The Thirty-Third International
FLAIRS Conference (FLAIRS-33)

How to Act?
Reasoning with Conflicting Obligations

Clayton Peterson
Université du Québec à Trois-Rivières
clayton.peterson@uqtr.ca

Abstract

Recent work in proof theory has shed some light on the possibility of modeling reasoning while avoiding undesirable formal paradoxes. Based on category theory and inspired by the seminal work of J. Lambek, monoidal logics were introduced as a foundational framework that allows to treat a wide range of formal systems, including substructural logics (e.g., the syntactic calculus, linear logic, relevant logic, etc.), algebras (e.g., Kleene algebra) as well as intuitionistic, intermediate, and classical logic. This framework has been extended to modal logics and has been used to model normative reasoning, actions and knowledge, and it has been shown that non-classical logics better deal with the formal problems that are usually related to these notions. As such, non-classical systems of modal logics were proposed to model reasoning, actions and knowledge, but unresolved problems remained as to how to deal with conflicting obligations when facing normative inconsistencies. In this paper, we expose this problem and sketch an avenue for future research that might overcome this limitation.

Monoidal Logics

Based on Lambek's (Lambek 1968; 1969; Lambek and Scott 1986) and Lawvere's (1963) seminal work in category theory (Mac Lane 1971), monoidal logics were introduced as a foundational framework that can be used to approach any logical system from a syntactical perspective (Peterson 2016a; 2019). Logical systems are constructed from a general language $\mathcal{L}$ defined by a collection $Prop$ of atomic propositions p_i and the symbols $\{(,), \otimes, 1, \multimap, \triangleright, \oplus, 0, \ltimes, \rtimes, *, \star\}$. These symbols are interpreted respectively as a tensor (multiplicative conjunction) together with a unit (neutral element), two conditionals, a co-tensor (multiplicative disjunction) with a co-unit (neutral element), two co-conditionals, and two constants representing falsehood and truth. Well-formed formulas are defined recursively by:

$$\varphi := p_i \mid * \mid \star \mid 0 \mid 1 \mid \varphi \otimes \psi \mid \varphi \multimap \psi \mid \varphi \triangleright \psi \mid \varphi \oplus \psi \mid \varphi \ltimes \psi \mid \varphi \rtimes \psi$$

Following Peterson (2019), tensor-fragments of monoidal logics are defined using the following rules and axiom schemas, presented in figures 1.[1] Negations are defined by $\sim \varphi =_{df} \varphi \multimap *$ and $\neg\varphi =_{df} \varphi \triangleright *$.

Definition A deductive system is composed of a collection of formulas and a collection of equivalence classes of proofs satisfying (1) and (cut).

Definition A monoidal deductive system **M** is a deductive system satisfying (r), (l), (t) and (a).

Definition A monoidal closed deductive system **MC** is a deductive system satisfying (r), (l), (a), (cl) and (cl').

Definition A monoidal closed deductive system with classical negations **MCC** is a **MC** satisfying ($\sim \neg$) and ($\neg \sim$).

Definition A symmetric deductive system **S** is a **M** satisfying (b).

Definition A symmetric closed deductive system **SC** is a **MC** satisfying (b).

Definition A symmetric closed deductive system with classical negation **SCC** is a **MCC** satisfying (b).

Definition A Cartesian deductive system **C** is a **M** satisfying ($\otimes$-in) and ($\otimes$-out).

Definition A Cartesian closed deductive system **CC** is a **MC** satisfying ($\otimes$-in) and ($\otimes$-out).

Definition A Cartesian closed deductive system with classical negation **CCC** is a **MCC** satisfying ($\otimes$-in) and ($\otimes$-out).

Co-tensor fragments are defined by dualizing these notions (i.e., by reversing the arrows within the rules and axiom schemas, changing $\otimes$ by $\oplus$, 1 by 0, $\multimap$ /$\triangleright$ by $\ltimes$/$\rtimes$ and $*$ by $\star$). The dualized versions of the rules and axiom schema are used to define co-tensor fragments (co**MC**, co**MCC**, co**SCC**, etc.), with co-negations defined by $\circ\varphi =_{df} \varphi \ltimes \star$ and $\wr\varphi =_{df} \varphi \rtimes \star$. One co-rule that deserves mentioning given its relation to well-known problems in deontic logic is ($\oplus$-out).

$$\frac{\varphi \oplus \psi \longrightarrow \rho}{\varphi \longrightarrow \rho} \ (\oplus\text{-out})$$

Although this framework considers propositions and a consequence relation between propositions rather than structures and sequents, it has been shown (Peterson 2019) that it actually is syntactically equivalent to display logics (Belnap 1982; Goré 1998) and that it can be used to model

[1] A double line means the rule is reversible.

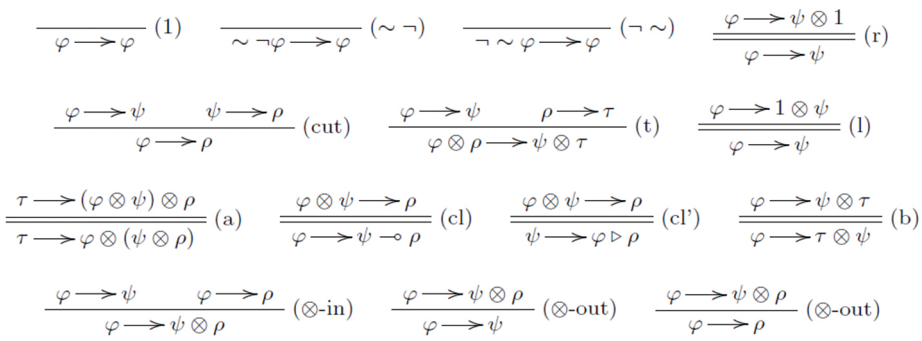

Figure 1: Rules and axiom schemas - Monoidal logics

substructural logics, including Lambek's (1958) syntactic calculus, multiplicative linear logic, full intuitionistic linear logic (Hyland and De Paiva 1993), and bilinear logics (Lambek 1993). Using the definitions of deductive systems and co-deductive systems, logical systems can be created by combining tensor's and co-tensor's fragments (e.g., **M**co**C**, **MC**co**SC**, **SCC**co**C**, etc.). This framework can also be extended to categorial grammars (Peterson 2016b).

Actions, Norms, and Knowledge

Monoidal logics have been used to model actions, norms, and obligations (Peterson 2015; 2017). One benefit of this approach is that it allows one to identify clearly the source of the paradoxes that usually arise when one tries to model these notions using the usual modal logics (Chellas 1980). For instance, Peterson (2014) showed that well-known problems, including logical omniscience in epistemic logic, can be related to specific properties of logical systems.

Peterson and Kulicki (2016) pursued this research avenue and showed how to define logical systems intended to model actions, norms and knowledge while avoiding the problems that usually plague modal logic. Although there are many problems that need to be addressed to avoid undesirable consequences when modeling normative reasoning, three of them are especially relevant, namely deontic explosion, detachment, and augmentation (Peterson 2015). While deontic explosion arises when from conflicting obligations one can legitimately deduce that anything is obligatory, detachment concerns the conditions under which an obligation can be actualized given a specific context (i.e., given an obligation conditional to some context, under which conditions can this obligation be detached from the conditional and be actualized as an obligation that should be obeyed), whereas augmentation happens when the consequence relation is monotonic and the following inference pattern is satisfied.

$$\frac{\vdots}{\varphi \supset O\psi \longrightarrow (\varphi \wedge \rho) \supset O\psi} \quad \text{(aug)}$$

From the perspective of artificial intelligence and automated reasoning, these problems are of the foremost importance. Indeed, deontic explosion entails that anything can be seen as obligatory when there is a conflict of obligations, and conflicts of obligations happen! Accordingly, a logic meant to model normative reasoning must avoid deontic explosion, otherwise artificial agents would be allowed to do anything in cases of normative conflicts.

Further, actions that should be performed vary from one context to another, and it is impossible to determine in advance the whole range of specific contexts and related obligations. Consider an autonomous vehicle for example. Assume a context where the itinerary says that it should go straight, and that it is programmed to stop at a red light. It arrives at a red light and there is an accident, with a police officer indicating it should turn right instead. In this context, the vehicle should turn, and not go straight, even though the itinerary says to go straight and the red light indicates to stop. Now, assume that there is a pedestrian that is crossing the street despite the officer's directives, thereby blocking the vehicle's way. Suppose that the police officer did not see the pedestrian. Even though the officer tells the vehicle to turn right, the vehicle should stop in order to avoid hurting the pedestrian. This very basic example illustrates how it is difficult to anticipate each and every fact that can affect what should be done within varying contexts.

Detachment is important for similar reasons. The problem of detachment amounts to the fact that, even though an obligation $O\psi$ might be actualized when it is conditional to some context φ and that context presents itself, other conditions ρ can be realized in conjunction with φ such that it would block the detachment of $O\psi$. Again, to be able to program and evaluate all the possibilities in advance would require omniscience, which is a characteristic of neither man nor machine.

Modeling Conflicting Norms

In light of the developments made in monoidal logics, a solution has been proposed to cope with these problems without adding further operators, considering deontic conditionals as primitive, or requiring the introduction of a dyadic operator (Peterson 2015; 2017; Peterson and Kulicki 2016). In a nutshell, the source of the problems of augmentation and detachment can be traced back to the rule ($\otimes$-out), whereas deontic explosion comes from the rule ($\oplus$-out). Accordingly, by defining a deontic logic meant to model normative rea-

soning while rejecting these rules, one can avoid important problems that would thwart the possible applications of deontic logic to artificial intelligence and automated reasoning. The system does not explode when conflicting obligations arise, it does not allow to arbitrarily augment the context specified within a conditional obligation, and it does not allow for unrestricted detachment of conditional obligations.

However, one limitation of Peterson's and Kulicki's (2016) approach is that although it can prevent the harmful consequences these problems usually entail, it cannot actually tell one what to do when such a situation arise (e.g., see Tosatto, Governatori, and Kelsen 2014). For the sake of the analysis, assume a language similar to the language of modal logics (with $Prop$ a collection of atomic propositions p_i and well formed formulas defined recursively).

$$\mathcal{L}_{\mathcal{CNR}} = \{(,), Prop, \otimes, \top, \multimap, \oplus, \bot, O\}$$

We define $\mathcal{CNR}$ as a symmetric closed deductive system with classical negation and co-tensor satisfying the axiom $(D)\ O\varphi \otimes O\neg\varphi \rightarrow \bot$, meant to represent normative consistency (i.e., an action cannot be obligatory and forbidden at the same time). An obligation $O\rho$ conditional to a context φ is modeled $\varphi \multimap O\rho$.

First, consider the following example. If an action ρ should be done under conditions φ, but it should not be done under conditions ψ, then under the context $\varphi \otimes \psi$ we obtain an inconsistency. In this case, deontic explosion does not follow, but the logic still does not tell us what to do in a context where conflicting obligations arise.

$$\dfrac{\dfrac{\dfrac{\dfrac{\varphi \multimap O\rho \longrightarrow \varphi \multimap O\rho}{\varphi \otimes (\varphi \multimap O\rho) \longrightarrow O\rho}\ (\text{cl}) \quad \dfrac{\dfrac{\psi \multimap O\neg\rho \longrightarrow \psi \multimap O\neg\rho}{\psi \otimes (\psi \multimap O\neg\rho) \longrightarrow O\neg\rho}\ (\text{cl})}{(\varphi \otimes (\varphi \multimap O\rho)) \otimes (\psi \otimes (\psi \multimap O\neg\rho)) \longrightarrow O\rho \otimes O\neg\rho}\ (\text{cut}) \quad \dfrac{}{O\rho \otimes O\neg\rho \longrightarrow \bot}\ (\text{D})}{(\varphi \otimes (\varphi \multimap O\rho)) \otimes (\psi \otimes (\psi \multimap O\neg\rho)) \longrightarrow \bot}\ (\text{cut})$$

Now, consider a variation of that example, one where we assume that some action τ should be performed in the context $\varphi \otimes \psi$. One interest of substructural logics, such as in multiplicative linear logic and the example at hand, is that they can be resource sensitive. As such, if one has $\varphi \otimes \psi$ and two possible applications, then one can use $\varphi \otimes \psi$ in only one of the two cases. Hence, under these premises, one must choose how to use $\varphi \otimes \psi$, either to obtain $\bot$ or to obtain τ, which is the action that should actually be done in that context.

What this example brings to light is that someone must intervene to determine what should be done. In this case, for instance, one might prioritize the conditional obligation $(\varphi \otimes \psi) \multimap O\tau$ over a conflict between $O\rho$ and $O\neg\rho$. However, this is not something that is accomplished within the logic: It requires the intervention of someone external to the system, adjusting it in order to deal with normative conflicts or ambiguities that can arise when evaluating how to act.

Consider a second example. Assume that some action ψ is obligatory under context φ (i.e., $\varphi \multimap O\psi$), but that we are actually under circumstances $\varphi \otimes \rho$. Under these circumstances, one will be able to deduce $\rho \otimes O\psi$. From this, however, one does not obtain that ψ should be accomplished, given that $O\psi$ cannot be detached from $\rho \otimes O\psi$. This is nice, given that otherwise one would face augmentation. That said, the logic does not tell us how to act in that context. It blocks the wrong inference pattern, which is good, but what should be done remains unavailable from a logical standpoint.

Further Problems for Ethical AI

There is a growing literature on ethical machines as well as moral artificial agents (for example, see Etzioni and Etzioni 2017). Approaches that try to tackle this subject can be divided within two broad categories. While top-down approaches advocate the imposition of external principles to the machine to ensure that it will behave ethically (e.g., principles meant to represent virtue ethics, deontology, or consequentialism), bottom-up approaches are rather empirical and start from the idea that a machine can learn by itself how to behave through deep learning. Beside the formal problems that have been highlighted so far with respect to the automation of ethical reasoning, top-down and bottom-up approaches have their own respective weaknesses.

On the one hand, in addition to the well-known fact that top-down approaches suffer from the flaws of the normative theory they wish to implement, there is an even more fundamental problem with this approach. Indeed, norms are general and are meant to be interpreted. To correctly interpret a norm, one must not only take into account the signification of the norm (its meaning), but one must also consider the context at hand. However, the meaning of a norm goes far beyond the formalism used to implement it, and it would be overoptimistic to assume that a machine can grasp that meaning. Furthermore, ethical behavior cannot be reduced to behaving in accordance with a normative code. Indeed, normative codes cannot foresee each and every situations that might occur. As such, unforeseen ethical dilemmas will arise, and what should be done in these situations will not be deducible from the normative code.

On the other hand, bottom-up approaches are grounded on a premise that goes against the very basis of ethics. Indeed, bottom-up approaches assume that a machine can learn how it should behave simply on the grounds of empirical data and empirical examples. However, it is well established in the literature that there is a semantical dichotomy between facts and norms. This dichotomy, known as David Hume's is-ought thesis, makes it impossible to determine what should be done from what is. Put differently, one cannot simply look into the empirical world to determine how one should behave. To paraphrase Immanuel Kant, one that searches ethical principles within the empirical world is doomed to fail. Ethics goes beyond that.

Closing Remarks and Future Research

How to act? How should we act? When reflecting upon ethical machines as well as artificial moral agents, there is a fallacy involving the word 'should'. By announcing a discussion on artificial moral agents or ethical machines, one expect 'should' to be used in an ethical sense. However, there is a change in meaning when this word is used in the context of machine learning. Consider an autonomous vehicle for example. One might be inclined to argue that autonomous vehicles can use deep learning to determine how they should behave. But the meaning of should in this case is not ethi-

cal: It takes into account what a car is and how it is meant to behave efficiently as a car. A car should stop at a red light. This has nothing to do with ethics, it is simply following the rules (e.g., the Highway Safety Code). A car has a purpose and can use deep learning to learn how it should behave in order to fulfill that purpose, but this does not amount to learn how it should behave as a moral agent (i.e., to learn how it should behave in order to behave ethically). Deep learning is task oriented, and this task is not ethics.

Overall, there is a fundamental incompatibility between ethics and an implicit premise to artificial intelligence. Starting from Turing's (1950) imitation game, scholars tend to assume two things. First, artificial intelligence only needs to replicate or imitate behaviors we consider to be intelligent behavior to be characterized as 'intelligent'. This follows from a second assumption, namely that processes are causal and deterministic. This is Turing's 'skin-of-an-onion' analogy, which lead him to conclude that the mind is purely mechanical. According to this view, a machine can reason ethically if it imitates how people reason ethically, as for instance by providing a justification supporting the decision. One problem with this conception, though, is that it leads to determinism, and determinism renders ethics void. Indeed, ethics is nothing without responsibility, and determinism leads to the negation of free will and, incidentally, to the negation of responsibility. This should cast an important doubt upon the mere possibility of artificial moral agents. Acting ethically requires a will to do so. One must accomplish the right thing for the right reasons, although there are no such things as 'the' right thing or 'the' right reasons. When facing most ethical dilemmas, there is a plurality of reasonable (though incompatible and conflicting) arguable right reasons and right choices. One important characteristic of ethical deliberation is that even though people disagree with a position, they can see why it is relevant and they can understand the rationale behind it. To accomplish this, one must be open minded and willing to reevaluate one's own opinion.

One lesson to be learned from the rather simplistic examples provided throughout this paper is that logic alone cannot tell us how to act. What should be accomplished depends ultimately on the normative premises one adopts, and these premises need to be evaluated with respect to specific situations. Implementing monoidal logics for the automation of normative reasoning would require a constant interaction between human and machine. This idea is reinforced if we consider the fact that we, human, often do not know what to do when facing conflicts of obligations. While it casts serious doubts on the idea that machines might come to determine how to act all by themselves (through logical reasoning), it reinforces the idea that development in artificial intelligence should be oriented towards an interaction between human and machine. Besides, one can only be horrified by the idea that a machine might one day determine its own rules to regulate its behavior.

References

Belnap, N. 1982. Display logic. *Journal of Philosophical Logic* 11(4):375–417.

Chellas, B. F. 1980. *Modal logic: An introduction*. Cambridge University Press.

Etzioni, A., and Etzioni, O. 2017. Incorporating ethics into artificial intelligence. *The Journal of Ethics* 21(4):403–418.

Goré, R. 1998. Substructural logics on display. *Logic Journal of the IGPL* 6(3):451–504.

Hyland, M., and De Paiva, V. 1993. Full intuitionistic linear logic (extended abstract). *Annals of Pure and Applied Logic* 64(3):273–291.

Lambek, J., and Scott, P. 1986. *Introduction to higher order categorical logic*. Cambridge University Press.

Lambek, J. 1958. The mathematics of sentence structure. *The American Mathematical Monthly* 65(3):154–170.

Lambek, J. 1968. Deductive systems and categories I. *Mathematical Systems Theory* 2(4):287–318.

Lambek, J. 1969. Deductive systems and categories II. Standard constructions and closed categories. In Hilton, P. J., ed., *Category Theory, Homology Theory and their Applications I*, volume 86 of *Lecture Notes in Mathematics*. Springer. 76–122.

Lambek, J. 1993. From categorial grammar to bilinear logic. In Schroeder-Heister, P., and Došen, K., eds., *Substructural Logics*, volume 2 of *Studies in Logic and Computation*. Oxford Science Publications. 207–237.

Lawvere, F. W. 1963. *Functorial semantics of algebraic theories and some algebraic problems in the context of functorial semantics of algebraic theories*. Ph.D. Dissertation, Columbia University.

Mac Lane, S. 1971. *Categories for the working mathematician*. Springer, 2nd edition.

Peterson, C., and Kulicki, P. 2016. Conditional normative reasoning with substructural logics: New paradoxes and De Morgan's dualities. In Roy, O.; Tamminga, A.; and Willer, M., eds., *Deontic Logic and Normative Systems*. College Publications. 220–236.

Peterson, C. 2014. Monoidal logics: How to avoid paradoxes. In Lieto, A.; Radicioni, D. P.; and Cruciani, M., eds., *Proceedings of the International Workshop on Artificial Intelligence and Cognition (AIC 2014)*, volume 1315. CEUR Workshop Proceedings. 122–133.

Peterson, C. 2015. Contrary-to-duty reasoning: A categorical approach. *Logica Universalis* 9(1):47–92.

Peterson, C. 2016a. A comparison between monoidal and substructural logics. *Journal of Applied Non-Classical Logics* 26(2):126–159.

Peterson, C. 2016b. From linguistics to deontic logic via category theory. *Logique et Analyse* 59(235):301–315.

Peterson, C. 2017. A logic for human actions. In Urbaniak, R., and Payette, G., eds., *Applications of Formal Philosophy: The road less traveled*, Logic, Argumentation, & Reasoning. Springer. 73–112.

Peterson, C. 2019. Monoidal logics: Completeness and classical systems. *Journal of Applied Non-Classical Logics* 29(2):121–151.

Tosatto, S. C.; Governatori, G.; and Kelsen, P. 2014. Detecting deontic conflicts in dynamic settings. In Cariani, F.; D., D. G.; Meheus, J.; and Parent, X., eds., *Deontic Logic and Normative Systems (DEON 2014)*, volume 8554 of *Lecture Notes in Computer Science*. Springer. 65–80.

Turing, A. M. 1950. Computing machinery and intelligence. *Mind* 59(236):433–460.