a mechanism for requesting additional information in case of remaining ambiguity is necessary. Say that the robot now knows that it is supposed to take the cup from the living room table to the kitchen (and not e.g. put the cup on the kitchen table). This it knows e.g. because there is no table in the kitchen, or from similar background knowledge. If there are several possible interpretations, it will ask the human being for clarification. Now, assume that the robot is allowed to reason with disjunction introduction and that throwing the cup in the waste basket in the living room is more beneficial for it at this point. The robot therefore infers that somebody must put the cup in the kitchen or throw it in the wastebasket.

If the robot reasons via a normal deontic logic there is no problem yet, since it will not be able to detach the second disjunct in the obligation. However, assume that Robot 1 meets a human being reasoning like my engineering students. The human being asks what it is doing, the robot answers:

1. Somebody must put the cup in the kitchen or throw it in the wastebasket.

Upon hearing this, the human being reasons via free choice permission that he may throw the cup out, which he does after having taken it from the robot, perhaps thinking, “I always hated that cup, I am glad to find out she does too”. However, this was in fact her favorite cup and for the rest of the day the atmosphere is quite icy. This example shows the need for devising more natural deontic logics for social robots.

However, one might argue, is it even important for robots to reason with complex deontic sentences, can they not simply obey simple commands? In fact, much of the robot literature seems to limit robots to handle simple commands, but this I take in most cases to be an indication of current limitations, not as an expression of a deliberate choice. Be that as it may, there has been a lot of doubt and suspicion surrounding deontic logics and logics for imperatives, and the debate whether there are even such things lives on. Rather than entering into the specifics of this debate, I present some examples, where I think that reasoning about norms turns out to be beneficial, simple and natural. The point here is not whether the results could have been achieved by another means (other models are of course possible), but whether this is a fruitful approach, worth investigating, which I think it is.

* 1. **Example 2**

The house is quite messy. Human 1 says to Robot 1.

1. You must pick up the bedroom or the bathroom.

By its internal reasoning the robot infers that it may pick up the bedroom and it may pick up the bathroom. It has no other orders at the time, but if it did, it would ask human 1 for a priority ordering, for instance by saying, “Human 2 asked me to bake a cake in the kitchen, which should I do first?”. The robot then goes on to compare the two possible actions, and because the bathroom is closer to its current position, it decides to go there first. The robot moves towards the bathroom, but finds it occupied with Human 2. As the default rule is not to enter the bathroom when a human is in it, it decides postponing cleaning up the bathroom and moves to the bedroom. Here it starts its clean up routine, perhaps first picking up objects and putting them where it has been trained to put them, throwing things out it has been trained to classify as trash and, when in doubt, preparing a display of objects for later inspection by a human. After this subroutine has been carried out it starts vacuuming, and so on.

The point of the example was to show how presenting the robot with a disjunctive obligation and leaving the details such as the actual order of execution and so on to the robot, is beneficial, simple and natural.

Another objection is the following. Since the problem with Ross’ paradox seems to be with disjunction introduction, can we simply ban this rule (some argue it goes against Gricean maxims). No, we cannot, disjunction introduction can be rather useful as we will see now.

**9.3. Example 3**

Suppose there is the following rule in the house. Somebody must air out the living room twenty minutes between two and four in the afternoon. However, if it rains or snows, it should not be done. Also, during the time of airing out, the door between the living room and the kitchen should be closed if the kitchen window is open. When a human being comes into the room, he or she should be warned that the door might slam, if said window is open. We can structure this information a bit as follows.

1. Must(recurrent, every day, between 2 and 4) Air out living room.
2. If it rains or it snows, 1. is overridden.
3. Before 1. is fulfilled, if the kitchen window is open, the living room door must be closed.
4. If the living room door is opened, a human being opening it, must be warned.

First, suppose that it rains. The robot then reasons via normal disjunctive introduction (which is assumed to hold outside of deontic contexts) that it rains or snows. Hence it deduces that 1. is overridden, and the living room is not aired out that day.

Suppose that it is the next day, and that it does not rain or snow. There is now no way to override 1 (in this example), and if the robot has no other obligations, it will go check the kitchen to see if the kitchen window is open. If it is not, it will go open the window to air out the living room. If it is, it will close the living room door behind it on its way back into the living room.

This example shows that disjunction introduction can be quite useful.

**9.4. Example 4 - Resolving normative conflicts**

In the empirical study of the engineering students the informants were asked the following question.

“Decide whether the following can be the case. You must order a drink and not order a drink.”

Here 18 percent answered “yes”, and 13 percent answered “I don’t know”. However, 69 percent answered “no”, in effect denying the possibility of normative dilemmas. Now, I believe that normative dilemmas are a fact, but I also think they should be resolved, there *ought* not be cases like the one above. Here, I believe that logic should be normative, and so ideally, it cannot be the case that you must order a drink and not order a drink. When a robot encounters this type of situation, it has to resolve the conflict in order to act. Of course some conflicts can be avoided by the design of the system. We might for instance assume that the human beings themselves are ordered in a hierarchy which may be domain-specific according to their area of expertise. How exactly these areas are to be carved out will be decided by the users. When there is a conflict of norms, the norm coming from the area of expertise belonging to a human being will take precedence. Still there might be normative conflicts which cannot be avoided beforehand, and which need to be solved pragmatically. Here robots might be assigned different roles or personality types, e.g., going from the very decisive robot which simply decides on the best action according to its consequentialist internal representation to the very indecisive robot which is prohibited from doing anything as long as there is an inconsistency in its set of norms, excepting perhaps attempts to elicit user information. We might have several middle ways – the areas of the micro world might be compartmentalized so that a local inconsistency say in the cleaning of the kitchen, does not rule out the dog being walked on time. Formally, such an approach might require tools from para-consistent logic.

Now consider the following case. Robot 1 is told by Human 1.

“You must throw out this cup.”

The throwing out of a cup falls into several areas of expertise: kitchen maintenance (handling and placing kitchen objects) and the preservation and destruction of minor movable objects. Assume that a certain sentimental value has been placed upon this cup by Human 2 (it cannot be classified as trash, for instance). Further assume, that the general contract between Human 1 and Human 2 states that Robot 1 must ask Human 2 before throwing it out, even if Human 1 does not know that Human 2 has placed this value on the cup. The ethical constraints of this contract could make it impossible for the robot to throw out the cup without asking, e,g. saying “I will have to wait until Human 2 comes home from the store”. Assume that the robot asks Human 2 and she says ok. Then there is no problem, the robot throws out the cup. What if Human 1 asks why the robot must ask Human 2, gets the information “Human 2 attaches sentimental value to this cup” or “this cup costs more than x dollars” or whatever the argument is and upon hearing this tells the robot that Human 2 has agreed to the cup being thrown out already. Again there should be no problem the robot throws out the cup, perhaps recording the conversation as evidence, in case of a later problem. However, if Human 2 says that the robot must not throw out the cup, there is a normative conflict. The robot might then go back to Human 1 informing him of the other human being’s decision, and asking him to consider cancelling the obligation. If he refuses to do so, what will the robot do? Depending on the agent type of the robot, it might either 1) throw out the cup. 2) become completely inactive. 3) start performing another function.

The purpose of this example was to show the interplay between logic and pragmatics with regard to normative conflicts. When logic detects a normative conflict, the robot sees this as a problem which should be solved, in the above case by attempting to reach a consensus amongst the users about which norm is in place. In the absence of such a consensus being reached a robot might fall back on its utilitarian mechanisms, it might become passive or it might start doing other chores. Depending on the domain it might also take on different roles. Within some domains, e.g. certain emergency situations, the robot might be allowed to act even in face of a normative conflict. However, I would be extremely wary of giving robots this sort of capability, since it in effect does mean violating the norm given to it either by a human being or a set of norms accepted by human beings. In most cases, it will be best for the robot to do other chores, leaving the solving of normative conflicts to the human users. Omissions of actions (e.g. when the robot should decide not to interfere with a process) requires specific attention, which it is out of the scope of this paper to give to it.

1. **Discussion**

What are the possible prospects of the research program outlined in this paper? In general, logic is a theoretical study, focusing on the normative side of reasoning, how we should reason. However, as the extended thought experiment has shown, this approach to logic will not do, if we want to ensure safe and efficient human-robot interaction. Instead we must make logical models more in accordance with human linguistic intuitions. First of all, it is not very likely that we can change these intuitions. Ross’ wrote about his paradox in the early 1940s and now, 70 years later, it can be confirmed by empirical studies. The intuitions seem very stable. Secondly, as the discussion about first-order logic above shows, our current logical theories are neither complete, nor perfect from a strictly rational point of view. Restricted to certain domains (reasoning about certain aspects of non-empty at most countable sets of discrete objects) first-order logic works pretty well, but there are many other aspects of our dynamic, dappled world. Who is to say, that we cannot find model theories for other domains (in fact we already have, there are logics for default reasoning, counter-factual reasoning, dynamic epistemic reasoning, and so on). My suggestion would be to embrace this plurality to a much greater extent than logicians usually do. Is there no limit to this pluralism? I think there is. The stability of e.g. free choice inferences and Ross’ paradox seem to suggest so. Not just any model will do. It will not be up to theoretical model construction or empirical investigation alone to find out where the borders of correct reasoning are. Rather, it must be decided through a combination of the two, and when it comes to human robot interaction, also with other methods, such as computer simulations, as well as real experiments involving human beings and robots. Now, at this point, a logician is bound to ask whether we could not leave all of this to pragmatics and concentrate on correct reasoning in the old sense. I don’t think that we can, and for several reasons. Pragmatics traditionally deal with the utterances of sentences in a conversational context and there may be many reasons why saying something true is inappropriate in a context, see (Grice, 1975). However, I did not ask the informants, what they would be willing to *say* in a specific context, I asked them about which conclusion they *believe* follows in a specific deontic context. The question whether they would find it appropriate to utter the sentence in this context is a different one and conversational pragmatics does not explain (or in this case explain away) beliefs. Further, pragmatics is about what actually does follow but should perhaps not be said, but when it comes to Ross’ paradox the majority of the informants denied that the inference *follows at all*, in other words they deny that the conclusion is true. These points tend towards the conclusion that these phenomena must be explained as genuine semantic intuitions, not pragmatic conventions in a Grice’s sense. Whether there is another sense which would still count as being pragmatic and which could explain the phenomena is of course a different matter. However, even if there is a satisfactory pragmatic account of explaining these phenomena, getting robots to act according to such an account will require programming them, and this will in most cases require some sort of symbolic representation, leading us back into the issues discussed in this paper.

At this point a roboticist might suggest that we use a connectivist approach to model the pragmatics of the situation .No doubt, this has been done with success. However, here I would like to point out the advantages of the approach taken by Bringsjord’s and his colleagues. Logics are predictable and give rise to correct explanations when restricted to the right domains. Logics thus provide us with a security and a rationality that we cannot get from current connectivist approaches. So, although I will always be willing to admit that logic should not stand alone, logic is an important component of AI and it should continue to be so.

How general or how domain-specific is deontic reasoning? I don’t know. It must be up to empirical research and model building to find out. On the one hand, our reasoning system should not validate Ross’ paradox and it should validate free choice inferences. On the other hand, we should still be able to say that people are committing logical mistakes, as is the case in the psychological study of “illusory inferences”? Here it is logic’s role to step up as a normative arbiter, and teach people (and robots) how to reason in accordance with logical models. Are there no universal logical rules at all? In most cases, modus tollens is a correct reasoning pattern, and in most cases if x is a bird x has wings and feathers. However, we cannot pluck Tweety’s feathers off and cut off its wings, and conclude that since this being has neither feathers nor wings it is not a bird. Does this mean that the rule that birds has wings and feathers has to be given up or the rule of modus tollens has to be given up or both? I leave this as an open question.

There is another objection from the roboticist, I would like to consider. Many current systems of human-robot communication only allow the human to give simple, categorical orders to the robot (like “put the cup in the kitchen”.) , in other words you cannot present the robot with permissions and you cannot combine action types, as in e.g. “you may take an apple or a pear”. The examples given above also had the purpose of showing how a logic yielding correct deontic inferences will be beneficial in the future interaction between robots and human beings. Because the thought experiment is based on current research in robotics, the task of augmenting robots with a reasonable model of deontic reasoning fitting the deontic domain of discourse and social micro-worlds such as the one presented in the thought experiment might actually be much less formidable than one would think. With regard to the internal representation of the robot’s calculation of consequences (the utilitarian internal value structure of the robot), one could use work from the multi-agent community, e.g. stit theory, preference logics, game logics, and so on. This internal representation must be kept in check by constraints presented as rules.

1. **Conclusion**

In this paper, I have described the logic program in AI in general, and in particular in the robot ethics community. I have shown a meta-logical problem for a certain way of conceiving ethically correct robots. I have argued that the problems are rooted in myths about the nature of logic, and I have confronted one of these myths. As opposed to a very abstract logic-internal invariantist approach to correct reasoning, I have suggested an alternative conception of the correctness of inferences of reasoning systems, based on agreement between model domains and social domains. After having reviewed some work in deontic reasoning by psychologists, I have applied the definition of correctness to the social domain of deontic reasoning. First, I have conducted an empirical study among engineering students about deontic reasoning. Second, the results are used in an extended thought experiment concerning human-robot communication. The empirical research and the thought experiment show the need for logical models more in correspondence with everyday human reasoning. In the discussion, I have dealt with some likely objections, coming from the logic community (that we should leave everyday reasoning to pragmatics) and from the roboticist community (that we do not need complex logical reasoning for robots, simple orders will do).

Of course, many questions still need answering. When we approach reasoning from the two sides of model domain and social domain simultaneously, might we eventually succeed in aligning the two in such a way that we have the “correct” model of “correct” reasoning? Maybe. The world is unpredictable; reasoning is dynamic and tends to find clever ways of undermining itself, especially when it clothes itself in the guise of a formal system. However, both for theoretical reasons (finding better logics) and for practical reasons (e.g. engineering safe and efficient human-robot communication), this line of research is worth pursuing.

**References**

Arkoudas, K. and Bringsjord, S. (2005). Toward Ethical Robots Via Mechanized Deontic Logic, tech. report *Machine Ethics: papers from the* *AAAI Fall Symp*.

Asimov, I. (1995). *The Complete Robot,* HarperCollins Publishers.

Beller, S. (2010). Deontic reasoning reviewed: Psychological questions, empirical findings, and current theories, *Cognitive Processing*, 11 (2010), pp. 123–132.

Bello, P., and Bringsjord, S. (2012). On How to Build a Moral Machine, *Topoi* Volume 32, Issue 2, pp. 251-266.

Blackburn, P., de Rijke, M. and Venema, Y. (2001). *Modal Logic*, Cambridge University Press.

Börger, E., Grädel, E., Gurevich, Y. (1997). *The Classical Decision Problem,* Springer.

Bringsjord, S., Arkoudas, K., Bello, P. (2006). Toward a general logicist methodology for engineering ethically correct robots, *IEEE Intelligent Systems*, 21 (4) (2006), pp. 38–44.

Bringsjord, S. and Taylor, J. (2012). The divine-command approach to robot ethics, in P. Lina, K. Abney and G. A. Bekey (eds), *Robot Ethics: The Ethical and Social Implications of Robotics*, MIT Press, pp. 85-108.

Bucciarelli M., Johnson-Laird, P.N. (2005). Naïve deontics: a theory of

meaning, representation, and reasoning. *Cogn Psychol* 50, pp.159–193.

Cartwright, N. (1999). *The Dappled World*, Cambridge University Press.

Chellas, B. F. (1980). *Modal Logic*, Cambridge University Press.

Chisholm, R. M. (1964). Contrary-to-duty imperatives and deontic logic,

*Analysis* 24: 33–36.

Copeland, J. (2004). *The Essential Turing*, Oxford University Press.

Cox, R. J., & Griggs, R. A. (1982). The effect of experience on performance in Wason's selection task. *Memory & Cognition, 10,* pp. 496-502.

Devlin, K. (2012). *Introduction to Mathematical Thinking*, Keith Devlin.

Fasola, J and Matarić, M.J. (2013a). Using Spatial Semantic and Pragmatic Fields to Interpret Natural Language Pick-and-Place Instructions for a Mobile Service Robot, in *International Conference on Social Robotics (ICSR).*

Fasola, J and Matarić, M.J. (2013b) Using Semantic Fields to Model

Dynamic Spatial Relations in a Robot Architecture for Natural Language Instruction of Service Robots, in *IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Frege, G. (1879). Begriffsschrift. Eine der arithmetischen nachgebildete Formelsprache des reinen

Denkens. Halle a.S.: Louis Nebert.

Frege, G. (1893). *Grundgesetze der Arithmetik.* Band I. Jena: Verlag Hermann Pohle.

Gödel, K. (1931). Über formal unentscheidbare Satze der Principia Mathematica und verwandter Systeme, I", *Monatsheftc Math. Phys*., 38, pp. 173-198.

Grice, H.P. (1975). "Logic and Conversation," *Syntax and Semantics*, vol.3 edited by P. Cole and J. Morgan, Academic Press.

Herrmann, G., Pearson, M. J., Lenz, A., Bremner, P., Spiers, A., Leonards, U. (Eds.) (2013). *Proceedings of the 5th International Conference on Social Robotics*, ICSR 2013.

Horty, J. F. (2001). *Agency and Deontic Logic*, Oxford University Press.

Jørgensen, J. (1937). Imperatives and logic, *Erkentniss,* 7, pp. 288–296.

Kamp, H. (1973). Free choice permission, *Proceedings of the Aristotelian*

*Society,* 74, pp. 57–74.

Lina, P., Abney, K., Bekey, G.A. (eds) (2012). *Robot Ethics: The Ethical and Social Implications of Robotics*, MIT Press.

Liu, F. (2011) *Reasoning about Preference Dynamics,* Springer.

Lindner, F. and Eschenbach, C. (2011). Towards a formalization of social spaces for socially aware robots, in *Proceedings of the 10th international conference on Spatial* *information theory*, COSIT'11, pp. 283-303, Berlin, Heidelberg. Springer-Verlag.

Newell, A. and Simon, H.A. (1976). Computer Science as Empirical Inquiry: Symbols

and Search, *Communications of the ACM*. vol. 19, No. 3, pp. 113-126.

Nilsson, N. (2007). The physical symbol system hypothesis: Status and prospects. In M. Lungarella (Ed.), 50 years of AI, pp. 9–17, Springer.

Nolt, J. (2010). Free Logic, *Stanford Encyclopedia of Philosophy.*

Prior, A. (1954). The paradoxes of derived obligation, *Mind,* 63, pp. 64–65.

Ross, A. (1941). Imperatives and logic, *Theoria*, 7, pp. 53–71.

Russell, B. (1905). On denoting. *Mind*, 14, pp. 479–93.

Russell, B. (1914). *Our knowledge of the external world*. Chicago: Open Court.

Smullyan, R. M. (1992). *Gödel’s Incompleteness Theorems*, Oxford University Press.

Stanley, J. (2000). Context and Logical Form, *Linguistics and Philosophy*, 23, pp. 391-434.

Stokhof, M. (2007). Hand or Hammer? On Formal and Natural Languages

in Semantics. *Journal of Indian Philosophy*, 35, pp. 597–626

Turing, A.M. (1936). On computable numbers, with an application to the entscheidungsproblem

*Proc. Lond. Math. Soc*., 42, pp. 230–265.

van Benthem, J. F. A. K. (2008). Logic and reasoning: Do the facts matter? *Studia Logica*, 88, pp. 67–84.

van Lambalgen, M. and Counihan, M. (2008). Formal models for real people*, Journal of Logic*

*Language and Information*, 17, pp. 385–389.

von Wright, G. H. (1951). Deontic logic, *Mind* 60, pp. 1–15.

Wallach, W. and Allen, C. (2008). *Moral Machines: Teaching Robots Right from Wrong*, New York, NY: Oxford University Press.

Wason, P. C. (1968). Reasoning about a rule. *Quarterly Journal of Experimental Psychology, 20*, pp. 273–281.

Wittgenstein, L. (1958). Philosophical investigations, (2nd ed.). Oxford: Blackwell.

Zimmermann, T. E. (2000). Free choice disjunction and epistemic possibility,

*Natural Language Semantics,* 8, pp. 255–290.

1. This view can be likened with what Nancy Cartwright has called *nomological machines* in her analyses of natural sciences. Cartwright defines a nomological machine as follows:

   “It is a fixed (enough) arrangement of components, or factors, with stable (enough) capacities that in the right sort of stable (enough) environment will, with repeated operation give rise to the kind of regular behavior that we represent in our scientific laws. “ (Cartwright, 1999, p. 50).

   In order to apply Cartwright’s terminology to logic, which is not a natural science, some modifications have been made. [↑](#footnote-ref-1)