
Logico-Mathematical Foundations

Selmer Bringsjord • Naveen Sundar G. • Mei Si

Department of Computer Science
Department of Cognitive Science
Lally School of Management & Technology
Rensselaer AI & Reasoning (RAIR) Lab
selmer@rpi.edu • govinn@rpi.edu
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA

Logico-Mathematical Foundations

Selmer Bringsjord • Naveen Sundar G. • Mei Si

Department of Computer Science
Department of Cognitive Science
Lally School of Management & Technology
Rensselaer AI & Reasoning (RAIR) Lab
selmer@rpi.edu • govinn@rpi.edu
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA

Logico-Mathematical Foundations

Selmer Bringsjord • Naveen Sundar G. • Mei Si

Department of Computer Science
Department of Cognitive Science
Lally School of Management & Technology
Rensselaer AI & Reasoning (RAIR) Lab
selmer@rpi.edu • govinn@rpi.edu
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA

Logico-Mathematical Foundations

Selmer Bringsjord • Naveen Sundar G. • Mei Si

Department of Computer Science
Department of Cognitive Science
Lally School of Management & Technology
Rensselaer AI & Reasoning (RAIR) Lab
selmer@rpi.edu • govinn@rpi.edu
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA
Hierarchy of Ethical Reasoning
Hierarchy of Ethical Reasoning
Hierarchy of Ethical Reasoning

U

DIARC
Hierarchy of Ethical Reasoning
Hierarchy of Ethical Reasoning

\[ \text{ADR}^M \]

\[ U \]

\[ \text{DIARC} \]
Hierarchy of Ethical Reasoning

\[ \text{DCEC}^* \]

\[ \text{ADR}^M \]

\[ \text{U} \]

DIARC

UIMA/Watson
Hierarchy of Ethical Reasoning

$DCEC_{CL}^*$

$DCEC^*$

$ADR^M$

$U$

DIARC

UIMA/Watson
Hierarchy of Ethical Reasoning

\[ DCEC_{CL} \]
\[ DCEC^* \]
\[ ADR^M \]
\[ U \]

DIARC

UIMA/Watson
Logic
Logic
MiniMaxularity

Logic

Propositional logic
MiniMaxularity

Logic

propositional logic

semantic-web logics

MiniMaxularity
Logic

- Propositional logic
- Semantic-web logics
- Description logics

MiniMaxularity
MiniMaxularity

Logic

propositional logic
semantic-web logics
description logics
fragments of FOL
Logic

- Propositional logic
- Semantic-web logics
- Description logics
- Fragments of FOL
- UIMA output

...
Art of Infallibility I
MiniMaxularity

Logic

propositional logic
semantic-web logics
description logics
fragments of FOL

temporal
temporal+epistemic
temporal+epistemic+deontic
+planning+arg semantics

epistemic

FOL

SOL

UIMA output

Art of Infallibility I
Logic

- propositional logic
- semantic-web logics
- description logics
- fragments of FOL

- temporal
- temporal+epistemic
- temporal+epistemic+deontic
- heterogeneous/visual
- epistemic

- UIMA output
- +planning+arg semantics
- Art of Infallibility I
MiniMaxularity

Logic

propositional logic
semantic-web logics
description logics
fragments of FOL
...

temporal
temporal+epistemic
temporal+epistemic+deontic
+planning+arg semantics

Art of Infallibility I

heterogeneous/visual
epistemic
SOL
MiniMaxularity

Logic

$\mathcal{L}_{\omega_1, \omega}$

DCEC*
Deontic Cognitive Event Calculus
(with Castañeda's *)

Art of Infallibility I

Infinitary (Aol 2)

Logic

FOL

SOL

propositional logic

semantic-web logics
description logics
fragments of FOL
UIMA output

epistemic
temporal
temporal+epistemic
temporal+epistemic+deontic
+heterogeneous/visual
+planning+arg semantics

...
DCEC*
Deontic Cognitive Event Calculus
(with Castañeda’s *)

1. natural language semantics (non-Montagovian)
2. higher-cognition tests (for Psychometric AI)
   (false-belief test, deliberative mind-reading
   mirror test for self-consciousness ...)
3. ethically correct robots
4. biz & econ simulation

Art of Infallibility I
**MiniMaxularity**

**Logic**
- propositional logic
- semantic-web logics
- description logics
- fragments of FOL
- temporal
- temporal+epistemic
- temporal+epistemic+deontic
- +planning+arg semantics
- heterogeneous/visual
- epistemic

**DCEC**
Deontic Cognitive Event Calculus  
(with Castañeda’s *)

**Art of Infallibility I**

**L_{\omega 1, \omega}**

**Infinitary (Aol 2)**
Logic

- propositional logic
- semantic-web logics
- description logics
- fragments of FOL
- temporal
- temporal+epistemic
- temporal+epistemic+deontic
- heterogeneous/visual
- epistemic
- SOL
- Infinitary (AoI 2)
- Art of Infallibility I
- UIMA output
- +planning+arg semantics
- MiniMaxularity

L_{\omega_1,\omega}
Hierarchy of Ethical Reasoning
Hierarchy of Ethical Reasoning
Hierarchy of Ethical Reasoning

$U$

DIARC
Hierarchy of Ethical Reasoning

U

UIMA/Watson

DIARC
Hierarchy of Ethical Reasoning

\[ \text{UIMA/Watson} \]

\[ \text{DIARC} \]
Hierarchy of Ethical Reasoning

- $DCEC^*$
- $ADR^M$
- $U$

DIARC

UIMA/Watson
Hierarchy of Ethical Reasoning

$DCEC^*_CL$

$DCEC^*$

$ADR^M$

$U$

UIIMA/Watson

DIARC
Hierarchy of Ethical Reasoning

$DCEC_{CL}^*$

$DCEC^*$

$ADR^M$

$U$

DIARC
Hierarchy of Ethical Reasoning

$DCEC^*_CL$

$DCEC^*$

$ADR^M$

$U$

DIARC

UIMA/Watson
Hierarchy of Ethical Reasoning

$DCEC^*_{CL}$

$DCEC^*$

$ADR^M$

$U$

DIARC

UIMA/Watson
Many experts to IBM: “Can’t be done!”
Many experts to IBM: “Can’t be done!”

No one asked me.
From computational logic for configuration and design to ...
David Ferrucci: Life After Watson

To the degree there was a human face of Watson, the “Jeopardy!” computer champion, it was David Ferrucci. He was the I.B.M. researcher who led the development of Watson, an artificial intelligence engine. The goateed computer scientist was always articulate and at ease in front of a camera or a microphone.

Dr. Ferrucci has left I.B.M. to join the giant hedge fund Bridgewater Associates. And the weight of the Watson-related fame, it seems, played a role. “I was so linked to the Watson achievement, and where I.B.M. was taking it, that I felt I was almost losing my identity,” he said in a recent interview.
Analytics bridge the Unstructured & Structured worlds

Unstructured Information
- Text, Chat, Email, Audio, Video
  - High-Value
  - Most Current Content
  - BUT...
  - Buried in Huge Volumes
  - Lots of Noise, Implicit Semantics
  - Inefficient Search

UIMA
- Identify Semantic Entities, Induce Structure
- Chats, Phone Calls, Transfers
- People, Places, Org, Events
- Times, Topics, Opinions, Relationships
- Threats, Plots, etc.

Structured Information
- Indices
- DBs
- KBs

Explicit Structure
- Explicit Semantics
- Efficient Search
- Focused Content
Analytics bridge the Unstructured & Structured worlds

Unstructured Information
- Text, Chat, Email, Audio, Video
- High-Value
- Most Current Content
- BUT...
  - Buried in Huge Volumes
  - Lots of Noise, Implicit Semantics
  - Inefficient Search

Structured Information
- Indices
- DBs
- KBs

UIMA
- Identify Semantic Entities, Induce Structure
- Chats, Phone Calls, Transfers
- People, Places, Org, Events
- Times, Topics, Opinions, Relationships
- Threats, Plots, etc.
Analytics bridge the Unstructured & Structured worlds

Unstructured Information
- Text, Chat, Email, Audio, Video

High Value
- Most Current Content
- BUT...
  - Buried in Huge Volumes
  - Lots of Noise, Implicit Semantics
  - Inefficient Search

Structured Information
- Indices
- DBs
- KBs
- Explicit Structure
  - Explicit Semantics
  - Efficient Search
  - Focused Content

UIMA
- Identify Semantic Entities, Induce Structure
- Chats, Phone Calls, Transfers
- People, Places, Org, Events
- Times, Topics, Opinions, Relationships
- Threats, Plots, etc.
Analytics bridge the Unstructured & Structured worlds

Unstructured Information

- Text, Chat, Email, Audio, Video
- High-value
- Most Current Content
- BUT...
- Buried in Huge Volumes
- Lots of Noise, Implicit Semantics
- Inefficient Search

$u \in \sum^*$

Structured Information

- Indices
- DBs
- KBs
- Explicit Structure
- Explicit Semantics
- Efficient Search
- Focused Content

- Identify Semantic Entities, Induce Structure
- Chats, Phone Calls, Transfers
- People, Places, Org, Events
- Times, Topics, Opinions, Relationships
- Threats, Plots, etc.
\[ u \in \sum^* \]

Analytics bridge the Unstructured & Structured worlds

Unstructured Information
- Text, Chat, Email, Audio, Video

Structured Information
- Indices
- DBs
- KBs

High Value
Most Current Content
BUT...
- Buried in Huge Volumes
- Lots of Noise, Implicit Semantics
- Inefficient Search

- Identify Semantic Entities, Induce Structure
- People, Places, Org, Events
- Times, Topics, Opinions, Relationships
- Threats, Plots, etc.
\[ u = (S, \ldots) \]

Unstructured Information

Structured Information

- Indices
- DBs
- KBs

High Value
- Most Current Content
- BUT...
- Buried in Huge Volumes
- Lots of Noise, Implicit Semantics
- Inefficient Search

Analytics bridge the Unstructured & Structured worlds

- Identify Semantic Entities, Induce Structure
- People, Places, Org, Events
- Times, Topics, Opinions, Relationships
- Threats, Plots, etc.
\( u \in \Sigma^* \quad \sum \rightarrow \Phi \quad \Phi = (S, \ldots) \)
\[ U : u \rightarrow \Phi \]
\[ U = (S, \ldots) \]
\[ u \in \Sigma^* \]
\[ A(v_1 \sqsubset u, R) \land A(v_2 \sqsubset u, R) \]

\[ \mathcal{U} : u \longrightarrow \Phi \]

\[ \mathcal{U} = (S, \ldots) \]

\[ u \in \Sigma^* \]
$(Ab(u) \land u \in \text{MedBase}) \rightarrow t(u) = \text{‘skin cancer’}

\begin{align*}
A(v_1 \sqsubseteq u, R) \land A(v_2 \sqsubseteq u, R)
\end{align*}

\mathcal{U} : u \longrightarrow \Phi

\mathcal{U} = (S, \ldots)

u \in \sum^*$
(\text{Ab}(u) \land u \in \text{MedBase}) \rightarrow t(u) = '\text{skin cancer}'

\mathcal{U}(v_1 \sqsubseteq u, R) \land \mathcal{A}(v_2 \sqsubseteq u, R)

\mathcal{U} : u \mapsto \Phi

\mathcal{U} = (S, \ldots)

u \in \Sigma^*
What is the “carry over” here?
Hierarchical Ethical Classifier (initial design)

- Preprocessing system for deciding whether a situation warrants deliberate ethical reasoning.
- Made up of atomic ethical classifiers (UIMA’s Analysis Engines)

\[ \text{term of sort } S \xrightarrow{} \text{atomic ethical classifier} \xrightarrow{} [\text{Yes, No, Delegate}] \]
Why?

• Not all situations need deliberate deontic reasoning.

• Need to quickly decide at every time instant whether the current situation requires deliberate, deontic reasoning.

• Need many heuristics to do so.

• The design provides a disciplined approach to organize and add new heuristics.
Hierarchical Ethical Classifier (UIMA-Style)

- More processing cost
- High-level classifiers
- Less processing cost
- Low-level classifiers

Semi-structured data (event calculus formulae and terms)

Sensors and low-level processors

World
Specification

- Processing goes to a higher-level classifier only if the corresponding lower classifier answers Delegate.
- Notion of *top-fired classifiers*.
- Systems answers:
  - **Yes**: If and only if any one of the top-fired classifiers answers Yes, or all the top-level atomic classifiers answer Delegate.
  - **No**: If and only if all the top-fired classifiers answer No.
Hierarchy of Ethical Reasoning

\[ D\mathcal{CEC}_{CL} \]
\[ D\mathcal{CEC}^* \]
\[ A\mathcal{DR}^M \]
\[ U \]

DIARC

UIMA/Watson
Hierarchy of Ethical Reasoning

\[ DCEC_{CL} \]
\[ DCEC^* \]
\[ ADR^M \]
\[ U \]

DIARC

UIMA/Watson
Analogico-Deductive Moral Reasoning (ADMR)
Analogico-Deductive Moral Reasoning (ADMR)

- Moral problem presented as *story* (in psychometric sense) and a *stem*, or *query*.
Analogico-Deductive Moral Reasoning (ADMR)

- Moral problem presented as *story* (in psychometric sense) and a *stem*, or *query*.

- A *stem* has correct answer $A$ and a set $P_i$ of correct proofs or arguments establishing $A$, relative to:
Analogico-Deductive Moral Reasoning (ADMR)

- Moral problem presented as *story* (in psychometric sense) and a *stem*, or *query*.

- A *stem* has correct answer $A$ and a set $P_i$ of correct proofs or arguments establishing $A$, relative to:
  - An associated implicit moral theory, and
Analogico-Deductive Moral Reasoning (ADMR)

- Moral problem presented as *story* (in psychometric sense) and a *stem*, or *query*.

- A *stem* has correct answer *A* and a set $P_i$ of correct proofs or arguments establishing *A*, relative to:
  - An associated implicit moral theory, and
  - A corresponding moral code
Analogico-Deductive Moral Reasoning (ADMR)

Input:
(story, query/stem)
Analogico-Deductive Moral Reasoning (ADMR)

Input: (story, query/stem)

- Analogy Source Cases
- Moral Theories and Codes
Analogico-Deductive Moral Reasoning (ADMR)

Input: (story, query/stem)

ADMR System

Analogy Source Cases

Moral Theories and Codes
Analogico-Deductive Moral Reasoning (ADMR)

Input: (story, query/stem)

ADMR System

Analogy Source Cases

Moral Theories and Codes
Analogico-Deductive Moral Reasoning (ADMR)

Input: (story, query/stem)

Output: \{(A_1, \text{proofs/arguments of } A_1), (A_2, \text{proofs/arguments of } A_2), \ldots\}

ADMR System

- Analogy Source Cases
- Moral Theories and Codes
The Heinz Dilemma (Kohlberg)

“In Europe, a woman was near death from a special kind of cancer. There was one drug that the doctors thought might save her. It was a form of radium that a druggist in the same town had recently discovered. The drug was expensive to make, but the druggist was charging ten times what the drug cost him to make. He paid $200 for the radium and charged $2,000 for a small dose of the drug.

The sick woman’s husband, Heinz, went to everyone he knew to borrow the money, but he could only get together about $1,000, which is half of what it cost. He told the druggist that his wife was dying and asked him to sell it cheaper or let him pay later. But the druggist said: “No, I discovered the drug and I’m going to make money from it.” So Heinz got desperate and broke into the man’s store to steal the drug for his wife. Should the husband have done that?”
Moral Problem $P_k$ \hspace{1cm} Solution to $P_{k-1}$

Moral Problem $P_3$ \hspace{1cm} Solution to $P_2$

Moral Problem $P_2$ \hspace{1cm} Solution to $P_1$

Moral Problem $P_1$

Moral Dilemma $D_k$ \hspace{1cm} Solution to $D_{k-1}$

Moral Dilemma $D_3$

Moral Dilemma $D_2$

Moral Dilemma $D_1$

Machine

Solution
Moral Problem $P_k$

Moral Dilemma $D_k$

...}

Solution to $P_k$

Solution to $D_k$

Machine

Solution to $P_{k-1}$

Solution to $D_{k-1}$

Solution to $P_{k-2}$

Solution to $D_{k-2}$

Solution to $P_1$
Moral Problem P_1
Moral Problem P_2
Moral Problem P_3
Moral Problem P_k

Moral Dilemma D_1
Moral Dilemma D_2
Moral Dilemma D_3
Moral Dilemma D_k

Solution to P_{k-1}
Solution to P_{k-2}
Solution to P_2
Solution to P_1

Solution to D_{k-1}
Solution to D_{k-2}
Solution to D_2
Solution to D_1
Fragment of Heinz in DCEC*

\[P_1\]
\[
\forall t : \text{Moment, } a : \text{Agent} \left(\text{holds(sick}(a), t) \land \left(\forall t' : \text{Moment } t' < T \implies \neg \text{happens(treated}(a), t + t')\right)\right)
\implies \left(\text{happens(dies}(a), t + T) \lor \text{holds(dead}(a), t + T)\right)
\]

\[P_2\]
\[
\text{holds(sick(wife(\text{I}^*)), t_0) } \land \left(\forall t' : \text{Moment } t' < T \implies \neg \text{happens(treated(wife(\text{I}^*)), t_0 + t')}\right)
\]

\[Q\]
\[
\text{happens(dies(wife(\text{I}^*)), t_0 + T) } \lor \text{holds(dead(wife(\text{I}^*)), t_0 + T)}
\]

Note: This adheres strictly to the syntax of DCEC*
Hierarchy of Ethical Reasoning

$DCEC^*_CL$

$DCEC^*$

$ADRM$

$U$

DIARC

UIMA/Watson
Hierarchy of Ethical Reasoning

\[ \text{DIARC} \]

\[ DCEC_{CL} \]

\[ DCEC^* \]

\[ ADR^M \]

\[ U \]

UIMA/Watson
DCEC
DCEC

Syntax

\[ S ::= \text{Object} | \text{Agent} | \text{Self} \sqcup \text{Agent} | \text{ActionType} | \text{Action} \sqcup \text{Event} | \text{Moment} | \text{Boolean} | \text{Fluent} | \text{Numeric} \]

\[ t ::= x : S \mid c : S \mid f(t_1, \ldots, t_n) \]

\[ p : \text{Boolean} \mid \neg \phi \mid \phi \land \psi \mid \phi \lor \psi \mid \phi \leftrightarrow \psi \mid \forall x : S. \phi \mid \exists x : S. \phi \]

\[ \phi ::= \]

\[ \forall x : S. \phi \]

\[ \exists x : S. \phi \]

\[ \neg \phi \]

\[ \phi \land \psi \]

\[ \phi \lor \psi \]

\[ \phi \leftrightarrow \psi \]

\[ \forall \phi \]

\[ \exists \phi \]

\[ \phi \land \psi \land \cdots \land \psi \]

\[ \phi \lor \psi \lor \cdots \lor \psi \]

\[ \neg \phi \]

\[ \phi \land \psi \land \cdots \land \psi \]

Rules of Inference

\[ \frac{C(t, \text{P}(a, t, \phi) \rightarrow \text{K}(a, t, \phi))}{R_1} \]

\[ \frac{C(t, \text{K}(a, t, \phi) \rightarrow \text{B}(a, t, \phi))}{R_2} \]

\[ \frac{C(t, \phi) \hspace{1em} t_1 \leq t \leq t_2 \leq t_n}{R_3} \]

\[ \frac{K(a_1, t_1, \ldots, K(a_n, t_n), \ldots)}{R_4} \]

\[ \frac{t_1 \leq t_2 \leq t_3}{R_5} \]

\[ \frac{C(t, K(a_1, t_1, \phi_1) \rightarrow \phi_2)}{R_6} \]

\[ \frac{K(a_1, t_2, \phi_1) \rightarrow K(a_1, t_3, \phi_2)}{R_7} \]

\[ \frac{t_1 \leq t_2 \leq t_3}{R_8} \]

\[ \frac{C(t, \forall x. \phi \rightarrow \phi[x \mapsto t])}{R_9} \]

\[ \frac{C(t, \phi_1 \leftrightarrow \phi_2 \rightarrow \neg \phi_2 \rightarrow \neg \phi_1)}{R_9} \]

\[ \frac{C(t, \phi_1 \land \cdots \land \phi_n \rightarrow \phi)}{R_{10}} \]

\[ \frac{C(t, \phi_1 \land \cdots \land \phi_n \rightarrow \phi)}{R_{10}} \]

\[ \frac{B(a, t, \phi) \rightarrow \psi}{R_{11a}} \]

\[ \frac{B(a, t, \phi) \rightarrow \psi}{R_{11b}} \]

\[ \frac{B(a, t, \phi) \rightarrow \psi}{R_{12}} \]

\[ \frac{B(a, t, \phi) \rightarrow \psi}{R_{13}} \]

\[ \frac{B(a, t, \phi) \rightarrow \psi}{R_{14}} \]

\[ \frac{B(a, t, \phi) \rightarrow \psi}{R_{15}} \]
### Syntax

- **Object** | **Agent** | **Self** | **Agent** | **ActionType** | **Action** | **Event** | **Moment** | **Boolean** | **Fluent** | **Numeric**

- $t ::= x : S \mid c : S \mid f(t_1, \ldots, t_n)$
- $p : \text{Boolean} \mid \neg \phi \mid \phi \land \psi \mid \phi \lor \psi \mid \phi \leftrightarrow \psi \mid \forall x : S. \phi \mid \exists x : S. \phi$
- $\phi ::= B(a,t,\phi) \mid D(a,t,\text{holds}(f,t')) \mid I(a,t,\text{happens}(a^*,\alpha,t'))$
- $O(a,t,\phi,\text{happens}(a^*,\alpha,t'))$

### Rules of Inference

- $C(t, P(a,t,\phi) \rightarrow K(a,t,\phi))$ [R1]
- $C(t, K(a,t,\phi) \rightarrow B(a,t,\phi))$ [R2]
- $C(t, \phi) \quad t_1 \leq t_1 \ldots \leq t_n$
- $K(a_1, t_1, \ldots, K(a_n, t_n, \ldots))$ [R3]
- $K(a,t,\phi)$ [R4]
- $t_1 \leq t_3, t_2 \leq t_3$
- $C(t, K(a_1, t_1, \phi_1) \rightarrow \phi_2)$ [R5]
- $B(a,t,\phi)$ [R6]
- $t_1 \leq t_3, t_2 \leq t_3$
- $C(t, B(a_1, t_1, \phi_1) \rightarrow \phi_2)$ [R7]
- $C(t, C(t_1, \phi_1) \rightarrow \phi_2)$ [R8]
- $C(t, \forall x. \phi \rightarrow \phi[x \mapsto t])$
- $C(t, \phi_1 \leftrightarrow \phi_2 \rightarrow \neg \phi_2 \rightarrow \neg \phi_1)$ [R9]
- $C(t, [\phi_1 \land \ldots \land \phi_n \rightarrow \phi] \rightarrow [\phi_1 \rightarrow \ldots \rightarrow \phi_n \rightarrow \psi])$ [R10]
- $B(h, t, B(s, t, \phi))$ [R11a]
- $B(a, t, \phi)$ [R11b]
- $I(a, t, \text{happens}(a^*, \alpha, t'))$ [R12]
- $P(a, t, \text{happens}(a^*, \alpha, t'))$ [R13]
- $B(a, t, \phi)$ [R14]
- $O(a, t, \phi, \text{happens}(a^*, \alpha, t'))$ [R15]

### DCEC

**Where are the emotions?**

- $\text{action} : \text{Agent} \times \text{ActionType} \rightarrow \text{Action}$
- $\text{initially} : \text{Fluent} \rightarrow \text{Boolean}$
- $\text{holds} : \text{Fluent} \times \text{Moment} \rightarrow \text{Boolean}$
- $\text{happens} : \text{Event} \times \text{Moment} \rightarrow \text{Boolean}$
- $\text{clipped} : \text{Moment} \times \text{Fluent} \times \text{Moment} \rightarrow \text{Boolean}$
- $\text{initiates} : \text{Event} \times \text{Fluent} \times \text{Moment} \rightarrow \text{Boolean}$
- $\text{terminates} : \text{Event} \times \text{Fluent} \times \text{Moment} \rightarrow \text{Boolean}$
- $\text{prior} : \text{Moment} \times \text{Moment} \rightarrow \text{Boolean}$
- $\text{interval} : \text{Moment} \times \text{Moment}$
- $*: \text{Agent} \rightarrow \text{Self}$
- $\text{payoff} : \text{Agent} \times \text{ActionType} \times \text{Moment} \rightarrow \text{Numeric}$
**DCEC**

**Syntax**

\[ S ::= \text{Object} \mid \text{Agent} \mid \text{Self} \mid \text{Agent} \mid \text{ActionType} \mid \text{Action} \mid \text{Event} \mid \text{Moment} \mid \text{Boolean} \mid \text{Fluent} \mid \text{Numeric} \]

\[ t ::= x : S \mid c : S \mid f(t_1, \ldots, t_n) \]

\[ p : \text{Boolean} \mid \neg \phi \mid \phi \land \psi \mid \phi \lor \psi \mid \phi \leftrightarrow \psi \mid \forall x : S. \phi \mid \exists x : S. \phi \]

\[ f ::= \text{P}(a,t,\phi) \mid \text{K}(a,t,\phi) \mid \text{C}(t,\phi) \mid \text{S}(a,b,t,\phi) \mid \text{S}(a,t,\phi) \]

\[ \phi ::= \text{B}(a,t,\phi) \mid \text{D}(a,t,\text{holds}(f,t')) \mid \text{I}(a,t,\text{happens}(\text{action}(a^*,\alpha),t')) \]

\[ \text{O}(a,t,\phi,\text{happens}(\text{action}(a^*,\alpha),t')) \]

**Rules of Inference**

\[ \frac{C(t,P(a,t,\phi) \rightarrow \text{K}(a,t,\phi))}{[R_1]} \]

\[ \frac{C(t,\phi) \quad t \leq t_1 \ldots \leq t_n}{[R_2]} \]

\[ \frac{K(a_1,t_1,\ldots,K(a_n,t_n,\phi) \ldots)}{[R_3]} \]

\[ \frac{t_1 \leq t_3, t_2 \leq t_3}{[R_4]} \]

\[ \frac{C(t,\text{K}(a_1,t_1,\phi) \rightarrow \phi)}{[R_5]} \]

\[ \frac{B(a_1,t,\phi) \quad B(a,t,\phi \rightarrow \psi)}{[R_6]} \]

\[ \frac{\text{S}(s,h,t,\phi)}{[R_7]} \]

\[ \frac{\text{I}(a,t,\text{happens}(\text{action}(a^*,\alpha),t'))}{[R_8]} \]

\[ \frac{\text{O}(a,t,\phi,\text{happens}(\text{action}(a^*,\alpha),t'))}{[R_9]} \]

\[ \frac{\text{action} : \text{Agent} \times \text{ActionType} \rightarrow \text{Action}}{} \]

\[ \frac{\text{initially} : \text{Fluent} \rightarrow \text{Boolean}}{} \]

\[ \frac{\text{holds} : \text{Fluent} \times \text{Moment} \rightarrow \text{Boolean}}{} \]

\[ \frac{\text{happens} : \text{Event} \times \text{Moment} \rightarrow \text{Boolean}}{} \]

\[ \frac{\text{clipped} : \text{Moment} \times \text{Fluent} \times \text{Moment} \rightarrow \text{Boolean}}{} \]

\[ \frac{\text{initiates} : \text{Event} \times \text{Fluent} \times \text{Moment} \rightarrow \text{Boolean}}{} \]

\[ \frac{\text{terminates} : \text{Event} \times \text{Fluent} \times \text{Moment} \rightarrow \text{Boolean}}{} \]

\[ \frac{\text{prior} : \text{Moment} \times \text{Moment} \rightarrow \text{Boolean}}{} \]

\[ \frac{\text{interval} : \text{Moment} \times \text{Boolean}}{} \]

\[ \frac{\text{payoff} : \text{Agent} \times \text{ActionType} \times \text{Moment} \rightarrow \text{Numeric}}{} \]
Step #1 (Selmer, Mei, Naveen): Integrate version of prior formalization of OCC with deontic concepts/operators.

Syntax

\[
S ::= \text{Object} \mid \text{Agent} \mid \text{Self} \mid \text{Agent} \mid \text{ActionType} \mid \text{Action} \mid \text{Event} \\
\text{Moment} \mid \text{Boolean} \mid \text{Fluent} \mid \text{Numeric}
\]

\[
t ::= x : S \mid c : S \mid f(t_1, \ldots, t_n)
\]

\[
p : \text{Boolean} \mid \neg \phi \mid \phi \land \psi \mid \phi \lor \psi \mid \phi \leftrightarrow \psi \mid \forall x : S. \phi \mid \exists x : S. \phi
\]

\[
\phi ::= \text{B}(a,t,\phi) \mid \text{D}(a,t,\text{holds}(f,t')) \mid \text{I}(a,t,\text{happens}(\text{action}(a^*,\alpha),t')) \mid \text{O}(a,t,\phi,\text{happens}(\text{action}(a^*,\alpha),t'))
\]

Rules of Inference

\[
\frac{C(t, \text{P}(a,t,\phi) \rightarrow \text{K}(a,t,\phi))}{[R_1]} \quad \frac{C(t, \text{K}(a,t,\phi) \rightarrow \text{B}(a,t,\phi))}{[R_2]}
\]

\[
\frac{C(t, \phi) \quad t_1 \leq t_1 \ldots t_3}{[R_3]} \quad \frac{\text{K}(a_1,t_1) \ldots \text{K}(a_n,t_n,\phi) \ldots}{[R_4]}
\]

\[
\frac{t_1 \leq t_1 \ldots t_3}{[R_5]} \quad \frac{C(t, \text{K}(a_1,t_1,\phi_1) \rightarrow \text{K}(a_2,t_2,\phi_1) \rightarrow \text{K}(a_3,t_3,\phi_2))}{[R_6]}
\]

\[
\frac{t_1 \leq t_3 \ldots t_3}{[R_7]} \quad \frac{C(t, C(t_1,\phi_1) \rightarrow \text{K}(a_2,t_2,\phi_1) \rightarrow \text{K}(a_3,t_3,\phi_2))}{[R_8]}
\]

\[
\frac{t_1 \leq t_3 \ldots t_3}{[R_9]} \quad \frac{C(t, \forall x. \phi \rightarrow \phi[x \mapsto t])}{[R_8]} \quad \frac{C(t, \phi_1 \leftrightarrow \phi_2 \rightarrow \neg \phi_2 \rightarrow \neg \phi_1)}{[R_9]}
\]

\[
\frac{C(t, \phi_1, \ldots, \phi_n, \phi_\psi)}{[R_{10}]} \quad \frac{B(a,t,\phi) \mid B(a,t,\phi \rightarrow \psi)}{[R_{11a}]} \quad \frac{B(a,t,\phi) \mid B(a,t,\psi)}{[R_{11b}]}
\]

\[
\frac{\text{S}(s,h,t,\phi)}{[R_{12}]} \quad \frac{I(a,t,\text{happens}(\text{action}(a^*,\alpha),t'))}{[R_{13}]} \quad \frac{P(a,t,\text{happens}(\text{action}(a^*,\alpha),t))}{[R_{14}]}
\]

\[
\frac{B(a,t,\phi) \mid B(a,t,\text{O}(a^*,t,\phi,\text{happens}(\text{action}(a^*,\alpha),t')))}{[R_{15}]}
\]

\[
\frac{O(a,t,\phi,\text{happens}(\text{action}(a^*,\alpha),t'))}{[R_{16}]}
\]

\[
\frac{\phi \leftrightarrow \psi}{[R_{17}]}
\]

\[
\frac{O(a,t,\phi,\gamma) \leftrightarrow O(a,t,\psi,\gamma)}{[R_{18}]}
\]
**DCEC**

**Syntax**

\[
S ::= \text{Object} \mid \text{Agent} \mid \text{Self} \rightarrow \text{Agent} \mid \text{ActionType} \mid \text{Action} \rightarrow \text{Event} \mid \text{Moment} \mid \text{Boolean} \mid \text{Fluent} \mid \text{Numeric}
\]

\[
t ::= x : S \mid c : S \mid f(t_1, \ldots, t_n)
\]

\[
p : \text{Boolean} \mid \neg \phi \mid \phi \land \psi \mid \phi \lor \psi \mid \phi \leftrightarrow \psi \mid \forall x : S. \phi \mid \exists x : S. \phi
\]

\[
\phi ::= \text{B}(a,t,\phi) \mid \text{D}(a,t,\text{holds}(f,t')) \mid \text{I}(a,t,\text{happens}(\text{action}(a^*,\alpha),t')) \mid \text{O}(a,t,\phi,\text{happens}(\text{action}(a^*,\alpha),t'))
\]

**Rules of Inference**

\[
\begin{align*}
\text{C}(t, \text{P}(a,t,\phi) \rightarrow \text{K}(a,t,\phi)) & \quad [R_1] \\
\text{C}(t, \text{K}(a,t,\phi) \rightarrow \text{B}(a,t,\phi)) & \quad [R_2] \\
\text{C}(t, \phi) & \quad [R_3]
\end{align*}
\]

\[
\begin{align*}
\text{K}(a_1,t_1,\ldots,\text{K}(a_n,t_n,\phi)) \quad [R_4] \\
\text{C}(t, \text{K}(a_1,t_1,\phi_1) \rightarrow \phi_2) \rightarrow (\text{K}(a_2,t_2,\phi_1) \rightarrow \text{K}(a_3,t_3,\phi_2)) & \quad [R_5] \\
\text{C}(t, \text{B}(a_1,t_1,\phi_1) \rightarrow \phi_2) \rightarrow (\text{B}(a_2,t_2,\phi_1) \rightarrow \text{B}(a_3,t_3,\phi_2)) & \quad [R_6] \\
\text{C}(t, \text{C}(t_1,\phi_1) \rightarrow \phi_2) \rightarrow (\text{C}(t_2,\phi_1) \rightarrow \text{C}(t_3,\phi_2)) & \quad [R_7]
\end{align*}
\]

\[
\begin{align*}
\text{C}(t, \forall x. \phi \rightarrow \phi[x \mapsto t]) & \quad [R_8] \\
\text{C}(t, \phi_1 \leftrightarrow \phi_2 \rightarrow \neg \phi_2 \rightarrow \neg \phi_1) & \quad [R_9]
\end{align*}
\]

\[
\begin{align*}
\text{C}(t, [\phi_1 \land \ldots \land \phi_n \rightarrow \phi] \rightarrow [\phi_1 \rightarrow \ldots \rightarrow \phi_n \rightarrow \psi]) & \quad [R_{10}] \\
\text{B}(a_1,t,\phi) \quad [R_{11a}] & \quad \text{B}(a_2,t,\phi) \quad [R_{11b}]
\end{align*}
\]

\[
\begin{align*}
\text{S}(s,h,t,\phi) \quad [R_{12}] & \quad \text{I}(a,t,\text{happens}(\text{action}(a^*,\alpha),t')) \quad [R_{13}] \\
\text{B}(h,t,\text{S}(s,t,\phi)) \quad [R_{13}] & \quad \text{P}(a,t,\text{happens}(\text{action}(a^*,\alpha),t)) \quad [R_{14}] \\
\text{B}(a,t,\phi) \quad [R_{15}] & \quad \text{O}(a,t,\phi,\text{happens}(\text{action}(a^*,\alpha),t'))
\end{align*}
\]

\[
\begin{align*}
\text{O}(a,t,\phi,\gamma) & \quad [R_{16}]
\end{align*}
\]

**Action** : Agent $\times$ ActionType $\rightarrow$ Action

**Initially** : Fluent $\rightarrow$ Boolean

**Happens** : Fluent $\times$ Moment $\rightarrow$ Boolean

**Clipped** : Moment $\times$ Fluent $\times$ Moment $\rightarrow$ Boolean

**Initiates** : Event $\times$ Fluent $\times$ Moment $\rightarrow$ Boolean

**Terminates** : Event $\times$ Fluent $\times$ Moment $\rightarrow$ Boolean

**Prior** : Moment $\times$ Moment $\rightarrow$ Boolean

**Interval** : Moment $\times$ Boolean

**Payoff** : Agent $\times$ ActionType $\times$ Moment $\rightarrow$ Numeric
A Logic of Emotions for Intelligent Agents

Bas R. Steunebrink
Department of ICS
Utrecht University
Utrecht, The Netherlands
bas@cs.uu.nl

Mehdi Dastani
Department of ICS
Utrecht University
Utrecht, The Netherlands
mehdi@cs.uu.nl

John Jules Ch. Meyer
Department of ICS
Utrecht University
Utrecht, The Netherlands
j jmeyer@cs.uu.nl

Abstract

This paper formalizes a well-known psychological model of emotions in an agent-specific language. This is done by introducing a logical language and its semantics which are used to specify an agent model in terms of mental attitudes including emotions. We show that our formalization renders a number of intuitions and plausible properties of emotions. We also show how this formalization can be used to specify the effect of emotions on an agent's decision making process. Ultimately, the emotions in this model function as heuristics as they assist agents in decision making.

Introduction

In psychological studies, the emotions that influence the deliberation and practical reasoning of an agent are considered as heuristics for preventing excessive deliberation (Damasio 1994). Meyer & Dastani (2004; 2006) propose a functional approach to describe the role of emotions in practical reasoning. According to this functional approach, an agent is assumed to execute domain actions in order to reach its goals. The effects of these domain actions cause and/or influence the appraisal of emotions according to a human-inspired model. These emotions in turn influence the deliberation operators of the agent, functioning as heuristics for decision making when specific domain actions have to be chosen next, which completes the circle.

Although heuristics for modeling the behavior of intelligent agents in real-time behavior is usually not considered, despite of their (arguably) positive contributions. Philosophical studies describing (idealized) human behavior have previously been formalized using one or more logics (often mixed or extended). For example, Bratman's BDI theory of belief, desire, and intention (Bratman 1987) has been modeled and studied in e.g. linear tense logic (Cohen & Levesque 1990) and dynamic logic (Meyer, Hoek, & Linde 1999).

We propose to model and formalize human emotions in logic. There exist different psychological models of emotions, of which we have chosen to consider the model of Ortony, Clore, & Collins (1988). The "OCC model" is suitable for formalization because it describes a concise hierarchy of emotions and specifies the conditions that elicit each emotion in terms of objects, actions, and events—concepts that can be captured in a formal language. In this paper, we introduce a logic for studying the appraisal, interrelations, and effects of the 22 emotions described in the OCC model. We take a computational approach, building not only a mathematically sound model but also keeping in mind its implementation in a multi-agent system. Multi-agent aspects of emotions, however, are not treated in this paper.

It should be noted that previous work on specifying and implementing emotions carried out by Meyer (2004) and Dastani (2006) follow Oatley & Jenkins' model of emotions (Oatley & Jenkins 1996) and comprises only four emotions: happy, sad, angry, and fearful. Emotions are represented as beliefs in an agent's cognitive state. Similar to our approach, the elaboration of an agent's appraisal of emotions in turn influence the agent's deliberation. Dastani & Meyer (2006) have defined transition semantics for their emotional model, which we also intend to do for our formalization of OCC. However, we intend to formalize the quantitative aspects of emotions as well, which were not considered in the purely logical model of Dastani & Meyer. Our work is also similar to other computational models of emotions, such as EMA (Gentz & Moulines 2004), CoMAdr (Stienen 2001), and the work of Picard (1997); however, our goal is not to design a specific computational model of emotions, but rather to develop a logic for studying emotional models, starting with the OCC model.

Language and Semantics

The OCC model describes a hierarchy that classifies 22 emotions. The hierarchy contains three branches, namely emotions concerning aspects of objects (e.g., love and hate), actions of agents (e.g., pride and admiration), and consequences of events (e.g., joy and pity). Additionally, these branches combine to form a group of compound emotions, namely emotions concerning consequences of events caused by actions of agents (e.g., gratitude and anger). Because the objects of these emotions (i.e., objects, actions, and events) correspond to notions commonly used in agent models (i.e., agents, plans, and goal accomplishments, respectively), this makes the OCC model suitable for use in the deliberation and emotional reasoning of multi-agent systems. It should be emphasized that emotions are not used to describe the entire cognitive state of an agent (as in the agent is

A logical formalization of the OCC theory of emotions

C. Adam (carole.adam.cmit@gmail.com)
KMITL University, Melbourne, VIC, Australia

A. Herzig (andreaz.hherzig@irit.fr)
University of Toulouse, INED, Institute de Recherche en Informatique de Toulouse, France

1. Introduction

There is a great amount of work concerning emotions in various disciplines such as philosophy (Gordon, 1987; Solomon and Carleton, 1984), economy (Elster, 1998; Loewenstein, 2000), neuroscience and psychology. In neuroscience, experiments have highlighted that individuals who do not feel emotions e.g. due to brain damage are unable to make rational decisions (see Damasio, 1994 for instance), refuting the commonsensical assumption that emotions prevent agents from being rational. Psychology provides elaborate theories of emotions ranging from their classification (Ekman, 1992; Darwin, 1872) to their triggering conditions (Lazarus, 1991, Ortony et al., 1990) and their impact on various cognitive processes (Fergus, 1995).

Computer scientists investigate the expression and recognition of emotion in order to design anthropomorphic systems that can interact with human users in a multi-modal way. Such systems are justified by the various forms of 'anthropomorphic behavior' that users ascribe to artifacts. This has lead to an increasing interest in Affective Computing, with particular focus on embodied agents (de Roos et al., 2003), ambient intelligence (Bartneck, 2002), intelligent agents (Steunebrink et al., 2007), etc. All these approaches generally aim at giving computer extended capacities for enhanced functionality or more credibility. Intelligent embodied conversational agents (IECA) use a model of emotions both to simulate the user's emotion and to draw their effective state and personality. Bates has argued for the importance of emo-
Automation of Reasoning
Automation of Reasoning

Denotational Proof Languages
Automation of Reasoning

Denotational Proof Languages
Automation of Reasoning

Denotational Proof Languages

Type-α DPL
Automation of Reasoning

Denotational Proof Languages

Type-α DPL

Proof checking.
Automation of Reasoning

Denotational Proof Languages

Type-α DPL

Proof checking.
Denotational Proof Languages

Type-α DPL

Type-ω DPL

Proof checking.
Automation of Reasoning

Denotational Proof Languages

Type-\(\alpha\) DPL

Proof checking.

Type-\(\omega\) DPL

Proof discovery (and checking).
Denotational Proof Languages

Type-α DPL
Proof checking.

Type-ω DPL
Proof discovery (and checking).

Automation of Reasoning

Denotational Proof Languages

Type-\(\alpha\) DPL
Proof checking.

Type-\(\omega\) DPL
Proof discovery (and checking).


Automation of Reasoning

Denotational Proof Languages

Type-α DPL

Proof checking.

Type-ω DPL

Proof discovery (and checking).

DPLs for $\mathcal{DCEC}^*$ under construction ...


Logicist NLP
Logicist NLP

**Two Major Approaches**
Logicist NLP

Two Major Approaches
Logicist NLP

Two Major Approaches

Deep Modeling
Logicist NLP

Two Major Approaches

Deep Modeling
Logicist NLP

Two Major Approaches

Deep Modeling

Controlled English
Logicist NLP

Two Major Approaches

Deep Modeling

Controlled English
Deep Modeling

Controlled English

On Deep Computational Formalization of Natural Language
Naveen Sundar Govindarajulu, John Licato and Selmer Bringsjord

Workshop on Formalizing Mechanisms for Artificial General Intelligence, 2013, AGI 2013
Deep Modeling
Deep Modeling
Utterance

Deep Modeling
Deep Modeling

Utterance

Syntactic Parser
Deep Modeling

Utterance

Syntactic Parser
Deep Modeling

Utterance → Syntactic Parser → Syntax Tree
Deep Modeling

Utterance

Syntactic Parser

Syntax Tree

Understanding System
Deep Modeling

Utterance

Syntactic Parser

Syntax Tree

Understanding System

Understanding Axioms
Deep Modeling

Utterance → Syntactic Parser → Understanding System → Understanding Axioms

Syntactic Tree

Meaning
Deep Modeling

Utterance → Syntactic Parser → Syntax Tree → Understanding System → Meaning → Conversation System

Understanding Axioms
Deep Modeling

- Utterance
- Syntactic Parser
  - Syntax Tree
- Understanding System
  - Meaning
  - Conversation System
- Understanding Axioms
- Reasoner
Deep Modeling

Utterance -> Syntactic Parser

Syntactic Parser -> Syntax Tree

Syntax Tree -> Understanding System

Understanding System -> Meaning

Meaning -> Conversation System

Conversation System -> Reasoner

Reasoner -> Understanding Axioms
Deep Modeling

Utterance → Syntactic Parser → Syntax Tree → Understanding System → Meaning → Conversation System → Reasoner → Response

Understanding Axioms
Controlled English
Controlled English

$DCEC_{CL}^*$ corresponds to a subset of English!
Controlled English

$\mathcal{DCEC}_{CL}^*$ corresponds to a subset of English!

RLCNL: RAIR Lab Controlled Natural Language
Controlled English

$DCEC^*_CL$ corresponds to a subset of English!

RLCNL: RAIR Lab Controlled Natural Language

K(ugv, now, holds(carrying(ugv, soldier), now))
Controlled English

\( \mathcal{DCEC}^* \) corresponds to a subset of English!

RLCNL: RAIR Lab Controlled Natural Language

\[ K(\text{ugv, now, holds}(\text{carrying}(\text{ugv, soldier}, \text{now})) \]

The ugv now knows that the fluent, 'the ugv is carrying the soldier,' holds now.
Controlled English

\( DCEC^*_{CL} \) corresponds to a subset of English!

**RLCNL:** RAIR Lab Controlled Natural Language

\[ K(ugv, \text{now}, holds(\text{carrying}(ugv, \text{soldier}), \text{now})) \]

The ugv now knows that the fluent, 'the ugv is carrying the soldier,' holds now.

\[ B(ugv, \text{now}, B(\text{commander}, t_1, \neg P(ugv, \text{anytime}, \text{happens}(\text{firefight}, \text{anytime})))) \]
Controlled English

\( DCEC^*_CL \) corresponds to a subset of English!

**RLCNL:** RAIR Lab Controlled Natural Language

\[ K(ugv, \text{now}, holds(carrying(ugv, soldier), \text{now})) \]

The ugv now knows that the fluent, 'the ugv is carrying the soldier,' holds now.

\[ B(ugv, \text{now}, B(commander, t_1, \neg P(ugv, \text{anytime}, \text{happens(firefight, anytime)}))) \]

The ugv now believes that the commander at moment \( t_1 \) believes that it is not the case that the ugv at any time perceives that a firefight happens at any time.
 Controlled English

\( \mathcal{DCEC}^*_\text{CL} \) corresponds to a subset of English!

**RLCNL:** RAIR Lab Controlled Natural Language

\[ K(\text{ugv}, \text{now}, \text{holds}(\text{carrying}(\text{ugv}, \text{soldier}), \text{now})) \]

The ugv now knows that the fluent, 'the ugv is carrying the soldier,' holds now.

\[ B(\text{ugv}, \text{now}, B(\text{commander}, t_1, \neg P(\text{ugv}, \text{anytime}, \text{happens}(\text{firefight}, \text{anytime})))) \]

The ugv now believes that the commander at moment \( t_1 \) believes that it is not the case that the ugv at any time perceives that a firefight happens at any time.

\[ K(l, \text{now}, O(l^*, \text{now}, \text{mission(main)}, \text{happens}(\text{action}(l^*, \text{silence}), \text{alltime}))) \]


Controlled English

\( DCEC^*_CL \) corresponds to a subset of English!

RLCNL: RAIR Lab Controlled Natural Language

\[ K(ugv, \text{now}, \text{holds}(\text{carrying}(ugv, \text{soldier}), \text{now})) \]

The ugv now knows that the fluent, 'the ugv is carrying the soldier,' holds now.

\[ B(ugv, \text{now}, B(\text{commander}, t_1, \neg P(ugv, \text{anytime}, \text{happens}(\text{firefight}, \text{anytime}))) \]

The ugv now believes that the commander at moment \( t_1 \) believes that it is not the case that the ugv at any time perceives that a firefight happens at any time.

\[ K(I, \text{now}, O(I^*, \text{now}, \text{mission(main)}, \text{happens}(\text{action}(I^*, \text{silence}, \text{alltime}))) \]

I now know that it is obligatory for myself under the condition that the main mission being carried out, that I myself should see to it that silence is maintained at all times.
**Controlled English**

\[ DCESC^*_CL \] corresponds to a subset of English!

**RCCLNL**: RAIR Lab Controlled Natural Language

\[ K(ugv, \text{now}, holds(carrying(ugv, soldier), now)) \]

The ugv now knows that the fluent, 'the ugv is carrying the soldier,' holds now.

\[ B(ugv, \text{now}, B(\text{commander}, t_1, \neg P(ugv, \text{anytime}, happens(firefight, anytime)))) \]

The ugv now believes that the commander at moment \( t_1 \) believes that it is not the case that the ugv at any time perceives that a firefight happens at any time.

\[ K(l, \text{now}, O(l^*, \text{now}, mission(main), happens(action(l^*, silence), alltime))) \]

I now know that it is obligatory for myself under the condition that the main mission being carried out, that I myself should see to it that silence is maintained at all times.

A Construction Manual for Robot’s Ethical Systems: Requirements, Methods, Implementations
Edited by Robert Trappl

Contents

Preface
Robert Trappl

Table of Contents

Introduction
Robert Trappl

PART I: REQUIREMENTS

1 Shall I Show You Some Other Shirts Too? The Ethics of Persuasive Robots
Jaap Ham and Andreas Spahn

2 Robot: Multi-Use Tool and Ethical Agent
Brigitte Krenn

3 Rude Robots, or: How to do Harm with Words
Sabine Puyr

4 Ethical Systems and Human-Robot Interaction - The case for authority sharing
Florian Gros and Catherine Tessier

PART II: METHODS

5 Non-Monotonic Resolution of Conflicts for Ethical Reasoning
Jean-Gabriel Ganascia

6 Grafting norms onto the BDI agent model
Mihnea Tufis and Jean-Gabriel Ganascia

7 Constrained Incrementalist Moral Decision Making for a Biologically Inspired Cognitive Architecture
Tamás Madl and Stan Franklin

8 Ethical Regulation of Robots is Not Optional: Ethical Reasoning Must be Embedded in Robot Operating Systems
Selmer Bringsjord and Naveen Sundar Govindarajulu

PART III: IMPLEMENTATIONS

9 Case-Supported Principle-Based Behavior Paradigm
Michael Anderson and Susan Leigh Anderson

10 Exploiting Logic Programming as a Computational Tool to Model Morality
Art Saptawijaya and Luis Moniz Pereira

Author Index
A Construction Manual for Robot’s Ethical Systems:
Requirements, Methods, Implementations
Edited by Robert Trapp

Contents

Preface
Robert Trapp

Table of Contents

Introduction
Robert Trapp

PART I: REQUIREMENTS

1 Shall I Show You Some Other Shirts Too? The Ethics of Persuasive Robots
   Jaap Ham and Andreas Spahn

2 Robot: Multi-Use Tool and Ethical Agent
   Brigitte Krenn

3 Rude Robots, or: How to do Harm with Words
   Sabine Payr

4 Ethical Systems and Human-Robot Interaction - The case for authority sharing
   Florian Gros and Catherine Tessier

PART II: METHODS

5 Non-Monotonic Resolution of Conflicts for Ethical Reasoning
   Jean-Gabriel Ganascia

6 Grafting norms onto the BDI agent model
   Mihnea Tuflis and Jean-Gabriel Ganascia

7 Constrained Incrementalist Moral Decision Making for a Biologically Inspired
   Cognitive Architecture
   Tamas Madl and Stan Franklin

8 Ethical Regulation of Robots is Not Optional: Ethical Reasoning Must be
   Embedded in Robot Operating Systems
   Selmer Bringsjord and Naveen Sundar Govindarajulu

PART III: IMPLEMENTATIONS

9 Case-Supported Principle-Based Behavior Paradigm
   Michael Anderson and Susan Leigh Anderson

10 Exploiting Logic Programming as a Computational Tool to Model Morality
    Ari Saptawijaya and Luis Moniz Pereira

Author Index
Most likely future — now:
Most likely future — now:

Only “obviously” dangerous higher-level AI modules have ethical safeguards.

Figure 1: Two Possible Futures

Figure 2: High-level architecture of a new system which augments the DIARC (Distributed Integrated A↵ect, Reflection and Cognition) robotic platform with ethical competence.

Ethical reasoning is implemented as a hierarchy of deontic logic systems which the DIARC system would call upon when confronted with a situation that the system feels is ethically charged. The ethical subsystem then attacks the problem with increasing levels of sophistication till it solves the problem, and then passes on the solution to DIARC. This state of affairs while seemingly satisfactory fails to meet our master requirement that all plans and actions should pass through the ethics system and all changes to the robot’s system (additions, deletions and updates to modules) pass through the ethical layer.

Under joint development by the HRI Lab at Tufts University and the RAIR Lab at RPI.
“Ethical Regulation of Robots is Not Optional: Ethical Reasoning Must be Embedded in Robot Operating Systems”

Most likely future — now:

Only “obviously” dangerous higher-level AI modules have ethical safeguards.

Higher-level cognitive and AI modules

Robotic Substrate

Figure 1: Two Possible Futures

The Situation Now

Figure 2 illustrates the high-level architecture of a new system which augments the DIARC (Distributed Integrated A↵ector, Reflection and Cognition) (Schermerhorn, Kramer, Brick, Anderson, Dingler & Scheutz 2006) robotic platform with ethical competence.

Ethical reasoning is implemented as a hierarchy of deontic logic systems which the DIARC system would call upon when confronted with a situation that the system feels is ethically charged. The ethical subsystem then attacks the problem with increasing levels of sophistication till it solves the problem, and then passes on the solution to DIARC. This state ofairs while seemingly satisfactory fails to meet our master requirment that all plans and actions should pass through the ethics system and all changes to the robot's system (additions, deletions and updates to modules) pass through the ethical layer.
Figure 2: The Situation Now

Robotic Stack

Moral/Ethical Stack

- $DCEC_{CL}$
- $DCEC^*$
- $ADR^M$
- $U$

References


“Ethical Regulation of Robots is Not Optional: Ethical Reasoning Must be Embedded in Robot Operating Systems”
• This situation not optimal. This leads to the “master requirement” proposed by us.

**Ethical Substrate:**

Every robot operating system should include an ethical substrate which sits between lower-level sensors and actuators and any higher-level cognitive system (whether or not that higher-level system itself is designed to enforce ethical regulation).
• This situation not optimal. This leads to the “master requirement” proposed by us.

**Ethical Substrate:**

Every robot operating system should include an ethical substrate which sits between lower-level sensors and actuators and any higher-level cognitive system (whether or not that higher-level system itself is designed to enforce ethical regulation).

---

“Ethical Regulation of Robots is Not Optional: Ethical Reasoning Must be Embedded in Robot Operating Systems”