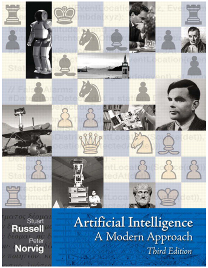
Discursive Input/Output Logic: Deontic Modals, and Computation

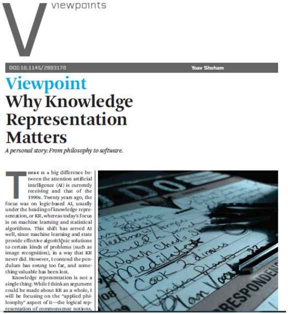
Ali Farjami
University of Luxembourg

PhD Defense

October 1, 2020

Logic-based AI





viewpoints

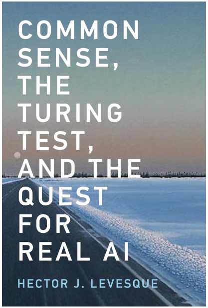
DOI:10.1145/2803170 Yoav Shoham

Viewpoint
Why Knowledge Representation Matters

A personal story: From philosophy to software.

THERE IS A big difference between the attention artificial intelligence (AI) is currently receiving and that of the 1990s. Twenty years ago, the focus was on logic-based AI, usually under the heading of knowledge representation, or KR, whereas today's focus is on machine learning and statistical algorithms. This shift has served AI well, since machine learning and stats provide effective algorithmic solutions to certain kinds of problems (such as image recognition), in a way that KR never did. However, I contend the pendulum has swung too far, and something valuable has been lost.

Knowledge representation is not a single thing. While I think an argument could be made about KR as a whole, I will be focusing on the "applied philosophy" aspect of it—the logical representation of commonsense notions,



- Logic in computer science (1980)
  - Relational database: formulas defines queries — *Codd 1981 (Turing award)*
  - Boolean satisfiability (SAT) — *SAT solvers*
  - Non-monotonic logic — *Common sense reasoning*
- Knowledge representation and reasoning (KR) (1990) — *Semantic Web*
- Machine learning + Symbolic logic (2020) — *Neuro-Symbolic AI*
- Trustworthy and responsible AI
- Deontic Logic — *Contrary-to-duty*

The Miners Example

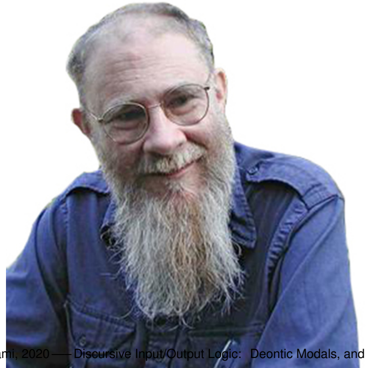
	in_A	in_B
$block\_A$	All live	All die
$block\_B$	All die	All live
$\neg(block\_A \vee block\_B)$	Nine live	Nine live

Deontic Models

Modal Logic Approach

Danielsson (1968), Hansson (1969), Føllesdal and Hilpinen (1970), van Fraassen (1973), and Lewis (1974), Kratzer (1977), ...

- A set of accessible worlds



Norm-based Approach

Makinson (1998), Makinson, van der Torre (2000,2001), Horty (2012), Hansen (2008), ...

- A set of norms
- Input/output logic
- Inference patterns

Research Objectives

Norms + Informational Modalities

- How can we use an algebraic setting such as Boolean algebras instead of a logical setting for building input/output logic on top of it?
- How can we introduce two groups of I/O operations similar to syntactical characterization of box and diamond in modal logic?
- How can we integrate conversational background informations, from Kratzerian framework, into input/output logic framework to build a more fruitful unified semantics for deontic modals?

Norms + Preferences

- How can we integrate input/output logic with Hansson and Lewis's conditional theory for building a new compositional theory about conditional deontic modals?
  - Resolving contrary-to-duty problems
  - Non-monotonic defeat mechanism within Hansson and Lewis's conditionals

Normative Reasoning in Computer Systems

- Providing a (faithful) embedding of some well-known deontic logics in HOL
- Encoding the logical embeddings in Isabelle/HOL

Research Objectives

Norms + Informational Modalities

- How can we use an algebraic setting such as Boolean algebras instead of a logical setting for building input/output logic on top of it?
- How can we introduce two groups of I/O operations similar to syntactical characterization of box and diamond in modal logic?
- How can we integrate conversational background informations, from Kratzerian framework, into input/output logic framework to build a more fruitful unified semantics for deontic modals?

Norms + Preferences

- How can we integrate input/output logic with Hansson and Lewis's conditional theory for building a new compositional theory about conditional deontic modals?
  - Resolving contrary-to-duty problems
  - Non-monotonic defeat mechanism within Hansson and Lewis's conditionals

Normative Reasoning in Computer Systems

- Providing a (faithful) embedding of some well-known deontic logics in HOL
- Encoding the logical embeddings in Isabelle/HOL

Research Objectives

Norms + Informational Modalities

- How can we use an algebraic setting such as Boolean algebras instead of a logical setting for building input/output logic on top of it?
- How can we introduce two groups of I/O operations similar to syntactical characterization of box and diamond in modal logic?
- How can we integrate conversational background informations, from Kratzerian framework, into input/output logic framework to build a more fruitful unified semantics for deontic modals?

Norms + Preferences

- How can we integrate input/output logic with Hansson and Lewis's conditional theory for building a new compositional theory about conditional deontic modals?
  - Resolving contrary-to-duty problems
  - Non-monotonic defeat mechanism within Hansson and Lewis's conditionals

Normative Reasoning in Computer Systems

- Providing a (faithful) embedding of some well-known deontic logics in HOL
- Encoding the logical embeddings in Isabelle/HOL

Research Objectives

Norms + Informational Modalities

- How can we use an algebraic setting such as Boolean algebras instead of a logical setting for building input/output logic on top of it?
- How can we introduce two groups of I/O operations similar to syntactical characterization of box and diamond in modal logic?
- How can we integrate conversational background informations, from Kratzerian framework, into input/output logic framework to build a more fruitful unified semantics for deontic modals?

Norms + Preferences

- How can we integrate input/output logic with Hansson and Lewis's conditional theory for building a new compositional theory about conditional deontic modals?
  - Resolving contrary-to-duty problems
  - Non-monotonic defeat mechanism within Hansson and Lewis's conditionals

Normative Reasoning in Computer Systems

- Providing a (faithful) embedding of some well-known deontic logics in HOL
- Encoding the logical embeddings in Isabelle/HOL

Research Objectives

Norms + Informational Modalities

- How can we use an algebraic setting such as Boolean algebras instead of a logical setting for building input/output logic on top of it?
- How can we introduce two groups of I/O operations similar to syntactical characterization of box and diamond in modal logic?
- How can we integrate conversational background informations, from Kratzerian framework, into input/output logic framework to build a more fruitful unified semantics for deontic modals?

Norms + Preferences

- How can we integrate input/output logic with Hansson and Lewis's conditional theory for building a new compositional theory about conditional deontic modals?
  - Resolving contrary-to-duty problems
  - Non-monotonic defeat mechanism within Hansson and Lewis's conditionals

Normative Reasoning in Computer Systems

- Providing a (faithful) embedding of some well-known deontic logics in HOL
- Encoding the logical embeddings in Isabelle/HOL

Methodology

Normative Reasoning

- Algebraic approach to input/output logic *Connection to modal logic*
  - Gabbay, Parent, and van der Torre: a geometrical view of I/O logic
  - *Upward-closed set of the infimum of* A instead of $Cn(A)$ *Compactness (?),...*
  - We use *upward-closed set* of A *Removing AND*
  - Reversibility of inference rules *Adding AND*
- Non-adjunctive input/output operations

$$\{\varphi_1, ..., \varphi_n\} \vdash \psi \Longrightarrow \begin{cases} \varphi_1 \wedge ... \wedge \varphi_n \vdash \psi \\ \\ \varphi_i \vdash \psi \qquad \varphi_i \in \{\varphi_1, ..., \varphi_n\} \end{cases}$$

- Semantical unification: Detachment + Conversational backgrounds
  - Syntactical unification *Adaptive logic*
- A semantical characterization of constrained input/output logic *Preferences*
  - Syntactical characterization *Adaptive logic characterizations of I/O logic*
  - No need to AND, SI and EQ required for syntactical translation

Logic Engineering

- *Shallow semantical embedding*
- Translating into Higher-order logic (HOL)
- Benchmark examples *CTD*

Kratzerian Framework

Conversational Backgrounds

Examples: knowledge, beliefs, relevant facts, desires, plans,...

Functions from evaluation worlds to sets of propositions

- *Modal base* determines the set of accessible worlds ($f(w)$)
- *Ordering source* induces the ordering on worlds ($g(w)$)

Quantification

$$[\![\text{be-allowed-to}]\!]^{w,f,g} = \lambda x\,(Best_{g(w)}(\bigcap f(w)) \cap x \neq \emptyset)$$

$$[\![\text{have-to}]\!]^{w,f,g} = \lambda x\,(Best_{g(w)}(\bigcap f(w)) \subseteq x)$$

where $Best_{g(w)}(\bigcap f(w))$ is given as follows:

$$\{w' \in \bigcap f(w) : \neg\exists w'' \in \bigcap f(w) \text{ such that } \exists y \in g(w) : w'' \in y \text{ and } w' \notin y\}$$

Quantification $\Longrightarrow$ Detachment

Conversational Backgrounds

Examples: knowledge, beliefs, relevant facts, desires, plans,...

Functions from evaluation worlds to sets of propositions

- *Modal base* determines the set of accessible worlds ($f(w)$)
- *Ordering source* induces the ordering on worlds ($g(w)$)

Quantification

Compatibility $[[\text{be-allowed-to}]]^{w,f,g} = \lambda x\,(Best_{g(w)}(\bigcap f(w)) \cap x \neq \emptyset)$

Entailment $[[\text{have-to}]]^{w,f,g} = \lambda x\,(Best_{g(w)}(\bigcap f(w)) \subseteq x)$

where $Best_{g(w)}(\bigcap f(w))$ is given as follows:

$$\{w' \in \bigcap f(w) : \neg\exists w'' \in \bigcap f(w) \text{ such that } \exists y \in g(w) : w'' \in y \text{ and } w' \notin y\}$$

Quantification $\Longrightarrow$ Detachment

Chisholm's Paradox

$$\text{DD}\ \frac{\bigcirc(t/g) \quad \bigcirc(g)}{\bigcirc(t)} \qquad\qquad \text{FD}\ \frac{\bigcirc(\neg t/\neg g) \quad \neg g}{\bigcirc(\neg t)}$$

Norm-based Semantics: Input/output logic

- "x is obligatory if a" $\rightsquigarrow$ "x can be detached in context a"
- **Output operation:** $x \in out(N^O, A)$ Normative system N^O
- *Detachment vs Quantification*

Detachment in Discursive Context : Out(N, Discursive Context)

- **Context** in a discourse — **Modal base** or **ordering source**
- Modal bases — **Factual**
- Ordering sources — Possible **inconsistency**

Out(N, modal base/ordering source)

Chisholm's Paradox

$$\text{DD}\ \frac{\bigcirc(t/g) \quad \bigcirc(g)}{\bigcirc(t)} \qquad\qquad \text{FD}\ \frac{\bigcirc(\neg t/\neg g) \quad \neg g}{\bigcirc(\neg t)}$$

Norm-based Semantics: Input/output logic

- "x is obligatory if a" $\rightsquigarrow$ "x can be detached in context a"
- **Output operation:** $x \in out(N^O, A)$ Normative system N^O
- *Detachment vs Quantification*

Detachment in Discursive Context : Out(N, Discursive Context)

- **Context** in a discourse — **Modal base** or **ordering source**
- Modal bases — **Factual**
- Ordering sources — Possible **inconsistency**

Out(N, modal base/ordering source)

Input/Output Logic: Proof system

On a Fundamental Problem of Deontic Logic

David Makinson
Les Etangs B2, La Ronce, 92410 Ville d'Avray, France
Email: d.makinson@unesco.org

Normative Systems

DAVID MAKINSON and LEENDERT VAN DER TORRE

INPUT/OUTPUT LOGICS

(Received on 16 November 1999; final version received on 13 March 2000)

- T: infer every $(\top, \top)$
- SI: from (a, x) and $\vdash b \to a$, infer (b, x)
- WO: from (a, x) and $\vdash x \to y$, infer (a, y)
- AND: from (a, x) and (a, y), infer $(a, x \wedge y)$
- OR: from (a, x) and (b, x), infer $(a \vee b, x)$
- CT: from (a, x) and $(a \wedge x, y)$, infer (a, y)

- **Unconstrained input/output logic**
- **Constrained input/output logic**

AGM theory/Contrary-to-duty

Input/Output Logic: Output operations

- Simple-Minded Output:

$$out_1(N,A) = Cn(N(Cn(A)))$$

- Basic Output:

$$out_2(N,A) = \bigcap\{Cn(N(V)) \mid A \subseteq V, V \text{ complete}\}$$

- Simple-Minded Reusable Output:

$$out_3(N,A) = \bigcap\{Cn(N(B)) \mid A \subseteq B = Cn(B) \supseteq N(V)\}$$

- Basic Reusable Output:

$$out_4(N,A) = \bigcap\{Cn(N(V)) \mid A \subseteq V \supseteq N(V), V\text{complete}\}$$

$deriv_i(N)$	Rules
$deriv_1(N)$	{$\top$, SI, WO, AND}
$deriv_2(N)$	{$\top$, SI, WO, OR, AND}
$deriv_3(N)$	{$\top$, SI, WO, CT, AND}
$deriv_4(N)$	{$\top$, SI, WO, OR, CT, AND}

Input/Output Logic: Output operations

-
  - Simple-Minded Output:

$$out_1(N,A) = \cancel{Cn}(N(\cancel{Cn}(A))) \qquad Up(N(Up(A)))$$

  - Basic Output:

$$out_2(N,A) = \bigcap\{\cancel{Cn}(N(V)) \mid A \subseteq V, V \text{ complete}\} \qquad \ldots$$

  - Simple-Minded Reusable Output:

$$out_3(N,A) = \bigcap\{\cancel{Cn}(N(B)) \mid A \subseteq B = Cn(B) \supseteq N(V)\} \qquad \cdots$$

$deriv_i(N)$	Rules
$deriv_1(N)$	{⊤, SI, WO, ~~AND~~}
$deriv_2(N)$	{⊤, SI, WO, OR, ~~AND~~}
$deriv_3(N)$	{⊤, SI, WO, ~~CT~~, ~~AND~~}

- I/O operations over Boolean algebras
- Stone's representation theorem — Possible world semantics
- $Up(X) = \{x \in B | \exists y \in X, y \leq x\}$ — $a \wedge b \notin Up(a, b)$

Non-adjunctive Logical Systems

Deriving the conjunctive formula $\varphi \wedge \psi$ from the set $\{\varphi, \psi\}$ fails

Discursive Systems

"[...] the joining of a thesis to a discursive system has a different intuitive meaning than has assertion in an ordinary system." jaskowski1969

$A \rightarrow_d B$ $\Diamond A \rightarrow B$

$$\boxed{Up(A \cup B) = Up(A) \cup Up(B); \; out_i(N, A) = \bigcup_{a \in A} out_i(N, a)}$$

Discursive Input/Output Logic: Output operations

- Simple-Minded Output :

$$out_1^{\mathcal{B}}(N, A) = Up(N(Up(A)))$$

- Basic Boolean I/O operation:

$$out_2^{\mathcal{B}}(N, A) = \bigcap\{Up(N(V)), A \subseteq V, V \text{ is saturated }\}$$

- Reusable Boolean I/O operation:

$$out_3^{\mathcal{B}}(N, A) = \bigcap\{Up(N(V)), A \subseteq V = Up(V) \supseteq N(V)\}$$

Discursive Input/Output Logic: Output operations

- Zero Boolean I/O operation: ($out_R(N,A) = Eq(N(A))$, $out_L(N,A) = N(Eq(A))$)

$$out_0^{\mathcal{B}}(N,A) = Eq(N(Eq(A)))$$

- Simple-I Boolean I/O operation:

$$out_I^{\mathcal{B}}(N,A) = Eq(N(Up(A)))$$

- Simple-II Boolean I/O operation:

$$out_{II}^{\mathcal{B}}(N,A) = Up(N(Eq(A)))$$

- Simple-Minded Output :

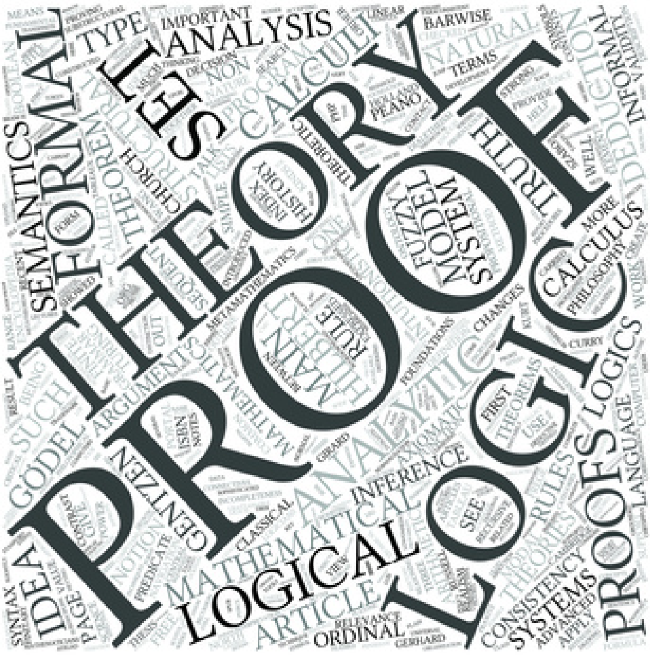
$$out_1^{\mathcal{B}}(N,A) = Up(N(Up(A)))$$

- Basic Boolean I/O operation:

$$out_2^{\mathcal{B}}(N,A) = \bigcap\{Up(N(V)), A \subseteq V, V \text{ is saturated }\}$$

- Reusable Boolean I/O operation:

$$out_3^{\mathcal{B}}(N,A) = \bigcap\{Up(N(V)), A \subseteq V = Up(V) \supseteq N(V)\}$$


PROOF
THEORY
LOGIC
LOGICAL
MATHEMATICAL
ANALYSIS
SET
FORMAL
SEMANTICS
PROOFS
CALCULUS
TRUTH
ARGUMENTS
GÖDEL
GENTZEN
INFERENCE
ARTICLE
ORDINAL
SYSTEMS
CONSISTENCY
TYPE
IMPORTANT
BARWISE
IDEA
SYNTAX
TERMS
FUZZY
MODEL
SYSTEM
HISTORY
INDEX
LOGICS
LANGUAGE
PHILOSOPHY
CHURCH
METAMATHEMATICS
FOUNDATIONS
CLASSICAL
CURRY
CHANGES
FIRST
THEOREMS
SEE
RULES
PEANO
PAGE
GERHARD
RELEVANCE

Discursive Input/Output Logic: Proof system

$(A, x) \in deriv_i(N)$ if $(a, x) \in deriv(N)$ for some $a \in A$

$deriv_i^{\mathcal{B}}$	Rules
$deriv_R^{\mathcal{B}}$	{EQO}
$deriv_L^{\mathcal{B}}$	{EQI}
$deriv_0^{\mathcal{B}}$	{EQI, EQO}
$deriv_I^{\mathcal{B}}$	{SI, EQO}
$deriv_{II}^{\mathcal{B}}$	{WO, EQI}
$deriv_1^{\mathcal{B}}$	{SI, WO}
$deriv_2^{\mathcal{B}}$	{SI, WO, OR}
$deriv_3^{\mathcal{B}}$	{SI, WO, T}

$$\text{EQO}\ \frac{(a,x) \qquad x = y}{(a,y)} \qquad \text{WO}\ \frac{(a,x) \qquad x \leq y}{(a,y)}$$

$$\text{EQI}\ \frac{(a,x) \qquad a = b}{(b,x)} \qquad \text{OR}\ \frac{(a,x) \qquad (b,x)}{(a \vee b, x)}$$

$$\text{SI}\ \frac{(a,x) \qquad b \leq a}{(b,x)} \qquad \text{T}\ \frac{(a,x) \qquad (x,y)}{(a,y)}$$

$\Diamond$ vs $\Box$ $\qquad (\Box A \wedge \Box B) \rightarrow \Box(A \wedge B)$

Discursive Input/Output Logic: Proof system

$(A, x) \in deriv_i(N)$ if $(a, x) \in deriv(N)$ for some $a \in A$

$deriv_i^{\mathcal{B}}$	Rules
$deriv_R^{\mathcal{B}}$	{EQO}
$deriv_L^{\mathcal{B}}$	{EQI}
$deriv_0^{\mathcal{B}}$	{EQI, EQO}
$deriv_I^{\mathcal{B}}$	{SI, EQO}
$deriv_{II}^{\mathcal{B}}$	{WO, EQI}
$deriv_1^{\mathcal{B}}$	{SI, WO}
$deriv_2^{\mathcal{B}}$	{SI, WO, OR}
$deriv_3^{\mathcal{B}}$	{SI, WO, T}

$$\text{EQO}\ \frac{(a,x) \qquad x = y}{(a,y)} \qquad \text{WO}\ \frac{(a,x) \qquad x \leq y}{(a,y)}$$

$$\text{EQI}\ \frac{(a,x) \qquad a = b}{(b,x)} \qquad \text{OR}\ \frac{(a,x) \qquad (b,x)}{(a \vee b, x)}$$

$$\text{SI}\ \frac{(a,x) \qquad b \leq a}{(b,x)} \qquad \text{T}\ \frac{(a,x) \qquad (x,y)}{(a,y)}$$

$\Diamond$ vs $\Box$ $\qquad (\Box A \wedge \Box B) \rightarrow \Box(A \wedge B)$

Adding Other Rules

$$\text{AND}\ \frac{(p,q) \qquad (p,r)}{(p, q \wedge r)} \qquad\qquad \text{CT}\ \frac{(p,q) \qquad (p \wedge q, r)}{(p,r)}$$

Reversibility of Inference Rules

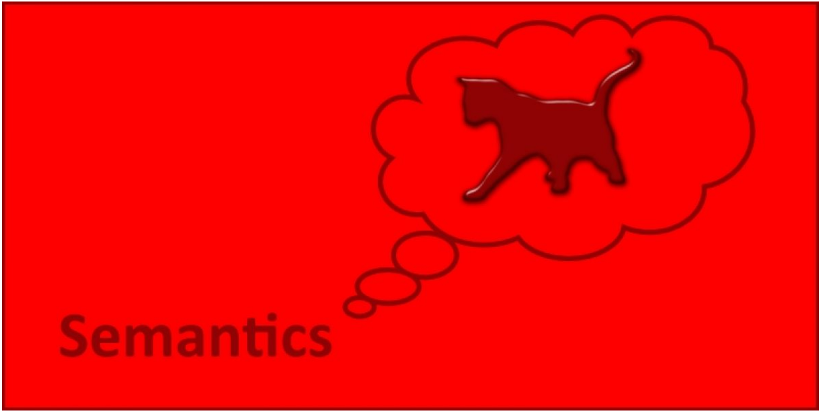
	SI	WO	CT		AND	OR
SI		√	none?	none?	none?	√
WO	√		SI, CT	√	√	none?
CT	√	√			SI, AND, CT	none?
AND	√	√	SI, CT	√		WO, OR, AND
OR	√	√	SI, CT, OR	SI, CT, OR	SI, AND, OR	

(Makinson and van der Torre 2000)

Adding AND: Output operations + Iteration of AND

$$
\begin{aligned}
out_i^{AND^0}(N,A) &= \mathbf{out}_\mathbf{i}^{\mathcal{B}}(N,A) \\
out_i^{AND^{n+1}}(N,A) &= out_i^{AND^n}(N,A) \cup \\
&\quad \{y \wedge z : y, z \in out_i^{AND^n}(N,\{a\}),\ \ a \in A\} \\
out_i^{AND}(N,A) &= \bigcup_{n \in N} out_i^{AND^n}(N,A)
\end{aligned}
$$

$deriv_i^X$	Rules
$deriv_{II}^{AND}$	{WO, EQI, AND}
$deriv_1^{AND}$	{SI, WO, AND}
$deriv_2^{AND}$	{SI, WO, OR, AND}


Semantics

Unification: Consistent premise sets

Suppose $\bigcap f(w) \neq \emptyset$

$$[[\text{be-allowed-to}]]^{w,f} = \lambda N^P \lambda x \ (x \in out(N^P, \{\bigcap f(w)\}))$$

$$[[\text{have-to}]]^{w,f} = \lambda N^O \lambda x \ (x \in out(N^O, \{\bigcap f(w)\}))$$

$$M, w \vDash \Box a_1 \wedge ... \wedge \Box a_n = \Box(a_1 \wedge ... \wedge a_n) \qquad a_i \in f(w)$$

Unification: Inconsistent premise sets

Suppose $\bigcap g(w) = \emptyset$, and

$$\text{Maxfamily}^{\cap}(g(w)) = \{\bigcap A | A \subseteq g(w) \text{ and } A \text{ is consistent and maximal}\}$$

$$[[\text{be-allowed-to}]]^{w,g} = \lambda N^P \lambda x\ (x \in out(N^P, \text{Maxfamily}^{\cap}(g(w))))$$

$$[[\text{have-to}]]^{w,g} = \lambda N^O \lambda x\ (x \in out(N^O, \text{Maxfamily}^{\cap}(g(w))))$$

$$M, w \vDash \Diamond(\overbrace{a_1 \wedge \ldots \wedge a_n}^{\text{Maxfamily}}) \wedge \ldots \wedge \Diamond(\overbrace{b_1 \wedge \ldots \wedge b_m}^{\text{Maxfamily}}) \qquad a_i, b_j \in g(w)$$

The Miners Example

1- Either the miners are in shaft A or in shaft B.

2- If the miners are in shaft A, we should block shaft A.

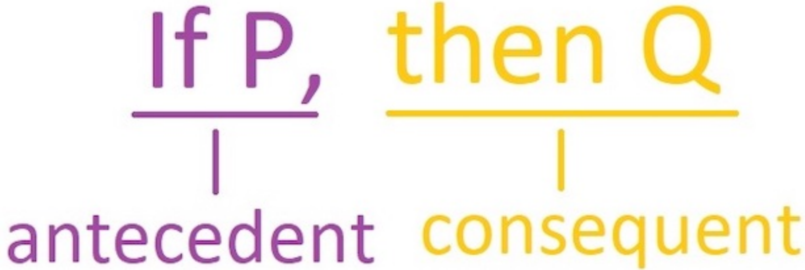
3- If the miners are in shaft B, we should block shaft B.

4- We should block neither shaft.

- Syntactical analysis (Not satisfactory)
- Not allowed by the baseline algorithm of Kratzerian framework (Cariani 2020)
- Kolodny and MacFarlane 2010 (Modus pones is invalid)

$$N = \{(ShA, blA), (ShB, blB), (\top, \neg blA \wedge \neg blB)\}$$

- $M, w \vDash \Box(shA \vee shB)$ — a set of **factual informations**
  $f(w) = \{shA \vee shB\}$ — $\neg blA \wedge \neg blB \in out(N^O, \{shA \vee shB\})$
- $M, w \vDash \Diamond shA \wedge \Diamond shB$ — a set of possible **inconsistent informations**
  $g(w) = \{shA, shB\}$ — $blA, blB, \neg blA \wedge \neg blB \in out(N^O, \{shA, shB\})$
- $C = \{blA\}$ — **constrained I/O logic**
  $f(w) = \{shA\}$ — $blA \in out_c(N^O, \{shA\})$


If P, then Q
antecedent
consequent

Syntax vs Semantics

$$\mathbf{Fm}(X) = \langle Fm(X), \wedge^{\mathbf{Fm}(X)}, \vee^{\mathbf{Fm}(X)}, \neg^{\mathbf{Fm}(X)}, \top^{\mathbf{Fm}(X)}, \bot^{\mathbf{Fm}(X)} \rangle$$

$\Gamma \vDash_{\mathbf{BA}} \varphi$ if and only if $\Gamma \vdash_{C} \varphi$

$$(p, q) \in derive_i^{\mathbf{Fm}(X)}(N)$$

if and only if

$V(q) \in out_i^{\mathcal{B}}(N^V, \{V(p)\})$ for every $\mathcal{B} \in \mathbf{BA}$, for every valuation V on $\mathcal{B}$

Neighborhood Characterization of I/O Operations $\qquad f : P(W) \to P(P(W))$

A Compositional Theory of Conditional Obligation and Permission

Input/output logic + Constraints (preferences)

$$\varphi > \bigcirc\psi \in derive_i^O(N^O)$$

if and only if

$(\varphi, \psi) \in derive_i^{\mathbf{Fm}(X)}(N^O)$ and

For every preference Boolean algebra $M = \langle \mathcal{B}, \mathcal{V}, \succeq_f \rangle$, for every valuation $V_i \in opt_{\succeq_f}(\varphi)$ we have $V_i(\psi) = 1_{\mathcal{B}}$

$$\varphi > P\psi \in derive_i^P(N^P)$$

if and only if

$(\varphi, \psi) \in derive_i^{\mathbf{Fm}(X)}(N^P)$ and

For every preference Boolean algebra $M = \langle \mathcal{B}, \mathcal{V}, \succeq_f \rangle$, there is a valuation $V_i \in opt_{\succeq_f}(\varphi)$ such that $V_i(\psi) = 1_{\mathcal{B}}$

- $\mathcal{B}$ is a Boolean algebra,
- $\mathcal{V} = \{V_i\}_{i \in I}$ is the set of valuations from $\mathbf{Fm}(X)$ on $\mathcal{B}$,
- $\succeq_f \subseteq \mathcal{V} \times \mathcal{V}$: $\succeq_f$ is a betterness or comparative goodness relation over valuations from $\mathbf{Fm}(X)$ to $\mathcal{B}$ such that $V_i \succeq_f V_j$ iff $(\{\varphi | V_i(\varphi) = 1_{\mathcal{B}}\}, \{\psi | V_j(\psi) = 1_{\mathcal{B}}\}) \in f$.

Chisholm's Paradox

It ought to be that a certain man go to help his neighbors.

It ought to be that if he goes he tell them he is coming.

If he does not go, he ought not to tell them he is coming.

$w_1 \bullet go, tell$

$w_2 \bullet go \quad w_3 \bullet$

$w_4 \bullet tell$

$$N^O = \{(\top, g), (g, t), (\neg g, \neg t)\}$$

- $\top > \bigcirc g \in derive_i^O(N^O)$
- $g > \bigcirc t \in derive_i^O(N^O)$
- $\neg g > \bigcirc(\neg t) \in derive_i^O(N^O)$

KR Tools: Higher-order logic theorem provers

- Church's simple theory of types — λ-calculus/Henkin models
- HOL provers
  - interactive: Isabelle/HOL, HOL4, Hol Light, Coq/HOL
  - automated: LEO-II, Satallax, Nitpick, Isabelle/HOL
- Isabelle/HOL
  - Bridges to external theorem provers — Sledgehammer tool
  - Model finders — Nitpick
  - Sophisticated user interaction

Deontic Logics in HOL

Semantical Embedding

Aligning Henkin models $\langle D, I\rangle$ with Kripke models $\langle S, R, V\rangle$

Possible worlds $s \in S$	Set of individuals $s_i \in D_i$
Acceptability relation R sRu	Binary predicates $r_{i\to i\to o}$ $Ir_{i\to i\to o}(s_i, u_i) = \top$
Propositional letters p^j Valuation function $s \in V(p^j)$	Unary predicates $p^j_{i\to o}$ Interpretation function $Ip^j_{i\to o}(s_i) = \top$

$$
\begin{aligned}
\lfloor p^j \rfloor &= p^j_\tau \\
\lfloor \neg\varphi \rfloor &= \neg_{\tau\to\tau} \lfloor\varphi\rfloor \\
\lfloor \varphi \vee \psi \rfloor &= \vee_{\tau\to\tau\to\tau} \lfloor\varphi\rfloor\lfloor\psi\rfloor \\
\lfloor \Box\varphi \rfloor &= \Box_{\tau\to\tau} \lfloor\varphi\rfloor
\end{aligned}
$$

$$
\begin{aligned}
\neg_{\tau\to\tau} &= \lambda A_\tau \lambda X_i \neg(A\,X) \\
\vee_{\tau\to\tau\to\tau} &= \lambda A_\tau \lambda B_\tau \lambda X_i (A\,X \vee B\,X) \\
\Box_{\tau\to\tau} &= \lambda A_\tau \lambda X_i \forall Y_i (\neg(r_{i\to i\to o} X\,Y) \vee A\,Y)
\end{aligned}
$$

- Modal translation of I/O operations in HOL $\Box$
- Åqvist dyadic deontic logic **E** in HOL $\bigcirc(/)$
- Dyadic deontic logic by Carmo and Jones in HOL $\Box_p, \Box_a, \bigcirc(/), O_p, O_a$

Isabelle/HOL: An infrastructure for deontic reasoning

```
theory IOBoolean
  imports Main

begin

typedecl i (* type for boolean elements *)
type_synonym τ = "(i⇒bool)"
consts     N :: "i⇒i⇒bool" ("N")    (* Nor
consts   dis :: "i⇒i⇒i" (infixr"∨"50)
consts   con :: "i⇒i⇒i"  (infixr"∧"60)
consts   neg :: "i⇒i"   ("¬_"[52]53)
consts   top :: i ("1")
consts   bot :: i ("0")

axiomatization where
  COMdis  : "∀X. ∀Y. (X ∨ Y) = (Y ∨ X)" and
  COMcon  : "∀X. ∀Y. (X ∧ Y) = (Y ∧ X)" and
  ASSdis  : "∀X. ∀Y. ∀ Z. (X ∨ (Y ∨ Z)) = (X ∨ (Y ∨ Z))" and
  ASScon  : "∀X. ∀Y. ∀ Z. (X ∧ (Y ∧ Z)) = (X ∧ (Y ∧ Z))" and
  IDEdis  : "∀X. (X ∨ 0) = X" and
  IDEcon  : "∀X. (X ∧ 1) = X" and
  COMPdis : "∀X. (X ∨ ¬X) = 1" and
  COMPcon : "∀X. (X ∧ (¬X) ) = 0" and
  Ddiscon : "∀X. ∀Y. ∀ Z. (X ∨ (Y ∧ Z)) = ((X ∨ Y) ∧ (X ∨ Z))" and
  Dcondis : "∀X. ∀Y. ∀ Z. (X ∧ (Y ∨ Z)) = ((X ∧ Y) ∨ (X ∧ Z))"
```

$$
\begin{aligned}
\lfloor p^j \rfloor &= p^j_i \qquad p^j \in X \\
\lfloor \top \rfloor &= \top_i \\
\lfloor \bot \rfloor &= \bot_i \\
\lfloor \neg\varphi \rfloor &= \neg_{i\to i}(\lfloor \varphi \rfloor) \\
\lfloor \varphi \vee \psi \rfloor &= \vee_{i\to i\to i} \lfloor \varphi \rfloor \lfloor \psi \rfloor \\
\lfloor \varphi \wedge \psi \rfloor &= \wedge_{i\to i\to i} \lfloor \varphi \rfloor \lfloor \psi \rfloor \\
\lfloor d_i(N)(\varphi,\psi) \rfloor &= (\bigcirc_i(N)_{\tau\to\tau}\{\lfloor \varphi \rfloor\}) \lfloor \psi \rfloor
\end{aligned}
$$

Isabelle/HOL: Output operations in HOL

$(\varphi, \psi) \in derive_i^{\mathbf{Fm}(X)}(N)$ iff $V(\psi) \in out_i^{\mathcal{B}}(N^V, \{V(\varphi)\})$ in all Boolean normative models $(\mathcal{N} = \langle \mathcal{B}, V, N^V \rangle)$

Faithful embedding

```
definition ordeIOB :: "i⇒τ" (infixr"≤"80) where "X ≤ Y ≡ ((X ∧ Y) = X)"
definition satuIOB :: "τ ⇒ bool" ("Saturated")  where
"Saturated V ≡ ∀X. ∀Y. (((V (X ∨ Y))⟶ (V X ∨ V Y)) ∧ ((V X ∧ (X≤Y)) ⟶ V Y))"
definition UpwardIOB :: "τ ⇒ τ" ("Up") where " Up V ≡ λX. (∃Z . (V Z ∧ Z ≤ X))"


definition outI :: "α ⇒ τ ⇒ τ" ("○I<_;_>")
  where "○I<M;A> ≡ λX. ∃U. (∃Y. (∃Z. (A Z ∧ (Z=Y) ∧ M Y U ∧ (U ≤ X) ) ) )"

definition outII :: "α ⇒ τ ⇒ τ" ("○II<_;_>")
  where "○II<M;A> ≡ λX. ∃U. (∃Y. (∃Z. (A Z ∧ (Z≤Y) ∧ M Y U ∧ (U = X) ) ) )"

definition out1 :: "α ⇒ τ ⇒ τ" ("○1<_;_>")
  where "○1<M;A> ≡ λX. ∃U. (∃Y. (∃Z. (A Z ∧ (Z≤Y) ∧ M Y U ∧ (U ≤ X) ) ) )"

definition out2 :: "α ⇒τ ⇒ τ" ("○2<_;_>")
  where "○2<M;A> ≡ λX. (∀V. ( (Saturated V) ∧ (∀U. (A U ⟶ V U ))
                  ⟶ (∃Y. (∃Z. ( (V Y) ∧ (M Y Z) ∧ (Z≤X)) )) ))"

definition out3 :: "α ⇒ τ ⇒ τ" ("○3<_;_>")
  where "○3<M;A> ≡ λX. (∀V. ( ((V = Up V) ∧ (∀U. (A U ⟶ V U ))∧(∀W.(∃Y.(V Y ∧ (M Y W)))⟶V W))
                  ⟶ (∃Y. (∃Z. ((Z≤ X)∧ N Y Z ∧ V Y) )) ))"
```

Isabelle/HOL: Output operations in HOL

$$\begin{cases} \text{Sub-rel}\, R\, Q \equiv \forall uv.\;\; Ruv \rightarrow Quv \\ \text{Close-AND}\, Q \equiv \forall uvw.\;\; (Quv \wedge Quw \rightarrow Qu(v \wedge w)) \\ \text{TCAND}\, R \equiv \lambda XY.\, \forall Q.\;\; \text{Close-AND}\, Q \;\rightarrow\; (\text{Sub-rel}\, R\, Q \rightarrow QXY) \end{cases}$$

```
theory outoperation imports IOBoolean
begin

definition Rout :: "α ⇒ τ⇒i⇒i⇒bool" ("Rout<_;_>")
  where "Rout<M;A> ≡ λZ. λX. ∃U. (∃Y. ( (A Z ∧ (Z≤Y) ∧ M Y U ∧ (U ≤ X)) ) )"
definition Sub_rel :: "α⇒α⇒bool" where "Sub_rel R Q ≡ ∀ u v. R u v ⟶ Q u v"

(* OUT1 orginal *)
definition Close_AND :: "α ⇒ bool" where "Close_AND Q ≡ ∀u v w.(Q u v ∧ Q u w ⟶ (Q u (v ∧ w)))"
definition TCAND :: "α ⇒ α" where "TCAND R ≡ λ X Y. ∀ Q. Close_AND Q ⟶ (Sub_rel R Q ⟶ Q X Y)"
definition outAND :: "α⇒ τ ⇒ τ" ("○AND<_;_>") where "○AND<M;A> ≡ λX. ∃Y. TCAND (Rout<M;A>) Y X"
(* OUT2 orginal *)
definition Close_OR :: "α ⇒ bool" where "Close_OR Q ≡ ∀ u v w. (Q v u ∧ Q w u ⟶ (Q (v ∨ w) u))"
definition TCOR :: "α ⇒ α" where "TCOR R ≡ λ X Y. ∀ Q. Close_OR Q ⟶ (Sub_rel R Q ⟶ Q X Y)"
definition outOR :: "α⇒ τ ⇒ τ" ("○OR<_;_>") where "○OR<M;A> ≡ λX. ∃Y. TCOR (Rout<M;A>) Y X"
definition outORAND :: "α⇒ τ ⇒ τ" ("○ORAND<_;_>")
  where "○ORAND<M;A> ≡ λX. ∃Y. TCAND (TCOR (Rout<M;A>))  Y X "
```

Isabelle/HOL: I/O proof systems in HOL

```
(*Derive2-Ob*)
definition derSIWOORAND :: "α⇒α" ("derSIWOORAND< >")
  where "derSIWOORAND<M> ≡ TCAND (TCOR (TCWO (TCSI (M))))"

(*Derive3-Ob*)
definition derSIWOCT :: "α⇒α" ("derSIWOCT< >")
  where "derSIWOCT<M> ≡ TCCT (TCWO (TCSI (M)))"
definition derSIWOCTAND :: "α⇒α" ("derSIWOCTAND< >")
  where "derSIWOCTAND<M> ≡ TCAND (TCCT (TCWO (TCSI (M))))"

(*Derive4-Ob*)
definition derSIWOCTORAND :: "α⇒α" ("derSIWOCTORAND< >")
  where "derSIWOCTORAND<M> ≡ TCAND (TCOR (TCCT (TCWO (TCSI (M)))))"
```

```
lemma "Close_AND (TCAND N)" unfolding Defst TCAND_def
  by metis

lemma "(M a b ∨ (∃ y. M y b ∧ (a ≤ y))) ⟶ derSI<M> a b"
using  Sub_rel_def Close_SI_def  TCSI_def
  unfolding Defst and Defs  derSI_def
  by metis

(*OUT1  completness*)
lemma "(○1<N;((λX. X = a))> y ⟶ derSIWO<N> a y)"
  using  Sub_rel_def Close_SI_def Close_WO_def  TCSI_def   TCWO_def
  unfolding Defst and Defs  derSI_def Sub_rel_def TCWO_def TCSI_def
  by metis
```

Normative Reasoning

- Discursive input/output logic: Detachment vs Quantification
  - Input/output logic for permission: Removing AND rule
  - Input/output logic for obligation: Adding AND rule
  - Semantic unification: Integrating input/output logic into Kratzerian framework
- Normative reasoning + Preferences/Normality
  - A compositional theory of conditional obligation and permission
- Algebraic method: I/O framework on top of any abstract logic
  - Input/output methodology: Secretarial assistant ($\mathcal{A} = \langle \mathcal{L}, C \rangle$)

Logic Engineering

- A dataset for normative reasoning: LogiKEy methodology
  - Faithful embedding of some deontic logics in HOL
  - Isabelle/HOL: An infrastructure for deontic reasoning

Normative Reasoning

- Discursive input/output logic: Detachment vs Quantification
  - Input/output logic for permission: Removing AND rule
  - Input/output logic for obligation: Adding AND rule
  - Semantic unification: Integrating input/output logic into Kratzerian framework
- Normative reasoning + Preferences/Normality
  - A compositional theory of conditional obligation and permission
- Algebraic method: I/O framework on top of any abstract logic
  - Input/output methodology: Secretarial assistant ($\mathcal{A} = \langle \mathcal{L}, C \rangle$)

Logic Engineering

- A dataset for normative reasoning: LogiKEy methodology
  - Faithful embedding of some deontic logics in HOL
  - Isabelle/HOL: An infrastructure for deontic reasoning

Normative Reasoning

- Discursive input/output logic: Detachment vs Quantification
  - Input/output logic for permission: Removing AND rule
  - Input/output logic for obligation: Adding AND rule
  - Semantic unification: Integrating input/output logic into Kratzerian framework
- Normative reasoning + Preferences/Normality
  - A compositional theory of conditional obligation and permission
- Algebraic method: I/O framework on top of any abstract logic
  - Input/output methodology: Secretarial assistant ($\mathcal{A} = \langle \mathcal{L}, C \rangle$)

Logic Engineering

- A dataset for normative reasoning: LogiKEy methodology
  - Faithful embedding of some deontic logics in HOL
  - Isabelle/HOL: An infrastructure for deontic reasoning

Normative Reasoning

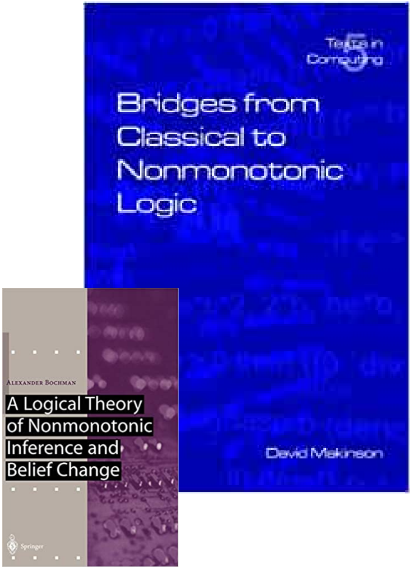
- Discursive input/output logic: Detachment vs Quantification
  - Input/output logic for permission: Removing AND rule
  - Input/output logic for obligation: Adding AND rule
  - Semantic unification: Integrating input/output logic into Kratzerian framework
- Normative reasoning + Preferences/Normality
  - A compositional theory of conditional obligation and permission
- Algebraic method: I/O framework on top of any abstract logic
  - Input/output methodology: Secretarial assistant ($\mathcal{A} = \langle \mathcal{L}, C \rangle$)

Logic Engineering

- A dataset for normative reasoning: LogiKEy methodology
  - Faithful embedding of some deontic logics in HOL
  - Isabelle/HOL: An infrastructure for deontic reasoning

Future Work: Normative reasoning

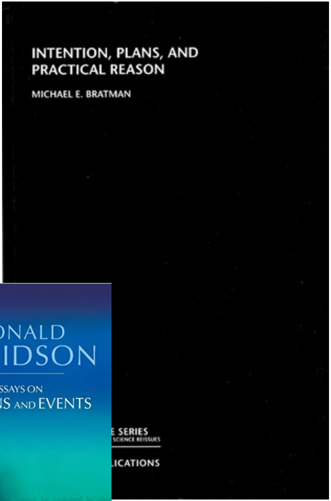
- Adding rules to logic:
  A full characterization
- Logic without reflexivity:
  Application in other domains such as causality
- Credulous (brave) inference:
  Belief change

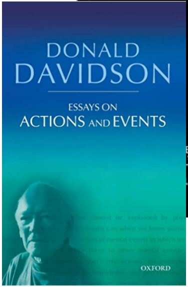


Future Work: Practical reasoning

How do norms interact with informational modalities such as beliefs and knowledge, and motivational modalities such as intentions and desires? (Ten Problems of Deontic Logic and Normative Reasoning in Computer Science)

- Anankastic conditionals (means-end reasoning)
- Rational architecture: BOID
- Human-Computer Interaction

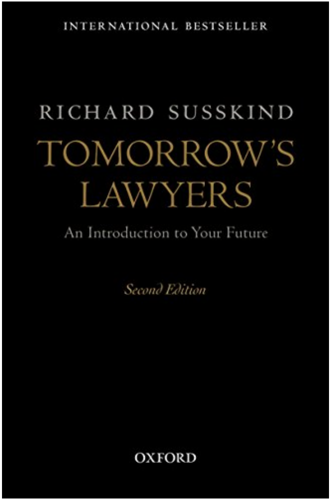




Future Work: Online legal guidance systems

“what if we, as lawyers, could make our knowledge and expertise available through a wide range of online legal services, whether for the drafting of documents or for the resolution of disputes?” (Susskind)

- Improving normative expressivity of the implemented logics in HOL
- A domain for individuals
- Logic and ontology
- Higher-order deontic logic



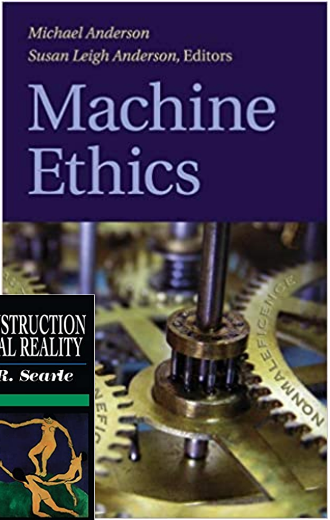
Future Work: Autonomous vehicles

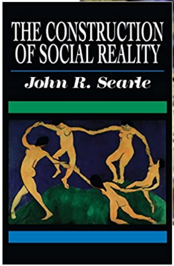
Can ethical frameworks and rules derived for human behavior be implemented as control algorithms in automated vehicles? (Implementable Ethics for Autonomous Vehicles)

- Cost Functions and consequentialism
- Constraints and deontological Ethics
- Norms and preferences
- How does the deontological approach fare with uncertainty?

(Normative) multi-agent systems

- Constitutive norms, Regulative norms
- Privacy policies and Knowledge management





Many Thanks!

سپاسگزارم#