Second Annual Conference on Advances in Cognitive Systems Poster Collection (2013) 1-18

The Importance of Two Cognitive Mechanisms in Analyzing Counterfactuals: An Implementation-Oriented Explication[1]

Ahmed M. H. Abdel-Fattah AHABDELFATTA@UNI-OSNABRUECK.DE
Ulf Krumnack KRUMNACK@UNI-OSNABRUECK.DE
Kai-Uwe Kühnberger KKUEHNBE@UNI-OSNABRUECK.DE
Institute of Cognitive Science, University of Osnabrück, Albrechtstr. 28, Osnabrück, Germany.

Abstract

We aim at covering three facets of the problem of analyzing counterfactual conditionals: emphasizing that the problem is crucial for cognitive systems to consider, investigating the cognitive mechanisms responsible for reasonable analyses of counterfactuals, and proposing how to computationally contribute to solving the problem by cognitive systems. Aiming at an implementation-oriented explication, the challenges that an artificial system may encounter in computationally solving this problem are discussed. The utilization of two particular computationally-plausible cognitive mechanisms is shown to be helpful in overcoming these challenges: we argue that the operational utilization of analogical mapping and conceptual blending is possible, and leads to reasonable analyses of counterfactual conditionals in artificial cognitive systems.

1. Cognitive Mechanisms and Essences of '*Smartness*'

The investigation of possible ways, in which the intelligence capacities of cognitive beings could be (philosophically, biologically, or physically) functioning, has always attracted the interest of numerous scholars over many centuries, and among various disciplines. But the scientific methodologies, developed for treating the related issues, have remarkably changed since the time the scholars started to focus on (*i*) the essences of the cognitive abilities that make humans smart, and (*ii*) how to artificially mimic them using formal, algorithmic, or computational processes. Alan Turing's information processing machine (a.k.a. the "Turing Machine") (Turing, 1936) may be considered a relatively recent breakthrough in this regard. Also, Newell and Simon's contributions laid out many conceptions that became standard in AI and aimed from their beginnings on implementing systems capable of 'general intelligent action' that indicates "the same scope of intelligence as we see in human action" (Newell & Simon, 1976, p. 116).

Over the past couple of decades, cognitive scientists have started applying their knowledge of technical systems to the study of the mind and behavior. There has always been

1. This paper is an extended version of (Abdel-Fattah, Krumnack, & Kühnberger, 2013).

a growing need to identify the various benchmark 'aspects' of what makes humans more intelligent than other cognitive beings. Proposing methods (or entire systems) that can abstractly or computationally model such aspects is of no less importance. Existing AI systems can already reason, plan, and perform actions, but their behavior may not be viewed as originally motivated by essential cognitive abilities that reflect one aspect of intelligence or another. Many of these systems neither focus on integrating human-comparable competencies nor on applying such competencies to a wider range of directions, but are usually designed to rather solve a specific task, and fail not only in solving another task but also in compatibly parsing that other task's input. For example, IBM's Watson and DeepBlue systems (Ferrucci et al., 2010; Hsu, 2002) can neither deal with each other's main functionalities nor even parse each other's input. Their impressive success obscures that these systems clearly lack 'general intelligent action' as suggested by (Newell & Simon, 1976).

Motivated by these ideas, we aim in this article at achieving three goals. We first attract the reader's attention to humans' cognitive competency of analyzing the reasonability of counterfactual conditionals. This sheds lights on the importance of a problem that has been maltreated in artificial cognitive systems, despite its wide importance and despite its long history across several fields (cf. Byrne, 2005; Lewis, 2001; Fauconnier, 1997, for example). Secondly, we discuss cognitive phenomena that could be responsible for this particular competency in humans. We argue that the ability to analyze this kind of conditionals is one of the essential aspects of smartness, which needs to be better treated and better understood when building cognitive systems. Finally, we show that the analyzability has the potential to be represented and computed by integrating the functionalities of analogy-making and conceptual blending; two of the fundamental cognitive mechanisms that have proved to play important roles in endowing cognitive systems with essences of cognition (cf. Abdel-Fattah et al., 2012; Martinez et al., 2011; Hofstadter, 2001, for example).

The rest of the article is structured as follows. The problem of counterfactuals is introduced in section 2. An elaboration on how a cognitive system might approach the problem is conceptually discussed from a high-level perspective in section 3. In section 4, we present our proposal on how to formally achieve this. A detailed worked-out example is given in section 5, before section 6 concludes the article with final remarks and future work.

2. Introducing Counterfactual Conditionals (CFC)

A *counterfactual conditional* (from here on CFC), is a conditional sentence in the subjunctive mood: an assumption-conclusion conditional that designates what would be (or could have been) the case if its hypothetical antecedent were true. CFCs are also known as subjunctive conditionals or remote conditionals. They are contrasted with both

1. "material conditionals": in which the antecedent and the consequent may have no relation in common yet the conditional itself can be true (because its truth value depends only on those of the antecedent and the consequent); and
2. "indicative conditionals": which can be thought of as operations given by statements of the form "If *antecedent*, (then) *consequent*".

Although indicative conditionals, too, may be sometimes seen as contrary-to-fact statements, a straightforward comprehension difference between indicative and subjunctive conditionals is classically shown by the well-known Oswald/Kennedy examples (Adams, 1970):

If Oswald did not kill Kennedy, someone else did. (1)

If Oswald had not killed Kennedy, someone else would have. (2)

Sentence 1 shows an indicative conditional, while sentence 2 is a subjunctive version. The majority of people would accept the former as reasonable yet reject the latter (cf. Pearl, 2011; Adams, 1975; Adams, 1970). Another difference is given in (Santamaría, Espino, & Byrne, 2005), where participants read presupposed facts very rapidly, indicating that a "priming effect" occurs when human participants read CFCs and not when they read indicative conditionals.

Table 1 gives a general form and some examples. A major part of the CFCs can be given in the general form of sentence 3. Other sentences, such as sentence 6, may also be paraphrased to agree with this form. The general form of sentence 3 has two parts: an antecedent (i.e. the assumption) and a consequent (i.e. the conclusion), which are both hypothetical statements. According to standard semantics, both parts could be 'false' (at least the assumption is a known falsehood). The concern, thus, is not about binary truth values of CFCs, like the case for material implications, but rather about analyzing and verifying them and their conditions for being considered meaningful or reasonable.

If it were the case that *antecedent*, then it would be the case that *consequent*.	(3)
If Ulf had left home early he would have caught the bus.	(4)
Ahmed would have cooked the dinner if Nashwa had not done so.	(5)
In France, Watergate would not have harmed Nixon.	(6)
If Mursi had not won the election, Egypt would not suffer a military coup.	(7)
If Julius Caesar was in command during the Korean war, then he would have used the atomic bomb.	(8)
If Julius Caesar was in command during the Korean war, then he would have used the catapult.	(9)

Table 1. A list of sentences that represent or paraphrase counterfactual conditionals.

In addition to the importance of their computational evaluation per se, CFCs situate themselves within entertaining scopes of end-to-end cognitive systems. Counterfactual reasoning is involved, and plays an important role (one way or another), in problems and puzzles of domains as diverse as learning, theory-of-mind, moral judgement, or decision-making under risk and uncertainty. In the field of theory-of-mind, for example, the children

in the famous muddy-children problem (cf. Shoham & Leyton-Brown, 2009, for instance) take actions because they fail to verify "what-if" situations that are contrary to their (common) knowledge. In the general case of the problem, n honest, logical-reasoner children commonly know that both (i) $1 \leq k \leq n$ of them is muddy, and (ii) the question: "*do you know whether you are muddy?*" have already been asked publicly for k rounds. A child would know she is muddy by first thinking to herself: "if I were not muddy, I would have known by the $(k-1)^{st}$ repetition of the question (the common knowledge) that the rest already know they are muddy". Then, and as the child fails to verify this CFC, the child concludes she must be muddy.

2.1 Analyzing CFCs by Humans and in Artificial Systems

We consider a CFC to be *verifiable* if its contrary-to-fact conclusion consistently follows from its contrary-to-fact assumption by a reasonable judgement. The analysis of a CFC is the reasoning process that leads to the judgment, which is assumed to hold in a (third) contrary-to-fact world that, in turn, depends on the reasoner's background and reasoning strategies. The verification of a CFC is a *judgement of reasonability* that involves the subjective importation of knowledge-based facts (Lee & Barnden, 2001, p. 8) and is weaker than logical validation. Yet this judgement can always be disputed (cf. Quine, 1960; Goodman, 1947), using CFCs like sentence 8 and sentence 9, for instance (cf. section 5).

We see the reasonable analysis of CFCs as a fundamental cognitive competency that may be used to designate, evaluate, and compare superior cognitive systems. It obviously requires a cognitive system to proficiently *create contrary-to-fact conceptions*, in order to reasonably analyze a given CFC. Humans, the ultimate exemplar of cognitive beings, are without any doubts the unique species that can perform such a reasonable analysis. They can do this because, in particular, they utilize logical reasoning, create alternatives to reality, communicate with language, hold rational beliefs, show rational behavior, as well as employ several cognitive capacities (cf. Abdel-Fattah et al., 2012; Abdel-Fattah, Besold, & Kühnberger, 2012). It is dazzling how humans smoothly analyze a given CFC and may convincingly estimate a rough truth degree, and even argue about it. In general terms, this can be achieved in humans by the imagination of a whole set of alternative conceptualizations that differ in certain aspects from their real world counterparts, but in which the CFC's antecedent holds. The reasoning process is then carried out in creatively-imagined worlds, yielding coherent results (cf. Byrne, 2005). We propose that this can be achieved in cognitive systems when the system is endowed with (computationally-plausible versions of) such abilities. In the following, a short literature overview identifies the most important ones of these abilities.

2.2 A Quick View of Some Treatments in Literature

The representation and verification of CFCs have always delivered debates within many disciplines, like philosophy, psychology, computer science, and linguistics. We mention important contributions in the literature that back up the ideas in the later discussion.

Philosophical treatments Beside Goodman's discussion of CFCs (Goodman, 1947), another classical line of work by David Lewis and Robert Stalnaker uses possible world semantics of modal logic to model CFCs based on a similarity relation between possible worlds. According to Lewis's account (Lewis, 2001), the truth type of a CFC in the form of sentence 3 can be either vacuously true, non-vacuously true, or false. This depends on the existence of a closely similar possible world to the real world, in which the antecedent and the consequent are true. The account is unclear as to what 'similarity' (or 'closeness') mean.

Psychological treatments Many cognitive scientists would agree that reasoning requires the creative production of mentally-constructed conceptual entities (Johnson-Laird, 1983). The creation and verification of CFCs, in particular, as alternatives to reality are widely explored in the pioneering work of Ruth Byrne (cf. Byrne, 2005), where many experiments about reasoning and imagination are carried out. Therefore, "a key principle is that people think about some ideas by keeping in mind two possibilities" (Santamaría, Espino, & Byrne, 2005) so that *two mentally-constructed domains* are needed in assessing the truth of a given CFC. These worlds (referred to as source and target domains below) are treated as conceptual spaces.

Linguistic treatments Classical approaches view language as consisting of statements that can be reasoned about in terms of their truth functions. But some linguists also deal with meaning construction in natural language by means of mentally-constructed spaces and their blending (cf. Coulson, 2006; Fauconnier, 1994). Of a particular interest is the analysis of CFCs in cognitive linguistics, presented in (Lee & Barnden, 2001), based on the mapping between different reasoning spaces and the drawing of analogies between these spaces. This analysis is implemented in an AI reasoning system and applied to the verification of certain CFCs (cf. Lee & Barnden, 2001), which shows that a form of the analysis can already be computed by artificial systems.

Algorithmic treatments Recently, an algorithmic approach towards CFCs was presented by Judea Pearl (cf. Pearl, 2011). Complete procedures for discerning whether a given counterfactual is 'testable' and, if so, expressing its probability in terms of experimental data are given in (Shpitser & Pearl, 2007)). Pearl's basic thesis of treating counterfactuals states that their generation and evaluation is done by means of "symbolic operations on a model". This model represents the beliefs an agent has about the "functional relationships in the world" (Pearl, 2011). In this way, Pearl views the procedure as a concrete implementation of Ramsey's idea (Ramsey, 1929), in which a conditional is accepted if its consequent is true after its *antecedent is (hypothetically) added to the background knowledge*, making whatever *minimal adjustments* that are required to *maintain consistency*.

3. A Tale of Two Cognitive Mechanisms

The modeling of counterfactual reasoning is not only highly disputed, but can also be considered to be AI complete: while seemingly easy for humans, the treatment of CFCs poses a hard problem for artificial systems. However, we think that the utilization of computationally-plausible cognitive mechanisms in the analysis of CFCs appears to be achievable in cognitive systems. As mentioned earlier (cf. section 2.1), the analysis of CFCs is a clear competency of humans, which obviously requires a high level of artificial intelligence if cognitive agents were to acquire this competency (or approximate it) in any cognitive system. Thus, and particularly when it comes to developing computational cognitive systems that can analyze the reasonability of CFCs, we consider this competency as a complex-structured mechanism, and propose that the verification could be possibly achieved by means of reducing this complex mechanism to simpler, rather essential, cognitively motivated, and computationally-plausible mechanisms, such as analogy-making and conceptual blending.

By abstracting the major ideas of the various treatments given in section 2.2, one can discover that 'similarity' between 'domain worlds' (or creatively imagined 'conceptions') plays a shared role in all the treatments. One can also note that a cognitive system may need to, at least, develop processes that consider 'conceptual domains' as inputs, compare the 'similarity' between these domains, and judge the reasonability of a given CFC by deciding whether or not a 'blend' of these concepts 'remain consistent' after 'adding the antecedent' of a given CFC to the background knowledge.

3.1 The Role of *"The Core of Cognition"*: Analogy Making

Analogy making is a cognitive ability that is important for many aspects of intelligence, and plays a significant role in a wide range of problem-solving contexts. Analogies are an important aspect of reasoning and "a core of cognition" (Hofstadter, 2001), so they can be used to explain various behavior and decisions (Abdel-Fattah et al., 2012). Analogy is important for concept learning and can also be seen as a framework for creativity (Abdel-Fattah, Besold, & Kühnberger, 2012; Hofstadter & the Fluid Analogies Research Group, 1995). The ability to see two dissimilar domains as similar, based on their common relational structure, is fundamental and ubiquitous for human cognition (Gentner, Holyoak, & Kokinov, 2001).

We will show that an analogy engine helps in analyzing CFCs in computational cognitive systems. In this paper, we will refer to *Heuristic-Driven Theory Projection* (HDTP) as an example of an analogy-making system for computing analogical relations between two domains. HDTP is a mathematically sound framework for analogy making, together with the corresponding implementation of an analogy engine for computing analogical relations between two logical theories, representing two domains. HDTP has been applied to different fields and extended in various directions (cf. Abdel-Fattah et al., 2012; Martinez et al., 2011; Schwering et al., 2009; Gust, Kühnberger, & Schmid, 2006, for example for more details about HDTP and an expanded elaboration of its application domains).

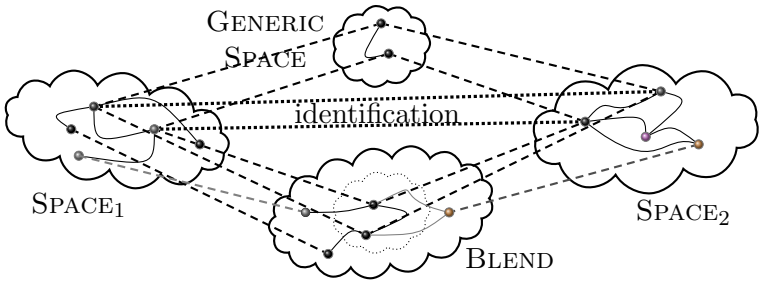

Figure 1. The prototypical four-space model of CB: common parts of the Space_1 and Space_2 concepts are identified, defining a Generic space and a Blend.

3.2 The Role of *"The Way We Think"*: Creation by Integration

Conceptual blending, or CB, is proposed as a powerful mechanism that facilitates the creation of new concepts by a constrained integration of available knowledge. CB operates by mixing two input knowledge domains, called *mental spaces*, to form a new one that basically depends on computed associations or alignments (which we call identifications) between the input domains. The new domain is called the *blend*, which maintains partial structures from both input domains and presumably adds an emergent structure of its own.

In the classical model of CB (cf. Fauconnier & Turner, 2002), two input concepts, the *source* and *target* spaces, represent two mental spaces. Common parts of the input spaces are matched by identification, where the matched parts may be seen as constituting a *generic space*. The blend space has an emergent structure that arises from the blending process and consists of some matched and possibly some of the unmatched parts of the input spaces. Figure 1 illustrates the prototypical four-space model of CB, in which the two general concepts, Space_1 and Space_2, represent *source* and *target* input spaces (the mental spaces).

CB has already shown its importance as a substantial part of expressing and explaining cognitive phenomena such as the interpretation of metaphors, counterfactual reasoning (Lee & Barnden, 2001), and a means of constructing noun-noun compounds (Abdel-Fattah & Krumnack, 2013) and new conceptions (Fauconnier & Turner, 2002).

3.3 The Combined Role: Analyzing CFCs by Employing Analogies and CB

Based on section 2.2, the treatments along the various directions appear to utilize humans' cognitive abilities of:[2]

1. conceptualizing hypothetical domains (as alternatives to reality) that contain the necessary background knowledge,

2. See also (Byrne, 2005), where many cognition experiments are given that further supports our proposed view.

2. intelligently drawing analogies between parts of the domains (and associating some of their constituting elements with each other), and
3. constructing a variety of possible consistent conceptualizations, in which the given CFC can be verified.

Therefore, the ideas of CB may be used, side by side with analogy-making, to analyze the reasonability of CFCs by blending two input mental spaces and constructing counterfactual blend spaces (cf. section 4.2), in which the analysis of CFCs can take place. We argue, then explain later in section 4, that the combination of (*i*) a powerful analogy engine and (*ii*) the ideas of conceptual blending (cf. Fauconnier & Turner, 2002), potentially endows cognitive systems with the ability to reasonably analyze (some) CFCs in an intuitive way. From an implementation-oriented perspective, this implies that artificial cognitive systems can analyze the reasonability of CFCs as long as computational versions of the aforementioned cognitive mechanisms, in particular, can be utilized.

4. Towards a Treatment Formalization: Constructing Counterfactual Blends

We now show that (at least a certain class of) CFCs can be analyzed by constructing appropriate blend spaces, using analogy between input domains that correspond to the antecedent and the consequent of a given CFC.

Our procedure is based on a structural mapping of two input domains, which correspond to the antecedent and the consequent of a given CFC. This gives rise to several *blend candidates*, which import major elements from one or the other input domain. The importation may render some blend candidates inconsistent, which reflects the non-reasonability of a given CFC. But those blend candidates that satisfy specific criteria (beside being consistent) will reflect a given CFC's reasonability. Therefore, a heuristics is formulated to choose the most plausible candidates, guided by the (logical) structure of the given CFC based on some fixed principles.

In our treatment[3], the analysis of a given CFC (in the general form of sentence 3) requires the creation of two mental domains for each of the involved parts (i.e. the antecedent and the consequent). In order to find similarities and suggest common background between the two parts, analogical mapping is used to compare the structural aspects in both domains. Associations between the two mentally-constructed domains can thus be found. Finally, a logically-consistent combination of the two domains can be suggested, as a newly-created blend of them, in which the reasoning process can occur. The reasoning process will take place in a blend space that forms the setting to verify the CFC. Some constraints could be imposed to give preference to one blend over another. Additionally, each conceptualization may be given a rank reflecting its relative plausibility.

3. Our approach may seem to have common characteristics with (Lee & Barnden, 2001)'s and (Fauconnier, 1997)'s, because all of them are more or less inspired by analogy and CB. However, we adopt a more general blending procedure and use a different method and heuristics to suggest the construction of blends.

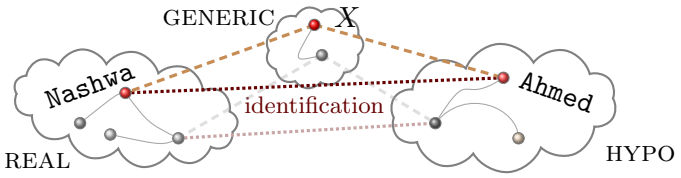

Figure 2. In the Generic Space, the element/term X generalizes/anti-unifies two elements/terms from Space_1 and Space_2 that play similar roles in their corresponding domains. This illustration is based on sentence 5, figure 1, as well as on the discussions in section 4.1.

To put these (and section 3's) ideas into a formal framework, the process will be split into two steps: the generalization of the given domains of a CFC (via analogy) and the construction of a counterfactual space (via blending).

4.1 Generalization and Structural Mapping

The mapping is based on a representational structure used to describe the two domains. In a computational system these descriptions may be given in a formal language, like first-order logic. The strategy applied here is based on the HDTP framework (Schwering et al., 2009), but we will use a schematic form of natural language for our examples to improve readability.

The basic idea is to detect *structural commonalities* in both domain descriptions by a generalization process. Then, based on this generalization, objects from both domains that have corresponding major roles can be identified. As an example consider the following statements about the real and a hypothetical world according to sentence 5 (see figure 2):

Ahmed cooked	(REAL)
Nashwa cooked	(HYPO)
X cooked	(GENERIC)

The statements (REAL) and (HYPO) can be generalized by keeping their common structure and replacing differing parts by variables in (GENERIC), where this generalization gives rise to the association:

$$X : \text{Ahmed} \triangleq \text{Nashwa}.$$

The richer the conceptualizations of the domains, the more correspondences may arise. However, an essential point in constructing the generalization is the principle of "coherence", which states that if a term occurs in multiple statements of a domain description, it should always be mapped to the same corresponding term of the other domain. Such a reusable mapping of terms is a good indicator for structural correspondence.

It is worth pointing out that the kind of generalization we mention here is usually referred to as anti-unification (cf. Plotkin, 1970). HDTP, in particular, applies restricted

higher-order anti-unification (Krumnack et al., 2007) to find generalizations of formulas and to subsequently propose analogical relations between the source and target domains (cf. figure 2).

4.2 Reasonability Principles for Counterfactual Blend Construction

The established mapping is used as a basis for constructing *counterfactual blend* candidates. (A 'counterfactual blend' will henceforth be denoted CFB.) Statements from both input domains can be imported, and the mapping is applied for merging them. But one should note that the objects, which are covered by the mapping, must play the same role in both input domains. Therefore, their simultaneous existence in a CFB is considered incompatible, although normal CB (cf. Coulson, 2006; Fauconnier & Turner, 2002) explicitly allows simultaneous occurrence of corresponding entities from both domains in the blend space. For each such object, thus, we have to reasonably choose one of the alternatives in a systematic way.

The following *reasonability principles* are proposed to guide the construction of CFBs:

(P1) Counterfactuality: A CFB candidate should satisfy the antecedent of the given CFC.

(P2) Choice: For every matching pair, one alternative is allowed to be imported into a CFB candidate.

(P3) Consistency: A CFB candidate should sustain (logical) consistency.

(P4) Maximality: A CFB candidate should contain as many imported instances of the original axioms as possible.

As it rules out many meaningless and unneeded possibilities from the beginning, (P1), the principle of counterfactuality, will be the starting point to achieve a reasonable CFB. It forces the antecedent of the CFC to hold in a CFB candidate and thereby provides the first reasonability criterion for selecting alternatives from the mapping pairs. In the next step, an initial description of a CFB candidate can be enriched by importing additional statements from any of the two input domains, keeping all the principles satisfied. During importation, all terms covered by the mapping have to be replaced coherently by the chosen alternative.[4] If no alternative for a term has been chosen yet, a choice has to be made and marked for *all subsequent occurrences* of that term. In general, the process should try to maximize the number of imported statements to allow for inferences of concern. One however has to assure that the constructed CFB stays consistent.

4. This discussion implies that P1, in particular, has a remarkable effect, not only on ruling out the importation of many meaningless and unneeded possibilities (e.g. inconsistent statements), but also on keeping a CFB candidate reasonable by enforcing modifications on some of the statements that may be imported: in this way, an inconsistent statement may have the potential to be modified to another version that can be imported. The imported version of the modified statement reasonably cohere with all the imported statements in the CFB candidate in hand (cf. footnote 5 on page 14).

It is obvious that these reasonability principles do not always lead to a unique CFB by allowing for multiple variants. This should be an essentially desirable feature in implementations of cognitive systems. Indeed, this feature may not be easily achieved in classical AI or cognitive systems without emphasizing the role that CB plays. Thanks to the ideas of CB, this feature allows for alternative verifications of a given CFC (cf. section 5), where the existence of multiple (reasonable) CFB spaces simulates the indecisiveness of humans in judging a given CFC. Remember that the judgement of a given CFC may always be disputed (cf. Quine, 1960, Goodman, 1947, sentence 8 and sentence 9), which means that a cognitive system may need to allow the possibility of having several (reasonable) CFB candidates for arguing about the same given CFC in several ways (a more concrete explanation is given at the end of the next example).

4.3 A Simple Example

To demonstrate these ideas, we first present a simple example here, leaving the thorough discussion to the more detailed example in section 5. The following is a simplified formalization of a metaphor discussed in (Turner & Fauconnier, 2003): "If Clinton were the Titanic, the iceberg would sink". The metaphor introduces two input domains, first the domain of political affairs in Washington:

Clinton hits the scandal. (W1)

Clinton does not sink. (W2)

and second the domain comprising the events around the Titanic:

The Titanic hits the iceberg. (T1)

The titanic sinks. (T2)

From this data, a generalization can be constructed, consisting of generalized facts that can be instantiated in both input domains. In our case, the generalization contains only one proposition:

X hits Y. (G1)

This generalization gives rise to an anlogical mapping:

$$X : \text{Clinton} \triangleq \text{the Titanic} \qquad Y : \text{the scandal} \triangleq \text{the iceberg}$$

Based on this analogy, we now construct a CFB using our four principles: According to the principle of counterfactuality (P1), the antecedent of the CFC has to be introduced to the blend:

Clinton hits the iceberg. (B1)

This instantiation allows already to choose alternatives according to the principle of choice (P2): $X \mapsto$ *Clinton* and $Y \mapsto$ *the Titanic*. The principle of maximality (P4) invites us

to introduce addional facts from the input domains into the CFB (substituting terms as necessary):

Clinton does not sink. (B2)

Clinton sinks. (B3)

Here (B2) is imported from (W2) and (B3) is imported from (T2) by applying the substitution. However, the resulting blend is inconsistent and the principle of consistency (P3) forces us to remove one of the conflicting facts (B2) and (B3) from the blend. For the intended interpretation, we would remove (B3). Assuming suitable background knowledge, like "if A hits B, then A or B sinks.", we can logically derive the the conclusion of the original CFC.

4.4 Some Remarks

We present our approach here in a very general way and avoid discussing issues of deeper philosophical nature. For instance, we give no explicit constraints on what counts as an admissible set of inputs, which allows us to anchor input domains in impossible or phantasy worlds (e.g. Star Wars). But in such cases it would be unclear whether to consider the conditionals as counterfactuals (with "factuality" being described by "impossible" worlds) or counterpossibles (with impossible antecedents). The latter counterpossibles are defined by Lewis as conditionals with impossible antecedents, and are always vacuously true regardless of the consequent.

Also, some approaches would argue in favour of a representation language for CFCs, that is different from the one our framework considers. One natural proposal would be to express axioms in the domain theory as weighted constraints capable of being broken at a cost (cf. Bello, 2012). Such questions are not particular to our work and are commonly known in the history of CFCs to be rigorous and painstaking. In the current paper, thus, we do not make any kind of distinction between kinds of CFCs, and only focus on considering the three stated facets of the problem (cf. section 1).

5. Example: The Caesar-Korean CFC

In this section, we give a worked out example that explains our procedure in more detail. The example also provides different lines of argumentation for verifying one given CFC.

Consider the following CFC (already introduced in section 2, based on (Quine, 1960, p. 222) and (Goodman, 1947)):

If Julius Caesar was in command during the Korean war,
then he would have used the atomic bomb. (8)

This conditional is to be interpreted in a hypothetical world, as it combines elements (Caesar and the Korean war) that do not belong together in the real world. This world is constructed by blending two domains, the Gallic Wars/*Roman Empire*, (RE), on the one hand and the *Korean war*, (KW), on the other. To formalize the example, we state the background

knowledge on the two domains that we believe is relevant to this discussion (N.B. temporal and tense aspects are disregarded in the given statements and representations). For the (RE) domain, this background knowledge can include the axioms:

Caesar is in command of the Roman army in the Gallic Wars, (RE1)
The catapult is the most devastating weapon, and (RE2)
Caesar uses the most devastating weapon. (RE3)

On the other hand, the (KW) domain can include the axioms:

McArthur is in command of the American army in the Korean War, (KW1)
The atomic bomb is the most devastating weapon, and (KW2)
McArthur does not use the atomic bomb. (KW3)

Based on these axiomatizations, and according to the ideas discussed in section 4, a generalization can be computed. The statements that will enter the generalization are only those, for which instances are present in both domains:

X is in command of the Y army in Z, and (G1)
W is the most devastating weapon. (G2)

From the generalization, a mapping of corresponding terms in both domains can be derived:

X : Ceasar $\triangleq$ McArthur Y : Roman $\triangleq$ American
Z : Gallic Wars $\triangleq$ Korean War W : catapult $\triangleq$ atomic bomb

CFB candidates can now be constructed by merging the two domains; identifying axioms and entities matched by the generalization, and keeping the four reasonability principles for CFB satisfied (cf. section 4.2). For example, the antecedent of the CFC in hand (sentence 8) must be satisfied in each reasonable CFB candidate (according to (P1), the principle of counterfactuality). So, one may start by insisting that the CFB candidate contains:

Caesar is in command of the American army in the Korean war, (B1)

then continue enriching the CFB candidate by importing further statements from the input domains, such as:

The atomic bomb is the most devastating weapon, (B2)
Caesar uses the most devastating weapon, (B3)
Caesar does not use the atomic bomb, (B4)

and so on. However, enriching a CFB may render it inconsistent or unreasonable. For example, a CFB that contains all of (B1), (B2), (B3), and (B4) violates the consistency

principle (P3) because (B4) contradicts what could be inferred from (B2) and (B3), namely (B5):

Caesar uses the atomic bomb. (B5)

In figure 3, depictions are given to illustrate (*i*) the Korean war domain, (*ii*) the Gallic Wars domain, (*iii*) the given generalization, and (*iv*) two (reasonable) CFB candidates (their construction is discussed below). For simplicity, figure 3 does not identify the terms but only the statements that are composed of those terms.

Note that the principle of counterfactuality can single-handedly prevent the importation of many (implausible) sentences into a CFB candidate.[5] In the current example, (P1) already enforces the choice for three of the mapped terms:

$$X \mapsto \text{Caesar}, \qquad Y \mapsto \text{American}, \qquad Z \mapsto \text{Korean War}.$$

Therefore, one can no longer import (implausible) statements such as:

McArthur does not use the atomic bomb, nor (KW3)

McArthur is in command of the Roman army in the Gallic Wars, (NOWAY1)

into any CFB candidate (otherwise the candidate is clearly unreasonable). Nevertheless, many CFB candidates can, in principle, still be constructed. A CFB candidate may import as many (plausible) statements as possible from any (or both) of the input domains (but perhaps not, simultaneously, all of them); sustaining its reasonability by making use of the guiding principles. However, the importation of one (plausible) statement or another may be found to cause reasonability problems. That is, a statement can have the potential to be imported into a CFB candidate, but may not be imported into such a candidate because this is practically prevented by the reasonability principles (otherwise the importation will render the CFB candidate unreasonable). For instance, consider (RE2) and (RE3) which infer (by classical deduction in the (RE) domain):

Caesar uses the catapult. (RE4)

An unmodified version of (RE4) can, in principle, be imported into a CFB candidate (unlike (KW3), which cannot be imported into any CFB candidate, unless it is modified by (P1)). But it is possible that one of the reasonability principles disallow the importation of (RE4) into a specific CFB candidate, especially when contradicting statements have already been imported into the same CFB candidate[6].

5. In fact, the principle of counterfacuality does more than that. One may have noticed that (B2) is an imported version of (KW2), and (B3) is an imported version of (RE3), whereas (B4) is not a directly-imported version of (RE4), nor of any other statement. (B4) is a rather restrictedly-imported version of (KW3), in which the term 'Caesar' replaces 'McArthur' (according to P1).

6. E.g. the candidate (CFB1) described below prevents (RE4) to be directly imported as it is, because the directly-imported versions of (RE2) and (RE3), namely (B2) and (B3), respectively, infer the statement (B5) that contradicts the imported version of (RE4) (namely, (B4')). Whilst, (RE4) can be directly imported, as it is, into another candidate, (CFB2), which does not include "both" imported versions of (RE2) and (RE3).

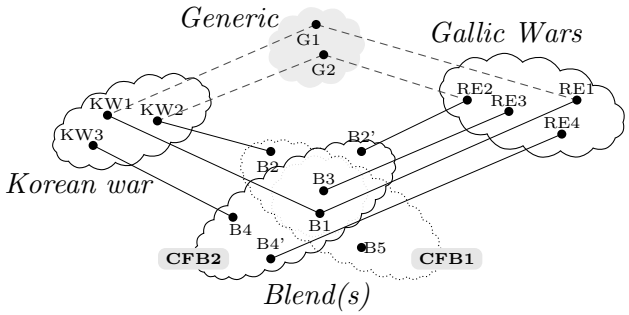

Figure 3. An illustration of two possible blend spaces for the CFC of sentence 8. For the sake of simplicity, the illustration does not show the mapped terms but rather depicts some of the representing sentences given in section 5.

One could in general get several, reasonable CFB candidates for the same CFC, but some of them may eventually be (logically) equivalent, according to our principles and heuristics. Two (non-equivalent) blend spaces for the CFC in hand are given in figure 3, and described in the next two paragraphs.

(CFB1): The main representational sentences of this blend candidate include (B1), (B2), (B3), and (B5) (see the right-hand blend in figure 3, with a dotted outline). This blend reasonably verifies the CFC because it implies that "Caesar is in command of the American army in the Korean war and uses the most devastating weapon, which is the atomic bomb". This CFB could be equivalent to another one that only contains (B1), (B2), and (B3) as axioms, since (B5) is (consistently) deducible from (B2) and (B3).

Note that (B1) is supported by (P1) and (P2). (B2) is imported using (P2); similarly (B3). Finally, (B5) is a direct inference of (B2) and (B3). Note that (P3) prohibits the inclusion of (B4) into (CFB1): (B4) is an instantiation of (KW3) in which 'Caesar' instantiates X, but (B4) has a potential clash with (B5).

(CFB2): This is an alternative blend space, which reflects the possibility that Caesar would use the catapult and not the atomic bomb. Its axioms include (B1), (B3), (B4), and the following sentence:

The catapult is the most devastating weapon, (B2')

which is a directly-imported version of (RE2). Also, in (CFB2):

Caesar uses the catapult, (B4')

results either as an inference from (B2') and (B3), or as a directly-imported version of (RE4) (which, itself, can already be inferred from (RE2) and (RE3) in the (RE) domain). In this (CFB2) blend (see the left-hand blend in figure 3, with a solid outline), (*i*) Caesar is in command of the American army according to (B1), (*ii*) the catapult is considered the

most devastating weapon according to (B2'), (*iii*) Caesar does not use the atomic bomb according to (B4), but rather (*iv*) Caesar uses the catapult according to (B4'). In addition, according to the proposed maximality principle[7], (CFB2) is more 'maximal' than (CFB1). According to the illustrations in figure 3, (B4') does not belong to (CFB1), which means that Caesar cannot use the catapult as an alternative in (CFB1).

6. Concluding Remarks and Future Works

The problem of analyzing CFCs has a long history in many disciplines, yet very few computational solution frameworks exist (especially as part of an artificial cognitive systems). In this article, we emphasize the importance and argue for the feasability of considering cognitive mechanisms in attacking this challenging problem. We focus on two particular cognitive mechanisms, analogical mapping and conceptual blending, and propose a computational strategy to contribute to solving the problem by cognitive systems. As a proof of concept, and to give a concrete example of applying the main ideas of our strategy, the presentation was based on the HDTP framework, using a schematic form of natural language to improve readability. However, the generality our ideas also allows for an application in other frameworks.

In our opinion, the general problem of analyzing CFCs deserves to be a benchmark problem for comparing and evaluating cognitive systems, by considering their proposals to analyzing different types of the CFCs. Not only are certain CFCs clearly quite harder to imagine or reason about than others, but the relationships of the entities appearing in these CFCs (among themselves and between their counterparts in the factual world) can be hard to handle as well. Nevertheless, a distinction in treating different CFCs needs to be reflected by cognitive systems in verifying such CFCs. This would require considering, among many other profound issues, a precise schematization of CFCs (cf. Lewis, 2001) combined with a cognitive architecture capable of handling natural language processing and dynamic outlooks on semantics (cf. Veltman, 2005).

In future work, we will focus on answering some related questions. For example, in the process of analyzing a CFC, the aspects in which the real and the hypothetical worlds differ may not be very obvious to identify. Even in his possible-world semantics treatment of CFCs, David Lewis did not give a precise definition of what a "miracle" is (Lewis, 2001). In any case, the setting of an adequate alternative CFB space calls for the construction of a (temporary) knowledge domain that may contain counterfactual knowledge entities. A construction–analysis process, like the outlined one, could be what one might expect from an artificial cognitive system. Also, we tried to restrict the form of the CFC to that of sentence 3, though it is still important to identify the characteristics of the CFCs, to which our approach can (or cannot) always be applied. No doubt that this is a completely non-trivial issue, in particular because a unified representational scheme may also be required. Moreover, actual computational models still need to be deeper investigated in order to get more practical insights into implementing the presented ideas.

7. As well as according to the currently given representations, of course.

References

Abdel-Fattah, A. M. H., Besold, T. R., Gust, H., Krumnack, U., Schmidt, M., Kühnberger, K.-U., & Wang, P. (2012). Rationality-Guided AGI as Cognitive Systems. *Proc. of the 34th annual meeting of the Cognitive Science Society* (pp. 1242–1247).

Abdel-Fattah, A. M. H., Besold, T. R., & Kühnberger, K.-U. (2012). Creativity, Cognitive Mechanisms, and Logic. *Proc. of the 5th Conference on Artificial General Intelligence, Oxford* (pp. 1–10). Springer.

Abdel-Fattah, A. M. H., & Krumnack, U. (2013). Creating analogy-based interpretations of blended noun concepts. *AAAI 2013 Spring Symposium: Creativity and (Early) Cognitive Development* (pp. 2–7). AAAI Press, Palo Alto, California.

Abdel-Fattah, A. M. H., Krumnack, U., & Kühnberger, K.-U. (2013). Utilizing Cognitive Mechanisms in the Analysis of Counterfactual Conditionals by AGI Systems. *Proc. of the 6th Conference on Artificial General Intelligence, Beijing* (pp. 1–10). Springer-Verlag Berlin Heidelberg.

Adams, E. (1970). Subjunctive and indicative conditionals. *Foundations of Language*, *6*, 89–94.

Adams, E. W. (1975). *The logic of conditionals*. Dordrecht: D. Reidel Publishing Co.

Bello, P. (2012). Pretense and cognitive architecture. *Advances in Cognitive Systems*, *2*, 43–58.

Byrne, R. (2005). *The rational imagination: How people create alternatives to reality*. MIT Press.

Coulson, S. (2006). *Semantic leaps: Frame-shifting and conceptual blending in meaning construction*. Cambridge University Press.

Fauconnier, G. (1994). *Mental spaces: Aspects of meaning construction in natural language*. Cambridge University Press.

Fauconnier, G. (1997). *Mappings in thought and language*. Cambridge University Press.

Fauconnier, G., & Turner, M. (2002). *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. New York: Basic Books.

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A., Lally, A., Murdock, J., Nyberg, E., Prager, J., Schlaefer, N., & Welty, C. (2010). Building Watson: An Overview of the DeepQA Project. *AI Magazine*, *31*, 59–79.

Gentner, D., Holyoak, K., & Kokinov, B. (Eds.). (2001). *The analogical mind: Perspectives from cognitive science*. MIT Press.

Goodman, N. (1947). The problem of counterfactual conditionals. *The Journal of Philosophy*, *44*, 113–118.

Gust, H., Kühnberger, K.-U., & Schmid, U. (2006). Metaphors and Heuristic-Driven Theory Projection (HDTP). *Theor. Comput. Sci.*, *354*, 98–117.

Hofstadter, D., & the Fluid Analogies Research Group (1995). *Fluid concepts and creative analogies. computer models of the fundamental mechanisms of thought*. New York: Basic Books.

Hofstadter, D. R. (2001). Epilogue: Analogy as the core of cognition. In D. Gentner, K. J. Holyoak, & B. N. Kokinov (Eds.), *The Analogical Mind: Perspectives from Cognitive Science*, 499–538. Cambridge, MA: MIT Press.

Hsu, F. H. (2002). *Behind deep blue: Building the computer that defeated the world chess champion*. Princeton Univ.

Johnson-Laird, P. N. (1983). *Mental models: towards a cognitive science of language, inference, and consciousness*. Cambridge, MA, USA: Harvard University Press.

Krumnack, U., Schwering, A., Gust, H., & Kühnberger, K.-U. (2007). Restricted higher-order anti-unification for analogy making. *Twenties Australian Joint Conference on Artificial Intelligence* (pp. 273–282). Springer.

Lee, M., & Barnden, J. (2001). *A computational approach to conceptual blending within counterfactuals* (Cognitive Science Research Papers CSRP-01-10). School of Computer Science, University of Birmingham.

Lewis, D. (2001). *Counterfactuals*. Wiley.

Martinez, M., Besold, T. R., Abdel-Fattah, A. M. H., Kühnberger, K.-U., Gust, H., Schmidt, M., & Krumnack, U. (2011). Towards a domain-independent computational framework for theory blending. *AAAI Technical Report of the AAAI Fall 2011 Symposium on Advances in Cognitive Systems* (pp. 210–217).

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: symbols and search. *Communications of the ACM*, *19*, 113–126.

Pearl, J. (2011). The algorithmization of counterfactuals. *Annals of Mathematics and Artificial Intelligence*, *61*, 29–39.

Plotkin, G. D. (1970). A note on inductive generalization. *Machine Intelligence*, *5*, 153–163.

Quine, W. V. (1960). *Word and object*. The MIT Press.

Ramsey, F. P. (1929). General propositions and causality. In D. H. Mellor (Ed.), *Philosophical papers*, 145–153. Cambridge University Press.

Santamaría, C., Espino, O., & Byrne, R. (2005). Counterfactual and semifactual conditionals prime alternative possibilities. *Journal of Experimental Psychology*, *31*, 1149–1154.

Schwering, A., Krumnack, U., Kühnberger, K.-U., & Gust, H. (2009). Syntactic Principles of Heuristic-Driven Theory Projection. *Journal of Cognitive Systems Research*, *10*, 251–269.

Shoham, Y., & Leyton-Brown, K. (2009). *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press.

Shpitser, I., & Pearl, J. (2007). What counterfactuals can be tested. *UAI* (pp. 352–359). AUAI Press.

Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, *42*, 230–265.

Turner, M., & Fauconnier, G. (2003). Metaphor, metonymy, and binding. In R. Dirven & R. Pörings (Eds.), *Metonymy and metaphor in comparison and contrast*, 469–487. Mouton de Gruyter.

Veltman, F. (2005). Making counterfactual assumptions. *Journal of Semantics*, *22*, 159–180.