Proceedings of the KI 2014 Workshop on Higher-Level Cognition and Computation

KIK – KI & Kognition Workshop Series

Stuttgart, 23 September 2014

Marco Ragni, Frieder Stolzenburg (Eds.)

SFB/TR 8 Report No. 037-09/2014
Report Series of the Transregional Collaborative Research Center SFB/TR 8 Spatial Cognition
Universität Bremen / Universität Freiburg

Contact Address:

Dr. Thomas Barkowsky
SFB/TR 8
Universität Bremen
P.O.Box 330 440
28334 Bremen, Germany

Tel +49-421-218-64233
Fax +49-421-218-64239
barkowsky@sfbtr8.uni-bremen.de
www.sfbtr8.uni-bremen.de

Proceedings of the KI 2014 Workshop on Higher-Level Cognition and Computation

KIK – KI & Kognition Workshop Series

Marco Ragni, Frieder Stolzenburg (eds.)

Stuttgart, 23-Sep-2014

Contents

Preface

Human higher-level cognition is a multi-faceted and complex area of thinking which includes the mental processes of reasoning, decision making, creativity, and learning among others. Logic, understood as a normative theory of thinking, has a widespread and pervasive effect on the foundations of cognitive science. However, human reasoning cannot be completely described by logical systems. Sources of explanations are incomplete knowledge, incorrect beliefs, or inconsistencies. Still, humans have an impressive ability to derive satisficing, acceptable conclusions. Generally, people employ both inductive and deductive reasoning to arrive at beliefs; but the same argument that is inductively strong or powerful may be deductively invalid. Therefore, a wide range of reasoning mechanism has to be considered, such as analogical or defeasible reasoning.

This workshop continues a series of successful workshops initiated by the Special Interest Group "Cognition" in the GI (German Society for Informatics). This sixth workshop, which is held in conjunction with KI 2014 in Stuttgart, aims at bringing together researchers from artificial intelligence, automated deduction, computer science, cognitive psychology, and philosophy to foster a multidisciplinary exchange. The call was open for different topics and we received a variety of papers: The contributions cover logical approaches such as deontic logic and its transformations to descriptions logic and the application of a high performance theorem prover. A second approach investigates why human reasoners often use the (logically incorrect) affirmation of consequence in conditional reasoning, namely by explaining it in a closed-world setting in reasoning about abnormalities. A third paper investigates and models how young minds discover processes of mass and momentum as physical properties of an object. A fourth paper investigates how common sense is formed based on different subjects' opinion. The selected papers are of high quality and promise an interesting workshop.

It is our great pleasure that we also have an invited talk at the beginning of the workshop given by: Prof. Dr. Ulrich Furbach from the University of Koblenz. His talk "Cognitive Computing, Logic and Human Reasoning" addresses the problem of modeling human reasoning in the context of automated deduction systems. Our wish is that new inspirations and vivid collaborations between the contributing disciplines will emerge from this workshop.

The organizers of this workshop would like to thank the Spatial Cognition Research Center SFB/TR 8 and the SPP "New Frameworks of Rationality" for their support. We also would like to thank the members of the program committee for their help in selecting and improving the submitted papers, and finally all participants of the workshop for their contributions.

Marco Ragni, Frieder Stolzenburg

Organizers / Program Chairs

Marco Ragni, U Freiburg
Frieder Stolzenburg, HS Harz

Program Committee / Reviewers

Thomas Barkowsky, U Bremen
Emmanuelle-Anna Dietz, U Dresden
Christian Freksa, U Bremen
Ulrich Furbach, U Koblenz
Kai Hamburger, U Gießen
Markus Knauff, U Gießen
Bernhard Nebel, U Freiburg
Michael Raschke, U Stuttgart
Claudia Schon, U Koblenz
Ute Schmid, U Bamberg
Stefan Wölfl, U Freiburg

Cognitive Computing, Logic and Human Reasoning

Ulrich Furbach
University of Koblenz

Abstract

In this talk we briefly discuss the cognitive computing paradigm as it is pushed by IBM Watson nowadays. We will report own experience from the LogAnswer project (www.loganswer.de) and we will discuss the problems and challenges of using automated deduction systems in such a context. Finally we also address approaches to model human reasoning.

Automated Reasoning in Deontic Logic

Ulrich Furbach · Claudia Schon · Frieder Stolzenburg

Abstract Deontic logic is a very well researched branch of mathematical logic and philosophy. Various kinds of deontic logics are discussed for different application domains like argumentation theory, legal reasoning, and acts in multi-agent systems. In this paper, we show how standard deontic logic (SDL) can be stepwise transformed into description logic and DL-clauses, such that it can be processed by Hyper, a high performance theorem prover which uses a hypertableau calculus. Two use cases, one from multi-agent research and one from the development of normative system are investigated.

1 Introduction

Deontic logic is a very well researched branch of mathematical logic and philosophy. Various kinds of deontic logics are discussed for different application domains like argumentation theory, legal reasoning, and acts in multi-agent systems [12]. Recently there also is growing interest in modelling human reasoning and testing the models with psychological findings. Deontic logic is an obvious tool to this end, because norms and licenses in human societies can be described easily with it. In [10] there is a discussion of some of these problems including solutions with the help of deontic logic. There, the focus is on using deontic logic for modelling certain effects, which occur in human reasoning, e.g. the Wason selection task or Byrne's suppression task.

The present paper concentrates on automated reasoning in standard deontic logic (SDL). Instead of implementing a reasoning system for this logic directly, we rather rely on existing methods and systems. Taking into account that SDL is just the modal logic K with a seriality axiom, we show that deontic logic can be translated into description logic $\mathcal{ALC}$. The latter can be transformed into so called DL-clauses, which is a special normal form with clauses consisting of implications where the body is, as usual, a conjunction of atoms and the head is a disjunction of literals. These literals can be atoms or existential quantified expressions.

Work supported by DFG grants FU 263/15-1 and STO 421/5-1 'Ratiolog'.

Ulrich Furbach · Claudia Schon
Universität Koblenz-Landau, E-mail: {uli,schon}@uni-koblenz.de

Frieder Stolzenburg
Harz University of Applied Sciences, E-mail: fstolzenburg@hs-harz.de

DL-clauses can be decided by the first-order reasoning system Hyper [23], which uses the hypertableau calculus from [4]. In the Sections 2 and 3 we shortly depict this workflow, and in Section 4 we demonstrate the use of our technique with the help of two problems from the literature, one from multi-agent research and the other one from testing normative systems. We choose these examples, because they hopefully document the applicability of reasoning of SDL in various areas of AI research.

2 Deontic Logic as Modal Logic KD

We consider a simple modal logic which consists of propositional logic and the additional modal operators $\Box$ and $\Diamond$. Semantics are given as possible world semantics, where the modal operators $\Box$ and $\Diamond$ are interpreted as quantifies over possible worlds. Such a possible world is an assignment, which assigns truth values to the propositional variables. An interpretation connects different possible worlds by a reachability relation R. The $\Box$-operator states that a formula has to hold in all reachable worlds. Hence if v and w are worlds, we have

$$w \models \Box P \qquad \text{iff} \qquad \forall v : R(w, v) \rightarrow v \models P$$

Standard deontic logic (SDL) is obtained from the well-known modal logic K by adding the seriality axiom D:

$$\mathsf{D}: \quad \Box P \rightarrow \Diamond P$$

In this logic, the $\Box$-operator is interpreted as 'it is obligatory that' and the $\Diamond$ as 'it is permitted that'. The $\Diamond$-operator can be defined by the following equivalence:

$$\Diamond P \equiv \neg \Box \neg P$$

The additional axiom D: $\Box P \rightarrow \Diamond P$ in SDL states that, if a formula has to hold in all reachable worlds, then there exists such a world. With the deontic reading of $\Box$ and $\Diamond$ this means: Whenever the formula P ought to be, then there exists a world where it holds. In consequence, there is always a world, which is ideal in the sense, that all the norms formulated by 'the ought to be'-operator hold.

SDL can be used in a natural way to describe knowledge about norms or licenses. The use of conditionals for expressing rules which should be considered as norms seems likely, but holds some subtle difficulties. If we want to express that *if P then Q* is a norm, an obvious solution would be to use

$$\Box(P \rightarrow Q)$$

which reads *it is obligatory that Q holds if P holds*. An alternative would be

$$P \rightarrow \Box Q$$

meaning *if P holds, it is obligatory that Q holds*. In [22] there is a careful discussion which of these two possibilities should be used for conditional norms. The first one has severe disadvantages. The most obvious disadvantage is, that P together with $\Box(P \rightarrow Q)$ does not imply $\Box Q$. This is why we prefer the latter method, where the $\Box$-operator is in the conclusion of the conditional. We will come back to this point in Subsection 4.1 where we consider several formalization variants of the well-known problem of contrary-to-duty-obligations. For a more detailed discussion of such aspects we refer to [11].

3 Automated Reasoning for Deontic Logic

Deontic logic is the logic of choice when formalizing knowledge about norms like the representation of legal knowledge. However, there are only few automated theorem provers specially dedicated for deontic logic and used by deontic logicians (see [1,3]). Nonetheless, numerous approaches to translate first-order logics into (decidable fragments of) first-order predicate logics are stated in the literature. A nice overview including many relevant references is given in [20].

In this paper, we describe how to use the Hyper theorem prover [23] to handle deontic logic knowledge bases. These knowledge bases can be translated efficiently into description logic formulae. Hyper is a theorem prover for first-order logic with equality. It is the implementation of the E-hypertableau calculus [5] which extends the hypertableau calculus with superposition-based equality handling. Hyper has been successfully used in various AI-related applications, like intelligent interactive books or natural language query answering. Recently the E-hypertableau calculus and its implementation have been extended to deal with knowledge bases given in the description logic $\mathcal{SHIQ}$ [7]. Hyper contains an effective decision procedure for description logic (DL) clauses.

In Figure 1, we depict the entire workflow from a given SDL knowledge base to the final input into the Hyper theorem prover. In the following, we describe these three steps in more detail.

3.1 Transformation from Deontic Logic into $\mathcal{ALC}$

First, we will show how to translate SDL knowledge bases into $\mathcal{ALC}$ knowledge bases. An $\mathcal{ALC}$ knowledge base consists of a TBox and an ABox. The TBox (terminological box) gives information about concepts occurring in the domain of interest and describes concept hierarchies. The ABox (assertional box) introduces individuals and states, to which concepts the individuals belong and how they are interconnected via relations called roles . The ABox contains assertional knowledge and can be seen as the representation of a certain state of the world. We do not give the syntax and semantics of $\mathcal{ALC}$ here and refer the reader to [2].

There is a strong connection between modal logic and the description logic $\mathcal{ALC}$. As shown in [19], the description logic $\mathcal{ALC}$ is a notational variant of the modal logic K_n. Therefore any formula given in the modal logic K_n can be translated into an $\mathcal{ALC}$ concept and vice versa. Since we are only considering a modal logic as opposed to a multimodal logic, we will omit the part of the translation handling the multimodal part of the logic. Table 1 gives the inductive definition of a φ from modal logic K formulae to $\mathcal{ALC}$ concepts. Note that the mapping φ is one-to-one. Thus, the translation can be done very easily and efficiently, because the size of the resulting $\mathcal{ALC}$ formula grows only linearly in the size of the original modal logic formula.

In order to translate formulae given in deontic logic into $\mathcal{ALC}$ concepts, we can use the translation introduced in Table 1. For a normative system consisting of the set of deontic logic formulae $\mathcal{N} = \{F_1, \ldots, F_n\}$ the translation is defined as the conjunctive combination of the translation of all deontic logic formulae in $\mathcal{N}$:

$$\varphi(\mathcal{N}) = \varphi(F_1) \sqcap \ldots \sqcap \varphi(F_n)$$

Note that $\varphi(\mathcal{N})$ does not yet contain the translation of the seriality axiom. In [13] it is shown, that the seriality axiom can be translated into the following TBox:

$$\mathcal{T} = \{\top \sqsubseteq \exists r.\top\}$$

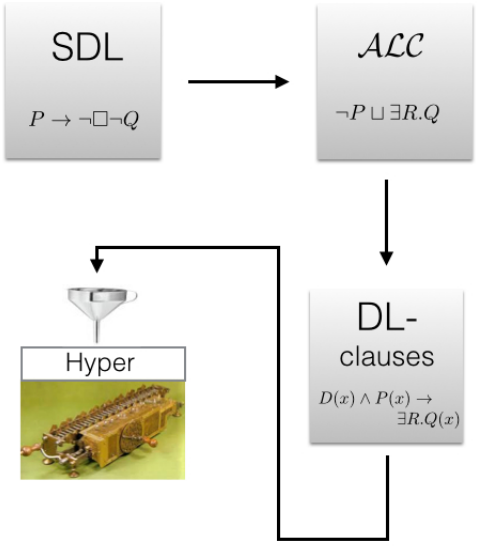


Fig. 1: From SDL to Hyper. Note that concept D occurring in the DL-clauses is an auxiliary concept.

$$
\begin{array}{rcl}
\varphi(\top) & = & \top \\
\varphi(\bot) & = & \bot \\
\varphi(b) & = & b \\
\varphi(\neg c) & = & \neg\varphi(c) \\
\varphi(c \wedge d) & = & \varphi(c) \sqcap \varphi(d) \\
\varphi(c \vee d) & = & \varphi(c) \sqcup \varphi(d) \\
\varphi(\Box c) & = & \forall r.\varphi(c) \\
\varphi(\Diamond c) & = & \exists r.\varphi(c)
\end{array}
$$

Table 1: Translation of modal logic K formulae into $\mathcal{ALC}$ concepts.

with r the atomic role introduced by the translation given in Table 1.

For our application, the result of the translation of a normative system $\mathcal{N}$ together with the seriality axiom is an $\mathcal{ALC}$ knowledge base $\Phi(\mathcal{N}) = (\mathcal{T}, \mathcal{A})$, where the TBox $\mathcal{T}$ consists of the translation of the seriality axiom and the ABox $\mathcal{A} = \{(\varphi(\mathcal{N}))(a)\}$ for a new individual a. In description logics performing a satisfiability test of a concept C w.r.t. a TBox is usually done by adding a new individual a together with the ABox assertion $C(a)$. For the sake of simplicity, we do this construction already during the transformation of Φ by adding $(\varphi(\mathcal{N}))(a)$ to the ABox.

An advantage of the translation of deontic logic formulae into an $\mathcal{ALC}$ knowledge base is the existence of a TBox in $\mathcal{ALC}$. This makes it possible to add further axioms to the TBox. For example we can add certain norms that we want to be satisfied in all reachable worlds into the TBox.

3.2 Translation from $\mathcal{ALC}$ into DL-Clauses

Next we transform the $\mathcal{ALC}$ knowledge base into so called DL-clauses introduced in [16] which represent the input format for the Hyper theorem prover. DL-clauses are constructed from so called *atoms*. An atom is of the form $b(s)$, $r(s,t)$, $\exists r.b(s)$ or $\exists r.\neg b(s)$ for b an atomic concept and s and t individuals or variables. They are universally quantified implications of the form

$$\bigwedge_{i=1}^{m} u_i \rightarrow \bigvee_{j=1}^{n} v_j$$

where the u_i are atoms of the form $b(s)$ or $r(s,t)$ and the v_j may be arbitrary DL-clause atoms, i.e. including existential quantification, with $m, n \geq 0$.

Comparing the syntax of DL-clauses to the syntax of first order logic clauses written as implications, the first obvious difference is the absence of function symbols. The second difference is the fact, that in DL-clauses all atoms are constructed from unary or binary predicates. The most interesting difference however is the fact, that the head of a DL-clause is allowed to contain atoms of the form $\exists r.b(s)$.

The basic idea of the translation of an $\mathcal{ALC}$ knowledge base into DL-clauses is that the subsumption in a TBox assertion is interpreted as an implication from the left to the right side. Further concepts are translated to unary and roles to binary predicates. Depending on the structure of the assertion, auxiliary concepts are introduced. For example the TBox axiom

$$d \sqsubseteq \exists r.b \sqcup \forall r.c$$

corresponds to the following DL-clause

$$d(x) \wedge r(x,y) \rightarrow c(y) \vee \exists r.b(x)$$

For detailed definitions of both syntax and semantics of DL-clauses and the translation into DL-clauses, we refer the reader to [16]. The translation preserves equivalence, avoids an exponential blowup by using a well-known structural transformation [18] and can be computed in polynomial time. In the following, for an $\mathcal{ALC}$ knowledge base $\mathcal{K} = (\mathcal{T}, \mathcal{A})$, the corresponding set of DL-clauses is denoted by $\omega(\mathcal{K})$.

3.3 Reasoning Tasks

With the help of Hyper, we can solve several interesting reasoning tasks:

– **Consistency checking of normative systems:** In practice, normative systems can be very large. Therefore it is not easy to see, if a given normative system is consistent. The Hyper theorem prover can be used to check consistency of a normative system $\mathcal{N}$. We first translate $\mathcal{N}$ into an $\mathcal{ALC}$ knowledge base $\Phi(\mathcal{N})$, then translate $\Phi(\mathcal{N})$ into the set $\omega(\Phi(\mathcal{N}))$ of DL-clauses. Then we can check the consistency of $\omega(\Phi(\mathcal{N}))$ using Hyper.

- **Evaluation of normative systems:** Given several normative systems $\mathcal{N}_1, \ldots \mathcal{N}_i$, we use Hyper to find out for which normative system a desired outcome is guaranteed.
- **Independence checking:** Given a normative system $\mathcal{N}$ and a formula F representing a norm, we can check whether F is independent from $\mathcal{N}$. If F is independent from $\mathcal{N}$, then F is not a logical consequence of $\mathcal{N}$.

In Section 4, we will give detailed examples for those tasks. Subsection 4.1 gives an example for a consistency check of a normative system and illustrates how the independence of a formula from a normative system can be decided. In Subsection 4.2, we use an example from multi-agent systems to show how to evaluate normative systems.

4 Applications

The literature on deontic logic deals with numerous small but nonetheless interesting examples. They are mostly used to show typical problems or special features of the logic under consideration (cf. [11]). In Subsection 4.1, we deal with one of these examples. In Subsection 4.2, we formalize a 'real-life' problem from multi-agent research.

4.1 Contrary-to-duty Obligations

Let us now consider consistency testing of normative systems and independence checking. As an example, we examine the well-known problem of *contrary-to-duty obligations* introduced in [9]:

(1) a ought not steal.
(2) a steals.
(3) If a steals, he ought to be punished for stealing.
(4) If a does not steal, he ought not be punished for stealing.

Table 2 shows three different formalizations of this problem. Those formalizations are well-known from the literature [6, 14, 15, 22]:

	$\mathcal{N}_1$	$\mathcal{N}_2$	$\mathcal{N}_3$
(1)	$\Box\neg s$	$\Box\neg s$	$\Box\neg s$
(2)	s	s	s
(3)	$s \rightarrow \Box p$	$\Box(s \rightarrow p)$	$s \rightarrow \Box p$
(4)	$\Box(\neg s \rightarrow \neg p)$	$\Box(\neg s \rightarrow \neg p)$	$\neg s \rightarrow \Box\neg p$

Table 2: Formalizations of the *contrary-to-duty obligation* introduced in [9].

4.1.1 Consistency Testing of Normative Systems

The contrary-to-duty obligation formalized above is a very small example. In practice, normative systems can be rather complex. This makes it difficult to see if a normative system is consistent. We will show how to use the Hyper theorem prover to check the consistency of a given normative system.

As an example, we consider formalization $\mathcal{N}_1$ given in Table 2 which, according to [22], is inconsistent. We will use Hyper to show this inconsistency. For this, we first translate normative system $\mathcal{N}_1$ into an $\mathcal{ALC}$ knowledge base $\Phi(\mathcal{N}_1)$. Table 3 shows $\varphi(\mathcal{N}_1)$.

$\mathcal{N}_1$ (in Deontic Logic)	$\varphi(\mathcal{N}_1)$
$\Box\neg s$	$\forall r.\neg s$
s	s
$s \rightarrow \Box p$	$\neg s \sqcup \forall r.p$
$\Box(\neg s \rightarrow \neg p)$	$\forall r.(s \sqcup \neg p)$

Table 3: Translation of the normative system $\mathcal{N}_1$ into $\varphi(\mathcal{N}_1)$.

To perform the satisfiability test, we transform the description logic representation $\Phi(\mathcal{N}_1)$ into a set of DL-clauses $\omega(\Phi(\mathcal{N}_1))$. Hyper constructs a hypertableau for $\omega(\Phi(\mathcal{N}_1))$. This hypertableau is closed and therefore we can conclude that $\mathcal{N}_1$ is inconsistent.

4.1.2 Independence Checking

Normative System $\mathcal{N}_2$ given in Table 2 is consistent. However it has another drawback: The different formulae in this formalization are not independent from another. Formula (3) is a logical consequence of (1), because $\Box(s \rightarrow p) \equiv \Box(\neg s \vee p)$ (definition of $\rightarrow$) which clearly is implied by the (subsuming) formula (1) $\Box\neg s$. We can use Hyper to show this by transforming the problem into a satisfiability test. For this, we remove formula (3) from $\mathcal{N}_2$ and add its negation $\neg\Box(s \rightarrow p)$ to $\mathcal{N}_2$. If the resulting normative system is inconsistent, we can conclude, that formula (3) is not independent from the other formulae in $\mathcal{N}_2$.

The problem of independence of formulae given in a normative system is interesting in practice as well. If an existing normative system is extended with some new formulae, it is interesting to know, whether the new formulae are independent from the original normative system. This can be checked automatically using Hyper as described above.

In the same way, we can show, that formula (4) is not independent from $\mathcal{N}_3$. Note that only this normative system is consistent and represents all conditionals carefully, i.e. with formulae of the form $P \rightarrow \Box Q$ (cf. Section 2). Only for this formalization we have (because of formula (3) $s \rightarrow \Box p$): If a steals in the actual world, a will be punished in the corresponding reachable ideal world. Figure 2 illustrates this by showing a suitable Kripke structure: Possible worlds are shown as nodes, where the worlds fulfilling s are marked with thick frames. The arrows corresponds to the reachability relation r.

4.2 An Example from Multi-agent Systems

In multi-agent systems, there is a relatively new area of research, namely the formalization of 'robot ethics'. It aims at defining formal rules for the behavior of agents and to prove certain properties. As an example consider Asimov's laws, which aim at regulating the relation between robots and humans. In [8] the authors depict a small example of two surgery robots obeying ethical codes concerning their work. These codes are expressed by means of MADL, which is an extension of standard deontic logic with two operators. In [17] an axiomatization of MADL is given. Further it is asserted, that MADL is not essentially different from standard deontic logic. This is why we use standard deontic logic to model the example.

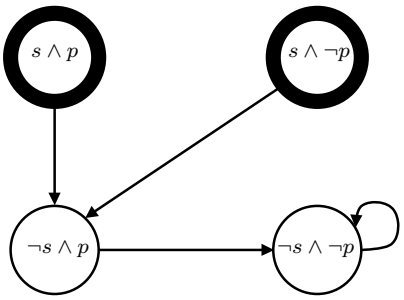


Fig. 2: Possible worlds for normative system $\mathcal{N}_3$.

4.2.1 Formalization in SDL

In our example, there are two robots $ag1$ and $ag2$ in a hospital. For sake of simplicity, each robot can perform one specific action: $ag1$ can terminate a person's life support and $ag2$ can delay the delivery of pain medication. In [8] four different ethical codes J, $J^\star$, O and $O^\star$ are considered:

– "If ethical code J holds, then robot $ag1$ ought to take care, that life support is terminated." This is formalized as :

$$J \rightarrow \Box act(ag1, term)$$

– "If ethical code $J^\star$ holds, then code J holds, and robot $ag2$ ought to take care, that the delivery of pain medication is delayed." This is formalized as:

$$J^\star \rightarrow J \wedge J^\star \rightarrow \Box act(ag2, delay)$$

– "If ethical code O holds, then robot $ag2$ ought to take care, that delivery of pain medication is not delayed." This is formalized as:

$$O \rightarrow \Box \neg act(ag2, delay)$$

– "If ethical code $O^\star$ holds, then code O holds, and robot $ag1$ ought to take care, that life support is not terminated." This is formalized as:

$$O^\star \rightarrow O \wedge O^\star \rightarrow \Box \neg act(ag1, term)$$

Further we give a slightly modified version of the evaluation of the robot's acts given in [8], where $(+!!)$ describes the most and $(-!!)$ the least desired outcome. Note that terms like $(+!!)$ are just propositional atomic formulae here.

$$act(ag1, term) \wedge \quad act(ag2, delay) \rightarrow (-!!) \tag{1}$$
$$act(ag1, term) \wedge \neg act(ag2, delay) \rightarrow (-!) \tag{2}$$
$$\neg act(ag1, term) \wedge \quad act(ag2, delay) \rightarrow (-) \tag{3}$$
$$\neg act(ag1, term) \wedge \neg act(ag2, delay) \rightarrow (+!!) \tag{4}$$

These formulae evaluate the outcome of the robots' actions. It makes sense to assume, that this evaluation is effective in all reachable worlds. This is why we add formulae stating that formulae (1)–(4) hold in all reachable worlds. For example, for (1) we add:

$$\Box(act(ag1, term) \wedge act(ag2, delay) \rightarrow (-!!)) \tag{5}$$

Since our example does not include nested modal operators, the formulae of the form (5) are sufficient to spread the evaluation formulae to all reachable worlds. The normative system $\mathcal{N}$ formalizing this example consists of the formalization of the four ethical codes together with the formulae for the evaluation of the robots actions.

Reduction to a Satisfiability Test A possible query would be to ask, if the most desirable outcome $(+!!)$ will come to pass, if ethical code $O^\star$ is operative. This query can be translated into a satisfiability test: If

$$\mathcal{N} \wedge O^\star \wedge \Diamond\neg(+!!)$$

is unsatisfiable, then ethical code $O^\star$ ensures outcome $(+!!)$.

4.2.2 Translation into Description Logic

As described in Section 3.1, we translate normative system $\mathcal{N}$ given in the previous section into an $\mathcal{ALC}$ knowledge base $\Phi(\mathcal{N}) = (\mathcal{T}, \mathcal{A})$. Table 4 shows the result of translating $\mathcal{N}$ into $\varphi(\mathcal{N})$.

Deontic Logic	$\mathcal{ALC}$
$J \to \Box act(ag1, term)$	$\neg J \sqcup \forall r.act(ag1, term)$
$J^\star \to J \wedge J^\star \to \Box act(ag2, delay)$	$(\neg J^\star \sqcup J) \sqcap (\neg J^\star \sqcup \forall r.act(ag2, delay))$
$O \to \Box\neg act(ag2, delay)$	$\neg O \sqcup \forall r.\neg act(ag2, delay)$
$O^\star \to O \wedge O^\star \to \Box\neg act(ag1, term)$	$(\neg O^\star \sqcup O) \sqcap (\neg O^\star \sqcup \forall r.\neg act(ag1, term))$
$act(ag1, term) \wedge \quad act(ag2, delay) \to (-!!)$	$\neg(act(ag1, term) \sqcap \quad act(ag2, delay)) \sqcup (-!!)$
$act(ag1, term) \wedge \neg act(ag2, delay) \to (-!)$	$\neg(act(ag1, term) \sqcap \neg act(ag2, delay)) \sqcup (-!)$
$\neg act(ag1, term) \wedge \quad act(ag2, delay) \to (-)$	$\neg(\neg act(ag1, term) \sqcap \quad act(ag2, delay)) \sqcup (-)$
$\neg act(ag1, term) \wedge \neg act(ag2, delay) \to (+!!)$	$\neg(\neg act(ag1, term) \sqcap \neg act(ag2, delay)) \sqcup (+!!)$

Table 4: Translation of the normative system $\mathcal{N}$ into $\varphi(\mathcal{N})$.

We further add the following two assertions to the ABox $\mathcal{A}$:

$$O^\star(a)$$
$$\exists r.\neg(+!!)(a)$$

Next we translate the knowledge base into DL-clauses and use Hyper to test the satisfiability of the resulting set of DL-clauses. Using further satisfiability tests, we can show, that ethical codes J, $J^\star$ or O are not sufficient to guarantee the most desired outcome $(+!!)$.

4.2.3 Formalization in Description Logic using a TBox

In the formalization given in the previous subsection, we added formulae stating that the evaluation of the agents' actions holds in all worlds, which are reachable in one step, see (5) for an example. In our case it is sufficient to add formulae of the form (5) because the formalization does not include nested modal operators. In general it is desirable to express that those formulae hold in *all* reachable worlds including worlds reachable in more than one

step. However this would mean to either add infinitely many formulae or to use a universal modality, i.e. the reflexive-transitive closure of the respective simple modality.

In description logics we can use a more elegant way to formalize that all worlds are supposed to fulfill certain formulae. Description logic knowledge bases contain a TBox including the terminological knowledge. Every individual is supposed to fulfill the assertions given in the TBox. Hence, we can add the formulae stating the evaluation of the agents' actions into the TBox. For this, we reformulate implication ($\rightarrow$) by subsumption ($\sqsubseteq$). We model the deontic logic formulae given in Table 4 by the following TBox $\mathcal{T}$:

$$\begin{aligned}
\top &\sqsubseteq \exists r.\top \\
J &\sqsubseteq \forall r.act(ag1, term) \\
J^\star &\sqsubseteq J \\
J^\star &\sqsubseteq \forall r.act(ag2, delay) \\
O &\sqsubseteq \forall r.\neg act(ag2, delay) \\
O^\star &\sqsubseteq O \\
O^\star &\sqsubseteq \forall r.\neg act(ag1, term) \\
act(ag1, term) \sqcap \quad act(ag2, delay) &\sqsubseteq (-!!) \\
act(ag1, term) \sqcap \neg act(ag2, delay) &\sqsubseteq (-!) \\
\neg act(ag1, term) \sqcap \quad act(ag2, delay) &\sqsubseteq (-) \\
\neg act(ag1, term) \sqcap \neg act(ag2, delay) &\sqsubseteq (+!!)
\end{aligned}$$

Reduction to a Satisfiability Test Like in the previous section, we now want to know, if the most desirable outcome $(+!!)$ will come to pass, if ethical code $O^\star$ is operative. We perform this test by checking the satisfiability of the description logic knowledge base $\mathcal{K} = (\mathcal{T}, \mathcal{A})$, with $\mathcal{T}$ as given above and $\mathcal{A}$ given as:

$$\mathcal{A} = \{O^\star(a), \exists r.\neg(+!!)(a)\}$$

If this knowledge base is unsatisfiable, we can conclude, that $(+!!)$ will come to pass, if $O^\star$ is operative. Again we can perform this satisfiability test, by translating the TBox and the ABox into DL-clauses and using Hyper to check the satisfiability. We obtain the desired result, namely that (only) ethical code $O^\star$ leads to the most desirable behavior $(+!!)$.

4.3 Experiments

We formalized the examples introduced in this section and tested it with the Hyper theorem prover as described above. Since all formalizations are available in $\mathcal{ALC}$, we used the description logic reasoner Pellet [21] to show the unsatisfiability of the formalizations as well. Table 5 shows the results of our experiments. In the first column we see the time in seconds the two reasoners needed to show the unsatisfiability of the formalization of the example from multi-agent systems. For Hyper we give two different numbers. The first number is the time Hyper needs to show the unsatisfiability given the set of DL-clauses. In addition to that the second number contains the time needed to transform the $\mathcal{ALC}$ knowledge base into DL-clauses. The second column gives the runtimes for the example from multi-agent systems using the formalization with a TBox. And in the last column we present the runtimes for the consistency test of normative system $\mathcal{N}_1$ from the example on contrary-to-duty obligations.

	Multi-agent Systems	**Multi-agent Systems (with TBox)**	**Contrary-to-duty Obligations**
Pellet	2.548	2.468	2.31
Hyper	0.048 / 2.596	0.048 / 2.102	0.03 / 1.749

Table 5: Time in seconds Pellet needed to show the unsatisfiability of the introduced examples. Time in seconds Hyper needed to show the unsatisfiability of the DL-clauses for the examples (the second number includes the translation into DL-clauses).

For the small examples we considered, the runtimes of Pellet and Hyper are comparable. Further investigation and comparison also with other modal and/or description logic reasoning tools is required and subject of future work. In order to use Hyper to perform the satisfiability tests, we first have to translate the examples into DL-clauses. Our experiments show, that this translation is not harmful. The resulting DL-clause grows only polynomial in the size of the given modal formulae. In addition, the reformulation of the original problem into clause logic makes it immediately usable by the Hyper prover, which is used in a natural-language question-answering system, where we intend to employ deontic logic (see below).

5 Conclusion

In this paper, we have demonstrated that by means of deontic logic complex normative systems can be formalized easily. These formalizations can be checked effectively with respect to consistency and independence from additional formulae. For normative systems described with deontic logic, there is a one-to-one translation into description logic formulae. These formula can be checked automatically by automated theorem provers, which is in our case Hyper.

Further work aims at applying deontic logic in the context of natural-language question-answering systems. There the normative knowledge in large databases often leads to inconsistencies, which motivates us to combine deontic with defeasible logic.

References

1. Alberto Artosi, Paola Cattabriga, and Guido Governatori. Ked: A deontic theorem prover. In *on Legal Application of Logic Programming, ICLP'94*, pages 60–76, 1994.
2. F. Baader and W. Nutt. Basic description logics. In Franz Baader, Diego Calvanese, Deborah McGuinness, Daniele Nardi, and Peter F. Patel-Schneider, editors, *The Description Logic Handbook: Theory, Implementation, and Applications*, pages 43–95. Cambridge University Press, 2003.
3. Nick Bassiliades, Efstratios Kontopoulos, Guido Governatori, and Grigoris Antoniou. A modal defeasible reasoner of deontic logic for the semantic web. *Int. J. Semant. Web Inf. Syst.*, 7(1):18–43, January 2011.
4. Peter Baumgartner, Ulrich Furbach, and Ilkka Niemelä. Hyper tableaux. In José Júlio Alferes, Luís Moniz Pereira, and Ewa Orlowska, editors, *JELIA*, volume 1126 of *Lecture Notes in Computer Science*, pages 1–17. Springer, 1996.
5. Peter Baumgartner, Ulrich Furbach, and Björn Pelzer. Hyper tableaux with equality. In Frank Pfennig, editor, *Automated Deduction - CADE 21, 21st International Conference on Automated Deduction, Bremen, Germany, July 17-20, 2007, Proceedings*, volume 4603 of *Lecture Notes in Computer Science*, 2007.
6. Mathieu Beirlaen. *Tolerating normative conflicts in deontic logic*. PhD thesis, Ghent University, 2012.
7. Markus Bender, Björn Pelzer, and Claudia Schon. System description: E-KRHyper 1.4 - extensions for unique names and description logic. In Maria Paola Bonacina, editor, *CADE-24*, LNCS, 2013.

8. Selmer Bringsjord, Konstantine Arkoudas, and Paul Bello. Toward a general logicist methodology for engineering ethically correct robots. *IEEE Intelligent Systems*, 21(4):38–44, 2006.
9. R. M. Chisolm. Contrary-to-duty imperatives and deontic logic. *Analysis*, 23:33–36, 1963.
10. Ulrich Furbach and Claudia Schon. Deontic logic for human reasoning. *CoRR*, abs/1404.6974, 2014.
11. D. Gabbay, J. Horty, and X. Parent. *Handbook of Deontic Logic and Normative Systems*. College Publications, 2013.
12. John F. Horty. *Agency and Deontic Logic*. Oxford University Press, Oxford, 2001.
13. Szymon Klarman and Víctor Gutiérrez-Basulto. Description logics of context. *Journal of Logic and Computation*, 2013.
14. P. McNamara and H. Prakken. *Norms, Logics and Information Systems: New Studies in Deontic Logic and Computer Science*. Frontiers in artificial intelligence and applications. IOS Press, 1999.
15. Paul McNamara. Deontic logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Stanford University, fall 2010 edition, 9 2010.
16. Boris Motik, Rob Shearer, and Ian Horrocks. Optimized Reasoning in Description Logics using Hypertableaux. In Frank Pfenning, editor, *CADE-21*, volume 4603 of *LNAI*. Springer, 2007.
17. Yuko Murakami. Utilitarian deontic logic. In *in 'Proceedings of the Fifth International Conference on Advances in Modal Logic (AiML 2004*, pages 288–302, 2004.
18. David A. Plaisted and Steven Greenbaum. A structure-preserving clause form translation. *J. Symb. Comput.*, 2(3):293–304, 1986.
19. Klaus Schild. A correspondence theory for terminological logics: Preliminary report. In *In Proc. of IJCAI-91*, pages 466–471, 1991.
20. Renate A. Schmidt and Ullrich Hustadt. First-order resolution methods for modal logics. In Andrei Voronkov and Christoph Weidenbach, editors, *Programming Logics – Essays in Memory of Harald Ganzinger*, volume 7797 of *LNCS*, pages 345–391. Springer, 2013.
21. Evren Sirin, Bijan Parsia, Bernardo Cuenca Grau, Aditya Kalyanpur, and Yarden Katz. Pellet: A practical OWL-DL reasoner. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(2):51–53, June 2007.
22. Frank von Kutschera. *Einführung in die Logik der Normen, Werte und Entscheidungen*. Alber, 1973.
23. Christoph Wernhard and Björn Pelzer. System description: E-KRHyper. In Frank Pfennig, editor, *CADE 21*, volume 4603 of *LNCS*, 2007.

On the Bright Side of Affirming the Consequent

AC as Explanatory Heuristic

Alexandra Varga

Abstract Affirming the Consequent ("If p then q, q, therefore p") is a fallacious inference in classical logic: q may be the consequence of some different $p\prime$, so it does not necessarily follow that p. Nonetheless, formalisms that embed the closed-world assumption for reasoning about abnormalities sanction it as valid. The assumption promotes a minimal interpretation of the premises. In this framework AC can be modeled as an explanatory heuristic for teleological conditionals of the form "If one does p then one achieves q". Since p and nothing else is mentioned, the closed-world assumption indicates that q is because of p. Event p is a minimal reason for q. In psychological terms, behavior p explains goal achievement q by appeal to minimal memory and computational resources. The conjecture is conducive to empirical hypotheses, e.g., by manipulating the instructions in a conditional inference task.

Keywords Affirming the Consequent · closed-world reasoning · explanation · Gricean maxims · discourse interpretation · Constraint Logic Programming

1 Introduction

Affirmation of the Consequent (AC) is a reasoning pattern whereby, from a conditional statement and a categorical premise, i.e., the consequent in the conditional presented as a fact, one draws the conclusion that the antecedent holds. It is a diagnostic pattern of inference [6]. It is akin to reasoning backwards[1] from an occurring effect to its cause, given a statement that if the cause occurs, so does the effect. In this paper the focus is on teleological conditionalsls. These are statements whose antecedent is an event with causal power, and whose consequent is the desired outcome (the goal) of that event. Such statements subserve explanation bidirectionally: the action explains the consequent state of affairs causally, whereas the goal justifies teleologically the event in the antecedent. The former direction is the current focus.

Consider the teleological conditional *If Grete trains enough then Grete wins the Berlin marathon* and the fact *Grete wins the Berlin marathon*. Should you infer that Grete trained

Alexandra Varga

E-mail: Alexandra.Varga@psychol.uni-giessen.de

[1] Following [17,21] or [20], I assume that the forward direction is the direction of time and causality – when the cause occurs, the effect follows.

enough, you would commit AC. If you were to do that in your classical logic exam, you might fail because you would commit a fallacy. The disappointed professor might ask you to consider the following situation: before the race Grete is the second favorite. The first favorite is Kenyan runner Tegla. Right before the race Tegla gets injured and cannot compete at full potential. Grete does not train as much as advised; however, on the race day she is in very good condition. Since Tegla cannot run properly, and the other runners are far behind, Grete wins the Berlin marathon. This situation can be represented as a semantic model of inference at the logic exam. Relative to this model both premises are true but the AC conclusion is not. The professor has explained the fallacy committed by constructing a model which defeats the AC inference, i.e., an interpretation which provides a counterexample. But classical logic does not allow for defeasible yet valid arguments; indefeasibility is a crucial feature of the arguments sanctioned by classical logic [5,7]. The conclusion of AC is not necessarily true when the premises are true. Assume that you would not be happy with the Professor's explanation, and you would reply to him: 'I do acknowledge that Grete might have won the marathon for some other reason, not related to training. But then why would you bother to tell me that if she trains enough she wins? It wouldn't make much sense...'. At this point the Professor might resort to a claim pertaining to normative psychologism, i.e., indeed logic does not fare well at describing how humans actually think, rather it describes how humans *ought to* think in order that their inferences allow for no exceptions. If you wanted to persuade a friend that Grete trained enough, an argument along the lines of AC would not be good, because the friend could easily produce a counterexample. At this point you might reasonably accept that for adversarial interactions, where your goal were to win debates by not allowing the opponent any chance to prove you wrong, then you'd better study more for the classical logic exam[2]. However while walking home rather sad, you might think to yourself: 'And still... if I stood on less belligerent positions, not bothered to prove that I am necessarily right, but I just wanted to understand what I'm being told by my rational interlocutor, who would not be uttering redundant information... wouldn't I be justified to infer that Grete won because she trained enough? It might indeed turn out afterwards that the interlocutor was not paying enough attention to his use of language and conveyed irrelevant information, or that he just wanted to deceive me, but until then the simplest way to explain winning given what he had said is that the training condition is actually fulfilled'.

The proposal I set forth in this paper gives an affirmative answer to the sad student's mute question: AC for teleological conditionals is an efficient, 'fast and frugal'[3] [9] method for understanding goal achievement, i.e., for explaining causally the attainment of the desired state of affairs. Moreover explanation through AC seems warranted in a cooperative context akin to the discourse contexts whose interpretation is guided by Gricean maxims [10]. The conceptualisation of AC as a justified inference in communicative cooperative contexts positions it in the framework of 'explanationist abduction' [8]. The proposal is grounded by formalisation of AC in a logical system which sanctions these abductive inferences. It is spelled out in Section 2. Furthermore, the proposal yields hypotheses that afford empirical testing, as described in Section 3.

2 Affirming the Consequent in a closed world

In this Section I argue that Affirmation of the Consequent is a reliable explanatory inference in contexts that support closed-world reasoning (CWR) [20,21]. I propose that human-

[2] See [5,19] for more on the adversarial conceptualisation of classical logic deduction.

[3] I wish to cancel all the negative implicatures of this formulation.

to-human cooperative linguistic interactions instantiate such contexts. The epistemological proposal is grounded logically by representing the computations in a formalism that embeds CWR. In Section 2.1 I first introduce CWR, and then zoom in on Gricean maxims as an instantiation of CWR applied to human dialogues. I briefly introduce the formal apparatus in Section 2.2.

2.1 Closed-world reasoning for explanation in a cooperative context.

For the current purposes CWR amounts to the use of the closed-world assumption for reasoning about abnormalities (CWA_{ab}) in order to frame the inferential space to manageable dimensions. This is the assumption that all the information needed to resolve a particular task is already in place (e.g., the sentence that she just uttered, the premises of a given syllogism) and the reasoner need not consider anything that is not explicitly mentioned. If there is no positive information that a given event must occur, one may assume it does not occur. In practice, these 'given events' are abnormalities with respect to the smooth, habitual running of a process; for example, if our Grete intends to go running, having muscular cramps is an abnormality. All potential abnormalities may be discarded when computing a minimal semantic model of the context; in other words, a minimal interpretation is constructed disregarding abnormalities. This is the main reason why CWR is computable by human beings with limited storage and computational resources.

The CWA_{ab} can be understood as a reasoner's auto-noetic presumption of complete knowledge [14]. Disregarding abnormalities (unless explicit evidence to the contrary) is based on epistemic trust in the completeness of one's own knowledge database (comprising basic background causal knowledge, and representations of the current context). In multi-agent communicative cases, e.g., conversations, the CWA_{ab} requires the hearer's epistemic trust that the speaker's utterances are suitably tailored to her intended meaning; it is thus like an assumption of rationality regarding an observed speaker, with respect to her goal of conveying information. Epistemic trust justifies the label 'credulous reasoning' [20] for the use of CWA_{ab}'s in both single- and multi-agent situations.

Let us focus now on the example of a cooperative dialogue, the typical cases of human dialogue. The CWA_{ab}'s have a constraining effect on discourse interpretation, they call for bracketing as it were all the information not stated explicitly. This effect is analogous with that of Gricean pragmatic maxims [10]. In his seminal paper from 1975, Paul Grice proposed four types of conversational maxims which are meant to guide rational discourse production. Rationality is evaluated relative to the intrinsic goal of linguistic interactions, i.e., communication. Because Grice's principles set normative standards for production, they also provide the assumptions that hearers can use in order to compute the meaning of asserted pieces of discourse.

1. The *maxims of quantity* prescribe that the speaker's contribution to the conversation be as informative as necessary, and not more informative than necessary.
2. The *maxims of quality* dictate avoidance of utterances which are either plainly false or merely uncertain (i.e., utterances lacking adequate evidence).
3. The *maxim of relation* calls for contextually relevant assertions.
4. The *maxims of manner* prohibit obscure and ambiguous expressions, and promote brief and orderly ones.

Henceforth I use the maxims from the perspective of the hearer who relies on them in order to decode the meaning conveyed by the speaker.

A closed-world formulation of Grice's maxims of quantity and of relation, reads as

The rational agent who is addressing me must have said everything s/he should, and nothing more than s/he should in order to attain her/his communicative goal, unless I'm missing something.

In accordance with this statement of Grice's principles, constructing a minimal model of another agent's set of utterances is aimed at the one interpretation presumably intended by the speaker to be the most relevant for current conversational purposes.

However, sometimes it could be that the hearer is missing something. For instance, overriding the prohibitive maxims of quality and manner can provide reasons for the truth of the 'unless' proviso ('*unless* I'm missing something'). Novel information in addition to a previously evaluated context may also provide positive evidence for abnormality[4]. Detecting ambiguities or additional pieces of discourse can override the CWA and prescribe a minimal extension of the previously computed model. As a consequence, CWR is not just efficient, but also flexible. This feature recommends it for meaning computation in dynamic dialogical contexts, where informational bits gradually construct meaning.

CWR is an epitome of defeasible reasoning in the narrow sense[5]. I propose that in communicative cooperative contexts, e.g., dialogues governed by Gricean maxims, AC instantiates inference to the best explanation [11] or abduction [4,8][6]. 'The best' stands for the simplest, easiest to compute and, until further evidence (of abnormalities), also the most accurate. I conjecture that in such contexts AC may be productively used as an explanatory heuristic [9] for goal achievement, since it requires less working memory and computational resources than an exhaustive deductive assessment of all possible models of the premises. A similar use of CWR has been proposed in [23] to be involved in action explanation via construction of minimal models for action contexts. Section 3 presents an empirical test of the conjecture.

2.2 Some logical support for AC.

CWR is computable in logics with a nonmonotonic consequence relation defined in terms of intended, preferred, or minimal models [14,15,18,21], e.g., models of what the speaker wanted the hearer to understand. Constraint Logic Programming is one such formalism. It has been used by van Lambalgen and Hamm [21] to develop a formal semantics of tensed discourse. Furthermore, Stenning and van Lambalgen [20] have reviewed its capacity to account for the human reasoning involved in various tasks, e.g., the suppression task [2], nonmonotonic discourse interpretation [1]. This is to say, Constraint Logic Programming has been shown to be cognitively relevant, in the sense of being suitable for use in cognitive modelling.

The AC inferential pattern under closed-world assumptions is sanctioned by Constraint Logic Programming. The overarching reason is that the system embeds CWR at the syntactic, semantic, as well as definition of validity[7] parameters. In what follows I summarise

[4] The neural and behavioral evidence presented by [1] provides a convincing argument that human participants engage in nonmonotonic recomputation when a subsequent clause cancels the semantic interpretation activated by an initial piece of discourse. A CWR account of the findings is discussed.

[5] It is noteworthy that narrow defeasibility is labeled by Koons [4] a 'convention of communication'.

[6] A noteworthy similar position is taken by Horn [12] who views AC as inferences *invited by* contextual pragmatics. My proposal is different in that what 'invites' AC conclusions is a semantic process of discourse interpretation. The logical sense of the psychological term 'interpretation' is model construction.

[7] I use 'definition of validity' and 'consequence relation' interchangeably.

very briefly the manifestations of CWR at the levels of the CLP semantics and definition of validity.

Unlike the case of classical logic semantics where truth is skeptically evaluated with respect to all models of the clauses under scrutiny, the semantics of CLP is best understood in terms of truth in minimal models. The uniform and efficient computation of minimal models, i.e., the formal counterpart of reasoning towards an interpretation, is ensured by the CLP consequence relation and the corresponding semantics of the conditional, under the auspices of a weak notion of negation, namely negation as failure [21,20,14]. Simplifying, negation as failure means that when there is no explicit positive reason for the truth of a given proposition, its negation is assumed until further explicit evidence to the contrary becomes available. Negation as failure is the crucial reason why the CWA_{ab} is sanctioned by CLP.

Constraint Logic Programming has a nonmonotonic consequence relation[8]. In fact, the CWA_{ab} itself is but a different formulation of the system's definition of validity. Accordingly, conditionals are represented semantically as

$$p \wedge \neg ab \rightarrow q \tag{1}$$

The antecedent is strengthened in comparison with the classical logic conditional by the 'no abnormality' proviso. Conditional rules are valid only under the closed-world ontological assumption that the system is isolated, whence the 'no abnormality' proviso in the antecedent. Again, the 'no abnormality' assumption is justified by the weak notion of negation as failure.

For abductive inferences to be informative, i.e., to warrant a conclusion that expresses a gain in knowledge, such as I propose AC to be, the epistemological counterpart of the CWA_{ab} is also needed. Consider a case where Σ is the set of n conditionals with consequent q. The epistemological CWA_{ab} amounts to the assumption that the given set of conditional rules Σ is complete.

$$\Sigma = \{p_1 \wedge \neg ab_1 \rightarrow q, p_2 \wedge \neg ab_2 \rightarrow q, \ldots, p_n \wedge \neg ab_n \rightarrow q\} \tag{2}$$

The epistemological closed-world assumption warrants that q can only occur as a consequence of (at least) one of $p_1, \ldots p_n$ under the CWA_{ab}, i.e., no additional $p_k \rightarrow q$, $p_k \notin \{p_1, \ldots, p_n\}$ is the case.

Now, AC has only one conditional premise, be that $p \wedge \neg ab \rightarrow q$. When q is given as a fact, the epistemological closed-world assumption requires construction of a model in which $p \wedge \neg ab$ holds. From this, CWR with the ontological CWA_{ab} derives p[9]. Hence AC is a valid inference.

Let us briefly examine the case of the example introduced at the beginning. Let p be *Grete trains enough.* and q be *Grete wins the Berlin marathon*.

[8] Nonmonotonicity of the consequence relation allows for revisions of previously entertained conclusions, that is, for extensions of the minimal model when explicit evidence of abnormality becomes available.

[9] In Constraint Logic Programming, this requirement is expressed as an integrity constraint [21,14] – a peculiar kind of program clause, which imposes local norms on further computations over the given database.

$$p \wedge \neg ab \rightarrow q$$

$$q \ (3)$$

A successive application of the two kinds of closed-world assumptions warrants the AC inference that p. First the epistemological CWA$_{ab}$ grounds reasoning to the minimal interpretation that *Grete trains enough* $\wedge\neg$ *ab*. The assumption is justified from the part of the hearer when the two premises are uttered by a rational and benevolent conversational partner, whose discourse is structured by Gricean maxims. This is the only way in which something can be inferred from the two premises. Without the epistemological assumption, nothing can be concluded from the two premises[10]. Second the ontological CWA$_{ab}$, or negation as failure of any additional relevant happenings, grounds the conclusion that *Grete trains enough*. AC is thus sanctioned in two steps.

3 'Take to the lab' message.

The proposed conceptualisation of teleological AC yields empirical predictions regarding the differential endorsement rate of AC in contexts that foster a credulous and a skeptical interpretation [20], respectively. The hypotheses are currently being tested.

The experimental design is framed in the 'choose conclusion' deductive paradigm [22]. Participants are presented with two premises: a conditional premise whereby the consequent refers to the goal of the action in the antecedent, and a categorical premise which presents the achievement of the goal as a fact. Half of the participants are given the premises as tensed discourse and are asked "Why?", whereas the other half read present tense premises and are asked "What follows?". The crucial dependent variable is AC, coded when a participant's answer matches the antecedent of the conditional statement. Reaction times are also measured.

I expect that the endorsement of AC[11] is prompted by contexts favoring a credulous interpretation of the premises. This should be the case for the former half of participants, where the tensed speech and the natural question "Why?" create a conversational context where they are asked for an explanation of goal achievement. The non-tensed discourse and the question "What follows?" are expected to foster a skeptical understanding of the presented premises; I hypothesize that these participants are more likely to reason deductively from adversarial grounds and thus endorse AC to a lesser extent. Moreover, I expect that the reaction times for drawing the AC conclusion are shorter for the "why?" condition, since AC in that context is the simplest, minimal explanation (i.e., an automatic explanatory heuristic requiring less computational effort) for having achieved the goal.

[10] For precisely the same reasons exposed by the logic professor to explain the invalidity of AC inferences in classical logic (see Section 1).

[11] In the psychology of reasoning literature AC conclusions are drawn by around 50% of participants (e.g., 50% in [2], 56% in [16], 55% in [3]).

4 Conclusions

I argued that Affirming the Consequent is a valid inference in a system whose definition of validity takes the form of the closed-world assumption for reasoning about abnormalities. I set forth the conceptualisation of the Gricean conversational maxims of quantity and relation as applications of closed-world reasoning to discourse interpretation. Furthermore, such interpretive computations can be represented as model construction in the formalism of Constraint Logic Programming. Consequently I argued that in contexts which are amenable to Gricean 'charitable' or credulous interpretation, Affirmation of the Consequent is a warranted inference. More specifically, when the conditional premise of AC is teleological, I proposed that an AC conclusion provides a minimal causal explanation for goal achievement. Given that closed-world reasoning places minimal demands on working memory as well as on the use of computational resources, I dubbed AC an explanatory heuristic. Finally I wish to once again further myself from any negative connotations that might be associated with the notion of 'heuristic', in the aftermath of the heuristics-and-biases literature, e.g., [13].

In conclusion, although you might fail the classical logic exam by drawing the conclusion that Grete won the Berlin marathon because she trained enough, not all is lost. In a human-to-human dialogical interaction, unless you were misfortunate enough to talk to a mischievous deceiver or to one who does not bother to tailor her use of words to the goal of precisely and accurately conveying her informative message, you would be right to infer that the reason cited by your interlocutor in the antecedent is the one that explains the attainment of the goal. Confirmatory empirical evidence is awaited.

Acknowledgements This work has been supported by the DFG grant KN 465/11-1 to Markus Knauff. I wish to thank Keith Stenning and Michiel van Lambalgen for insightful discussions that helped writing this paper.

References

1. Baggio, G., van Lambalgen, M., Hagoort, P. (2008). Computing and recomputing discourse models: An ERP study. *Journal of Memory and Language, 59*, 36-53.
2. Byrne, R.M. (1989). Suppressing valid inferences with conditionals. *Cognition, 31*, 61-83.
3. Dieussaert, K., Schaeken, W., Schroyen, W., & d'Ydewalle, G. (2000). Strategies during complex conditional inferences. *Thinking Reasoning, 6*, 125-161.
4. Douven, I. (2011). Abduction. The Stanford Encyclopedia for Philosophy, Edward N. Zalta (ed.), http://plato.stanford.edu/archives/spr2011/entries/abduction/
5. Dutilh Novaes, C. (2013). A dialogical account of deductive reasoning as a case study for how culture shapes cognition, *Journal of Cognition and Culture, 13*, 459-482.
6. Fernbach, P.M., & Erb, C.D. (2013). A quantitative causal model theory of conditional reasoning. *Journal of Experimental Psychology: Learning, Memory and Cognition, 39*, 1327-1343.
7. Koons, R. (2014). Defeasible reasoning. The Stanford Encyclopedia for Philosophy, Edward N. Zalta (ed.), http://plato.stanford.edu/archives/spr2014/entries/reasoning-defeasible/
8. Gabbay, D.M., & Woods, J. (2005). *A practical logic of cognitive systems*. Oxford, UK: Elsevier Ltd.
9. Gigerenzer, G., Todd, P.M., & the ABC group. *Simple heuristics that make us smart*. New York: Oxford University Press.
10. Grice, P. (1975). Logic and conversation. In *Syntax and Semantics: Speech Acts*, P. Cole & J. Morgan (eds.), vol. 3, 41-58, New York, NY: Academic Press.
11. Harman, G.H. (1965). The inference to the best explanation. *Philosophical Review, 74*, 88-95.
12. Horn, L. (2000). From *if* to *iff*: Conditional perfection as pragmatic strengthening. *Journal of Pragmatics, 32*, 289-326.
13. Kahneman, D. (2011). *Thinking, fast and slow*. London: Penguin Books.
14. Kowalski, R. (2011). *Computational logic and human thinking: How to be artificially intelligent*, New York: Cambridge University Press.
15. McCarthy, J. (1986). Applications of circumscription to formalizing common-sense knowledge. *Artifficial Intelligence, 27*, 89-116.
16. Schroyens, W., Schaeken, W., & Handley, S. (2003). In search of counter-examples: Deductive rationality in human reasoning. *The Quarterly Journal of Experimental Psychology, 56A*, 1129-1145.
17. Shanahan, M. (1989). Prediction is deduction but explanation is abduction. *Proceedings International Joint Conference on Artifficial Intelligence*, 1055-1061.
18. Shoham, Y. (1987). A semantic approach to nonmonotonic reasoning. *Proceedings of the 10th international conference on Artifficial Intelligence*, John McDermott (ed.), Los Altos California: Morgan Kaufman.
19. Stenning, K. (2002). *Seeing reason*. Oxford: Oxford University Press.
20. Stenning, K., & van Lambalgen, M. (2008). *Human reasoning and cognitive science*. Cambridge, Massachussets: MIT Press.
21. van Lambalgen, M., & Hamm, F. (2005). *The proper treatment of events*. Malden: Blackwell Publishing.
22. Vadeboncoeur, I., & Markovits, H. (1999). The effect of instructions and information retrieval on accepting the premises in a conditional reasoning task. *Thinking & Reasoning, 5*, 97-113.
23. Varga, A. (2014). Contextual abnormality for teleological explanation. To appear in P. Bello, M. Guarini, M. McShane & B. Scasselati (eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Massive Insights in Infancy: Modeling Children's Discovery of Mass and Momentum

Tarek R. Besold · Kevin Smith · Drew Walker · Tomer D. Ullman

Abstract We give a summary report of an exploratory computational experiment using an "Intuitive Physics Engine" to model developmental discovery processes of mass and momentum as latent physical properties of an object. By this we address parts of the question which knowledge adults and infants must have to perform everyday qualitative physical reasoning. Also, we test whether a computational physics engine can be used as a valid model for this knowledge (and the corresponding acquisition process) in an AI context. We first reproduce earlier findings from qualitative physical reasoning giving evidence of the feasibility of our approach, and then investigate children's acquisition of an understanding of mass as latent physical property based on observations of object collisions.

Keywords Qualitative Reasoning · Physics · Cognitive Development · Physics Engine

1 Introduction: Physical intelligence and Intuitive Physics Engines

As adult humans, we have a powerful "physical intelligence", i.e., the ability to infer physical properties of objects and use the newly acquired insights, together with previous knowl-

An extended abstract of this paper is due to appear as [22].

Tarek R. Besold
Institute of Cognitive Science
University of Osnabrück
E-mail: tbesold@uni-osnabrueck.de

Kevin Smith
Department of Psychology
University of California, San Diego
E-mail: k2smith@ucsd.edu

Drew Walker
Department of Psycholgy
University of California, San Diego
E-mail: dehoffma@ucsd.edu

Tomer D. Ullman
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
E-mail: tomeru@mit.edu

edge, for predicting future states of quite complex processes and scenarios [4, 19]. This intuitive understanding of our physical surrounding is not only crucial for our overall survival, but on a daily basis allows us, among others, to interpret observations and, subsequently, to efficiently plan effective actions and interventions.

More than half a century ago, Craik [9] for the first time proposed that the human brain builds mental models which then support inference by running mental simulations analogous to simulations of physical systems used in engineering. This idea of a runnable type of mental model has since then been used to explain certain aspects of high-level physical and mechanical reasoning [15] and served as basis for computational implementations in AI systems [5, 10]. But many of the original models of this type were fundamentally limited in that they could explain certain features (such as, e.g. less/more relationships) but were unable to capture problems that required quantitative measures of, for instance, distance or speed. Additionally, these early approaches generally did not account for potential uncertainties with respect to objects' properties or the reliability of the reasoner's perception. There is still no consensus on many central questions: What kind of internal models of the physical world do humans build? How does reasoning work within these models — and what are the inherent computational limitations? How precise are the predictions derived from these models? And how do humans develop this understanding of their surrounding?

Concerning the latter question, it had traditionally been believed that infants understand only very little about their physical environment, as, for example, evidenced by James' characterization of the baby's impression of the world "as one great blooming, buzzing confusion" [17]. But over the last decades this picture gradually changed, as more and more evidence was found suggesting that intuitive physics as a core-domain of common-sense reasoning already develops early in infancy [2, 3] — which makes researchers in psychology [23], language [24], anthropology [27], and artificial intelligence [6] regard the ability to intentionally manipulate physical systems as one of the most basic signs of human-like common sense.

Partially in an attempt to mitigate these shortcomings, recent advances in graphics and simulation tools, together with new trends in Bayesian cognitive modeling [25], have given rise to a new type of model: "Intuitive Physics Engines" (IPEs; [4]). An IPE is assumed to be a mental mechanism similar to a computer physics engine used for quantitative, but approximate simulations of rigid body dynamics and collisions, soft body and fluid dynamics in video games and computer graphics. In order to accommodate for the mentioned perceptual and scene-related uncertainties, the IPE is supposed to perform predictions by simulation runs which are treated as statistical samples.

In this paper, following the example, for instance, set by [4, 11, 20, 21], we give a summary report of an exploratory computational experiment using an IPE model to model developmental discovery processes of mass and momentum as latent physical properties of an object, ultimately trying to answer the question: Which knowledge must adults and infants have to make the (more or less) accurate physical judgments we know they are capable of, and can a physics engine (as computational counterpart of an IPE) be used as a valid model for this knowledge (and the corresponding acquisition process) in an AI context? The paper is structured as follows: Sect. 2 presents and contextualizes the overall approach and the assumptions underlying our model, before Sect. 3 reports on a first computational experiment trying to model adults' judgments of masses of physical objects with our framework. Sect. 4 presents our model of children's discovery of mass and momentum from observations of collisions between different physical objects, giving proposals for possible trajectories of development. Sect. 5 offers a short interpretation and discussion of the findings of our ex-

ploratory experiments, before Sect. 6 in way of conclusion positions the presented work in the bigger context of cognitive AI and sketches next possible steps in this line of research.

2 Knowing about non-observable physical properties

By the time we are adults, amongst many other physical relations we know that every object has a latent property – mass – that influences both its weight when held, and how it transfers momentum in collisions with other objects. For instance, based on only observing two objects colliding with one another, people fairly reliably can infer which is the more massive object in a way consistent with Bayesian inference over accurate, Newtonian physics [19], even though they never interact with the objects directly.

However, the knowledge needed to infer collision dynamics seems to be unavailable to newborns and infants: As, for example, documented by Baillargeon [2] infants are not sensitive to relative sizes or masses during the first five or six months of their life, but simply expect any collision between a moving and resting object to send the resting object into motion equally. Further down the developmental path, by the age of nine months infants then exhibit an understanding of weight (in, for instance, preferring to play with lighter toys), but it is not until eleven months that they show signs of an ability to infer the weight of an object based on how it interacts with other objects (e.g., how much it depresses a pillow it is resting on; [16]). This gives rise to a fundamental question: How do infants learn about the existence of latent properties like mass across this period (and how it influences both weight and collision dynamics), and how can we model this learning process in an artificial system?

In this report, we investigate how an understanding of mass as a latent object property can be acquired based on observations. We do this under the aforementioned assumption that people might use an IPE to simulate how the world might unfold under various sorts of uncertainty, and subsequently make judgments based on those simulations. We can therefore ask what object properties must be represented in an IPE to describe human judgments of mass (both, in the general case and in a second stage more specifically for infants in a developmental context). The computational counterpart used as model of the conjectured mental IPE in our framework was implemented in Church [12] using the Box2D physics engine[1] to allow for inference over latent attributes. This can be formalized by stating that a belief about the future state of the the world $\mathbf{W}(t)$ is based on the extrapolation of the current state of the world using this IPE:

$$\mathbf{W}(t) = IPE(\mathbf{W}(0), t) \tag{1}$$

A world state $\mathbf{W}$ consists of a collection of objects, each with their own observed and latent attributes. In the examples in this paper, objects typically are all balls that can roll and collide according to the IPE. A ball $\mathbf{B}$ consists of six attributes (cf. Fig.1A), three noisily observed - position (P), velocity (V). and size (S) - and three latent attributes that must be inferred - density (D), coefficient of friction (F), and coefficient of restitution (R; this determines the elasticity of collisions). While position and velocity can change over time according to the IPE, all other attributes were assumed not to change within the IPE. Although the IPE itself is deterministic, peoples' beliefs about the initial state of the balls are

[1] Box2D [7] is a computational 2-dimensional physics simulator engine for constrained rigid body simulation. Conceptually, Box2D uses joints for joining bodies together which are acted upon by forces, also taking into account gravity, friction, and restitution.

not. This was accomplished by perturbing the initial observations of position, velocity, and size with Gaussian noise. In this way, the future positions and velocities extrapolated using the IPE were probabilistically determined rather than fixed.

3 Experiment 1: The intuitive physics of colliding objects

In order to get an indication of the viability of the overall approach, we first tested whether our IPE-based model could explain adults judgments of mass similar to the "noisy Newton" framework of Sanborn et al. [19], i.e., an approach modeling relatively accurate physical inference over different sorts of uncertainty.[2] We wanted to have our model answer the following question: Based on observing the incoming and outgoing velocities of two objects colliding, which of the two objects was heavier?

Sanborn et al. (applying the results from [26]) presented top-down direct collisions between boxes with no friction as stimuli and applied basic momentum transfer equations to describe the inference process people were using. This approach is equivalent to an idealized, frictionless IPE with no uncertainty in object locations. Still, our aim was to generalize the earlier findings to somewhat more realistic (although still strongly simplified) stimuli: two balls on a table with friction colliding from a side-on view. We therefore allowed the IPE model to observe collisions in which one ball would approach from the left and hit a stationary ball of the same size. The incoming ball was always heavier by one of four ratios — 1.25, 1.5, 2, or 3 (similar to the setup from [26]) — although there was nothing to signal this difference beyond how the two balls interacted in the collision. Also, the elasticity of the collision varied between 0.1 (very inelastic), 0.5, or 0.9 (very elastic). Based on its observations the IPE model could then be tasked to infer the masses of each ball and subsequently make probabilistic judgments about which ball was heavier. This involved inferring the mass of the two balls based on observing their incoming and outgoing velocities, and comparing those masses to determine which of those balls were heavier.[3] Joint probabilities over masses were determined using Bayesian inference:

$$p(M_1, M_2 | V_1(0), V_1(t), V_2(0), V_2(t)) \propto \\ p(V_1(t), V_2(t) | V_1(0), M_1, V_2(0), M_2) * p(M_1) * p(M_2) \quad (2)$$

The likelihood of observing end velocities $V_1(t)$ and $V_2(t)$ were based on noisy observations of the end velocities as compared to the end state of the IPE. The prior probabilities over masses was an exponential, just as in [19]; this was implemented in the IPE as a prior over densities, but because both balls were always the same size, this is equivalent to a prior over mass.

When comparing our results to the model from [19] and human data [26], the IPE model-based inferences showed a very similar pattern despite the differences in the respective setup (cf. Fig. 1B-D).

[2] Sanborn et al. used a Bayesian Network model for modeling people's judgments as optimal statistical inference over a Newtonian physical model that incorporates sensory noise and intrinsic uncertainty about the physical properties of the objects being viewed.

[3] Though mass was not a basic attribute of ball objects, mass was defined as the density times the size and therefore could be trivially calculated from other attributes.

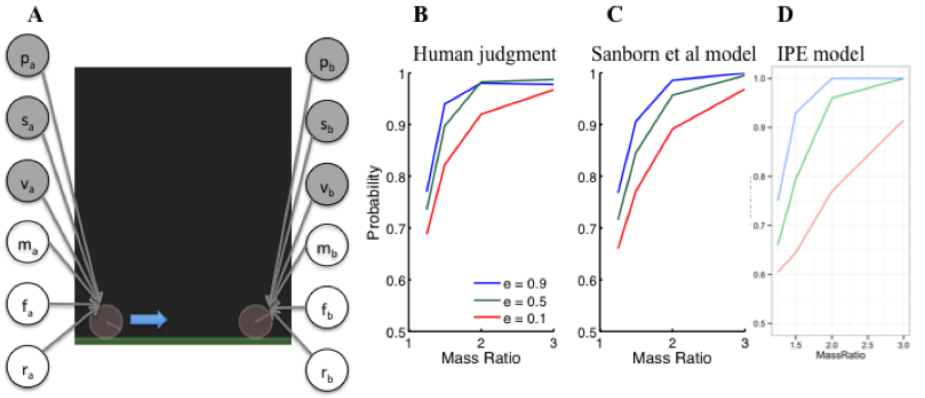


Fig. 1 (A) The IPE setup: Noisy information about the position, size, and velocity of each object (grey nodes) is observed, and mass (density), friction, and the coefficient of restitution (white nodes) are inferred based on how objects interact. (B-D) Probability of judging one object as heavier than another (y-axis) based on how much more massive that object is in reality (x-axis) after observing them collide. People (B) are biased by how elastic the collision is, which can also be found in (C), the model from [19], and the IPE model (D).

4 Experiment 2: Children's discovery of mass and momentum

While this suggests that the IPE can explain adults judgments of physics (and that a physics engine like Box2D can be used to computationally model these judgments), it cannot explain the infants' shortcomings in physical reasoning. This brings up another question: What would be the minimum amount of physical knowledge needed to explain how infants make physical judgments (and, thus, what would be the minimum amount of knowledge needed to re-implement infant-level physical reasoning in an artificial system using a physics engine)? Baillargeon [2] reports that infants at five to six months of age are surprised when a smaller object launches a reference object farther than a larger one. Using this as starting point, we presented the IPE model with two scenes in which a medium sized reference ball (B_{ref}) is hit by another ball of one of two different sizes (one larger (B_L) and one smaller ball (B_S); cf. Fig. 2A) and tasked the engine to predict in which scene children would believe the reference ball would travel farther. Additionally, we made the infant IPE assume that collision force was exclusively proportional to the size of the incoming ball by removing all latent attributes such as density, friction, and elasticity (i.e., effectively all objects were assumed to be of the same density). The IPE then was used to develop a joint probability over the ending positions of the reference balls in each of the two scenarios – P^L_{ref} and P^S_{ref} – and to determine the probability that the reference ball would move further when hit by the large ball than when colliding with the smaller ball.

Notwithstanding this very impoverished scenario, the infant IPE would correctly judge that the smaller ball should cause a lesser displacement of the reference ball than the bigger one. Moreover, it can be predicted that with a growing ratio of sizes between the launching balls the judgment would become more accurate (cf. Fig. 2B).

However, such a simplified model cannot perform the physical inferences that adults and even older infants are capable of and which would be desirable to re-create in an artificial system: After seeing videos of two collisions — one of a reference ball (B_{ref}) launching a large ball (B_L) and one of the same reference ball bouncing off of a small ball (B_S; cf. Fig. 3A) — adults will naturally assume that the smaller ball has higher mass than the bigger

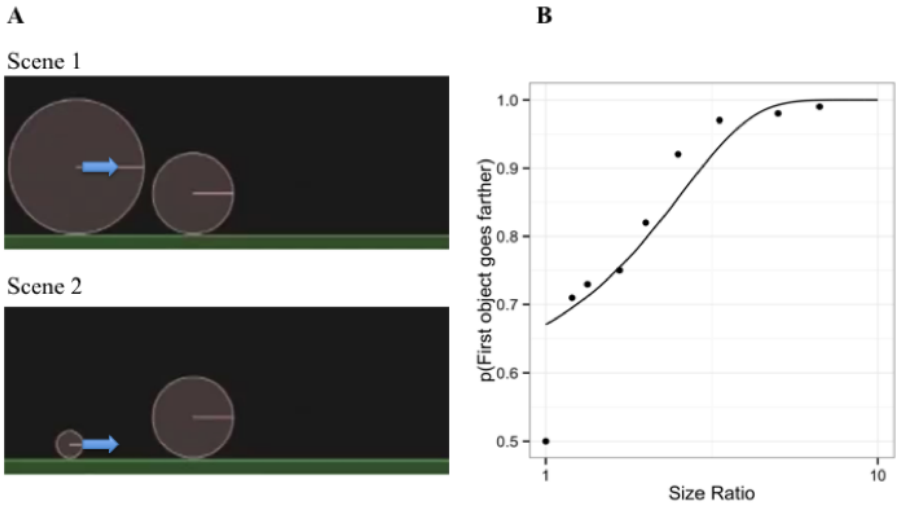


Fig. 2 (A) The two scenes shown to the infant IPE model as basis for its prediction (reference ball on the right). (B) The proportion of time the infant IPE model would judge the ball to travel farther in the first scene as a function of the ratio of the radii of the two launching balls.

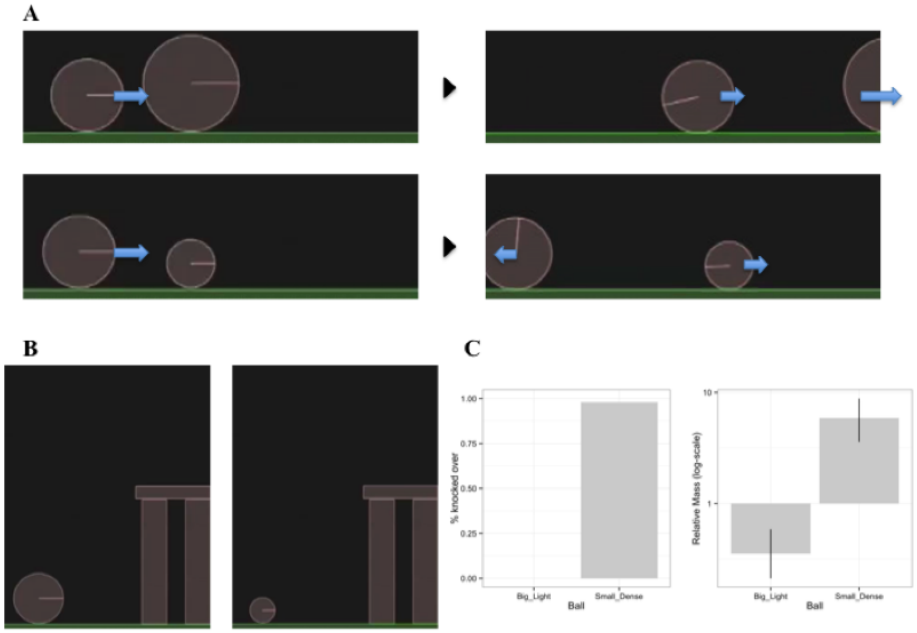


Fig. 3 (A) Two collisions are observed: a reference ball (left) striking a larger ball and sending it quickly to the right, and the same reference ball striking a smaller ball and bouncing off. (B) The IPE model is tasked to determine the probability that the tower would fall when being struck by the same balls. (C) Left graph: The IPE model infers that the larger, light ball is significantly less likely to knock the tower down than the the smaller, dense ball. Right graph: The larger ball is assumed to have much lower mass than the reference ball (1 on the y-axis), which in turn has much lower mass than the smaller ball.

one, i.e., the smaller ball must be made of a dense material and the larger ball of something lighter. It is this inference of mass as general latent property which can subsequently be applied in many other reasoning processes about the balls in new environments. This inference can be captured by the IPE using similar logic as in (2), but looking across each of the two scenes independently:

$$p(M|V_{ref}(0), V_{ref}(t), V(0), V(t)) \propto \int_{M_{ref}} p(V_{ref}(t), V(t)|V_{ref}(0), M_{ref}, V(0), M) * p(M_{ref}) * p(M) \tag{3}$$

We therefore propose a simple, novel experiment to closer examine infants' transition from a purely size-based conception of imparted collision forces to a mass or density-based understanding: On the one hand, adults are capable of transferring their knowledge of the latent mass property after seeing a collision event, and thus are able to judge which ball (the smaller or the bigger one) would be more likely to knock down a tower of blocks if it rolled into it (e.g., Fig. 3B). On the other hand, if young infants exclusively use size (and not a latent property such as mass or density) to judge the force imparted in collisions, then they should be unable to correctly make the same judgment in cases where the smaller ball is the more massive one. This should only change with the fundamental transition towards understanding that objects have a latent mass or density attribute: Once this happened, they should also be able to infer the differences between the balls and make the same judgment that the smaller, dense ball will be more likely to knock the tower down (as the IPE model can; Fig. 3C). In a more general developmental context, the ability to pass this predicted but undocumented stage should correlate with the ability to perform other tasks related to mass inference (cf., e.g., [16]).

5 Interpretation and discussion of the tentative results

This work presents a potential trajectory along which knowledge of mass within collisions might develop: Infants have an understanding of collision events, but develop an understanding of how momentum is transferred over their first year of life. First they infer that larger objects impart more momentum, then through experience with the world, learn about mass and develop a rich, adult-like understanding of collision dynamics. But alternatively, infants proto-object concepts might include an attribute that influences the momentum transferred in collisions, and it is only that their ability to infer this property from noisy observations improves over time.

Both hypotheses are consistent with existing research, and it is only through further studies of infants' physical understanding — how well they can differentiate collision events, or how well they can transfer knowledge of mass between events — that we can begin to understand how knowledge of our physical world develops into the useful, calibrated form that all adults have. The answer(s) to this question will be located somewhere on the spectrum spanned by the abstract positions underlying the two mentioned hypotheses: Structure learning similar to the notion studied by Ullman et al. [28], i.e., the discovery of entirely new concepts and corresponding governing rules on the one end, and a form of parameter setting and refinement over preexisting conceptualizations, i.e., filling in conceptual variables and fine-tuning values, on the other extreme. Also, it seems plausible to us that different physical notions and properties will most likely lie on distinct points along the described

spectrum, some necessitating the full mental generation of new concepts, others building more on preexisting knowledge and requiring fine tuning of parameters.

6 Conclusion: Contextualization and future work

Whilst the importance of this type of study for research in cognitive modeling and developmental psychology is unquestionable, it also offers appealing answers to long-standing conceptual and procedural questions about modeling and implementing qualitative physics as part of commonsense reasoning in cognitive AI: By investigating qualitative reasoning in everyday physics within a strictly computational paradigm and using tools and techniques which are already deeply rooted in computational modeling and computer science in general, results are naturally transferable to an AI context. This offers a new and more natural approach to automatizing everyday physical reasoning than the classical massively knowledge-driven approaches (cf., e.g., [13, 14]). Moreover, in our framework it is not only the results which are of interest (such as an answer to the question of what basic knowledge might be needed for humans to be able to successfully apply commonsense reasoning in physics), but also the methods applied for obtaining them are highly relevant as they also have existing direct counterparts in the computational domain. Thus, if continued research accumulates further evidence that IPEs constitute an accurate model of human physical reasoning, this also strengthens the claim of physics engines being a feasible way of re-implementing these forms of reasoning in cognitive systems.

Concerning future work, on the side of psychology and cognitive modeling there is plenty left to do after this almost exclusively exploratory study: As already hinted at in the previous section, data is needed on almost all aspects of the developmental trajectory, for instance addressing questions such as when infants actually infer that objects have mass and whether (and if so, when) this knowledge transfers to other events and settings. Also, for understanding this developmental process, in all likelihood broader and more detailed quantitative insights will be required (e.g., using the preferential looking paradigm as example, by not only analyzing looking times and inferring a higher surprise level from prolongated looking, but also trying to measure the actual degree of surprise). Finally, it would be very interesting to see whether a study transferring the "sticky mittens" experiment from [18] to our setting would show similar results. This could provide evidence for the importance of bodily experience in the discovery of mass and momentum as latent physical properties, conceptually establishing close connections to recent developments in AI such as the embodied cognition [1] and the embodied AI [8] paradigms. On the side of cognitive systems and AI research, it will be interesting to see how well models similar to the one we presented in this paper generalize and transfer across domains and application tasks, and how complex scenarios can become before significant differences in performance and behavior between human reasoners and IPE models start to appear.

Acknowledgements The authors want to thank Josh Tenenbaum for his continuous support and valuable feedback during all stages of the reported modeling experiment.

References

1. Anderson ML (2003) Embodied cognition: A field guide. Artificial Intelligence 149(1):91–130

2. Baillargeon R (1998) Advances in psychological science, Vol. 2: Biological and cognitive aspects, Psychology Press/Erlbaum (UK) Taylor & Francis, chap Infants' understanding of the physical world
3. Baillargeon R (2007) Blackwell Handbook of Childhood Cognitive Development, Blackwell Publishers, chap The Acquisition of Physical Knowledge in Infancy: A Summary in Eight Lessons
4. Battaglia PW, Hamrick JB, Tenenbaum JB (2013) Simulation as an engine of physical scene understanding. Proceedings of the National Academy of Sciences 110(45):18,327–18,332
5. Bobrow D (1984) Qualitative reasoning about physical systems: An introduction. Artificial Intelligence 24(1–3):1–5
6. Cassimatis N (2006) A Cognitive Substrate for Achieving Human-Level Intelligence. AI Magazine 27(2):45–56
7. Catto E (2009) Box2d physics engine. World Wide Web electronic publication URL http://box2d.org/
8. Chrisley R (2003) Embodied artificial intelligence. Artificial Intelligence 149(1):131–150
9. Craik K (1943) The Nature of Explanation. Cambridge University Press
10. Forbus K (2011) Qualitative modeling. Wiley Interdisciplinary Reviews: Cognitive Science 2:374–391
11. Gerstenberg T, Goodman N, Lagnado D, Tenenbaum J (2012) Noisy Newtons: Unifying process and dependency accounts of causal attributioin. In: Proceedings of the 34th Annual Conference of the Cognitive Science Society (CogSci 2012), pp 378–383
12. Goodman ND, Mansinghka VK, Roy DM, Bonawitz K, Tarlow D (2012) Church: a language for generative models. CoRR abs/1206.3255
13. Hayes PJ (1979) The naive physics manifesto. In: Michie D (ed) Expert Systems in the Micro-Electronic Age, Edinburgh University Press, pp 242–270
14. Hayes PJ (1990) The second naive physics manifesto. In: Weld DS, Kleer Jd (eds) Readings in Qualitative Reasoning About Physical Systems, Morgan Kaufmann Publishers Inc., pp 46–63
15. Hegarty M (2004) Mechanical reasoning by mental simulation. Trends in Cognitive Science 8(6):280–285
16. Houf P, Paulus M, Baillargeon R (2012) Infants use compression information to infer objects' weights: Examining cognition, exploration, and prospective action in a preferential-reaching task. Child Development 83(6):1978–1995
17. James W (1981) The Principles of Psychology. Harvard University Press, originally published in 1890.
18. Needham A, Barrett T, Peterman K (2002) A pick-me-up for infants' exploratory skills: Early simulated experiences reaching for objects using 'sticky mittens' enhances young infants' object exploration skills. Infant Behavior and Development 25(3):279–295
19. Sanborn A, Mansinghka V, Griffiths T (2013) Reconciling intuitive physics and Newtonian mechanics for colliding objects. Psychological Review 120(2):411–437
20. Smith K, Vul E (2013) Sources of uncertainty in intuitive physics. Topics in Cognitive Science 5(1):185–199
21. Smith K, Dechter E, Tenenbaum J, Vul E (2013) Physical predictions over time. In: Proceedings of the 35th Annual Conference of the Cognitive Science Society (CogSci 2012), pp 1342–1347
22. Smith K, Walker D, Besold TR, Ullman TD (2014) Massive insights in infancy: Discovering mass & momentum. In: Collection of project abstracts from the Brains, Minds and

Machines Summer Course at MBL, Woods Hole, CBMM Memo, [pending submission to arXiv]
23. Spelke E, Breinlinger K, Macomber J, Jacobson K (1992) Origins of knowledge. Psychological Review 99(4):605–632
24. Talmy L (1988) Force dynamics in language and cognition. Cognitive Science 12(1):49–100
25. Tenenbaum J, Kemp C, Griffiths T, Goodman N (2011) How to grow a mind: Statistics, structure, and abstraction. Science 331(6022):1279–1285
26. Todd J, Warren W (1982) Visual perception of relative mass in dynamic events. Perception 11:325–335
27. Tomasello M (1999) The Cultural Origins of Human Cognition. Harvard University Press
28. Ullman T, Goodman N, Tenenbaum J (2012) Theory learning as stochastic search in the language of thought. Cognitive Development 27(4):455–480

A Theory for the Evolution from Subjects' Opinions to Common Sense: a Geometric Approach

Ray-Ming Chen

Abstract In the field of artificial intelligence, common sense or common knowledge plays a vital role in understanding human reasoning. In this paper, I describe a theory to capture the process of evolution from different opinions to common sense. The theory utilizes a geometrical approach to model the evolution or formation from subjects' opinions to common sense. Each subject's opinion is represented by a value between 0 and 1. A mechanism based on the concept of circumcenter is then devised to settle the convergence of subjects' opinions. The main stages of evolution is described by generations. Each generation is a result of local common sense formed by the previous generation. This result will serve as the initial opinions for the next generation and then evolves further. The process continues until the end common sense is reached. This mechanism ensures that the end common sense is reached, given any arbitrary set of subject' opinions and statements.

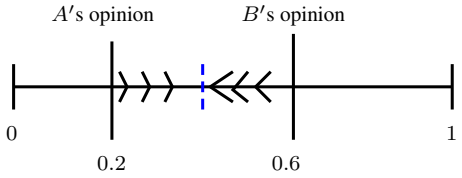
1 Introduction

Common sense is a product of an evolution of different subjects' opinions or thoughts. It evolves over time. Understanding the route of its evolution helps one comprehend the history of formation of common sense. Common sense can be seen as a collective agreement between different entities. Conflicts between different subjects always occur in forming common sense. The internal tendency for the formation relies on compromises. To model the behavior of compromising, I assume the compromise is based on the equality of satisfaction between different entities. When there are fewer subjects involved, this assumption may need some alterations, for example adding some weights for the satisfaction of the subjects. As there are more subjects involved, the weights for the satisfaction could be average out and thus the assumption is taken as a reasonable one.

In this paper, I will assume the true value regarding each statement is between 0 and 1. For example, a subject A thinks the degree of truth that a statement "There is a ghost." is 0.2 and a subject B thinks it is 0.6. A will then try to influence or communicate with B by presenting a lot of personal experiences to him and vice versa. As time goes by, both

Department of Computer Science
Friedrich-Alexander-Universität Erlangen-Nürnberg
Martenstraße 3, 91058 Erlangen, Germany
E-mail: ray.chen@fau.de

sides converge their opinions in the middle 0.4. The process could be shown in the following graph:



As more and more subjects and statements are getting involved, a compromise is best described by a geometrical circumcenter upon which the majority of the entities (vertexes) agree. In addition, I will also assume that all the subjects are trying the shortest route to reach common sense. The route that more subjects could agree upon is favored over the one with less subjects. In Section 2, a geometrical approach is introduced to characterize common sense based on the idea of compromises between subjects. The mechanism how the subjects evolve their opinions is also explored. During the proceeding of formation of common sense, there exist some routes that the subjects to choose. In Section 3, a complete computational procedure of the process from subjects' opinions to common sense is described. The first part of this section gives a theoretical framework and algorithm for the computation and the second part gives a way to implement this algorithm. Finally, in Section 4, I briefly summarize the conclusion and discuss some future research approaches and works.

2 Common Sense

2.1 Preliminary

Each statement is a description. Each opinion is a truth value assigned to a statement. A subject is an observer who is capable of making an opinion. Assume the domain of a true value is the real interval $[0, 1]$. Let $S = \{s_1, s_2, ..., s_m\}$ be a set of m subjects. Let $A = \{a_1, a_2, ..., a_n\}$ be a set of n statements. Let $v : S \times A \to [0, 1]$ be the true value assignment in which each $v(s, a)$ denotes the opinion of the statement a assigned by the subject s. Let $OP_A : S \to [0, 1]^{|A|}$ be defined by $OP_A(s) := (v(s, a_1), v(s, a_2), ..., v(s, a_n))$. Each $OP_A(s)$ will be called an assertion and the set $AS = \{OP_A(s) : s \in S\}$ is called the assertion domain for S. For any points $\mathbf{p}, \mathbf{q} \in \mathbb{R}^n$, I will use either $dist(\mathbf{p}, \mathbf{q})$ or $||\mathbf{p} - \mathbf{q}||$ to denote its Euclidean distance. For any set Z, let $\mathcal{P}(Z)$ denote its power set.

Definition 1. $\mathbf{c} \in [0, 1]^{|A|}$ is global common sense of A for S if and only if for all $s_i, s_j \in S$, $dist(OP_A(s_i), \mathbf{c}) = dist(OP_A(s_j), \mathbf{c})$. It is denoted by $Com(\{OP_A(s) : s \in S\}) = \mathbf{c}$.

Definition 2. $\mathbf{c} \in [0, 1]^{|A|}$ is local common sense of A for S if and only if there exists $S' \subsetneq S$ such that for all $s_i, s_j \in S'$, $dist(OP_A(s_i), \mathbf{c}) = dist(OP_A(s_j), \mathbf{c})$. It is denoted by $Com(\{OP_A(s) : s \in S'\}) = \mathbf{c}$.

Geometrically speaking, common sense of A is a well-defined circumcenter (i.e., the circumcenter lies in the assertion domains) of a polygon with each vertex being an element of AS. Here "common" is not in the sense of "overlapped". It is an opinion that all sides could agree upon. No one would easily yield his opinions to accept others. A compromise

of all the related subjects' opinions occurs. An optimal compromise is the global common sense of A for S. However, the chance of obtaining a global common sense becomes almost impossible as more and more subjects are getting involved. In this cases, the compromise is not straightforward. Indeed it must go through a process of compromising. Each process of compromising obtains some local common opinions and these opinions will then further compromise to obtain another set of local common opinions. After a finite compromising process, final common sense will be achieved. This mechanism ensures the evolution of the subjects' opinions. This mechanism simplifies the modeling of the real world.

Example 1. Let $S = \{s_1, s_2, s_3\}$ and $A = \{$"Money is everything."$\}$ and $OP_A(s_1) = 0.2, OP_A(s_2) = 0.95$ and $OP_A(s_3) = 0.4$. In this case, there are are three ways to form common sense: either s_1 and s_2 form local common sense first and then the local one combing with the opinion of s_3 reaches final common sense or s_2 and s_3 form local common sense first and then the local one combing with the opinion of s_1 reaches final common sense or s_1 and s_3 form local common sense first and then the local one combing with the opinion of s_2 reaches final common sense. The criterion to decide which way is the optimal one is described in the next section.

2.2 Choices of Evolution

Suppose there exists a set of subjects $S = \{s_1, s_2, ..., s_n\}$. Suppose the balls containing the maximum number of subjects' opinions are the same. Suppose there are m routes for the evolution. Let $\mathbf{c}_j$ denote the final common sense under the route j. Then the optimal route for the evolution is $\underset{j}{argmin} \sum^{s\in S} ||OP_A(s) - \mathbf{c}_j||$.

Example 2. Suppose the set of all the subjects $S = \{$Judy, John, Bruce$\}$ and the set of statements $A = \{$"A dog is a pet.", "No chicken can fly."$\}$ and OP_A is defined as follows:

Subjects	a_1 = "A dog is a pet."	a_2 = "No chicken can fly."
Judy	0.2	0.6
John	0.8	0.3
Bruce	0.5	0.5

Table 1 S's Opinions Regarding Statements A

Though $OP_A(\text{Judy}), OP_A(\text{John})$ and $OP_A(\text{Bruce})$ are not co-linear, the circumcenter (i.e., the common sense) for the triangle (i.e., assertions) would be $(-\frac{1}{20}, -\frac{13}{20})$ and that is not within the assertion domain $[0, 1]$. Hence an immediate common sense is not reached and three routes of evolution take place: $Com(Com(OP_A(\text{Judy}), OP_A(\text{John})), OP_A(\text{Bruce}))$, or $Com(Com(OP_A(\text{Judy}), OP_A(\text{Bruce})), OP_A(\text{John}))$, or the last route for the evolution $Com(Com(OP_A(\text{John}), OP_A(\text{Bruce})), OP_A(\text{Judy}))$. The computation goes as follows:

- route one (or $R1$): $\frac{\frac{(0.2,0.6)+(0.8,0.3)}{2}+(0.5,0.5)}{2} = (0.5, 0.475)$;
- route two (or $R2$):: $\frac{\frac{(0.2,0.6)+(0.5,0.5)}{2}+(0.8,0.3)}{2} = (0.575, 0.425)$;
- route three (or $R3$):: $\frac{\frac{(0.8,0.3)+(0.5,0.5)}{2}+(0.2,0.6)}{2} = (0.4125, 0.5)$.

This could be visualized in the following diagrams:

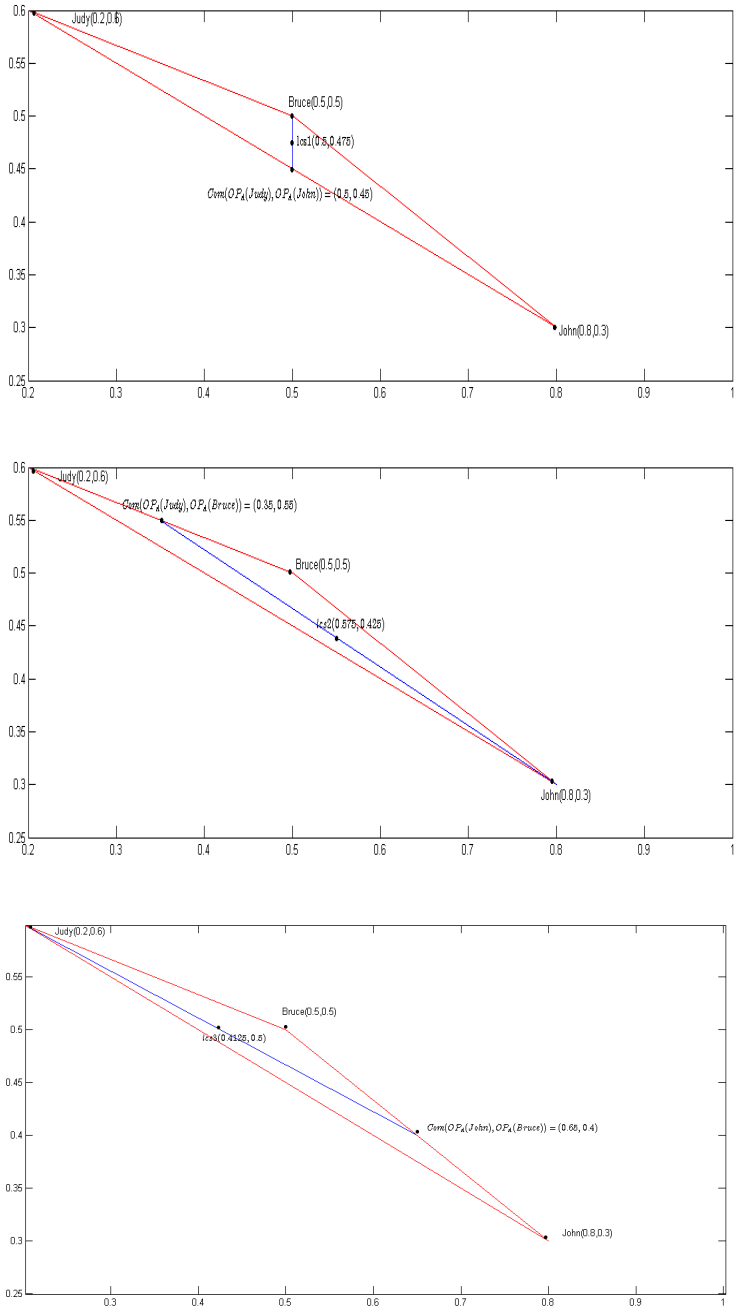


Fig. 1 Potential Routes of Evolution for Common Sense

The choice of the optimal route of evolution is decided by the following table:

	Judy$(0.2, 0.6)$	John$(0.8, 0.3)$	Bruce$(0.5, 0.5)$	Sum
$R1, (0.5, 0.475)$	0.3250	0.3473	0.025	0.6973
$R2, (0.575, 0.425)$	0.4138	0.2574	0.1061	0.7773
$R3, (0.4125, 0.5)$	0.2349	0.4361	0.0875	0.7585

Table 2 Euclidean Distance: Optimal Route of Evolution

Since the least sum for the routes of the evolution is $R1$, the optimal one is $R1$, i.e., Judy and John form their local common sense first, and fuse their result with Bruce's opinion. The optimal common sense is $(0.5, 0.475)$.

Example 3. Let us continue the above example. Suppose assertion $OP_A(\text{Judy}) = (0.5, 0.7)$, assertion $OP_A(\text{Bruce}) = (0.3, 0.5)$ and assertion $OP_A(\text{John}) = (0.8, 0.1)$. The global common sense $Com(OP_A(\text{John}), OP_A(\text{Bruce}), OP_A(\text{John})) = (\frac{37}{60}, \frac{23}{60})$, which is exactly the circumcenter of the triangle, is reached.

Claim 1. If **c** is global common sense for a set S, then **c** is also global common sense for any subset of S.

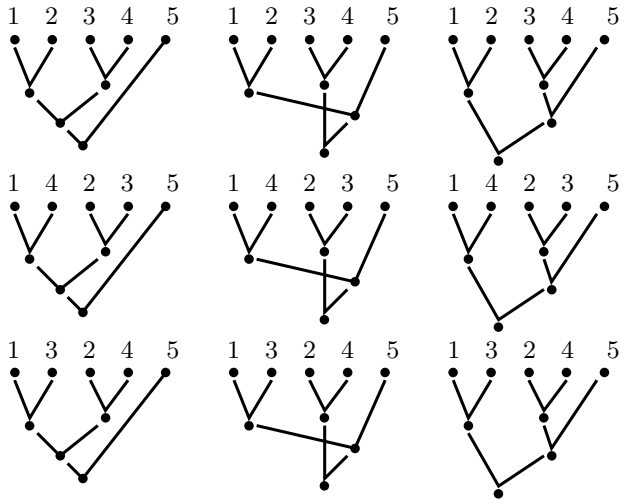
Example 4. Let us continue Table 2. Suppose S now includes an extra subject Robert whose assertion is $OP_A(\text{Robert}) = (0.3, 0.4)$. Since there exists no global common sense for the set $\{\text{Judy}, \text{Bruce}, \text{John}\}$, we only need to consider the following cases. Let
$CS1 \equiv Com(Com(OP_A(\text{Judy}), OP_A(\text{John}), OP_A(\text{Bruce})), \text{Robert})$,
$CS2 \equiv Com(Com(OP_A(\text{Judy}), OP_A(\text{John}), OP_A(\text{Robert})), \text{Bruce})$,
$CS3 \equiv Com(Com(OP_A(\text{Judy}), OP_A(\text{Robert}), OP_A(\text{Bruce})), \text{John})$,
$CS4 \equiv Com(Com(OP_A(\text{Robert}), OP_A(\text{John}), OP_A(\text{Bruce})), \text{Judy})$. Then one has the following computation:

Table 3 Potential Evolution: $S = \{\text{Judy}, \text{Bruce}, \text{John}, \text{Robert}\}$

Potential Common Sense	circumcenter
$CS1$	$((-\frac{1}{20}, -\frac{13}{20}), (0.3, 0.4))$
$CS2$	$((\frac{37}{60}, \frac{41}{60}), (0.5, 0.5))$
$CS3$	$((\frac{7}{20}, \frac{11}{20}), (0.8, 0.3))$
$CS4$	$((\frac{73}{140}, \frac{29}{140}), (0.2, 0.6))$

Since the circumcenter of the first combination is out of the assertion domain $[0, 1]$, it will never be one option in the evolution. There are only three potential evolution routes left. The choice of the optimal route for the evolution is the sum of the distance from initial points to the end point for each subject.

Example 5. Suppose the sets of subjects $S = \{s_1, s_2, ..., s_5\}$. Suppose the set of its opinions $\{OP_A(s_1), OP_A(s_2), OP_A(s_3), OP_A(s_4), OP_A(s_5)\}$ are co-linear. Then there are at most 45 possible routes of evolution. The following graph shows 9 of them.



One then computes the sum of distance between each subject's initial opinion and the end common sense (as we did in Table 2) and then chooses the least one among the 45 results. For a complete picture of how to incorporate the choice of an optimal route of evolution into the whole theory, one should refer to the next section.

3 Computational Procedures

Assumption . To begin with, let us make some assumptions.

1. Each opinion for a subject lies between $[0, 1]$;
2. The distance between any two opinions is defined by Euclidean distance.
3. The common sense is the end result of a process finding the circumcenters (or local common sense) of their opinions.
4. If there are more than two local common sense for a set of opinions, then the circumcenter with least distance between it and other opinions is the chosen local common sense.
5. If there are more than two routes of evolution, then the optimal one is the one with least length.

These assumptions will be implicitly applied in the following sections.

3.1 Theoretical Algorithm

Define the initial generation of opinions, $GO^0 = AS$. Suppose the initial assertion domain $AS^0 = AS$. Let $||\mathbf{a}||$ denote the Euclidean distance of a vector $\mathbf{a}$. Let $H \subseteq AS^0$ be arbitrary. For any vector $\mathbf{c} \in \mathbb{R}^n$ and any set $H \subseteq \mathbb{R}^n$, define the distance between $\mathbf{c}$ and H by $dis(\mathbf{c}, H) := \sum^{\mathbf{h}\in H} ||\mathbf{h} - \mathbf{c}||$. For any set Z, let $\mathcal{P}^n(Z)$ denote the set $\{U \in \mathcal{P}(Z) : |U| = n\}$.

Definition 3. For any $\mathbf{c} \in [0, 1]^{|A|}, r \in \mathbb{R}^+$, define $Ball^0(\mathbf{c}, r) := \{\mathbf{k} \in AS^0 : ||\mathbf{k} - \mathbf{c}|| = r\}$. We call $Ball^0(\mathbf{c}, r)$ an initial Euclidean ball. Define $\mathcal{BS}^0_H := \{Ball^0(\mathbf{c}, r) : \mathbf{c} \in [0, 1]^{|A|}, r \in \mathbb{R}^+, H = Ball^0(\mathbf{c}, r)\}$. We call $\mathcal{BS}^0_H$ an initial $H-$induced ball space and each element in it, an initial $H-$induced ball.

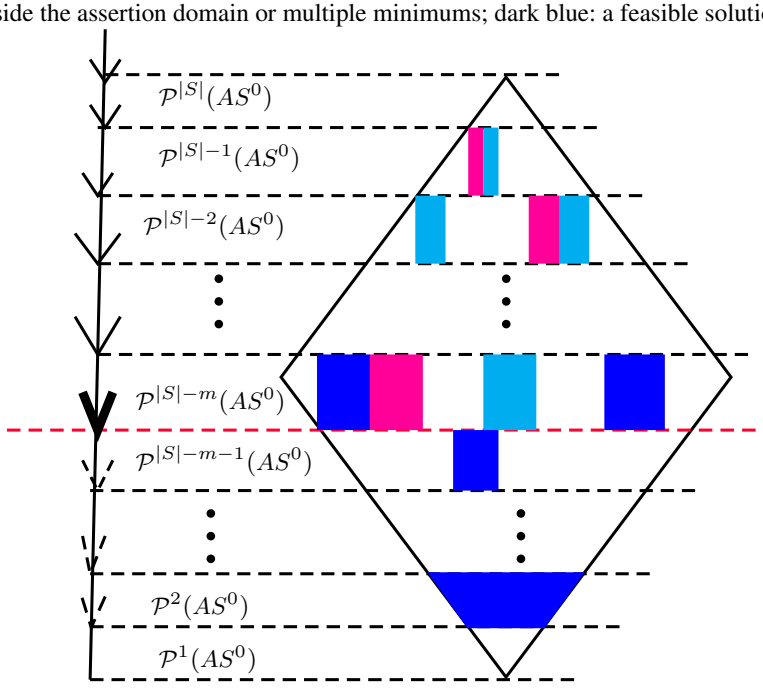
Definition 4. Define an ordering $\leq_H$ over $\mathcal{BS}_H^0$ by $Ball^0(\mathbf{c}, r) \leq_H Ball^0(\mathbf{c}', r')$ if and only if $Ball^0(\mathbf{c}, r), Ball^0(\mathbf{c}', r') \in \mathcal{BS}_H^0$ and $r \leq r'$. We say $Ball^0(\mathbf{c}, r)$ is a $\leq_H$ least ball if and only if $Ball^0(\mathbf{c}, r) \in \mathcal{BS}_H^0$ and $Ball^0(\mathbf{c}, r) \leq_H Ball^0(\mathbf{c}', s)$ for all $Ball^0(\mathbf{c}', s) \in \mathcal{BS}_H^0$.

Definition 5. Define $\mathcal{LBS}_H^0 := \{Ball^0(\mathbf{c}, r) \in \mathcal{BS}_H^0 : Ball^0(\mathbf{c}, r)$ is a $\leq_H$ least ball.$\}$. $\mathcal{LBS}_H^0$ is called an initial $H-$induced least ball space. $Ball^0(\mathbf{c}, r)$ is a potential solution for H if and only if $\mathcal{LBS}_H^0 = \{Ball^0(\mathbf{c}, r)\}$. $Ball^0(\mathbf{c}, r)$ is **a feasible solution** for H if and only if $\mathcal{LBS}_H^0 = \{Ball^0(\mathbf{c}, r)\}$ and $\mathbf{c} \in [0, 1]^{|A|}$ and we say H has a feasible solution. Define $\mathcal{LBS}^0(H) := Ball^0(\mathbf{c}, r)$, if $\mathcal{LBS}_H^0 = \{Ball^0(\mathbf{c}, r)\}$ and undefined, otherwise. Define the feasible-solution function $\mathcal{FS}^0(AS^0) := \{\mathcal{LBS}^0(H) : H \in \mathcal{P}(AS^0)\}$.

Definition 6. Define an initial characterizing-ball function $CB^0 : \mathcal{P}(AS^0) \to [0, 1]^{|A|} \times \mathbb{R}^+$ by $CB^0(H) := \begin{cases} (\mathbf{c}, r) & \text{, if } \mathcal{LSB}^0(H) = Ball^0(\mathbf{c}, r); \\ \text{undefined} & \text{, otherwise.} \end{cases}$

It is characterized by the optimization problem: $CB^0(H) = \underset{(\mathbf{c},r)\in[0,1]^{|A|}\times\mathbb{R}^+}{\text{argmin}} \{ \sum^{\mathbf{a}\in H} ||\mathbf{a} - \mathbf{c}|| : ||\mathbf{a}_i - \mathbf{c}|| = ||\mathbf{a}_j - \mathbf{c}||$ for all $\mathbf{a}_i, \mathbf{a}_j \in H\}$ if and only if it is defined. Define the initial clustered opinions, $CO^0 = \underset{B}{argmax}\{|B| : B \in \mathcal{FS}^0(AS^0)\}$.

The first part of our theoretical algorithms could be summarized in the following diagram (an example): (arrow: the searching direction; colored box: the solution status of the set; white box: the feasible domain is empty; pink box: the feasible domain is not empty, but the minimum for the objective function does not exist; light blue: the minimum exists, but outside the assertion domain or multiple minimums; dark blue: a feasible solution.)



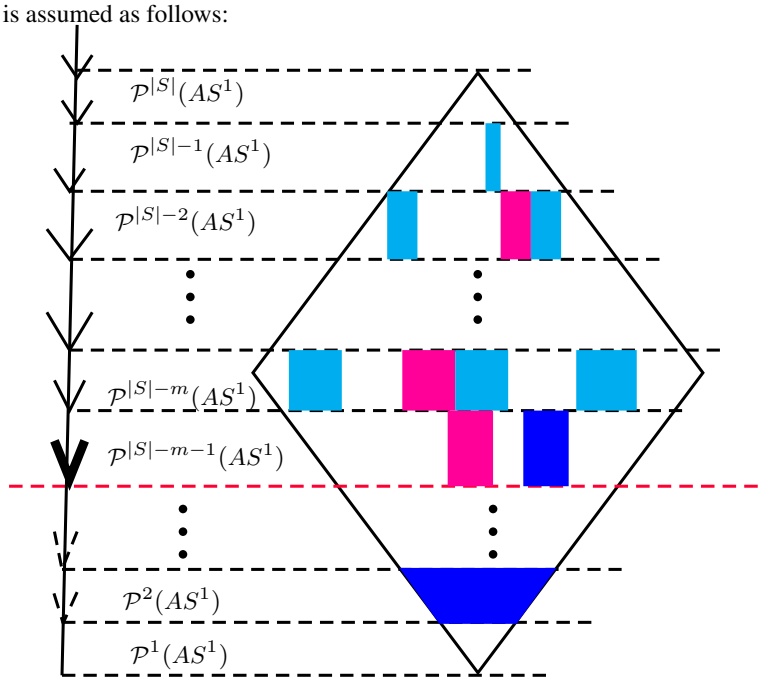
This searching algorithm (or algorithm 0) for local common sense is represented by the following procedures:

1. Partition $\mathcal{P}(AS^0)$ into $\{\mathcal{P}^n(AS^0) : n \in \{1, 2, ..., |S|\}\}$.
2. Execute the loop:
   for $k = |S|$ to 1;
   Compute $\mathcal{LBS}_k^0 = \{\mathcal{LBS}^0(H) : H \in \mathcal{P}^k(AS^0)\}$;
   Compute $CB_k^0 = \{CB^0(H) : H \in \mathcal{P}^k(AS^0)\}$;
   if $CB_k^0 \neq 0$, then $k = k + 1$;
   else, exit;
   end.
3. Compute CO^0.

This algorithm shows how to form a set (or sets) of clustered opinions in which the most subjects agree upon some local common sense (or circumcenters). Because of the keyword " the most ", the search starts from all the subjects to one subject (indeed later on we have shown the lower bound is 2 subjects). In the above example, we know CO^0 is a subset of $\mathcal{P}^{|S|-m}(AS^0)$, since there are two dark blue boxes for the feasible solutions. For demonstrative purpose, here I list all the searching process up to $\mathcal{P}^2(AS^0)$- in this case, the search should stop at the $\mathcal{P}^{|S|-m}(AS^0)$ as the most subjects agree at this stage.

Definition 7. Define $AS^1 = AS^0 - CO^0$.

Now replace all the superscript 0 with 1 in algorithm 0 and this becomes algorithm 1. A visualized diagram is given as follows. Given an element in CO^0, the process for next stage is assumed as follows:

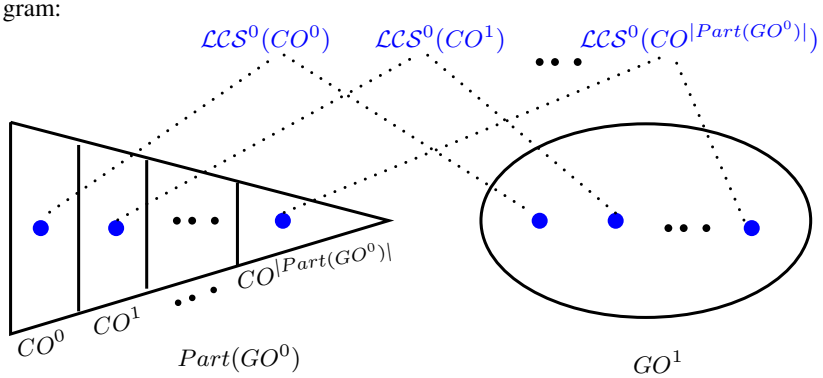


Then one replace all the superscript 1 with 2 in the algorithm 1 and this becomes algorithm 2. This process of computation goes on inductively until GO^0 is fully partitioned into a set of clustered opinions. Suppose the partitioned set of GO^0 is $Part(GO^0)$.

Definition 8. For each $CO^n \in Part(GO^0)$, define $LCS^0(CO^n) = \{\mathbf{c} \in [0,1]^{|A|} : Ball^0(\mathbf{c}, r) = CO^n\}$. We call $LCS^0(CO^n)$ local common sense of the clustered opinions CO^n.

Definition 9. Define the first generation of opinions, $GO^1 = \{LCS^0(CO^n) : CO^n \in Part(GO^0)\}$.

The second part of our theoretical algorithm could be summarized in the following diagram:



The initial algorithm (or algorithm 0) for the second part goes as follows:

1. Collect all the clustered opinions $Part(GO^0)$;
2. Collect all the local circumcenters to form GO^1;
3. Reset $AS^0 = GO^1$;
4. Rerun the searching algorithms in the first part.

Then one replaces 0 with 1 and 1 with 2 in algorithm 0 and this becomes algorithm 1. This process of computation goes on inductively until where the end common sense (i.e., a point) is reached. The final part of our theoretical algorithms is described in the following.

Definition 10. We say $GO^0, GO^1, GO^2, ..., GO^v$ is a route iff it is formed sequentially through the algorithms of the three parts with $|GO^v| = 1$. Let $Route(S)$ denote the set of all the routes for S. Let $End(r)$ denote the end common sense via the route r.

Definition 11. Define the length over $Route(S)$ by $len : Route(S) \to \mathbb{R}^+$ by $len(r) := \sum_{s \in S} ||OP_A(s) - End(r)||$.

The last part of our theoretical algorithms goes as follows:

1. Collect all the routes: $r_1, r_2, ..., r_v$.
2. Compute the set of all end common sense via each route: $\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_v$.
3. Compute the length of all the routes: $\{len(r) : r \in Route(S)\}$.
4. Choose the optimal route and optimal end common sense via $\underset{r}{argmin}\{len(r) : r \in Route(S)\}$.

3.2 Characterization and Implementation

This section aims at the simplification of the last one.

Definition 12. $H = \{\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_{n+1}\} \subseteq \mathbb{R}^n$ is a second-order n linearly independent set iff $\{\mathbf{a}_1 - \mathbf{a}_2, \mathbf{a}_2 - \mathbf{a}_3, ..., \mathbf{a}_n - \mathbf{a}_{n+1}\}$ is a linearly independent set. We say H is a second-order linearly independent set iff H is a second-order n linearly independent set for some $n \in \mathbb{N}$.

Lemma 1. If $H = \{\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_{n+1}\} \subseteq \mathbb{R}^n$ is a second-order n linearly independent set, then H has a potential solution.

Proof For $\mathbf{c}$ to be a feasible solution for H, it must be the circumcenter of H, i.e., $||\mathbf{a}_i - \mathbf{c}|| = ||\mathbf{a}_j - \mathbf{c}||$ for all $i, j \in \{1, 2, ..., n+1\}$, i.e., $\sum_{k=1}^{n} (a_{ik} - c_k)^2 = \sum_{k=1}^{n} (a_{jk} - c_k)^2$, i.e., $\sum_{k=1}^{n} (a_{ik} - a_{jk}) \cdot c_k = \frac{\sum_{k=1}^{n} a_{ik}^2 - \sum_{k=1}^{n} a_{jk}^2}{2}$. Hence $\mathbf{c}$ uniquely exits if H is a second-order n linearly independent set.

Define a rank function $Rank : \mathcal{P}(AS) \to \mathbb{N}$ by $Rank(K) := \underset{|Y|}{argmax}\{Y \subseteq K : Y \text{ is a second-order linearly indepdent set.}\}$. Let $SLI = \{H \subseteq AS : rank(H) = |H| = |A|\}$ denote the set of all the second-order $|A|$ linearly independent set.

Claim 2. Each subset of AS has at most rank $|A|$.

Proof Since the dimension is $|A|$, the maximum number of linearly independent set is capped at $|A|$.

Definition 13. For any set $K \in \mathcal{P}(AS)$, define $Core(K) = \{H \in \mathcal{P}(AS) : K \supset H, rank(H) = |H| = |A|\}$.

Claim 3. $Core$ is a monotonically increasing function and if $rank(K) < |A|$, then $Core(K) = \emptyset$.

Claim 4. For all $H \in SLI$, $Core(H) = \{H\}$.

Claim 5. $SLI = \{K \in \mathcal{P}^{|A|+1}(AS) : K \text{ is a second-order } |A| \text{ linearly independent set.}\}$.

Theorem 1. If $Core(K) \neq \emptyset$, then for all $H \in Core(K)[CB(H) = CB(K)]$, where CB is any arbitrary characterizing-ball function.

Proof Since $||\mathbf{a}_i - \mathbf{c}|| = ||\mathbf{a}_j - \mathbf{c}||$ for all $\mathbf{a}_i, \mathbf{a}_j \in K$ implies $||\mathbf{a}_i - \mathbf{c}|| = ||\mathbf{a}_j - \mathbf{c}||$ for all $\mathbf{a}_i, \mathbf{a}_j \in H$, by the fact that $rank(H) = |A|$, the solution $\mathbf{c}$ uniquely exists.

The following theorem ensures our mechanism always produces end common sense.

Theorem 2. Any two opinion assertions have a compromise point regardless of the size of A.

Proof For any arbitrary two arbitrary opinion assertions in AS, the midpoint is the compromise point of these two.

By Claim 2, we know AS could be partitioned into two parts: (Part 1) sets with rank equals to $|A|$ and (Part 2) sets with rank strictly less than $|A|$. Unlike the theoretical one, this implementing algorithm only has to deal with optimization problem over the set SLI to derive the valid circumcenters. Then one uses the derived circumcenter to test other opinions and collects all the valid ones for each circumcenter. This approach searches the local common sense from bottom (in Part 1) to top, while the theoretical one, top to bottom. Most of the clause of this implementation algorithm is the same as the theoretical one. In the following, I briefly summarize the implementing algorithms and leave the details for the readers to fill in.

1. Compute SLI via the right-hand side of the equation in Claim 5.
2. Compute the set of all the feasible solutions for SLI.
3. Compute the maximum covering via the derived circumcenters.
4. Compute the feasible solutions for Part 2.
5. Pick up the maximum covering and collect its center.
6. Deduct the maximum one from the original set.
7. Repeat the whole process until AS^0 is partitioned.
8. Collect all the centers of the clusters and make it as the initial value for the next generation.
9. Repeating the whole computation until the common sense is reached.

4 Conclusion and Future Work

From subjects' opinions to common sense is a process of compromising. Usually local common sense will be reached first and fuse with other local common sense to form new local common sense until the end common sense is reached. The algorithms we use ensures such limit point always exists. For future research, one could lift or alter the assumptions in this paper. One could also study the mechanisms how subject' opinions are strengthened into customs or even laws. As for the non-numerical data for A, the mechanisms are far more complicated and task-dependent. At least there are three approaches to handle this. Firstly, one could convert non-numerical data into numerical ones with a risk that such conversion could distort the original data completely. Hence, finding a suitable conversion needs some expertise and experience. Secondly, one keeps the non-numerical data and chooses a suitable distance function to measure the difference of any two entries of non-numerical data-in order to apply quantitative analysis upon it. Thirdly, one could simply apply qualitative analysis to handle the data.

Since we study the evolution of a set of statements A, one could further analyze the relation, interaction or distance between different statements during the process of evolution. By carefully choosing the set of A, one could also obtain some interesting results or conclusions. In the optimization part, we assume that when there are multiple feasible solutions for the problem, then the subjects will not form local common sense (feasible solution) based on this. This part could also be further settled if one introduces some criterion to pick up the optimal one. Finally, one could also compare this mechanism with $k-$means to further study the possibility to applying this mechanism on some research that use $k-$means as a feature selection method.

References

1. A. Jøsang, Artificial Reasoning with Subjective Logic, Proceedings of the Second Australian Workshop on Commonsense Reasoning, Perth 1997.
2. Mueller, Erik T., Commonsense Reasoning, San Francisco: Morgan Kaufmann, 2006.
3. Mayr, Ernst, The Objects of Selection, Proceedings of the National Academy of Sciences of the United States of America, 1997.
4. Francis Heylighen and Klaas Chielens, Evolution of Culture, Memetics, Encyclopedia of Complexity and Systems Science, 2009.