Utilizing Cognitive Mechanisms in the Analysis of Counterfactual Conditionals by AGI Systems

Ahmed M.H. Abdel-Fattah, Ulf Krumnack, and Kai-Uwe Kühnberger

University of Osnabrück, Albrechtstr. 28, Germany,
{ahabdelfatta,krumnack,kkuehnbe}@uni-osnabrueck.de

Abstract. We give a crisp overview of the problem of analyzing counterfactual conditionals, along with a proposal of how an artificial system can overcome its challenges, by operationally utilizing computationally-plausible cognitive mechanisms. We argue that analogical mapping, blending of knowledge from conceptual domains, and utilization of simple cognitive processes lead to the creative production of, and the reasoning in, mentally-created domains, which shows that the analysis of counterfactual conditionals can be done in computational models of general intelligence.

Keywords: General Intelligence, Analogy, Conceptual Blending, Counterfactual Conditionals, Cognitive Mechanisms.

In artificial general intelligence, AGI, it is extremely important to identify the various benchmark aspects of general intelligence, GI, and propose methods or systems that can computationally model such aspects. Existing applications may not be accepted as having GI, not because they lack the ability to e.g. reason, plan, or perform actions, but rather because their behavior is not viewed as originally motivated by essential cognitive abilities that reflect one GI aspect or another. Such applications neither integrate human-comparable competencies nor apply such competencies in various fields. They are usually designed to solve a specific task, and fail not only in solving another task but also in compatibly parsing that other task's input (cf. IBM's Watson and DeepBlue systems).

Motivated by this idea, our goal in the present article is twofold. We first aim at discussing the competency of human beings in analyzing the reasonability of counterfactual conditionals; a problem that has been maltreated in computational systems, despite its wide importance and long history (see [1, 2] for example). We introduce the ability to analyze counterfactual conditionals as one of the specific GI aspects that needs to be better treated and more understood in AGI systems. Secondly, we investigate the cognitive phenomena that could be responsible for this particular competency, and show that they could be represented and computed by integrating the functionality of basic mechanisms, such as analogy-making, (mental) concept invention, and conceptual blending.

The problem of counterfactuals is quickly introduced in section 1, and an elaboration on how an AGI system might approach the problem are conceptually discussed from a high-level perspective in section 2. In section 3 we present our ideas of how to formally do this. A detailed worked-out example is given in section 4, before section 5 concludes the paper with some final remarks.

K.-U. Kühnberger, S. Rudolph, and P. Wang (Eds.): AGI 2013, LNAI 7999, pp. 1–10, 2013.

1 Counterfactual Conditionals (CFC)

A *counterfactual conditional*, henceforth CFC, is a conditional sentence in the subjunctive mood: an assumption-conclusion conditional that designates what would be (or could have been) the case when its hypothetical antecedent were true. Table 1 gives a general form and some examples.

While the majority of CFCs is given in the general form of sentence 1, others (e.g. sentence 4) may also be paraphrased to agree with this form. The general form has two parts: an antecedent (i.e. the assumption) and a consequent (i.e. the conclusion), which are contrary-to-fact (hypothetical) statements. According to standard semantics, both parts could have the truth value 'false' (the assumption, at least, is usually a known falsehood). Thus, the concern is not about binary truth values of CFCs, like the case for material implications, but rather about analyzing and verifying them and their truth conditions.

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

If it were the case that *antecedent*, then it would be the case that *consequent*.	(1)
If the current president had not won the last election, then Ulf would have won it.	(2)
Ahmed would have cooked the dinner if Nashwa had not done so.	(3)
In France, Watergate would not have harmed Nixon.	(4)
If Julius Caesar was in command during the Korean war, then he would have used the atom bomb.	(5)
If Julius Caesar was in command during the Korean war, then he would have used the catapult.	(6)

We consider a CFC *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 degree of intelligent thinking. The verification of a CFC is a judgement of reasonability that involves the subjective importation of knowledge-based facts [3, p. 8] and is weaker than logical validation. Yet this judgement can always be disputed (cf. [4, 5]), using CFCs like sentence 5 and sentence 6, for instance.

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. The classical works of David Lewis and Robert Stalnaker use possible world semantics of modal logic to model CFCs based on a similarity relation between possible worlds. According to Lewis's account [1], the truth type

of a CFC in the form of sentence 1 depends on the existence of close possible worlds 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. The creation and verification of CFCs as alternatives to reality are widely explored in the pioneering work of Ruth Byrne (cf. [2]), 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" [6] so that two mentally-created domains are needed in assessing the truth of a given CFC. These domains – referred to as source and target below – are treated as conceptual spaces.

Linguistics Treatments. Some linguists deal with meaning construction in natural language by means of mentally-created spaces and their blending (cf. [7, 8]). Of a particular interest is the analysis of CFCs in cognitive linguistics, presented in [3], which is 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. [3]), which shows that a specific form of the analysis can already be computed in some systems.

Algorithmic Treatments. Judea Pearl has recently presented an algorithmic approach towards CFCs (cf. [9]). 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" [9] (which can be altered).

2 Cognitive Mechanisms and Counterfactual Conditionals

The modeling of counterfactual reasoning is not only highly disputed, but can also be considered AI complete. While seemingly easy for humans, the treatment of CFCs poses a hard problem for artificial systems. The utilization of computationally-plausible cognitive mechanisms in the analysis of CFCs appears, however, to be achievable in cognitive systems.

We will explain how the combination of an analogy engine, with an implementation of the ideas of conceptual blending, potentially endows AGI systems the ability to reason about CFCs in an intuitive and easy way. Our ideas (cf. section 2.2 and section 3.2) propose that (at least a certain class of) CFCs can be evaluated by constructing appropriate blend spaces. This is similar in spirit to [3, 10] and [11], but we adopt a more general blending procedure and use a different method to construct the blends. Our procedure is based on a structural mapping of two domains and gives rise to several blend candidates. A heuristics is formulated to choose the most plausible ones from these candidates, guided by the logical structure of the CFC based on some fixed principles.

2.1 Two Core Mechanisms

The analysis of CFCs is a competency of cognitive agents and obviously requires a high level of GI. Human beings, as the ultimate exemplar of such agents, can proficiently imagine sane situations and smoothly create contrary-to-fact conceptions in

order to reason about CFCs and verify them. This is achieved by imagining alternative conceptualizations that differ in certain aspects from their real-world counterparts, but in which the CFC's antecedent is imposed.

When it comes to developing computational AGI systems that can reason about CFCs, we consider the analysis of CFCs 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.

Analogy Making: The Role of the Core Cognitive Mechanism. Analogy making is a cognitive ability that is important for many aspects of GI: It is important for concept learning and can also be seen as a framework for creativity [12, 13]. Analogies are an important aspect of reasoning and "a core of cognition" [14], so they can be used to explain some behavior and decisions [15]. A CFC like sentence 4 illustrates how CFCs clearly employ analogical mapping [16].

We discuss later how an analogy engine helps in analyzing CFCs in computational systems. In this discussion, we use *Heuristic-Driven Theory Projection* (HDTP) as an example of an analogy-making system for computing analogical relations between two domains. HDTP has originally been developed for metaphor and analogy-making, and has been applied to different fields and extended in various directions (cf. [17,18, for example] for more details about HDTP). In HDTP, conceptualizations can be represented as *domains*, where intra-domain reasoning can be performed with logical calculi.

Conceptual Blending (CB): 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 the mapping 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. [16, e.g.]) two input concepts, *source* and *target*, 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. CB has already shown its importance as a substantial part of expressing and explaining cognitive phenomena such as metaphor-making, counterfactual reasoning [3], and a means of constructing new conceptions [16].

2.2 Utilizing Analogies and CB for Analyzing CFCs

Side by side with analogy-making, the ideas of CB may be used to analyze CFCs by means of blending two input *mental models* [19] to create a counterfactual blend world, in which the analysis of CFCs can take place. From a computational perspective, this means that AGI systems can be built to analyze CFCs by utilizing a computational version of the aforementioned cognitive mechanisms in particular.

As we may now view[1] it, the treatments along the various directions can be reduced to the utilization of the humans' cognitive abilities: (1) of conceptualizing hypothetical domains (as alternatives to reality) that contain the necessary background knowledge, (2) of intelligently drawing analogies between parts of the domains (and associating some of their elements with each other), and (3) of imagining a variety of possible consistent conceptualizations, in which the CFC can be verified.

In our treatment, the analysis of a given CFC (in the general form of sentence 1) 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 aspects in both domains. Associations between the two mentally-created 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 A Blending-Based Formal Treatment

To put the ideas of the previous section 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).

3.1 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 [18], but here 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 roles can be identified. As an example consider the following statements about the real and a hypothetical world according to sentence 2:

The current president won the last election	(REAL)
Ulf won the last election	(HYPO)
X won the last election	(GENERAL)

The statements (REAL) and (HYPO) can be generalized by keeping their common structure and replacing differing parts by variables in (GENERAL). This kind of generalization is usually referred to as anti-unification. This generalization gives rise to the association:

[1] Beside the earlier elaborations, many experiments of cognition are given in [2] that support the proposed view.

$$X : \text{Ulf} \triangleq \text{The current president}$$

The richer the conceptualizations of the domains, the more the correspondences that 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.

3.2 Counterfactual Blend Construction (CFB)

In a second step, the mapping is used as a basis for constructing a *counterfactual blend* space, henceforth CFB. Statements from both domains are imported, and the mapping is applied for merging them. Objects covered by the mapping play the same role in both domains and their simultaneous existence is therefore considered incompatible[2] in a CFB space. For each such object, thus, we have to choose one of the alternatives in a systematic way. The following principles are proposed to guide the construction process:

(P1) Counterfactuality: A CFB should satisfy the antecedent of the CFC.
(P2) Choice: For every matching pair, one alternative should be selected consistently.
(P3) Consistency: A CFB should remain logically consistent.
(P4) Maximality: A CFB 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. It forces the antecedent of the CFC to hold in a CFB and thereby provides the first criterion for selecting alternatives from the mapping pairs. In the next step, this initial description can be enriched by importing additional statements from the two input domains. During importation, all terms covered by the mapping have to be replaced coherently by the chosen alternative. 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.

These principles do not lead to a unique CFB by allowing for multiple variants. This is in fact a desirable feature, as it allows for alternative verifications of the CFC. The existence of multiple (consistent) CFB spaces simulates the indecisiveness of humans in judging a given CFC (remember that the judgement of a given CFC may always be disputed). In the following section, we give a worked out example that explains this procedure and provides two different lines of argumentation for verifying a given CFC.

4 The Caesar-Korean Blends: A CFC Example

We explain our approach using a well-known example (cf. [4] and [5, p. 222]), already introduced above in sentence 5. This conditional is to be interpreted in a hypothetical

[2] This differs slightly from normal CB [7, 16], which explicitly allows for simultaneous occurrence of corresponding entities from both domains.

world, as it combines elements (Caesar and the Korean war) which 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 hand. To formalize the example, we state the background knowledge on the two domains that we believe are relevant to this discussion (we disregard temporal and tense aspects in the given representation). For the (RE) domain this can be:

Caesar is in command of the Roman army in the Gallic Wars. (RE1)

The catapult is considered the most devastating weapon. (RE2)

Caesar uses the most devastating weapon. (RE3)

From this we might infer (by classical deduction) that:

Caesar uses the catapult. (RE4)

On the other hand, the (KW) domain can be described by the axioms:

McArthur is in command of the American army in the Korean War. (KW1)

The atom bomb is considered the most devastating weapon. (KW2)

McArthur does not use the atom bomb. (KW3)

Based on these axiomatizations, 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. (G1)

W is considered the most devastating weapon. (G2)

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

X : Ceasar $\triangleq$ McArthur $\quad$ Y : Roman $\triangleq$ American

Z : Gallic Wars $\triangleq$ Korean War $\quad$ W : catapult $\triangleq$ atom bomb

Now CFB spaces can be constructed by merging the two domains, identifying axioms and entities matched by the generalization. See figure 1 for a possible depiction of the Korean war domain, the Gallic wars domain, the given generalization, and two CFB spaces (discussed below). We start by the principle of counterfactuality and obtain:

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

This step already enforces the choice for three of the mapped terms:

$X \mapsto$ Caesar, $\quad$ $Y \mapsto$ American, $\quad$ $Z \mapsto$ Korean War.

We now try to continue enriching the current CFB by importing further statements, such as:

The atom bomb is considered the most devastating weapon. (B2)

Caesar uses the most devastating weapon. (B3)

Caesar did not use the atom bomb. (B4)

Though it may obey (P1), (P2), and (P4), a CFB in which (B1), (B2), (B3), and (B4) are all imported violates the consistency principle (P3) because (B4) contradicts what could be inferred from (B2) and (B3):

Caesar uses the atom bomb. (B5)

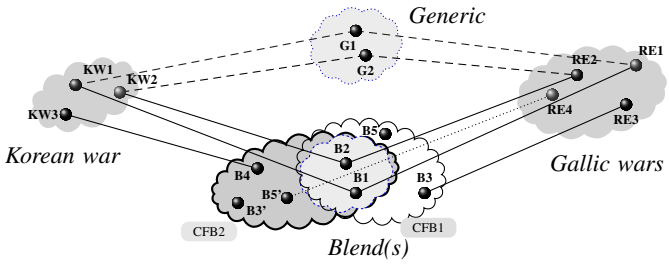

Fig. 1. An illustration of two possible blend spaces for the given CFC

Nevertheless, blend spaces can still be constructed by using the guiding principles specified above. One could in general get several CFB spaces for the same CFC, but some of them may eventually be equivalent according to these principles. Two (non-equivalent) blend spaces for the CFC in hand are given in figure 1:

(CFB1) Its axioms include (B1), (B2), (B3), and (B5) (see the outlined CFB (right) in figure 1). This blend 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 considered the atom bomb". However, this CFB could be equivalent to another one that only contains (B1), (B2), and (B3), since (B5) is (consistently) deducible from (B2) and (B3). (B1) is supported by (P1) and (P2); (B2) is imported using (P2), similarly the statement (B3); Finally, (B5) is a direct inference of (B2) and (B3). Note again that (P3) prohibits the importation of (B4), which is an instantiation with 'Caesar' replacing X, because its potential clash with (B5).

(CFB2) This is an alternative blend space, which reflects the possibility that Caesar would use the catapult. Its axioms include (B1), (B2), and the following fact:

The catapult is considered the most devastating weapon [import] (B3')

And, as an inference, one could get:

Caesar uses the catapult. [import] (B5')

In this blend (shown as a gray-filled blend (left) in figure 1), Caesar is in command of the American army, the atom bomb is considered the most devastating weapon, and Caesar does not use the atom bomb, rather the catapult. According to the proposed maximality principle (and the current representation), (CFB2) is more 'maximal' than (CFB1).

5 Concluding Remarks

The problem of analyzing CFCs has a long history in many disciplines, yet very few computational solution frameworks exist (especially in AGI). We wanted in this paper to emphasize the importance and feasibility of considering the utilization of cognitive mechanisms in attacking this challenging problem.

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[3]. In any case, the setting of an adequate alternative CFB space calls for the creation of a (temporary) knowledge domain that may contain counterfactual beliefs. A creation–analysis process, like the outlined one, could be what one might expect from an AGI system. In our opinion, the general problem of analyzing CFCs deserves to be a benchmark problem for comparing and evaluating AGI systems, by considering their proposals to analyzing CFCs. No doubt that this is a completely non-trivial issue, in particular because a unified representational scheme should be also used. Moreover, actual computational models still need to be investigated in order to get more practical insights into the solution of the problem.

References

[1] Lewis, D.: Counterfactuals. Library of philosophy and logic. Wiley (2001)
[2] Byrne, R.: The Rational Imagination: How People Create Alternatives to Reality. Bradford Books, MIT Press (2005)
[3] Lee, M., Barnden, J.: A computational approach to conceptual blending within counterfactuals. Cognitive Science Research Papers CSRP-01-10, School of Computer Science, University of Birmingham (2001)
[4] Goodman, N.: The problem of counterfactual conditionals. The Journal of Philosophy 44, 113–118 (1947)
[5] Quine, W.V.: Word and Object. The MIT Press (1960)
[6] Santamaría, C., Espino, O., Byrne, R.: Counterfactual and semifactual conditionals prime alternative possibilities. Journal of Experimental Psychology 31(5), 1149–1154 (2005)
[7] Coulson, S.: Semantic Leaps: Frame-Shifting and Conceptual Blending in Meaning Construction. Cambridge University Press (2006)
[8] Fauconnier, G.: Mental Spaces: Aspects of Meaning Construction in Natural Language. Cambridge University Press (1994)
[9] Pearl, J.: The algorithmization of counterfactuals. Annals of Mathematics and Artificial Intelligence 61(1), 29–39 (2011)
[10] Lee, M.: Truth, metaphor and counterfactual meaning. In: Burkhardt, A., Nerlich, B. (eds.) Tropical Truth(s): The Epistemology of Metaphor and other Tropes, pp. 123–136. De Gruyter (2010)
[11] Fauconnier, G.: Mappings in Thought and Language. Cambridge University Press (1997)
[12] Hofstadter, D., The Fluid Analogies Research Group: Fluid Concepts and Creative Analogies. Computer Models of the Fundamental Mechanisms of Thought. Basic Books, New York (1995)
[13] Abdel-Fattah, A.M.H., Besold, T., Kühnberger, K.-U.: Creativity, Cognitive Mechanisms, and Logic. In: Bach, J., Goertzel, B., Iklé, M. (eds.) AGI 2012. LNCS, vol. 7716, pp. 1–10. Springer, Heidelberg (2012)

[3] Even in his possible-world semantics treatment of CFCs, David Lewis did not give a precise definition of what a “miracle” is [1].

[14] Hofstadter, D.: Epilogue: Analogy as the core of cognition. In: Gentner, D., Holyoak, K.J., Kokinov, B.N. (eds.) The Analogical Mind: Perspectives from Cognitive Science, pp. 499–538. MIT Press (2001)
[15] Abdel-Fattah, A., Besold, T.R., Gust, H., Krumnack, U., Schmidt, M., Kühnberger, K.U., Wang, P.: Rationality-Guided AGI as Cognitive Systems. In: Proc. of the 34th Annual Meeting of the Cognitive Science Society, pp. 1242–1247 (2012)
[16] Fauconnier, G., Turner, M.: The Way We Think: Conceptual Blending and the Mind's Hidden Complexities. Basic Books, New York (2002)
[17] Gust, H., Kühnberger, K.U., Schmid, U.: Metaphors and Heuristic-Driven Theory Projection (HDTP). Theor. Comput. Sci. 354, 98–117 (2006)
[18] Schwering, A., Krumnack, U., Kühnberger, K.U., Gust, H.: Syntactic Principles of Heuristic-Driven Theory Projection. Journal of Cognitive Systems Research 10(3), 251–269 (2009)
[19] Johnson-Laird, P.N.: Mental models: towards a cognitive science of language, inference, and consciousness. Harvard University Press, Cambridge (1983)