Belief ascription and the Ramsey test*

Karolina Krzyżanowska

k.krzyzanowska@gmail.com

Abstract

In this paper, I analyse a finding by Riggs et al. (1998) that there is a close connection between people's ability to reason with counterfactual conditionals and their capacity to attribute false beliefs to others. The result indicates that both processes may be governed by one cognitive mechanism, though false belief attribution seems to be slightly more cognitively demanding. Given that the common denominator for both processes is suggested to be a form of the Ramsey test, I investigate whether Stalnaker's semantic theory of conditionals, which was inspired by the Ramsey test, may provide the basis for a psychologically plausible model of belief ascription. The analysis I propose will shed some new light on the developmental discrepancy between counterfactual reasoning and false belief ascription.

Although other people's beliefs are never directly accessible to us, we seem to be quite successful "mindreaders". We do not only reason about what others think, know, or desire, but we aslo often base our own decisions on what we take to be other people's states of mind. This ability, known in the literature as "theory of mind"[1] or "mindreading",[2] seems to be one of the most fundamental skills human beings have developed. The reasoning process we go through while interpreting other people's behaviour in terms of their states of mind often seems to be routine and effortless, if not automatic. Yet understanding that an agent may have a belief that does not correspond to reality seems to be inherently difficult for a developing child. Despite the results of non-verbal tests suggesting that infants can already anticipate that

*The final version appeared in *Synthese* (2013) 190:21–36, DOI 10.1007/s11229-012-0160-5. I am most indebted to Igor Douven for the discussions which helped me to develop the ideas put forward here and for many extremely helpful comments on all the previous versions of this paper. Other friends and colleagues whose ideas and comments I made use of include Richard Dietz, David Etlin, Piotr Kaźmierczak and Jakub Szymanik. All errors and inconsistencies remain my own.

[1] See e.g. Premack and Woodruff (1978), Carruthers and Smith (1996).

[2] E.g. Baron-Cohen (1995), Apperly (2011).

an agent will act on the basis of a false belief,[3] it is not before the age of four that children are able to answer an explicit question about an agent's false beliefs correctly.[4] In the so-called standard false belief task, known also as an unexpected transfer task, a child observes an agent hiding a toy in a container, and, later, another agent transferring that toy to another container while the first agent is absent. The child is then asked where the first agent thinks the toy is or where she will look for it (Wimmer and Perner (1983), Baron-Cohen et al. (1985)).

What the standard false belief task seems to measure is an ability to form belief reports, i.e. explicit representations of other people's beliefs. Belief reports are sentences of the form: "Agent A believes that p," where p is a proposition meant to express one of A's beliefs. I understand belief ascription to be a process of forming such belief reports; a process that people engage in whenever they want to explain or to predict other people's behaviour.[5]

For instance, when I see Thomas buying a recording by Paolo Pandolfo, I may come to think that Thomas likes baroque music. On the other hand, if I already know that he does not like baroque music I may think that he intends to give the recording to someone else. For instance, if our common friend, Steve, has a birthday and I know that Thomas is going to participate in the birthday party, then I may come to believe that it is meant to be a present for Steve and that:

(1) Thomas believes that Steve likes baroque music.

This might be the best explanation of Thomas' behaviour that I have found.[6] But how exactly do I proceed when I formulate a sentence like (1)? Even though over thirty years of psychological investigations have not brought an answer that everyone would agree upon, there are certain well-established findings, of which one is going to be the main concern of this paper. I will discuss an interesting result first obtained by Riggs et al. (1998), who found that young children's difficulty in attributing false beliefs is related to their

[3] It has been shown, for instance, that children look longer when an agent behaves in a way that is inconsistent with his own beliefs. This is believed to be a sign of surprise, and consequently interpreted as an evidence for early mindreading abilities. See Baillargeon et al. (2010) for a review of literature on false belief understanding in infants; and Bloom and German (2000) for a discussion of the problems with the standard false belief task as a measure of theory of mind.

[4] The discrepancy between infant's apparent false-belief understanding and older children's difficulties with false belief task may be plausibly explained in terms of "two system" account Apperly (2011) that distinguishes between "low-level" visual perspective-taking and "higher-level" explicit reasoning about other people's states of mind.

[5] In fact, there are data suggesting that adults do not form explicit belief reports spontaneously; they need to have a reason to do so (Apperly et al. 2006). Additionally, a study by Amsterlaw and Wellman (2006) indicates that explanation can facilitate children's learning to understand false beliefs.

[6] This is obviously not the only way of constructing belief reports. If, for instance, Thomas tells me that Steve likes baroque music, and I take him to be honest, I may formulate (1) straightaway—it suffices that I interpret what Thomas says as a testimony of what he believes. This paper's only concern though is belief ascription that relies crucially on the process of mindreading.

inability to reason about counterfactual situations. This result has been taken to indicate that there might be one mechanism that enables both abilities. Yet attributing false beliefs seems to be slightly more difficult for children than counterfactual reasoning. If indeed the two processes hinge on some common cognitive skills, what is responsible for the developmental discrepancy noted in the literature? Given that the common denominator for counterfactual reasoning and belief ascription has been suggested to be a mechanism exceedingly similar to what is known in the philosophical literature as the *Ramsey test* (Peterson and Riggs 1999), I investigate whether an analysis of belief attribution based on a semantic theory of conditionals, which was inspired by the Ramsey test, will shed some new light on the relationship between the two processes.

1 False belief reports and counterfactual conditionals

Riggs et al. found that there is a significant correlation between children's performance in, on the one hand, tasks in which children are supposed to answer a question about a protagonist's mistaken beliefs, and on the other hand, tasks in which children are asked a conditional question about a counterfactual situation. On this basis, they argue that there is a close connection between our understanding of other people's false beliefs and our ability to reason with counterfactual conditionals.

In order to test the relation between these two capacities, Riggs et al. used a version of a standard false belief task (Wimmer and Perner 1983) called "the Post Office Story." The story begins in Sally and Peter's home. Peter, a fireman, was ill, so he went to bed. Sally then went to a shop to buy some medicine. While Sally was shopping, somebody called and asked Peter for help putting out a fire at the Post Office. Peter got out of bed and went to the Post Office. Then children were asked the test questions:

Counterfactual situation: If there had been no fire, where would Peter be?

False belief: Where does Sally think Peter is?

Both questions have the same answer, namely "at home," and children who make a mistake do so, in both cases, by indicating the protagonist's actual location instead of the intended, counterfactual one—they commit a so-called "realist error." Riggs and colleagues reported a significant correlation in children's performance in the two tasks. Children younger than 4 years of age have difficulty in both task, older children succeed to answer both counterfactual and false belief question, although there are some children who find the question concerning the counterfactual situation easier than the false belief task[7] (see pp. 77–79 and pp. 81–82 of Riggs' paper for statistical details). Similar observations were made by e.g. German and Nichols (2003); Guajardo and

[7] A more significant difference in difficulty of the two tasks has been reported by Perner

Turley-Ames (2004); Guajardo et al. (2009); Drayton et al. (2011). Moreover, Grant et al. (2004) obtained similar results when they tested autistic children who are believed to have difficulties in acknowledging other people's states of mind (Baron-Cohen et al. 1985). On the basis of this and a couple of similar experiments, Riggs et al. claim that in order to pass the false belief test children have to develop an ability to understand counterfactual situations. This leads to the hypothesis that there is one mechanism underlying both the ability to reason counterfactually and the ability to attribute false beliefs.

In a subsequent study, Peterson and Riggs (1999) suggest that a common denominator for both processes is a mechanism of, as they call it, "modified derivation," which is a form of reasoning based on hypothetically modified sets of beliefs, as opposed to the "standard derivation," that is reasoning based on our own beliefs or knowledge. The idea of *modified derivation* is strikingly similar to the Ramsey test (Ramsey 1929/1990) according to which we decide whether to accept a conditional statement, "If p, then q", by adding hypothetically the antecedent p to our stock of beliefs, making minimal changes to maintain consistency (if necessary), and then deciding whether q is acceptable in the resulting hypothetical belief state. The reasoning strategy described by Ramsey can be conceived as a form of "simulation heuristics" (Evans and Over 2004, p. 119) for the interpretation of conditionals, since, as in the case of belief ascription, it involves projecting ourselves in an imaginary situation. An agent *pretends* that a certain proposition is true or, in other words, that a certain state of affairs holds, in order to find out what happens or what is true in this imaginary, pretend world.

In the context of belief attribution, the mechanism of mental simulation is construed as an ability to put oneself in someone else's shoes, that is, to project oneself into another person's situation. Traditionally, the simulation-driven account of mindreading has been contrasted with the so-called theory–theory account, according to which our understanding of other people's states of mind is underpinned by a quasi-scientific folk-psychological theory[8]. According to Peterson and Riggs (1999, p. 92), what is necessary for a child to pass the false belief task is an ability to simulate, that is, to project oneself into an imagined situation, or a situation from another person's point of view, which is achieved by means of modified derivation. Yet, as already mentioned, a small number of children who answer the counterfactual question correctly still fail the false belief test, which suggests that the latter is slightly more difficult than the former. In order to account for this fact the authors put forward that mindreading additionally requires an "adequate theoretical understanding of the mental," and hence in their view belief ascription is a simulation-theory

et al. (2004) who examined how the complexity of a scenario affects children's responses. They found that in simple scenarios counterfactual questions are much easier than in more complex scenarios, whereas children's performance in the false belief task did not depend on the complexity of a scenario. Even though later results cast doubts on whether their easiest scenarios really require counterfactual reasoning (Rafetseder and Perner 2010), the discrepancy between the two tasks still calls for an explanation.

[8] See e.g. Carruthers and Smith (1996) for an overview of various positions in the mental simulation vs. theory–theory debate

hybrid.

Although it seems natural and tempting to explain the link between counterfactual reasoning and belief ascription in terms of simulation theory, it is not the only interpretation compatible with this finding. Some authors, for instance Leslie (1987) or Perner (1991), are of the opinion that children's difficulty in the false belief task is a matter of their underdeveloped metarepresentational capacity, and this is believed to account for the developmental discrepancy between the two processes. Although this claim seems to account mainly for the developmental discrepancy between the two processes, rather than their kinship, it can also be reconciled with the correlation of the two tasks, as well as, which may be less obvious, with (a version of) the modified derivation hypothesis. According to the representational view on theory of mind, Sally's thought that Peter is at home is a primary representation. After Peter goes to the post office, Sally keeps representing Peter as being at home. Her belief becomes a misrepresentation. Young children's difficulties in the false belief task are explained in terms of their having problems with metarepresentations, that is with representing other agents' (mis)representations. In order to understand the latter, a person needs to be aware of a distinction between *what is represented* and *what it is represented as*, which mirrors the Fregean distinction between *reference* and *sense* (1892/1948). Metarepresentations encode both types of information: what an agent aims at representing and as what he or she represents it. In order to pass the false belief test, a child must be able to understand that in Sally's mind Peter's location is represented *as* home, although it really is the post office.

Pretence as well as thinking about past or future situations, non-existing objects, and hypothetical reasoning all involve representations of something which is not actual reality. But those representations are *decoupled* from reality (Leslie 1987; Perner 1991)—the link between what is represented and what it is represented as, if there is any at all, is here very loose. A two-year old child has no troubles in pretending that a banana is a telephone, and is able to switch between two representations. Contrary to more difficult cases of false belief understanding or counterfactual reasoning, a pretend world, which is in fact a representation of an alternative reality, is relatively unconstrained by the actual state of affairs. What is decoupled does not conflict with the actual state of affairs, and hence it does not contradict a child's knowledge.

Belief reports and other metarepresentations are *about* those decoupled representations, but they are not decoupled themselves. A person ascribing a false belief does not only construct a misrepresentation, that is a representation of an object or a situation as it is from the other person's mistaken perspective, but she also keeps in mind certain information concerning this representation, namely that it is someone else's and that it is mistaken.[9] Although counterfactuals are not usually considered as metarepresentations,

[9] The difference between pretence and belief ascription or counterfactual reasoning can be also easily explained within the simulation theory paradigm: the pretended world is what is simulated, and belief reports and counterfactuals are about the pretended, simulated worlds. For an interesting account based on similar observations see Recanati (2000).

they are structurally very much alike. An antecedent of a counterfactual conditional does not only express a proposition about a certain non-actual state of affairs, but it also conveys the information that this proposition is false. In other words, whenever a person reasons with counterfactuals, she represents what is expressed by the antecedent *as* conflicting with current reality. Moreover, the representation of a counterfactual state of affairs, both in the case of counterfactual conditionals and in the case of false beliefs, strongly depends on current reality. More precisely, the actual world imposes certain constraints on what the counterfactual representation of a given situation may be like. The constraints must be taken into account while constructing the counterfactual representation. The plausible way to construct such a representation is the mechanism of modified derivation that has been discussed above. Within the metarepresentational account this mechanism may be construed as a process of making hypothetical modifications of the agent's own representation, with everything else remaining the same. The output is a representation of the same object or situation, but as something different—as it would look like from another person's perspective or as if what is expressed by the antecedent of a counterfactual was true.[10]

Given that I am interested in belief ascription as a conscious process of generating explicit representations of other agent's states of mind, I am going to look at the modified derivation as a process of modifying one's primary representations. The modified derivation construed as such is still a kind of simulation, because it is meant to utilise our own cognitive resources for the purpose of modelling alternatives to reality or other people's beliefs. In view of the aforementioned connection between belief ascription and reasoning with counterfactual conditionals, it is worth investigating whether a theory of conditionals originating in the idea of Ramsey test, which is analogous to the concept of modified derivation, may elucidate the mechanism underlying the process of belief ascription.

2 The Ramsey test

Although the Ramsey test for conditionals was relegated to a footnote, it attracted a great deal of attention among philosophers and logicians. In his 1968 paper, Stalnaker developed Ramsey's idea into a full-fledged semantics for both indicative and subjunctive conditionals, based on the kind of possible worlds semantics proposed by Kripke (1963). According to Stalnaker's semantics, a conditional statement, "If p, then q", is true if and only if q holds in the nearest possible world in which p is the case. The nearest possible world differs from the actual one only with respect to the truth value of p and, if necessary, the values of other propositions that need to be changed in order to avoid inconsistency. Formally, there is a selection function that takes

[10] The role of the constraints that the actual world imposes on the possible alternatives in the process of counterfactual reasoning has been experimentally investigated in Rafetseder et al. (2010), Rafetseder and Perner (2010) and Perner and Rafetseder (2011).

as input an antecedent of a conditional and a world where the conditional is evaluated and outputs the nearest possible world or worlds in which the antecedent holds.

Its correspondence with Ramsey's idea is for present purposes the key advantage of Stalnaker's semantics, but there are also other, more pragmatic reasons to choose this kind of framework.[11] If we want to investigate the link or similarities between two phenomena or to provide a unified account for them, we need a framework suitable for both. In philosophy and philosophical logic there is already a long tradition of identifying propositions, and hence also beliefs, with sets of possible worlds. But the notion of a possible world has also been recognised as a handy tool in psychological inquiry, especially when it concerns people's hypothetical and counterfactual reasoning, that is their ability to think about or to mentally represent alternatives to reality (Rafetseder et al. 2010).

The theory of conditionals needs to be slightly adjusted so it will suit the purpose of modelling the process of belief ascription. Finding out what exactly must be changed should give us some insight into the relationship between counterfactual reasoning and belief ascription. It may help to elucidate to what extent they are alike, and why one of them seems to be easier than the other. The core of Stalnaker's semantics, the selection function picking out the nearest possible world, remains the same. Since we do not deal with conditionals any more, the crucial modification concerns the first argument of the selection function. In the case of belief reports, it is a set of propositions that an ascriber has to revise his own belief set with to take the observed agent's perspective. The main difference is not that it is a set, since if it is a finite set of propositions, we can represent it as a (sometimes very long) conjunction of these propositions.

The more significant difference between interpreting conditionals and belief reports is that in the second case we may need to undertake different epistemic actions with respect to different propositions in the aforementioned set. In the belief change literature it is customary to distinguish between three kinds of change in epistemic states, to wit: expansions, revisions and contractions (Gärdenfors 1988). An agent chooses an action depending on a relation between an input, that is some proposition p, and her prior belief set denoted here by $\mathbf{B}$. Expansion takes place whenever an agent simply learns a piece of information p which is consistent with $\mathbf{B}$. Contraction is a removal of a proposition p from $\mathbf{B}$ in a way that p cannot be re-inferred. Revision occurs whenever the new piece of information p is inconsistent with $\mathbf{B}$. In that case, the agent is supposed to accept p and make only those minimal changes in $\mathbf{B}$ that are necessary to maintain consistency. Hypothetical belief change taking place in the process of belief ascription may require revising $\mathbf{B}$ with respect to some propositions, other propositions may have to be added to $\mathbf{B}$, and still some others may have to be contracted from $\mathbf{B}$.

[11] In fact, Stalnaker's theory is not the only one that draws on the Ramsey test; see Bennett (2003) for an overview.

Different kinds of changes typically result in different epistemic states afterwards. We may have different attitudes towards a proposition p. Both before and after the belief change we may either accept p, reject p (that is accept $\neg p$), or we may have no opinion—be agnostic—about p. Moreover, propositions that belong to an agent's belief set may have different degrees of epistemic entrenchment. Whenever an agent must give up some of his previously accepted beliefs he can do that in many different ways. What determines which propositions he will give up easier than others is the ordering of epistemic entrenchment. More precisely, a belief φ is more entrenched than ψ in an agent's belief set $\mathbf{B}$ if the agent is more inclined to give up ψ rather than φ if he cannot stick to both. Logically stronger sentences are epistemically less entrenched, therefore they will be given up before the weaker ones. Tautologies will never be given up, hence they are the most entrenched sentences. What is less obvious, epistemic entrenchment of a sentence is independent from the probability assigned to it. Gärdenfors (1988, p. 87) explains it in the following way:

> If I have full belief in a sentence, that is, if it is accepted in my belief set, then I judge it to be maximally probable; but I do not regard all sentences that I accept as having equal epistemic entrenchment… Rather than being connected with probability, the epistemic entrenchment of a sentence is tied to its explanatory power and its overall informational value within the belief set.

For instance, an agent may be equally confident that his perception is reliable and also that inanimate objects do not levitate by themselves. If he then sees then a levitating table, he may give up his belief in the reliability of his perception more readily than the belief in the laws of physics. If so, then the latter is more entrenched than the former with respect to the agent's belief state, and the reason for that may be the role it plays in the agent's belief system.

When reasoning from another person's point of view, it may not suffice to hypothetically accept her beliefs and pretend they are all true, just because two belief sets may differ with respect to the orderings of epistemic entrenchment over the very same propositions. We may both believe that John is very knowledgable about the geography of Europe and that Eindhoven is a city in the Netherlands. But once we learn that according to John, Eindhoven is in Belgium and not in the Netherlands, I may give up my prior belief about the location of Eindhoven, and you may give up the belief about John's knowledge of geography. This is the reason why, especially in more complicated contexts, the ordering of epistemic entrenchment needs to be captured in an ascriber's hypothetically modified belief set. Again, some situations may not require such a sophisticated simulation, and the false belief task discussed in this paper is in fact one of them.

In the light of the above considerations, it should be clear that a factor responsible for simplicity or complexity of belief attribution is the context in which it occurs. Different contexts impose different constraints on the set of propositions with which we hypothetically revise our beliefs in order to "simulate" somebody else's perspective. This set, denoted here as $\mathbf{P}(c)$, is then a function of a context c in which beliefs are attributed.

In contrast to reasoning with conditionals, where our hypothetical belief change is based just on the antecedent of the given conditional, it is not always entirely obvious what propositions should fall into $\mathbf{P}(c)$ and how we learn about them. In simple scenarios, like the Post Office story, the set $\mathbf{P}(c)$ may be a singleton and contain the same proposition that is also expressed by an antecedent of a corresponding counterfactual conditional. Realising the constraints that a context imposes on the content of $\mathbf{P}(c)$ is a crucial part of the task. It cannot be just any set, but the smallest set of the propositions that will allow us to take the other person's perspective in the particular situation. When I am considering Anna's beliefs about the political situation in Poland, I am not interested in whether what she believes about dolphins is true. Even if we disagree with respect to, say, how intelligent dolphins are, I do not have to revise those beliefs in order to "simulate" her point of view on Polish politics. In addition, I do not even need to know whether we share any other beliefs. On top of that, Anna may have many beliefs that could have affected her opinion on the subject matter but are entirely obscure to me. The question then may arise how ascribing beliefs to others is even possible. Needless to say, belief ascription is not always accurate. However, if we had bothered ourselves about it every time we think about other people's beliefs, we would neither be able to interpret their behaviour nor simply communicate with them.

On that account, what has to underlie any theory of belief ascription that draws from the notion of a simulation is an assumption that our beliefs cannot be massively false, and hence that people share most of their beliefs with others. There are reasons to accept this assumption as uncontroversial. Donald Davidson pointed out that

> a correct understanding of the speech, beliefs, desires, intentions, and other propositional attitudes of a person leads to the conclusion that most of a person's beliefs must be true, and so there is a legitimate presumption that any one of them, if it coheres with most of the rest, is true (Davidson 2001, p. 146).

Everyday communication is grounded in the (supposedly shared) assumption that other people have roughly the same background knowledge as we do. We believe that our beliefs are (mostly) true and we tend to assume other people's beliefs to be true as well. What governs our interpretation of other people's speech or behaviour is, as Davidson puts it, the Principle of Charity.[12] We do not ascribe false beliefs to others as long as we have no evidence that they

[12] The Principle of Charity was conceived as a methodological postulate. However, data on belief ascription and modified derivation may be construed as providing empirical support for the claim that charity is a fairly accurate explanation of how people manage to understand each other.

are actually wrong. For the same reason—as I shall argue in the following section—we are disinclined to attribute contradictory beliefs to anyone.

3 Towards a formal analysis of belief ascription

As already mentioned in the previous section, what counterfactual reasoning and reasoning about other people's mistaken beliefs[13] have in common is that they both involve thinking about different possibilities, that is different ways the world might be or might have been. The antecedent of a counterfactual conditional usually explicitly expresses what a speaker believes to be false. Therefore, a counterfactual statement makes a claim about a certain non-actual state of affair, that is about something that holds in another possible world. What might seem less obvious, thinking about another person's false beliefs is considering an alternative to our own view on how the world is. Even if we are convinced that the person's beliefs are entirely mistaken, usually we should not have troubles with acknowledging that what is compatible with her beliefs is a way the world might have been. It is worth noting that despite the metaphysical associations they seem to bear, possible worlds can be conceived simply as different ways *the* world might be or might have been. In Stalnaker's view, they are

> what people distinguish between in their rational activities. To believe in possible worlds is to believe only that those activities have a certain structure, the structure which possible worlds theory helps to bring out (Stalnaker 1984, p. 57).

Analysing belief ascription in terms of possible worlds semantics is hence a natural way to investigate the structure that reasoning about beliefs and reasoning about counterfactual situations most likely share.

The possible worlds model for belief ascription that I want to propose is a quadruple $\mathcal{M} = \langle W, R, s, v \rangle$ where:

- W denotes a set of all possible worlds;
- R is a set of relative possibility relations $R_a \in W \times W$ for each agent $a \in A$, where A is a finite set of agents to whom beliefs can be attributed;
- s is a partial world selection function that maps worlds and propositions $p \in \wp(W)$, that is sets of worlds, into subsets of W;
- v is a valuation function that assigns truth values to propositions in particular worlds.

[13] In the present paper, wherever I use the notion of false beliefs, I mean beliefs that are false or wrong from the ascriber's perspective. This qualification is taken as read from here on.

The set of truth values consists of three elements: $\{1, 0, u\}$, where the third value should be read as "undefined" or "undecided."[14] It reflects the possibility that an agent may be agnostic about a proposition.

The notion of relative possibility deserves some explanation. Roughly speaking, an agent's beliefs determine a set of possible worlds that are compatible with his belief state, that is worlds an agent considers to be candidates for the actual world. An agent believes that one of those worlds is the actual one, but since his beliefs are incomplete, some worlds are indistinguishable for him. This is the standard interpretation of the accessibility relation in semantics for epistemic logic: a world w' is accessible from a world w if and only if the two worlds w and w' are indistinguishable (from the agent's perspective). In the context of counterfactual reasoning another notion of accessibility is needed. Here, even if an agent knows that p, he may hold that it might have been the case that $\neg p$. Therefore, if w is one of the worlds compatible with an agent's belief state, then $\langle w, w' \rangle \in R$ means that a world w' is a possible alternative to w. In other words, even if w' is excluded as a candidate for the actual world, it *might have been* the case that the actual world was w'. Given that reasoning about other agents' beliefs requires considering what might have been the case from *their* perspective and given that, as mentioned in the previous section, different agents may have different orderings of the epistemic entrenchment, each agent a must be simulated by means of a designated R_a. The only constraint that has to be imposed on R_a is reflexivity: every possible world w is possible with respect to itself.

In his 1968 paper Stalnaker introduces a notion of an absurd world in which every sentence is true. This is an impossible world, which Stalnaker needed for the interpretation of conditional sentences with impossible antecedents, for instance: "If circles were square, I could win the Olympic Games." Nute and Cross (2002) point out that this complication is not necessary since we may define s as a partial world selection function. Then if for some antecedent p of a conditional and a world w the value of the selection function is not defined, then p is not true in any world possible with respect to w.

A partial world selection function s in the model of belief ascription plays a role which is analogous to the role it plays in Stalnaker's account of conditionals except that it takes as its input a set of propositions $\mathbf{P}(c)$ and a possible world, and outputs a set of possible worlds. For every agent a, every context c, and every world $w \in W$, there is an S such that $s(\mathbf{P}(c), w) = S \subseteq W$ and for all $w' \in S$: $\langle w, w' \rangle \in R_a$. S is a selected set of nearest possible worlds that are compatible with the ascriber's hypothetically revised belief set. If an ascriber does not need to revise his own beliefs to take another agent's perspective, the set $\mathbf{P}(c)$ is empty and hence $s(\mathbf{P}(c), w) = \mathbf{B}$, where $\mathbf{B}$ is the set of ascriber's beliefs.

We say that the belief report "a believes that p" is true in a world w if and only if p is true in all possible worlds w' in the selected set S that is the

[14] For the present purpose, it is irrelevant whether we think of u as a third value or as indicating the lack of a truth value.

value of $s(\mathbf{P}(c), w)$. Formally:

$$v(\text{“}a \text{ believes that } p\text{”}, w) = 1 \quad \text{iff} \quad \text{for all } w' \in S\colon v(p, w') = 1.$$

Conversely, we say that the belief report “a believes that p” is false in a world w if and only if p is false in all possible worlds w' in the selected set S that is the value of $s(\mathbf{P}(c), w)$:

$$v(\text{“}a \text{ believes that } p\text{”}, w) = 0 \quad \text{iff} \quad \text{for some } w' \in S\colon v(p, w') = 0.$$

A belief report is evaluated from the perspective of an ascriber, who may be ignorant about many of a's beliefs and also well aware of his own ignorance. He may, for instance, be undecided as to whether p holds in the closest possible worlds selected in the process of simulation. As long as we are interested in actual people, who by no means are omniscient and perfect reasoners, an option of the truth-value being undefined is essential. When it is the case depends on some extralogical factors. One possibility is that v outputs u as a value when the partial selection function is undefined for a given set of propositions in a world w, that is when the ascriber cannot make sense of someone's beliefs:

$$v(\text{“}a \text{ believes that } p\text{”}, w) = u \quad \text{iff} \quad s(\mathbf{P}(c), w) = u.$$

In Stalnaker's theory, conditionals with impossible antecedents are always trivially true.[15] In a model of belief ascription a corresponding clause would not be a desirable feature; even if we find out that some of an agent's beliefs are inconsistent, we usually do not want to claim that for any proposition he believes both this proposition and its negation. We would rather try to make sense of his beliefs, so the contradiction would turn out to be only apparent. Let me explain it using an example.

Anna has been accused of plagiarism. She copied someone's text, pasted it into her own article without any quotation marks or a reference, and published the article. After being caught, she denied that she plagiarised, although she admitted that she copied and pasted a text that she was not an author of and that she did not refer to the source. Assume that I am convinced that she is not lying (say, she was examined by a lie detector) and that she truly believes what she says. Since I am also convinced that copying and pasting someone else's work is an obvious case of plagiarism, it seems that from my point of view Anna holds contradictory beliefs:

(2) Anna believes that she committed plagiarism and that she did not commit plagiarism.

In order to interpret a belief report of the form “Anna believes that p and that not p” in the most straightforward way, I would need to make all the

[15] Within a simplified version of Stalnaker model theory, the truth conditions for a conditional are the following: $v(\text{“If } p\text{, then } q\text{”}, w) = 1$ if and only if $v(q, s(p, w)) = 1$ or $s(p, w)$ is undefined (Nute and Cross 2002).

necessary changes in the stock of my beliefs that will let me hypothetically believe the contradiction. But if I do that, I am allowed to attribute to Anna virtually any belief I can think of. This is not really a desirable result. I do not want to say that, for any proposition, Anna believes both that proposition and its negation. Even if she has some inconsistent beliefs, she may still maintain that Warsaw is the capital of Poland, and deny the opposite claim at the same time. Therefore, keeping Davidson's plea for charity in mind, we should try to explicate why she claims that she did and did not commit plagiarism. For instance, it might be simply the case that Anna understands the word "plagiarism" differently, that is, her definition of "plagiarism" is different from mine, and, given that definition, the belief she would express by the sentence "I did not commit plagiarism" is not inconsistent with her belief that she committed something that I call plagiarism. There is also a possibility that when asked about the definition of plagiarism, Anna will give an answer that is correct in my view: "plagiarism is identical to P." In order to rationalise or explain her beliefs in a charitable way, we need to refer to something else than the mistaken definition of plagiarism. The best candidate is then an inference failure. Normally, from the premises: "plagiarism is identical to P" and "I commited P," by substitution of identicals we infer "I committed plagiarism." But human beings are never perfect reasoners; we sometimes fail to see relations between our own beliefs, and thus we are not aware of all their consequences. On the other hand, if we see that other people's reasoning is incorrect, we may be able to understand where they made mistakes. In daily practice, we do not take seemingly contradictory beliefs readily at face value. It is plausible then that s will never be undefined. We may always be able to find a way to avoid attributing a contradiction.

Nonetheless, it may also happen that despite our best efforts we are not able to make sense of Anna's inconsistent beliefs. It still does not seem psychologically plausible to claim that she believes everything. If I totally fail to understand her, I may be inclined to refuse to think about her beliefs at all. I simply do not know how to interpret her, thus it may be reasonable to say that for any proposition p, $v(\text{“Anna believes that } p\text{”}, w) = u$. As I have already mentioned, evaluation of some belief reports depends on various extralogical factors. The formal account reaches its limits notably when beliefs to be attributed are inconsistent. Then we can reason only case by case.

4 What is happening in the Post Office story?

The analysis is supposed to capture the process of belief ascription from the first person perspective, or, in other words, it is supposed to model what is happening in the ascriber's mind. In the Post Office story, this is a subject participating in the experiment who ascribes beliefs to Sally, and whose ascription is modelled. If the ascriber is wondering what Sally believes about Peter's location, or whether it is true that Sally believes that Peter is at home, she needs to identify those propositions—relevant to the one at issue—towards which she has a different attitude than Sally. Arguably, there may be many different ways to do the hypothetical revision suitable for the given purpose. For instance, since the ascriber knows that Sally was not at home when there was a call about the fire, he may start the hypothetical revision with the proposition:

(3) Peter was called and asked for help with the fire at the Post office.

But (3) presupposes:

(4) There is a fire at the Post office.

which would have to be contracted as well, for if Sally knows that there is a fire, she may expect the fireman Peter to be called and asked for help. Thus she may no longer be entirely sure, that he is in bed, and so the ascriber will not be entirely sure about her beliefs.

In the present context—call it *post*—what definitely differentiates the ascriber's and Sally's perspectives is the proposition (4). Although it is not the only information that does not belong to Sally's belief set, the hypothetical contraction of it from the ascriber's own beliefs is sufficient for the simulation of her perspective, because it will trigger the removal of other relevant ones. Hence (4) is the only information that must fall into the set $\mathbf{P}(\mathit{post})$. The ascriber has to hypothetically contract (4) from her own beliefs, together with all those propositions that, if not removed, may be used to reconstruct the ascriber's primary belief set. For the ascriber knows that:

(5) Peter was called and asked for help, went out of bed and went to the Post Office, *because* of the fire.[16]

in order to maintain consistency the propositions:

(6) Peter was called and asked for help with putting out the fire.

(7) Peter went out of bed and went to the Post Office.

must be removed as well. The resulting, hypothetical belief state is compatible with a certain subset of W, namely, the set of worlds in which it is not the case that there is a fire at the post office. Those worlds are not equally plausible though. We can easily imagine a situation in which there is no fire, but ill Peter was kidnapped and hidden at the post office. Nevertheless, the statement "If there had been no fire, Peter would have been kidnapped and kept at the post office" does not sound very convincing. We would not like to attribute to Sally a belief in such a scenario, and the reason for that is the fact that of the worlds in which there is no fire the "kidnapping-worlds"

[16] Counterfactual and causal reasoning are closely related (Byrne 2002), and in fact counterfactual conditionals play an important role in philosophical analyses of causality (Lewis 1973). The statement: "If there had been no fire, Peter would not go out of bed" seems to follow directly from (5), and hence a person accepting the former should also accept the latter.

are not the closest.[17] In the closest possible worlds in which (4) is not the case, there is no reason for Peter to go out of bed. Taking into account the possibility of kidnapping would require expanding the set $\mathbf{P}(post)$ with some additional propositions entirely *ad hoc*, and then making many more hypothetical changes in the belief set. We may also be able to find a story that would explain Sally's knowledge about Peter actually being at the Post Office. Obviously, someone could have told her about the fire at the Post Office, but this possibility might be easily ruled out if the scenario used in the experiment is specific enough. She might also have, for instance, some sort of telepathic talent, which allows her to track Peter's location regardless where she is herself. Yet, this is not a scenario we would like to treat seriously.[18]

The table below illustrates the process we need to go through when, trying to predict Sally's behaviour or simply answering the questions asked in the experiment, we are thinking about her beliefs or about the counterfactual situation they concern.

<table>
<tr><th>In the actual world w</th><th>In the closest possible worlds $s(\mathbf{P}(post), w)$</th></tr>
<tr><td colspan="2">Peter is ill.
Peter is in bed when Sally goes out.</td></tr>
<tr><td>There is a fire at the Post Office.
Peter was called and asked for help with putting out the fire.
Peter went to the Post Office.</td><td></td></tr>
<tr><td>Peter is at the Post Office.</td><td>Peter is in bed.</td></tr>
</table>

Belief ascription and counterfactual reasoning may be seen as a matter of a correct information retrieval. As long as I do not have any reason to think that my own and Sally's beliefs are different, I assume that we share the perspective. This is why, until the moment when Sally goes out, there is no distinction between the w-column and $s(\mathbf{P}(post), w)$-column in the table. Since Sally has no idea about the things I am learning about the fire, starting with "There is a fire at the Post Office" the representation of the situation from my perspective is different than the representation of it from Sally's perspective. While I am filling out the w-column of the table with new pieces of information, Sally's $s(\mathbf{P}(post), w)$-column remains the same. In order to answer the test questions a subject needs to look into the proper column of such a table.

The fact that the counterfactual task seems to be easier than the false belief task, at least for some children, does not have to be explained, as Peterson and Riggs suggested, in terms of a theoretical understanding of mental states that belief ascription would additionally require. On the basis of the analysis

[17] Although the notion of "closeness" or "similarity" of worlds is blatantly imprecise, as Evans and Over (2004) report, people do have strong intuitions about which possibilities are closer than the others to the way things actually are (p. 117).

[18] In the terminology of Lewis' (1996) this would be one of those possibilities that we are entitled to ignore.

I presented above, it should now be clear where the difference might come from. Although the two processes share the underlying simulation mechanism, this simulation requires an input which might be identified differently in different tasks. In the counterfactual task, the input for the simulation is provided by the antecedent of the counterfactual question that a child is asked. In the false belief task, though, the input for the process of simulation is much less evident, and a child needs to identify it herself on a basis of contextual cues. To put it differently, the two tasks diverge with respect to the role of context, which does not interfere with the counterfactual task, but does strongly affect belief ascription. Belief attribution may require increased processing capacities, because the child does not only have to keep in mind different possibilities, but also identify the reason for those different possibilities to occur, that is the propositions that she has to hypothetically revise her own beliefs with. In simple, well-delineated situations like the one used in the experiment, it is typically easy to enumerate those propositions, although they are not explicitly given. In the Post Office story, this is only one proposition; the very same proposition is expressed by the antecedent of the related counterfactual conditional. Therefore, most subjects who are able to answer the counterfactual question correctly have also little trouble detecting the differences between their own and the protagonist's perspective, and, as a result, they give the right answer to the false belief question as well. Nevertheless, input for the simulation in the counterfactual case is prompted by the very question a child is supposed to answer, whereas, in the false belief task, a child needs to figure it out herself, which accounts for the increased complexity of the task.

The analysis sketched above does not pretend to be a full-fledged model of belief ascription. Yet, I hope it makes it easier to see that the additional level of difficulty in the false belief ascription as opposed to the counterfactual reasoning can be explained in a simple and modest way, without appealing to anything specifically mental. Moreover, I believe that the considerations contained in this paper could serve as a departure point for a formal and psychologically plausible theory of belief ascription that could possibly yield some interesting empirical predictions as well as give new insight into many aspects of mindreading that are not as yet fully understood.

References

Amsterlaw, J. and Wellman, H. M. (2006), 'A microgenetic study of the development of false belief understanding', *Journal of Cognition and Development* **7**(2), 139–172.

Apperly, I. A. (2011), *Mindreaders. The Cognitive Basis of "Theory of Mind"*, Psychology Press.

Apperly, I. A., Riggs, K. J., Simpson, A., Chiavarino, C. and Samson, D. (2006), 'Is belief reasoning automatic', *Psychological Science* **17**(10), 841–844.

Baillargeon, R., Scott, R. M. and He, Z. (2010), 'False-belief understanding in infants', *Trends in Cognitive Sciences* **14**(3), 10–18.

Baron-Cohen, S. (1995), *Mindblindness. An Essay on Autism and Theory of Mind*, The MIT Press.

Baron-Cohen, S., Leslie, A. M. and Frith, U. (1985), 'Does the autistic child have a "theory of mind"?', *Cognition* **21**, 37–46.

Bennett, J. (2003), *A Philosophical Guide to Conditionals*, Oxford University Press, Oxford.

Bloom, P. and German, T. P. (2000), 'Two reasons to abandon the false belief task as a test of theory of mind', *Cognition* **77**, 25–31.

Byrne, R. M. (2002), 'Mental models and counterfactual thoughts about what might have been', *Trends in Cognitive Sciences* **6**(10), 426–431.

Carruthers, P. and Smith, P. K., eds (1996), *Theories of Theories of Mind*, Cambridge University Press.

Davidson, D. (2001), A coherence theory of truth and knowledge, *in* 'Subjective, Intersubjective, Objective', Oxford University Press, pp. 137–153.

Drayton, S., Turley-Ames, K. J. and Guajardo, N. R. (2011), 'Counterfactual thinking and false belief: The role of executive function', *Journal of Experimental Child Psychology* **108**, 532–548.

Evans, J. S. B. T. and Over, D. E. (2004), *If*, Oxford University Press, Oxford.

Frege, G. (1892/1948), 'Sense and reference', *The Philosophical Review* **57**(3), 209–230.

Gärdenfors, P. (1988), *Knowledge in Flux. Modelling Dynamics of Epistemic States*, The MIT Press.

German, T. P. and Nichols, S. (2003), 'Children's counterfactual inferences about long and short causal chains', *Developmental Science* **6**, 514–523.

Grant, C. M., Riggs, K. J. and Boucher, J. (2004), 'Counterfactual and mental state reasoning in children with autism', *Journal of Autism and Developmental Disorders* **34**(2), 177–188.

Guajardo, N. R., Parker, J. and Turley-Ames, K. (2009), 'Associations among false belief understanding, counterfactual reasoning, and executive function', *British Journal of Developmental Psychology* **27**, 681–702.

Guajardo, N. R. and Turley-Ames, K. (2004), 'Preschooler's generation of different types of counterfactual statements and theory of mind understanding', *Cognitive Development* **19**, 53–80.

Kripke, S. A. (1963), 'Semantical consideration on modal logic', *Acta Philosophica Fennica* **16**, 83–94.

Leslie, A. M. (1987), 'Pretense and representation: The origins of "theory of mind"', *Psychological Review* **94**(4), 412–426.

Lewis, D. (1973), 'Causation', *The Journal of Philosophy* **70**(17), 556–567.

Lewis, D. (1996), 'Elusive knowledge', *Australasian Journal of Philosophy* **74**(4), 549–567.

Nute, D. and Cross, C. B. (2002), Conditional logic, *in* D. M. Gabbay and F. Guenthner, eds, 'Handbook of Philosophical Logic', 2nd edn, Vol. 4, Springer, pp. 1 – 98.

Perner, J. (1991), *Understanding the Representational Mind*, The MIT Press.

Perner, J. and Rafetseder, E. (2011), Counterfactual and other forms of conditional reasoning: Children lost in the nearest possible world., *in* C. Hoerl, T. McCormack and S. Beck, eds, 'Understanding Counterfactuals, Understanding Causation. Issues in Philosophy and Psychology', Oxford University Press, in press.

Perner, J., Sprung, M. and Steinkogler, B. (2004), 'Counterfactual conditionals and false belief: a developmental dissociation', *Cognitive Development* **19**, 179–201.

Peterson, D. M. and Riggs, K. J. (1999), 'Adaptive modelling and mindreading', *Mind & Language* **14**(1), 80–112.

Premack, D. and Woodruff, G. (1978), 'Does the chimpanzee have a theory of mind?', *The Behavioral and Brain Sciences* **1**, 512–26.

Rafetseder, E., Cristi-Vargas, R. and Perner, J. (2010), 'Counterfactual reasoning: Developing a sense of "nearest possible world"', *Child Development* **81**(1), 376–389.

Rafetseder, E. and Perner, J. (2010), 'Is reasoning from counterfactual antecedents evidence for counterfactual reasoning?', *Thinking & Reasoning* **16**(2), 131–155.

Ramsey, F. P. (1929/1990), General propositions and causality, *in* D. H. Mellor, ed., 'Philosophical Papers', Cambridge University Press, pp. 145–163. (original publication, 1929).

Recanati, F. (2000), *Oratio Obliqua, Oratio Recta. An Essay on Metarepresentation*, The MIT Press.

Riggs, K. J., Peterson, D. M., Robinson, E. J. and Mitchell, P. (1998), 'Are errors in false belief tasks symptomatic of a broader difficulty with counterfactuality?', *Cognitive Development* **13**, 73–90.

Stalnaker, R. C. (1968), A theory of conditionals, *in* N. Rescher, ed., 'Studies in Logical Theory, American Philosophical Quarterly', Monograph Series, 2, Oxford: Blackwell, pp. 98–112.

Stalnaker, R. C. (1984), *Inquiry*, The MIT Press.

Wimmer, H. and Perner, J. (1983), 'Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception.', *Cognition* **13**(1), 103–128.