Some Empirical Results Concerning Deontic Reasoning: Models, Schema, or Both?

Yingrui Yang (yangyri@rpi.edu)
Rensselaer Polytechnic Institute
Department of Cognitive Science; 110 8th Street
Troy, NY 12180 USA
Paul Bello (Paul.Bello@rl.af.mil)
Air Force Research Laboratory - Information Directorate
Information Systems Research; 525 Brooks Rd.
Rome, NY 13441 USA

Abstract

Herein, we explore the psychology of deontic reasoning through the presentation of a heterogeneous natural logic combining inference schemas with a preference-based model-theoretic semantics such as those typically found in various formalisms for nonmonotonic reasoning. We conjecture that the heterogeneous approach is a generalization of various other hypotheses concerning deontic reasoning, and provides a robust framework for explaining semantic intricacies which are present in so-called "deontic paradoxes." As an initial investigation, two theories were tested: The first hypothesis states that people represent an obligation as a conditional statement which explicitly includes the concept of violation, and the other postulates that people not only prefer deontically perfect situations to less-than-perfect situations, but also have preference between these sub-ideal situations. Two sets of experiments were conducted in order to gain some insight regarding these two ideas, and the results show strong evidence supporting our initial intuitions.

Introduction

The psychology of reasoning has generally been divided on a number of issues which are germane to the development of an adequate theory of human thinking. While it is almost universally accepted that there is a deviance between human performance on various reasoning tasks, and normative solutions to the same; the cause of this deviance is still under debate. On the one hand, the *model theory* of reasoning (Johnson-Laird 1983) proposes that when a human subject interprets a reasoning problem, they utilize a set of procedures for modeling the initial relationships between propositions throughout the problem and various methods for manipulating the contents of the models in order to reach putative conclusions concerning possible states of affairs. The construction of a model depends on the interpretation of various linguistic terms (quantifiers, in particular), and the retrieval of general knowledge from memory. The model theory is domain-independent, and is related to the formal semantics of both first-order logic, and alethic modal logic, which is the logic of possibility and necessity. On the other hand, there exist a number of *mental logics* (Rips 1990; Braine 1978) which propose that humans possess a repertoire of inference schemas similar to those found in various formal systems of natural deduction. Inclusion of specific schemas varies between systems, as does the property of domain specificity. While these theories have been successful at describing human performance data on a number of different inferential tasks, including syllogistic reasoning, they have yet to provide a sufficiently general explanation for the frequent *content effects* which occur during the reasoning process.

While we are particularly interested in deontic[1] effects on inference, it has been shown that both causal and relevance (thematic relationships between antecedents and consequents) effects are abundant in everyday reasoning. In this paper, we present a natural logic based on the formal framework in (Tan & van der Torre, 1994) which provides a generalization of previous results on the psychology of deontic reasoning, and harnesses model-theoretic ideas to neatly represent and reason about deontic paradoxes.

Previous Research

Perhaps the most hotly debated of these content effects is given by deontic interpretations of Wason's Selection Task (WST) (Wason & Johnson-Laird, 1972), which asks subjects to perform a meta-inferential analysis of the material conditional. Briefly, in the classical non-deontic version of WST, subjects are presented with four cards, each one showing the letter A, the letter D, the number 2, and the number 7 respectively[2]. The subjects are also given a rule stating "if there is a vowel on one side of a card, then there is an even number on the other side of it", and are asked which cards would be necessary to flip in order to demonstrate the validity of the rule. As it turns out, performance on the WST is not very good. Less than 10% of untrained reasoners are able to give the answer: A and 7, which is the only combination usable for falsifying the conditional.

However, it has been shown in a number of studies that when the WST is presented with thematic content which expresses a deontic character, facilitation occurs, and subjects manage to correctly identify the correct items (Cheng & Holyoak 1985). A number of explanations have been put forth in order to account for this strange phenomena, perhaps the most famous of these being the social contract algorithm (Cosmides & Tooby 1989) in the context of evolutionary psychology, and the

[1] Deontic reasoning concerns norms, obligations, permissions, and forbiddance.

[2] Any combination of a vowel, a consonant, an even number and an odd number may be used in the presentation of the task.

pragmatic reasoning schemas theory (Cheng & Holyoak 1985).

We hypothesize that the mechanisms underlying deontic reasoning are necessarily both model-theoretic and proof-theoretic, and that there are specific interactions which take place between these two modes of representation which facilitate complex reasoning about obligations. This hypothesis is consistent with the theory of *Mental MetaLogic*, developed in (Yang & Bringsjord 2005).

A Brief Introduction to Deontic Reasoning

Deontic reasoning concerns the representation and formal manipulation of obligations, permissions, and prohibitions. Traditionally, deontic logic has been developed against the backdrop of classical modal logic, reasoning about what ought to be the case; with obligations and permissions as analogues of necessities and possibilities. However, in its original form, as given by Georg Henrik von Wright in his seminal 1951 article (von Wright 1951), deontic logic was developed as a logic of agency, being concerned with what an agent ought to do. This issue aside, we shall briefly introduce standard deontic logic, and an appropriate possible world semantics, describing the classical model-theoretic interpretation of deontic statements. The general form of statements in standard deontic logic (SDL) is as follows:

- Obligation: p is obligatory for agent a if and only if p is necessary for $a's$ being a good person. Formally: $O_a p$ iff $\Box(G(a) \rightarrow p)$.
- Permission: p is permissable for agent a if and only if p is possible and a is a good person. Formally: $P_a p$ iff $\Diamond(G(a) \wedge p)$
- Forbiddance: p is forbidden if and only if an agent is obligated to $\neg p$. Formally: $F_a p$ iff $O_a \neg p$.

The concept of permission is related to obligation in the following intuitive way: Pp iff $\neg O \neg p$. Through the rest of this section, we will briefly present so-called "standard deontic logic" in order to introduce two notorious paradoxes which have been the catalyst for almost all of the work done on more advanced deontic logics since their inception. The impetus, of course, is to show that these deontic paradoxes are far from being esoteric constructions born of mind of philosophers; rather, they represent commonplace dilemmas in which all of us have found ourselves in, and which deserve an accounting for from the standpoint of the psychology of reasoning.

Standard Deontic Logic: Syntax

Briefly, SDL is composed of the smallest $\mathbf{S} \subseteq \mathbf{L}$[3] such that it contains all of propositional logic, and the following four axiom schemas:

1. (**K**): $O(p \rightarrow q) \rightarrow (Op \rightarrow Oq)$
2. (**D**): $\neg O\bot$ or equivalently $\neg(Op \wedge O\neg p)$
3. (**MP**): from p and $p \rightarrow q$, derive q
4. (**NEC**): from p derive Op

From these four basic axioms, all of the machinery of SDL can be built. We now move our discussion onward into the realm of deontic semantics.

Standard Deontic Logic: Semantics

Deontic semantics can be interpreted as the standard possible-worlds semantics of normal modal logic (Chellas 1980). A Kripkean interpretation of deontic semantics is a triple $M = \langle W, I, R \rangle$, where W is the universe of possible worlds, I is an interpretation such that it assigns a subset of W to each sentence (all possible worlds at which the sentence is true, written $w \models p$ if w is a subset of possible worlds, and p is the sentence), and R is a binary relation among the worlds. A deontic sentence is valid if and only if it is true at every world $w \in W$ for any interpretation M. A sentence q is a logical consequence of another sentence p if and only if there is no interpretation M and world w such that $w \models p$ and not $w \models q$ for any interpretation M. In the style of normal modal logic, necessitation (obligation) of a sentence is understood as truth of that sentence in every accessible world (via the binary relation R, and possibility (permissibility) of a sentence is understood as truth of that sentence in at least one accessible world. Formally:

- $w \models Op$ iff $w' \models p$ for every $w' \in W$ such that wRw' holds.
- $w \models Pp$ iff $w' \models p$ for some $w' \in W$ such that wRw' holds.
- Additionally, for every $w \in W$, there is a $w' \in W$ such that wRw' holds (serial property of R).

It should be made clear that SDL only makes the distinction between ideal and non-ideal states of affairs. As we shall see, the proper treatment of deontic paradoxes[4] requires an intuitive semantics capable of distinguishing between ideal and sub-ideal worlds along with adapting techniques from non-monotonic logics to deal with ordering these worlds in a reasonable way to generate a representation of the differences between sub-ideal worlds which are necessary in resolving the paradoxes we shall present. We have chosen to investigate DIODE (Tan & van der Torre 1994), which is a framework for deontic evaluation.

[3] Where **L** is comprised of an infinite number of propositional variables, together with the usual connectives, defined in their usual way: $\neg$, $\rightarrow$, and O

[4] The two most famous deontic paradoxes, Forrester's Paradox and Chisholm's Paradox, will be presented in the next section, and shown to be representable within the formal framework which we are investigating.

DIODE: A Diagnostic Framework for Defeasible Deontic Evaluation

Let us begin by formally specifying the logical language to be used as the basis of the DIODE theory. DIODE is especially designed for the formulation of conditional obligations (which are premises in the paradoxes) through the following construction: "*if α is the case then β ought to be the case* $\equiv \alpha \wedge \neg V_i \rightarrow \beta$. The constant V_i represents a unique *violation constant*, indexed specifically to obligation i. The conditional obligation presented above can be read as "if α is the case, and this obligation is not violated, then β is the case. Unconditional obligations can be represented in this manner as well. For example, the obligation to not kill $O(\neg k)$ is represented as $\neg V_i \rightarrow \neg k$, stating that in the absence of the violation of this obligation, it ought to be that no killing occurs.

Definition 1: DIODE Language: Let L be a propositional logic. L_V is extended with a number of violation constants V_i. We use $\models$ to represent entailment.

Definition 2: Deontic Theories: Let T be a deontic theory of L. T is a collection of factual sentences F (referring to what is actually the case), a set of background knowledge sentences of L, and a set of conditional and unconditional norms of the form $\alpha \wedge \neg V_i \rightarrow \beta$ or $\neg V_i \rightarrow \beta$ respectively where $\alpha, \beta \in L$.

We now introduce the preferential semantics which defines a partial pre-ordering on the models of T. This semantics is used for ordering all of the ideal and subideal situations according to how many violations occur within each model.

Definition 3: Partial Pre-Order: Let T be a theory of L_V and M_1 and M_2 two models of T. M_1 is preferred over M_2, written $M_1 \sqsubseteq M_2$, if and only if $M_1 \models V_i$ then $M_2 \models V_i$ for all i. We write $M_1 \sqsubset M_2$ (M_1 is strictly preferred over M_2) for $M_1 \sqsubseteq M_2$ and not $M_2 \sqsubseteq M_1$.

Definition 4: Preferential Satisfaction: A model M preferentially satisfies A (written as $M \models_{\sqsubset} A$) if and only if $M \models A$ and there is no other model M' such that $M' \models A$ and $M' \sqsubset M$. This grants M the status of a *preferred model* of A.

Definition 5: Preferential Entailment: A preferentially entails B if and only if for any model M, if $M \models_{\sqsubset} A$ then $M \models B$.

The notion of preferential entailment can be used to identify *minimal violation sets* for a given deontic theory.

Definition 6: Minimal Violation Sets: Let T be a theory of L_V and M a preferred model of T, i.e. $M \models_{\sqsubset} T$. The set $\{V_i | M \models V_i\}$ is a preferred violation set of T.

DIODE defines a notion of *contextual obligation* as those sentences of L which are true in preferred models of a deontic theory T. More clearly, if some fact of the matter ($f \in F$) defined in the deontic theory T induces a preference ordering among models of T, those sentences of L which are true in the preferred models of T become obliged in the deontic context induced by f.

Definition 7: Contextual Obligations in DIODE: Let T be a theory of L_V. T provides a contextual obligation for α if and only if $T \models_{\sqsubset} \alpha$ and $\alpha \in L$.

Let's examine these semantic principles at work by taking a look at some of the deontic paradoxes which motivated our discussion in chapter three.

Forrester's Paradox

Forrester's paradox is easily represented within the DIODE framework. Recall that the symbolization of the paradox amounts to the following:

1. It is obligated that one doesn't kill: $O(\neg k)$.
2. If one kills, it is obligated that one kill gently: $k \rightarrow O(g)$.
3. Gently killing logically implies killing: $g \rightarrow k$[5].
4. One kills: k.

A problem arises here which is caused by the status of premise number three as a theorem. In particular, standard deontic logic admits the inference rule $\frac{\vdash p \rightarrow q}{\vdash O(p) \rightarrow O(q)}$, which states that a tautologous conditional yields a conditional consisting of an obligated antecedent and an obligated consequent. From premises 2 and 4, we derive $O(g)$, which taken with $O(g) \rightarrow O(k)$ yields $O(k)$ which contradicts premise number 1.

The Forrester paradox represented as a deontic theory in the DIODE language is composed of the following sets:

- A set of facts $F = \{k\}$.
- A set of background knowledge sentences $B = \{g \rightarrow k\}$.
- A set of norms $N = \{\neg V_1 \rightarrow \neg k, \neg V_2 \wedge k \rightarrow g\}$.

The reasoning process is relatively simple once the problem has been converted to propositional form. It requires no special knowledge of deontic inference rules, and allows for a relatively easy-to-understand presentation of results. In general, there are 2^n models generated by n propositions in a particular reasoning problem. Forrester's paradox only contains two propositions: g and k. Four models are generated for these propositions, but in our graphical presentation, we only depict three of

[5]This is the interesting caveat which makes the paradox work. The background fact that gentle killing logically implies killing is taken to be an analytical truth (tautology) here.

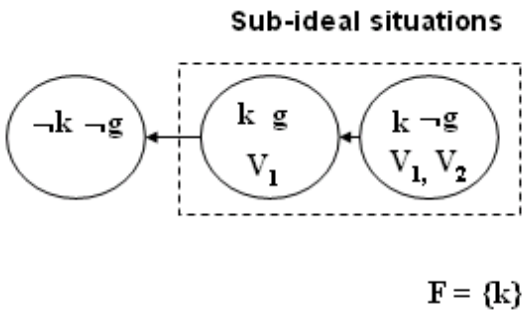

Figure 1: The Forrester Paradox

them[6]. The circles given in the visual depictions of the paradoxes denote equivalence classes of models. Only models which are preferred for some factual situation are given. The dashed box collects those equivalence classes which pertain directly to the fact of the matter (given by the satisfaction of k within a class), and is informally called "zooming in" on the facts. The preference ordering runs from left to right, with the leftmost set of models denoting ideal situations, just like those discussed in standard deontic logic. Instead of all non-ideal situations being clumped together into one equivalence class, they are separated in two ways: by what the facts are, and by the number of violations which are entailed by the facts. As can be seen in figure 1, taking propositions k and g to be the case yields violation V_1 through a simple application of *modus tollens*. Taking k and $\neg g$ to be the case yields the same violation of premise 1, but also a violation of premise 2, allowing us to always prefer the situation where if it's the case that I am forced to kill, I kill gently, to a situation where if I am forced to kill, I kill savagely. In fact, g becomes a contextual obligation under the theory T which represents Forrester's paradox.

Chisholm's Paradox

As in the case of Forrester's paradox, we are able to represent Chisholm's paradox intuitively and examine its' features. Recall that the Chisholm set consists of the the following premises:

1. It ought to be that Jones helps his neighbor: $O(h)$
2. It ought to be that if Jones goes to help his neighbors, that he tells them he is coming: $O(h \rightarrow t)$
3. If Jones does not help his neighbors, he ought to not tell them he is coming: $\neg h \rightarrow O(\neg t)$
4. Jones does not help: $\neg h$

The Chisholm paradox analyzed in standard deontic logic yields counterintuitive results. Since SDL admits the inference $\frac{\vdash O(\alpha), O(\alpha \rightarrow \beta)}{\vdash O(\beta)}$, from premises 1 and 2 we derive $O(t)$. From premises 3 and 4, we derive $O(\neg t)$. The D axiom in SDL states that $\neg(O(\alpha) \wedge O(\neg\alpha))$, thus leaving us in a bit of a conundrum.

Representing the Chisholm set within DIODE is natural, as well. Following our algorithm for the generation of a deontic theory, we obtain:

- A set of facts $F = \{\neg h\}$.
- A set of background knowledge sentences $B = \{\ \varnothing\ \}$.
- A set of norms $N = \{\neg V_1 \rightarrow h,\ \neg V_2 \rightarrow (h \rightarrow t),\ \neg h \wedge \neg V_3 \rightarrow \neg t\}$.

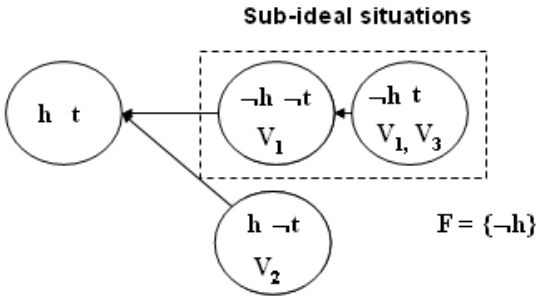

Figure 2: The Chisholm Paradox

Similar to our treatment of Forrester's paradox, we first develop an ordering on the equivalence classes of the models generated by the two propositions h and t, followed by "zooming in" on the fact of the matter, represented by $\neg h$. Any violation of premise 1 produces the violation V_1. Furthermore, if Jones tells his neighbors that he is coming, but never shows up, a violation of the third premise is generated, resulting in V_3. What's interesting is that if Jones helps, but doesn't tell his neighbors he is coming, a violation of premise 2 ensues, satisfying V_2. This violation holds the same deontic status as the state of affairs represented by not helping and not telling, yet is represented in a different preference order relative to the ideal situation, since h is contrary to the fact of the matter. It should be clear that a contextual obligation to not tell ($\neg t$) is generated in the case where $\neg h$ is the fact of the matter.

Some Testable Features of DIODE

In order to demonstrate that deontic reasoning consists of both model-theoretic and proof-theoretic components, we have chosen two particular features of the DIODE framework to investigate. The first feature concerns the representation of obligations, and the second concerns the semantic ordering principles which have been utilized in the presentation of the examples we have looked at thusfar.

In DIODE, obligations are represented via the explicit representation of a *violation constant* assigned to each statement capturing a normative proposition. While the particular version of the DIODE framework which we

[6]Models containing $\neg k$ and V_1 are not shown because those models would never be preferred in any circumstance, according to definition 3 of the partial pre-ordering over models.

have chosen to investigate does not provide explicit recommendations for what a reasoner ought to do, it does allow the reasoner to make judgments concerning what ought to be the case. In this respect, DIODE gives an analogous explanation of so-called "cheater detection," or the phase of reasoning concerned with looking for violations, and therefore predicts similar treatment of the deontic conditional presented in both (Cosmides & Tooby 1989) and (Cheng & Holyoak 1985).

Moreover, we predict that reasoners presented with paradoxical situations are able to both grossly classify worlds as being ideal or non-ideal, and in the case of non-ideal worlds, are also able to consistently order them based on the semantic principles presented in the DIODE framework. The interaction between these two phenomena is due to a relationship between the schema-like inference of violation, and the model-based ordering principles which result after such violations have been inferred.

Experiments

Two sets of experiments were conducted to test the hypotheses set forth in the previous section. Twelve experimental items were used to substantiate these claims, and can be provided upon request. Experiments 1-4 used between-subjects design, while experiments 5 and 6 used within-subjects design.

Materials: Twelve experimental problems are given. Problems 1-4 are propositional deontic problems, which were designed to test if subjects represent an obligation in the conditional form which we have described. Two items are of the form *modus ponens* (if p then q, p; therefore, q) and two of the form *modus tollens* (if p then q, not q; therefore, not p). For each form, the correct answer for one problem is "true", and for the other is "false". Problems 5-8 are quantified counterparts to items 1-4. We predict that the quantified versions function as their propositional deontic counterparts. Problem 9 is the Forrester Paradox, and problem 10 is Chisholm's paradox, both of which we have previously analyzed. These two problems were used to test the prediction that people not only prefer the ideal situation to sub-ideal situations, but also make preference between sub-ideal situations. Again, problems 10 and 11 are the quantified counterparts of problems 9 and 10. We also predict that the quantified versions function as their propositional deontic counterparts.

Subjects: 163 Rensselaer undergraduates participated in experiments to earn extra course credits.

Results and Discussion:

Data Set 1: 18 subjects did problems 1-4. The overall accuracy is 92%, which is consistent with psychological literature in two ways. First, this high accuracy is consistent with that of ordinary *modus ponens*, as presented by a number of researchers. Second, though people have difficulty with *modus tollens*, deontic content may suppress these errors, and lead to facilitation. Thus, this result supports the hypothesis that people do mentally represent obligations in conditional form.

Data Set 2: 17 subjects did problems 5-8. The overall accuracy is 93%. This result supports our prediction that the quantified versions function as their deontic counterparts. In comparison with result of Data Set 1-1, the correlation of performances between two versions is significant ($r = .89$), accounting for about 79.5% of variance.

Data Set 3: 15 subjects did problems 9 and 10, of which, problem 10 has a more complex structure, yielding unclear results. The result from Problem 9 is informative. 53% of the answers made the preference order 1>2>3, which is the predicted ordering. Note that this percentage is reliably more frequent, in comparison with 13% for the second frequent chosen order (Wilcoxon test, $z = 3.22$, $p < .01$).

Data Set 4: 17 subjects did problems 11 and 12. For problem 11 (which is the quantified version of problem 9), 47% of answerers made the preference order 1>2>3, which is dominant in comparison with 18% for the second most frequently chosen order order (Wilcoxon test, $z = 2.49$, $p < .05$). This result supports again the prediction that the quantified version functions as its deontic counterpart. Similar to problem 10, the complexity inherent in problem number 12 forces further analysis, due to unclear results.

Data Set 5: 50 subjects did Problems 1-8, which is a within-subjects design. Similar to the results from Data Sets 1-1 and 1-2, the overall accuracy on Problems 1-4 is 94%, and accuracy on problems 5-8 is 94.5%. As in the case of between-subject design, the correlation between the performances of the versions is again significant ($r = .91$), accounting for about 80% of variance.

Data Set 6: 46 subjects did problems 9-12. For problem 9 and its quantified counterpart (problem 11), 63% answers made the predicted preference order 1>2>3, which is significantly more frequent than the second chosen order (16%) (Wilcoxon test, $z = 3.71, p < .001$). Interestingly, 83% of answerers made the same preference order between the models given in problem 9 and problem 11.

General Discussion: The results from Data Sets 1, 2, and 5 have provided empirical evidence for the hypothesis that people are likely to present obligations as conditional statements. The results from Data Sets 3, 4, and 6 support the hypothesis that not only do people prefer the ideal situation to sub-ideal situation, but they also make preference between sub-ideal situations when the obligation principles are violated: they prefer a situation with less violations than more violations. These results seem to indicate that the DIODE framework is likely to be psychologically plausible. As for the problems requiring further study, it has been repeatedly stated within the literature on mental models that subjects often have difficulty reasoning about more than three models at once. In the cases of problems 10 and 12, subjects must reason about 4 models, which are semantically separated by two different preference orderings, further complicating matters. We believe that more detailed instruction will yield

a set of responses similar to those generated in problems 9 and 11.

Future Work

There are several other features of the DIODE framework which we feel to be psychologically plausible, and informative in coming to an account of a natural logic for deontic reasoning. Firstly, we wish to determine if contextual obligations are consistently picked out by subjects, after having identified the best non-ideal situation. This insight will provide us a clue as to whether certain kinds of conditional obligations may be derived only through semantical (model-based) reasoning, further supporting our contention that deontic reasoning is necessarily heterogeneous. Secondly we wish to lend empirical support to the extensive analysis in (Tan & van der Torre 1994) concerning exceptions, and their relationship to obligations. As a logical framework, DIODE is designed for *defeasible* deontic reasoning, clarifying notions of when obligations are overridden by facts (as shown in this presentation) or when they are overridden by more specific/important obligations (which are called "exceptions").

Acknowledgments

The work presented in this paper was funded by the Air Force Office of Scientific Research through the auspices of the National Research Council under a visiting summer faculty appointment for Dr. Yang at Rome Laboratory.

References

Johnson-Laird, P.N. (1983). *Mental Models*. Cambridge: Cambridge University Press.

Rips, L.J. (1990). Reasoning. *Annual Review of Psychology*, *41*, (pp. 321–353).

Braine, M.D.S. (1978). On the relation between the natural logic of reasoning and the standard logic. *Psychological Review*, *85*, (pp. 1–21).

Tan, Y-H., & van der Torre, L.W.N. (1994). Multi-Preference Semantics for a Defeasible Deontic Logic. *Proceedings of JURIX'94*, Amsterdam, Holland.

Wason, P.C. & Johnson-Laird, P.N. (1972). *Psychology of Reasoning: Structure and Content*. Cambridge MA: Harvard University Press.

Cheng, P. & Holyoak, K. (1985). Pragmatic reasoning schemas. *Cognitive Psychology*, *17*, (pp. 391–416).

Cosmides L. & Tooby, J. (1989). Evolutionary psychology and the generation of culture, Part II. Case study: A computational theory of social exchange. *Ethology and Sociobiology*, *10*, (pp. 51–97).

Yang, Y. & Bringsjord, S. (2005). *Mental MetaLogic: A New Unifying Theory of Human and Machine Reasoning*. Mahwah NJ: Erlbaum.

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

Chellas, B. (1980). *Modal Logic: An Introduction*. Cambridge: Cambridge University Press.