Editor's Introduction

Marta Jorba and Sergi Oms
University of Barcelona

This special issue of *Disputatio* has its origins in the thirteenth *Taller d'Investigació en Filosofia* (TIF) which took place from January 12th to 14th at the University of Barcelona. This annual conference was first held at the University of Barcelona in 1999, and each year it has been hosted by a different university: the Autonomous University of Barcelona, University of Barcelona, University of Girona, University Rovira i Virgili and University of València. The main distinguishing features of the TIF are its organization by a committee of graduate and postgraduate philosophers from the universities mentioned and the will to create a philosophy conference for students who are either doing their PhD or have read their thesis in the last three years. The initial objective of the TIF was to bring together graduate students from Catalan or Spanish universities in order to discuss their work; since then it has grown considerably, receiving participants from all Europe and, in the last few years, from all over the world.

The XIII TIF was held over three days and covered a wide range of topics: epistemology, philosophy of language, history of philosophy, aesthetics, philosophical logic and philosophy of mind. A remarkably large number of high quality submissions were received from many different universities, an indication of the recognition that the TIF currently enjoys.

This volume contains four of the ten papers presented in the conference. In 'Do honeybees have concepts?' Bernardo Aguilera, from the University of Sheffield, presents Peter Carruthers' version of the view that animals can think and argues that this view fails to provide convincing grounds for accepting concept possession in animals. He suggests that conceptual thought involves two constraints, namely, concept individuation and the generality constraint, which are not satisfied in Carruthers' account.

Marilia Espirito Santo, from the University Federal do Rio Grande do Sul, is the author of the paper 'On the transcendental deduction in

Kant's *Groundwork* III'. She deals with the third section of Kant's *Groundwork,* which aims to prove the possibility of the categorical imperative. She presents the argument as a transcendental deduction and discusses Henry Allison's reconstruction of it, arguing that this reconstruction could not have been accepted by Kant himself.

In 'Might-counterfactuals and the principle of conditional excluded middle', Ivar Hannikainen, from the University of Sheffield, shows the compatibility of ontic might-counterfactuals and the principle of conditional excluded middle. He does so by arguing for the semantic underdetermination of the antecedent of a might-counterfactual with respect to the counterfactual worlds it selects for evaluation.

Thomas Raleigh, from the National Autonomous University of Mexico (UNAM), provides the paper 'Visual Experience and Demonstrative Thought', where he presents a problem for common-factor theories of experience concerning the account they give of demonstrative thought. Building an argument based on a claim by Snowdon 1992, he concludes that such theories are committed to attributing quite widespread referential errors or failures amongst the non-philosophical population. After addressing some possible objections, he argues that the argument applies to any common-factor theory.

This is the second time the contributions of speakers at TIF workshops have found their way to publication. Some of the papers from the twelfth edition were edited by Mireia López and published in *Theoria. An international Journal for Theory, History and Foundations of Science* in 2010, vol. 25, 2. We wish to thank her for her bold decision to publish the proceedings of this conference in a philosophy journal. Her success in doing so and our desire to continue the tradition, led us to take on the present project.

We are indebted to many people. We would like to thank Miguel Ángel Sebastián, who co-organized the conference with us, for all his efforts to ensure that the conference was a success. We also acknowledge the contributions of all the people who have been in the TIF since its foundation in 1999 – the organizing and scientific committees and the participants in the workshop in previous years – who helped to create a highly supportive atmosphere. Our special thanks go to this year's scientific committee, for agreeing to assess the high volume of submissions we received, and, in some cases, for their kindness in reviewing extra papers. Thank you to all the members of LOGOS for giving support and continuity to this graduate conference and for their readiness to help. We also appreciate the commitment

of all the speakers and commentators at the conference, who contributed to the lively philosophical atmosphere over the three days. Special thanks go to the contributors of this issue for agreeing to cooperate in the project and for their patience. We also acknowledge the valuable comments of the reviewers which helped to improve the papers presented in this volume.

This conference received financial support from the following organisms, which we are also pleased to acknowledge: the LOGOS group (Logic, Language and Cognition Research Group), the Faculty of Philosophy of the University of Barcelona, the University of Barcelona and the Spanish Ministry of Science and Innovation.

We would like to conclude thanking the editorial committee of *Disputatio,* and Teresa Marques in particular, for their willingness to publish this issue.

Do honeybees have concepts?

Bernardo Aguilera Dreyse
University of Sheffield

Abstract
Can animals think? In this paper I address the proposal that many animals, including insects such as honeybees, have genuine thoughts. I consider one prominent version of this view (Carruthers 2004; 2006) that claims that honeybees can represent and process information about their environments in a way that satisfies the main hallmarks of human conceptual thought. I shall argue, however, that this view fails to provide convincing grounds for accepting that animals possess concepts. More precisely, I suggest that two important aspects of conceptual thought, viz., concept individuation and the generality constraint, are not satisfied.

Keywords
Animal cognition, concepts, modularity, concept individuation, generality constraint.

1. Introduction

The long standing debate about whether animals can think has been fuelled in recent years by scientific research that provides strong evidence that the cognitive processes of many animal species are computational. Behaviours that at first sight could appear as immediate responses to environmental contingencies, after careful observation and experimental procedures have shown to be governed by structured representations following complex computational algorithms. Examples come from diverse parts of the animal kingdom, including insects, birds and mammals.

These findings have been received with enthusiasm by advocates of a particular view of animal cognition (Carruthers 2006), which embraces a version of the computational theory of mind (CTM) and the massive modularity hypothesis (MMH). In short, this view holds that animals process information about their environment by means of computation, and that their cognitive architecture is mainly modularised into domain specific processors. In this paper I shall discuss the

claim, put forward by proponents of this view, that most animals have concepts[1], and that those concepts can combine forming (propositional) thoughts. In particular, I deal with the case of honeybees, for two reasons. First, because their behaviour has been extensively studied and there is general consent that they carry out computations, and secondly because it is particularly provocative to suggest that they can think. If it turns out that honeybees have concepts and thoughts, it appears convincing that this capacity is widespread in the animal kingdom.

In this paper I shall assume that CTM gives a plausible account of human concepts and thoughts and, likewise, that there are no reasons in principle to reject the idea that cognitive architecture is massively modular. However, I will argue that when these ideas are deployed for the case of honeybee cognition, in the way put forward by Carruthers, they fail to provide convincing grounds for the possession of concepts. More precisely, I shall claim that two main features of conceptual thought, i.e. concept individuation and the generality constraint, are not always satisfied.

This paper is structured in the following way. In section 2 I sketch the basic tenets of CTM and MMH, and in section 3 I explain how CTM gives a plausible account of conceptual thought, addressing some common objections. These sections are intended to give a general background about the views that have inspired the proposal about animal cognition that is criticised further on in the paper. Section 4 gives a brief exposition of current research in honeybee behaviour in order to make clear how it strongly suggests that they are computational systems. In section 5 I explain the claims put forward by authors who interpret this symbolic processing as a form of conceptual thought, and then in sections 6 and 7 I present my arguments against that view. Lastly, section 8 gives some final remarks.

[1] A terminological note: In philosophical usage, 'concepts' are generally understood as abstract entities, however in psychology the term is used to designate mental representations (Margolis & Laurence 2007). In this paper I will follow the psychological usage, but understanding particular mental representations as concept *tokens*, that instantiate mental representation *types*. Whenever I refer to concept tokens I shall use italics, and when referring to concept types I shall write them in capitals.

2. Concepts, Computation and Modularity

The mind provides us with a meaningful perspective of the world, and much of that job is carried out through our capacity to conceptualise what our perceptual systems bring to our minds (cf. Crane 2001). Concepts are the main building blocks of our world view and therefore they are generally considered as one of the main features of the mind. Concepts also make up thoughts, which allow us to have mental states that interact causally to produce intelligent behaviour. I leave open the possibility that a mind could be defined by states that are non-conceptual, such as perceptions or emotions. However, for the purposes of this paper I shall be interested in the common view that minded creatures can think, and that thoughts are constituted by concepts.

This view is compatible with CTM, which has been the dominant approach of how the mind works over the past four decades. The CTM has two basic tenets. One is that the mind is a representational system. That means that the mind picks up information about the environment and encodes it as mental representations. This information is made available by perceptual systems, and can be stored in memory for future processing (Sterelny 1990). The second tenet is that mental representations are processed following computational (i.e. algorithmic) steps. This means that these processes are performed in ways only responsive to the formal properties of the representational states, whilst their contents (i.e. what they mean) are preserved along the computational steps without playing any causal role in the process (Haugeland 1981). One of the most influential articulations of CTM has been by Fodor (1975; 1987). Since he is often quoted by the proponents of animal cognition, I will focus on his account of CTM for the rest of this paper.

One of the main contributions by Fodor to CTM was to make explicit the idea that the mind must have an inner medium of representation that carries out the computations. He also claimed that the properties of productivity and systematicity of thought could only be explained if this inner medium has the compositional structure of a language. For that reason, he proposed that the mind has a language of thought (LOT). According to this view, thinking consists in entertaining sentences in LOT. Words in LOT express concepts, and sentences express propositions. LOT is where thought and its properties (i.e. syntactic and semantic) are situated. Its basic structure is

supposed to be innate, and thus not dependent on learning a language. As noted above, this idea has important implications for animal cognition, since it states that thinking is not a capacity derived from the possession of a natural language, leaving open the possibility that non-linguistic creatures could think.

However, it is important to be careful when attributing concepts to animals. For example, pigeons can be trained to sort pictures into categories of tree or person, but these findings do not warrant the conclusion that they have concepts. Pigeons may be just grouping together common visual elements into a single internal representation, without being able to make further recognitional distinctions and inferences that are characteristic of possessing abstract concepts such as those of a tree or a person (Allen & Hauser 1996). I shall say more about these capacities and the individuation of concepts in the next section. For the moment, it suffices to say that the view criticised in this paper claims that some of the internal representations of honeybees are not just trivial forms of information processing, but also meet some of the relevant criteria for concept attribution.

The MMH is a claim about cognitive architecture. The main idea is that the mind does not work as a single, domain-general system, but has several functionally distinguishable modules that process domain-specific information and work quickly and rather isolated from one another. Initial accounts of cognitive modules restricted their processing to perceptual and motor information. However, proponents of MMH have proposed that mental processes involving thoughts and reasoning are also modular (Cosmides & Tooby 1994, Pinker 1997).

According to MMH, the modular parcellation of cognitive capacities constitutes an extremely common evolutionary feature that enhanced the adaptability of organisms by permitting them to deal more efficiently with their environments (Carruthers 2006). That explains why the animal mind is supposed to be massively modular. Some empirical evidence has been put forward to defend this claim. To give one example, the navigational capacities of many animals, including rats and birds, have been shown to be modular (see Shettleworth 1998 for a review). They have been studied in artificial environments that offer limited kinds of information that can be used by them to orientate. Animals proved to be able, not only to use these different environmental clues to navigate, but to deploy them in a way that requires computation, such as vector addition or template matching. However, some kinds of information appear to be per-

ceived and used independently, without the capacity to integrate it with other visual clues. All this suggests that they process the various kinds of spatial information by dedicated cognitive modules, that exhibit the hallmarks of domain-specificity, computational processing and isolation.

3. Conceptual thought in CTM: content and individuation of concepts

According to CTM, thoughts are sentences in LOT and concepts are the elements from which they are constructed. When an agent is thinking, chains of propositions are tokened in her mind, one leading to the other following algorithmic steps that are sensitive to the syntactic properties of LOT. So, for example, an agent could think:

When it's raining, there are no rabbits in the meadow
Now it's raining
So, there are no rabbits in the meadow

Here, the propositions have a syntactic structure that can be recognised by the system (i.e. the brain) as an instance of *modus ponens*, and then processed in a way that mirrors the logical structure of the argument. The thought can be carried out mechanically, independently of the content of the concepts involved in it. This suggests that thought can be viewed as a purely syntactic procedure.

This idea of mechanised thinking is at the core of CTM. It has many advantages, one of them being how the logical structure of reasoning could be implemented in a digital computer. However, it has the counterintuitive consequence that what-the-thought-is-about does not appear to play any causal role in the thinking process. In the previous example, we could replace *rabbits* for *foxes*, and the thinking process will still be the same (i.e. an instance of *modus ponens* specifiable purely by syntax), however, of course it is not the same to think about rabbits as it is to think about foxes. Moreover, how could a creature build up a perspective on the world with a representational system that is purely syntactic? How could concepts be regarded as meaningful then?

The response of CTM to these questions is that they simply never said that semantics could be ignored, or that it could be reduced to syntax (Horst 2009). Most proponents of CTM have divorced them-

selves from an extreme view about the 'autonomy' of a syntactic language, perhaps advocated by early developers of artificial intelligence, a view that would allow concepts to become meaningful just in virtue of local intra-linguistic manipulations. Instead, they claim that the mind is a large syntactic system[2] capable of interacting causally with the environment, in a way that can be described as denoting objects and properties in the world (Rey 1997, Haugeland 2003). On his view about animal cognition, Carruthers (2004; 2006) seems to agree with these constraints on being a genuine thinker, adding the idea that this syntactic system should mirror a belief/desire cognitive architecture.

Returning to the issue about the meaning of a concept, CTM claims that it is determined by its content, which is fixed from 'outside' the domain of thought by the input and output causal relations that concept has with the external world. For example, what makes an agent to instantiate the concept of RABBIT is that she has been caused to think about rabbits every time there has been a causal connection between rabbits and her perceptual systems. In other words, the interaction of the agent with the word gives the concepts meaning, which is then preserved along the computational processes where the concept takes part. This account of conceptual content is usually called 'causal theory of content'. There is controversy about how to precisely determine content, and several theories are available. However, for the purposes of this paper, suffice it to say that causal factors in determining content are dominant among theorists of CTM (Rey 1997).

It is important to note that even though conceptual content is independent from the syntactic structure of thought, this is not the case with concept individuation. Two concepts may share their contents (i.e. have the same extension), but differ in two further aspects. One of them is the expression in LOT where they are instantiated, what is usually called the 'mode of presentation'. For example, an agent could think about water both tokening the expression *water* and H_2O[3], which constitute two modes of presentation for the same content. A

[2] In fact, this can be considered a version of what Searle 1980 calls 'the systems reply' to his famous Chinese Room argument.

[3] To simplify the exposition, I am using English words as expressions of LOT. But LOT, at least according to CTM, does not correspond to any natural language.

second aspect is the inferential role of these expressions. Two concepts can share the same content, though differing in their causal effects on other thoughts and behaviour. So when an agent thinks tokening the LOT expression *water*, she may be lead to think about drinking, but when tokening H_2O may be caused to think about chemistry. Both concepts have the same meaning, but differ in their modes of presentation and causal roles.

To sum up: when concepts are instantiated in mental states (i.e. as concept-tokens), they are individuated by their content, mode of presentation and inferential role. There is controversy about how to specify causal roles (cf. Fodor 1992: ch.6 and Margolis & Laurence 2007), but that should not bother us here. The main point I want to present is that when an agent instantiates a concept (-type) in her mind, she has a mental representation with a content that express the same meaning, tokened in a particular expression in LOT, and with a particular inferential role over the rest of LOT. Any account about the cognitive significance that is implied with the possession of a concept should consider these three aspects of concept individuation.

Though the picture of the mind given by CTM has some controversial issues, it still appears suitable for the purposes of cognitive psychology at least for three reasons. First, through the idea of an internal language and mechanisms to fix content, it provides a plausible account of how the mind could make up a meaningful perspective on the world. Second, it explains how this perspective could be realistically constructed, by showing how mental states could be instantiated in an internal medium of representation. This allows us to treat the agent as a genuine thinker with real and casually efficacious mental states, and also helps to solve the 'mode of presentation problem' (i.e. explaining how an agent could have two different thoughts about the same thing, by using different LOT expressions).

Finally, a third advantage of this view is that it can serve the purposes of a scientific psychology, by giving an account of how the contents of structured mental states can take part in the mental life of an organism. LOT provides the cognitive vehicles for causally efficacious sequences of thoughts, whilst their inferential roles can describe computational patterns that instantiate principles of rationality. In other words, it allows a scientific explanation of how cognitive agents behave in virtue of their intentional mental states (i.e. beliefs, desires, etc.). This is particularly important for this essay, since philosophers of cognitive ethology have embraced a similar view for animal psy-

chology (see Allen & Bekoff 2006 and Carruthers 2004; 2006). They claim that a mentalistic framework like that used by cognitive psychology (i.e. folk-psychology) can be applied to explain animal behaviour, attributing many animals with structured thoughts that interact causally according to rational patterns.

4. Honeybees as computational systems

In this section I will give a brief review of some complex behaviours that have been studied in honeybees in order to show how plausible it seems to claim that they are endowed with computational states and processes.

As is well known, honeybees have notable navigational capacities that make them able to fly from their hives to sources of food and then return. Sometimes they rely on landmarks to orientate, whilst they also use dead reckoning (calculate their position by estimating the direction and distance travelled). They exploit the solar azimuth as a directional referent, being able to estimate its position in the sky at different times of the day in order to set and hold a compass course (Collett & Collett 2002). More surprisingly, they can integrate this information and use it flexibly. For example, in some experiments bees were captured after feeding and carried in a dark box to an unfamiliar releasing point. When released, they initially continued to fly the course they were on when captured, but they soon recognised they were lost, and began an extensive search until they found a familiar landmark. Then, they were able to fly straight to their hives in a vector they have never flown before (Menzel et al. 2000). These experiments suggest that honeybees can represent many features of their environments and integrate them with stored information about distance and direction relative to their hives.

Another remarkable fact about honeybees is their communicational capacities. Foraging bees transmit information about food resources to other bees via different kinds of dances they perform inside the hive (Gould & Gould 1996). Some features of the dance such as the angle of movement as measured from the vertical, and the number of 'waggles' they made at some point of the dance, convey information about the expected direction of the sun for the time of the day, and the distance of the food source. The bees in the hive are not just able to integrate the communicated information and fly to the

food, but can also evaluate it along a number of dimensions. For example, they are less likely to fly to distant sources of food, and show preference for rich sources of food.

These findings suggest that the behaviour of honeybees cannot rely exclusively on fixed action patterns, or be conditioned responses to stimuli. Instead, they seem able to form complex and structured representations of their environments, including information about distance, time, direction and location. They can also transmit this information and use it in a rather flexible and systematic way. Plausibly, many authors have claimed that the best explanation for these complex behaviours is that honeybees can carry out computational processes over causally efficacious and structured representations (Carruthers 2006, Gallistel 2009, Tetzlaff & Rey 2009).

5. Honeybees as thinking creatures

Through several writings, Carruthers (2004; 2006; 2009) has given a detailed defence of the computational capacities and the massive modularity of animal cognition. Among them are writings on honeybees, whose striking behavioural complexities I summarised above. He also moves a step forward in claiming that honeybees have conceptual thought, according with the framework of CTM. I shall summarise his view and then present my arguments against it in the following sections.

Carruthers argues that the capacity of certain animals to represent specific features of their environment and to process them following algorithmic steps, constitutes a genuine form of means/ends reasoning. He contrasts it with forms of associative conditioning or innate releasing mechanisms, which cannot explain the flexibility and complexity of certain behaviours (e.g. those of the honeybees presented above). On the contrary, many animal behaviours are mediated by cognitive processes that involve explicit representations and purposeful reasoning. He goes on to claim that these processes can be characterised in terms of belief-states and desire-states that are discrete, structured, and causally efficacious in virtue of their structural properties.

A great part of the force of Carruthers' argument rests on two assumptions. The first is that CTM works for human beings, an assumption that I shall take for granted for the purposes of this paper. Sec-

ond, and more importantly for present purposes is that the difference between human and animal cognition is basically a matter of complexity and not of kind. He claims that there are no reasons a priori to impose human standards to define 'mindness', and that what matters for having a mind is to possess the right cognitive architecture: a compositional medium of representation able to structure internal states that can interact causally through basic practical inferences to select and guide behaviour. But, why regard these internal states as genuine beliefs and desires?

The response given by Carruthers 2004 is that, after studying animal behaviour, we can infer that those internal states have a structure that resembles that of beliefs and desires, both in their propositional form and in the way they interact following the characteristic roles of beliefs and desires in practical reasoning. So, by virtue of their resemblance to our own internal states and folk-psychology, he argues that these internal states can be externally (and realistically) characterised as beliefs and desires, and therefore structured by concepts.

As previously noted, Carruthers' idea of animal cognition also involves claims about massive modular architecture. The computational systems of animals are supposed to be organised in cognitive modules, and that implies that the representations and computations the animal carries out are distributed into separated, domain-specific and rather isolated units. For example, honeybees seem to use distinct modular systems to navigate inside or outside their hives (Carruthers 2009). When inside the hive they orientate themselves using gravity-based and olfactory cues, whilst they rely on solar bearings when outside it. It is not that they choose between one navigational system or another. According to MMH, honeybees do not represent those systems as alternative sets of spatial representations[4], but they activate one or another when the relevant input is present. So, if honeybees are thinkers, they deploy different ways of thinking about their environment depending on which module they are using. Those differences are principally related with the representational vehicles they deploy (e.g. from distinct perceptual formats), and the inferential roles they occupy (i.e. directed to different domain-specific tasks).

[4] Because they lack second order beliefs. In Carruthers (2002), the author defends his view that cross-modular cognitive integration of thoughts is restricted to linguistic creatures.

6. First argument against Carruthers' proposal: concept individuation

Now I shall argue that the view that honeybees have conceptual thought, as explained in the preceding section, has several problems. In particular, I argue that it becomes implausible to state that honeybees can individuate and use the concept HIVE. Since this concept denotes an essential feature of their environment, it is hard to see how they could be treated as genuine thinkers if they lack it.

My arguments outline some problems associated with the idea that animals with a simple cognitive system and a massively modular architecture could individuate concepts. Recall the two navigation modules that honeybees have for orientating inside or outside their hives. They should be able to entertain the concept of HIVE in both cases, presumably in different representational formats, one based on gravitational and olfactory cues, whilst the other based on visual cues. Also, they should be able to combine this concept with others, in order to form propositional thoughts. According with the framework of MMH, instances of the concept of HIVE should coexist within each module with other concepts that concern the specific computations that the module was designed to perform. For example, suppose that the navigational module for outside the hive has the concept of BLUE, whilst the module for inside does not. Instead the module for navigating inside the hive has the concept of WAX-ODOUR, absent in the other module. This observation leads to the conclusion that the honeybee would only be able to think *the hive has wax-odour* when it is inside the hive, whilst the thought *the hive is blue* could only be entertained when outside it.

Does that mean that the honeybee has two different concepts of HIVE, one for each module? If we recall how CTM defined the semantic properties of LOT, we could give a tentative response: they are instances of the same concept, since they both have the same extension, i.e. they are both about hives. However, their mode of presentation and inferential role in the propositions they constitute must be different, since each module has a domain-specific set of representations and causal roles. They respond to specific input and output channels and carry out the computations specified by the function of the module. So the honeybee appear to be instantiating

two different concepts, and this could be seen as a problem when individuating them as tokens of the same concept of HIVE.

However, Carruthers sees this situation as unproblematic, suggesting that animals do not have a single LOT, but several LOTs, one for each module. That means that the honeybee may turn out to have many modules which can think about hives, but do it in radically different ways, as if there were different languages that cannot understand each other. However, I believe that this idea does not work. Let me restate some ideas about concept individuation to then expose my arguments.

What fixes the content of HIVE is its extension. Two agents share the content of HIVE if hives in the world causally co-vary with instances of HIVE in their heads. But, as I explained in section 3, concept-tokens are not individuated just by their contents. *Content* attribution depends on their extension, whilst *concept* attribution also depends on their LOT expression and inferential role. So, to some extent the meaning of a concept is independent of its inferential role, but that does not imply that inferential roles are irrelevant to determine whether a system is really instantiating a concept. According to CTM, LOT tokens realise concepts thanks to their place in a causal network that connects them to the world in the appropriate way, in the sense that the semantic properties of the tokens are preserved along the computational processes of the system. This imposes a constraint on the internal coherence on the system, which makes possible the fact that not any interpretation of the semantic properties of concepts could 'make sense' (Haugeland 1981). Otherwise, if any interpretational scheme could make sense of what the system's tokens are about, concept attribution would become something trivial.

This idea of internal coherence is what makes the individuation of concepts problematic in the present case. Even if an agent could think about the same thing in different ways, and thus individuate the same concept-type in several concept-tokens, there should be a consistent relation between their inferential roles, at least making it plausible to justify concept possession instead of simple forms of categorisation (cf. section 2). If an agent happens to think about hives in radically different ways, we may be justified to doubt whether this agent could possess the concept of HIVE in any relevant sense. And this seems to be the case of honeybees, since each module processes information about hives through representational vehicles and computations that were designed for specific tasks, that may be different and even

contravene one another. If nothing ensures some degree of internal coherency within the conceptual system of an agent, to treat modular processes as instantiating genuine concepts strikes me at least as problematic.

It is important to remark that I do not intend to raise general scepticism about how concepts could be instantiated in a massively modular mind, or be shared among different people. For instance, it is plausible to suggest that in the case of human beings the faculty of language provides a medium for conceptual identity (or similarity) within the mind and also among people who share a natural language. But this is not, of course, the case of honeybees, and so it is unlikely that their cognitive architecture could have the resources to deal with the concerns about concept individuation presented in this paper.

7- Second argument against Carruthers' proposal: the Generality Constraint

My second argument is also related to some consequences of MMH in the individuation of concepts, but this time I focus on the generality constraint (GC). The GC is often assumed as an essential characteristic of conceptual thought, and was first stated by Evans as follows:

> We cannot avoid thinking of a thought about an individual object x, to the effect that it is F, as the exercise of two separable capacities; one being the capacity to think of x, which could be equally exercised in thoughts about x to the effect that it is G or H; and the other being a conception of what it is to be F, which could be equally exercised in thoughts about other individuals, to the effect that they are F. (Evans 1982: 75)

The main idea is that genuine thinkers should be capable of producing and entertaining an unbounded set of novel well-formed combinations of concepts. This capacity is closely related with what has been called the systematicity and productivity of thought, which have been proclaimed by CTM theorists as elemental features of thought. In Fodor's words:

> Productivity and systematicity are also universal features of human thought (and, for all I know, of the thoughts of many infra-human crea-

> tures). There is no upper bound to the number of thoughts that a person can think. (Fodor 1994: 106-7)

Moreover, CTM offers one of the most compelling explanations about the cognitive mechanisms that underlie these features, based on the compositional nature of LOT (see section 2). So the GC can be safely regarded as a hallmark of thought that honeybees should fulfil if they have genuine concepts.

Now suppose that the perceptual apparatus of the honeybee is able to discriminate between three colours: green, yellow and red, and that this capacity is deployed in a module for flower recognition which does not have *hive* among its repertoire of concepts (this example is fictional, but serves to exemplify some possible consequences of the MMH). Also, suppose that the navigational module for outside the hive mentioned earlier includes among its domain-specific repertoire of representations for colour just *green* and *yellow*, but not *red*. So, among the operations of this module the honeybee might be able to combine the concept of HIVE with the concepts of GREEN and YELLOW, forming the thoughts *the hive is green* and the *hive is yellow*. However, she will not be able to think *the hive is red*. This appears to violate the GC.

Carruthers (2004; 2009) has defended the conceptual capacities of animals, based on their apparent capacity to form thoughts with compositional structure. He acknowledges that given the restrictions derived from a modular architecture, honeybees may be unable to meet the GC. A clear case is that they have limited productivity, since their cognitive architecture prevents their concepts to combine with others outside their own modules. Carruthers plausibly makes the point that the capacity to creatively form an unbounded set of new combinations between concepts appear to be a particular human capability, that does not seem to be necessary for concept possession. But he does claim that genuine thought must be at least partially systematic. He states his position with what he calls a 'weak' version of the GC, defined as follows:

> If a creature possesses the concepts *F* and *a* (and is capable of thinking *Fa*), then for some other concepts *G* and *b* that the creature could possess, it is metaphysically possible for the creature to think *Ga*, and in the same sense possible for it to think *Fb*. (Carruthers 2009: 97)

Following this definition, for a creature to satisfy the weak GC we should expect it to be able to make at least *some* combinations between the concepts it possesses, as opposed to a 'strong' version of the GC where it should be capable to think *all* possible combinations. So in the previous example, even though honeybees cannot deploy their concept of HIVE to think *the hive is red*, the fact that they can think *the hive is green* and *the hive is yellow* shows that they have the capacity to recombine their concepts, at least in a modest way that satisfies the weak GC.

But, why should we accept this weak version of the GC? Is it too modest? Carruthers argues that it satisfies what he takes to be the core of concept possessing: compositionality. This is the capacity to have thoughts that are structured in a way in which its components can be detached from their current form to re-structure at least some other thoughts. To have the capacity to make all possible combinations of thoughts constitutes an ideal, he suggests, that perhaps only humans can get close to achieving.

I find his defence of the weak GC unconvincing. Even though it works as a constraint on compositional structure, it is too weak as a constraint on genuine thinking as is the purpose of the GC. A creature who is able to entertain concepts should be able to detach them from their current inferential roles, in order to then deploy them in new compositional thoughts. But let us imagine a module with a fixed architecture consisting on a few combinable concepts. It would satisfy the weak GC, however nothing implies that its concepts could be detached from their current roles. If the algorithms carried out by its cognitive machinery are innately specified and thus hardwired within the margins of the module, the inferential roles of its concepts are not modifiable, even if the concepts appear mirroring certain combinations.

However, perhaps a more serious problem with Carruthers' proposal comes again as a consequence of the massive modular architecture of honeybee cognition. Following its original formulation, the GC is intended to ensure that when a creature really has the concept *F*, we are committed to the view that when it has any thought that deploys this concept (e.g. *Fa*, *Fb*, etc.) it is exercising the same conceptual capacity (see Evans 1982: 101-105). However, this is does not seem work with honeybees. Let me explain this with an example.

Recall the previous example of the two modules for flower recognition and for navigation. The honeybee would be able to think *the*

flower is yellow in the first module, whilst *the hive is yellow* in the second. Contrary to what the GC proclaims, the conceptual capacities deployed to think about the concept of YELLOW in both cases are different, thus raising doubts about whether the insect is really able to entertain the concept of YELLOW. It could be argued that both modules share the same conceptual capacities, but the nature of cognitive modules seems to count against this idea. Cognitive modules are often conceived as 'mental organs' in analogy with the organs of the body, since they evolved functionally specialised mechanisms in same way as the heart or the lungs (Pinker 1997). It is a natural consequence of this specialisation that the functions performed by these organs correspond to distinct biological capacities, and in the same sense the functions performed by each module can be regarded as the product of distinct cognitive capacities.

8. Final remarks

As I have stated since the beginning, my purpose in this paper has not been to criticise the main tenets of CTM or MMH. They could be perfectly true, and some version of them suitable for animal cognition. My point has been to argue that the requirements for conceptual thought are not fulfilled in a model that simply conjoins both views without further refinements. It has also been my purpose to direct my arguments to a simple cognitive architecture, as seems to be the case of honeybee cognition. More sophisticated versions of MMH, which may incorporate massive conceptual networks and/or mechanisms for cognitive integration, may well be immune to the arguments raised in the present paper.

It is always tempting to attribute a belief-desire psychology to animals, and, without doubt, the evidence of their computational capacities make them good candidates to be thinkers. However, whether this evidence alone is enough to account for the conceptual nature of their representations is far from clear. It has been the purpose of this paper to show that Carruthers' account of honeybee concepts presents at least two problems. A conclusion could be, to put it roughly, that there is good evidence to regard honeybees as sophisticated computers, however not good reasons to regard them as having thoughts and concepts.

Perhaps one of the main limitations of MMH to give a plausible picture of a mind is that it goes against the intuitive view of the mind as a unitary perspective on the world. As noted in section 3, in order to sound plausible CTM needs to claim that the meaning of concepts does not reside in local computational pathways, but in processes that are part of a whole computational system connected with the external world. It is hard to see how concepts that are enclosed in modules and therefore cannot interact with those from other modules could be regarded as part of such a unitary conceptual system.

A plausible alternative, I believe, could still hold that animal cognition is massively modular, but claim that genuine minds emerged when second-order representations (or metarepresentations) evolved in animals. This could have provided a cross-modular medium to detach the split repertoire of representations contained in modules, and integrate them into a unitary representational system, that gets closer to an intuitive picture of what a mind is. Some authors have suggested that metarepresentational capacities are present in just a few highly intelligent animals, such as some primates (Sperber 2000). However, others have proposed that metarepresentations could be wide spread in the animal kingdom, probably under a non-propositional representational format (Bermúdez 2009, Proust 2009). The issue about whether metarepresentations are a necessary condition for having a mind (and therefore genuine concepts), or when they appeared in phylogeny, goes beyond the purposes of this paper, however.

Bernardo Aguilera Dreyse
bedobardo@gmail.com

9. References

Allen, Colin & Bekoff, Marc. 1997. Species of Mind: The Philosophy and Biology of Cognitive Ethology. Cambridge MA: MIT Press.

Allen, Colin & Hauser, Marc. 1996. Concept Attribution in Nonhuman Animals: Theoretical and Methodological Problems in Ascribing Complex Mental Processes. In Readings in Animal Cognition, edited by M. Bekoff & D. Jamieson. Cambridge MA: MIT Press, 47-62.

Bermúdez, José Luis. 2009. Mind reading in the animal kingdom. In The Philosophy of Animal Minds, edited by R. Lurz. Cambridge: Cambridge University Press, 145-164.

Carruthers, Peter. 2002. The cognitive functions of language. The Behavioral and brain sciences, 25(6): 657-74.

Carruthers, Peter. 2004. On Being Simple Minded. American Philosophical Quaterly, 41(3): 205-220.

Carruthers, Peter. 2006. The Architecture of the Mind. New York: Oxford University Press.

Carruthers, Peter. 2009. Invertebrate concepts confront the generality constraint (and win). In The Philosophy of Animal Minds, edited by R. Lurz. Cambridge: Cambridge University Press, 89-107.

Collett, T. S., & Collett, M. 2002. Memory use in insect visual navigation. Nature reviews. Neuroscience, 3(7): 542-52.

Cosmides, Leda & Tooby, John. 1994. Origins of domain specificity: the evolution of functional organization. In Mapping the mind: Domain specificity in cognition and culture, edited by L. A. Hirschfeld and S. A. Gelman. Cambridge: Cambridge University Press, 85–116.

Crane, Tim. 2001. Elements of Mind. New York: Oxford University Press.

Evans, Gareth. 1982. The Varieties of Reference. Oxford: Oxford University Press.

Fodor, Jerry. 1975. The Language of Thought. Cambridge MA: Harvard University Press.

Fodor, Jerry. 1987. Psychosemantics: The Problem of Meaning in the Philosophy of Mind. London: Bradford.

Fodor, Jerry. 1992. A Theory of Content and Other Essays. Cambridge MA: MIT Press.

Fodor, Jerry. 1994. Concepts: a potboiler. Cognition, 50(1-3): 95-113.

Gallistel, Charles Ransom. 2009. The foundational abstractions. In Of Minds and Language: A Dialogue with Noam Chomsky in the Basque Country, edited by M. Piattelli-Palmirini, J. Uriagereka, & P. Salaburu. New York: Oxford University Press, 58-73.

Gould, James & Gould, Carol Grant. 1998. The Honey Bee. New York: Scientific American Library.

Haugeland, John. 1981. Semantic Engines: An Introduction to Mind Design. In Mind Design: Philosophy, Psychology, and Artificial Intelligence, edited by J. Haugeland. Cambridge MA: MIT Press, 1-34.

Haugeland, John. 2003. Syntax, Semantics, Physics. In Views Into the Chinese Room: New Essays on Searle and Artificial Intelligence, edited by J. M. Preston & M. A. Bishop. New York: Oxford University Press, 379–392.

Horst, Steven. 2009. The Computational Theory of Mind. In The Stanford Encyclopedia of Philosophy, edited by E. Zalta. Retrieved from http://plato.stanford.edu/archives/win2009/entries/computational-mind.

Margolis, Eric & Laurence, Stephen. 2007. The Ontology of Concepts—Abstract Objects or Mental Representations? Noûs, 41(4): 561-593.

Menzel, R., Brandt, R., Gumbert, a, Komischke, B., & Kunze, J. 2000. Two spatial memories for honeybee navigation. Proceedings. Biological sciences / The Royal Society, 267(1447): 961-8.

Pinker, Steven. 1997. How the Mind Works. New York: Norton.

Proust, Jöelle. 2009. The representational basis of brute metacognition: a proposal. In The Philosophy of Animal Minds, edited by R. Lurz. Cambridge: Cambridge University Press, 165-183.

Rey, Georges. 1997. Contemporary Philosophy of Mind: A Contentiously Classical Approach. Oxford: Blackwell.

Searle, John. 1980. Minds, Brains and Programs. The Behavioral and Brain Sciences, 3: 417-24.

Shettleworth, Sara J. 1998. Cognition, Evolution, and Behavior. New York: Oxford University Press.

Shettleworth, Sara J. 1998. Cognition, Evolution, and Behavior. New York: Oxford University Press.

Sperber, Dan. 2000. Metarepresentations in an Evolutionary Perspective. In Metarepresentations: A Multidisciplinary Perspective, edited by D. Sperber. New York: Oxford University Press, 117-138.

Sterelny, Kim. 1990. The Representational Theory of Mind: An Introduction. Oxford: Basil Blackwell.

Tetzlaff, Michael & Rey, Georges. 2009. Systematicity and intentional realism in honeybee navigation. In The Philosophy of Animal Minds, edited by R. Lurz. Cambridge: Cambridge University Press, 72-88.

Might-counterfactuals and the principle of conditional excluded middle*

Ivar Hannikainen
University of Sheffield

Abstract
Owing to the problem of inescapable clashes, epistemic accounts of might-counterfactuals have recently gained traction. In a different vein, the might argument against conditional excluded middle has rendered the latter a contentious principle to incorporate into a logic for conditionals. The aim of this paper is to rescue both ontic might-counterfactuals and conditional excluded middle from these disparate debates and show them to be compatible. I argue that the antecedent of a might-counterfactual is semantically underdetermined with respect to the counterfactual worlds it selects for evaluation. This explains how might-counterfactuals select multiple counterfactual worlds as they apparently do and why their utterance confers a weaker alethic commitment on the speaker than does that of a would-counterfactual, as well as provides an ontic solution to inescapable clashes. I briefly sketch how the semantic underdetermination and truth conditions of might-counterfactuals are regulated by conversational context.

Keywords
Inescapable clashes, counterfactuals, Lewis-Stalnaker, possible worlds, semantic underdetermination.

1. Introduction

Consider the following conjunction:

$$(\varphi \Diamond\!\!\rightarrow \psi) \,\&\, (\varphi \Box\!\!\rightarrow \chi \vee \varphi \Box\!\!\rightarrow \sim\chi)$$

* Thank you to Marta Jorba, Sergi Oms, Miguel Ángel Sebastián and all the attendees for fostering such a productive and stimulating environment at the 13th Taller d'Investigació en Filosofia. I am grateful also to Javier Vilanova Arias for his terrific guidance in writing this paper and Antonio Blanco Salgueiro, Luis Fernández Moreno, Gonçalo Santos and an anonymous reviewer for precious feedback on diverse versions of this paper.

Lewis famously upheld, as a consequence of his account of the comparative similarity relation between possible worlds, that the second conjunct need not be true. In so doing, he denied the *principle of conditional excluded middle* (CXM) and committed to saying things like:

> It is not the case that if Bizet and Verdi were compatriots, Bizet would be Italian; and it is not the case that if Bizet and Verdi were compatriots, Bizet would not be Italian (1973: 80).[1]

For Stalnaker (who understands might-counterfactuals as expressions of epistemic possibility), the first conjunct cannot be an ontic claim since whether 'If φ, then it might be the case that χ' is true or false depends on the speaker's epistemic status. Additionally, on account of Stalnaker's selection function, there is a single antecedent world by which to evaluate the truth of the consequent; and, therefore, there are no matters of fact about what *might* (or might not) have been the case, only about what *would* (or would not) have been the case. To me, this counts against Stalnaker's analysis: there must be matters of fact about what might counterfactually have been the case which might-counterfactuals serve to describe. Whether ψ might have been the case if it were the case that φ is an objective matter, and this being so is compatible with a semantics for conditionals (Stalnaker 1968, 1981) according to which χ either would have been the case if it were the case that φ or, if it were the case that φ, it would not have been the case that χ. The purpose of this essay, then, is to outline and defend an account of counterfactuals according to which CXM holds and might-counterfactuals express ontic, rather than epistemic, possibilities.

In Sections 2 and 3, I will introduce the topic by way of a historical review, looking at Lewis's and Stalnaker's views with regard to both might-counterfactuals and CXM, and then I will present the most developed epistemic account of might-counterfactuals (DeRose 1991, 1999). Next, in Section 4, I will lay out the major problem for ontic accounts that DeRose has furthered, the so-called *problem of inescapable clashes*, as well as his own solution. In the following section, Section

[1] Moreover, he claimed that $\sim(\varphi \mathrel{\Box\!\!\rightarrow} \chi)$ & $\sim(\varphi \mathrel{\Box\!\!\rightarrow} \sim\chi)$ does not contradict $\varphi \mathrel{\Box\!\!\rightarrow} (\chi \vee \sim\chi)$. To be sure, for Stalnaker, on the other hand, 'Either if Bizet and Verdi were compatriots Bizet would be Italian, or Bizet would not be Italian if Bizet and Verdi were compatriots' is true.

5, I will advance my alternative explanation of the phenomenon of inescapable clashes. Lastly, in Sections 6 and 7, I will flesh out the account of might-counterfactuals that underlies my solution to the problem of inescapable clashes, try to answer some of the preliminary worries it could raise, and show how my ontic account of might-counterfactuals is compatible with CXM.

2. Might-counterfactuals and CXM in Lewis and Stalnaker

It'll be helpful to begin by briefly reviewing Lewis's and Stalnaker's semantics for conditionals paying special attention to how might-counterfactuals are defined and how CXM fares.

2.1. Lewis's duality thesis

Lewis's 1973 theory of conditionals is formulated in terms of a *comparative similarity relation*. Let the comparative similarity relation $C_i(j, k)$ mean that j is more similar to i than k is to i. A would-counterfactual, $\varphi \mathrel{\Box\!\!\rightarrow} \psi$, is true iff there is a φ–world j such that ψ is true in j, and in all φ–worlds which are at least as similar to i as j. Crucially, the comparative similarity relation determines a weak total ordering which allows comparative similarity ties. In virtue of this feature, considering again the famous Bizet-Verdi example, the possible world(s) in which Bizet is French and the possible world(s) in which Verdi is Italian are tied in terms of comparative similarity. That is, there is a compatriot-world, c_1, in which 'Bizet is Italian' is true, but there is another compatriot-world, c_2, which is at least as similar to the actual world as c_1, in which 'Bizet is Italian' is false. Thus, the counterfactual 'If Bizet and Verdi were compatriots, Bizet would be Italian' is false. These very considerations make the counterfactual 'If Bizet and Verdi were compatriots, Bizet would not be Italian' false too. By conjoining these two false counterfactuals, we get Lewis's rejection of CXM.

Lewis put forth a straightforward and highly intuitive definition of might-counterfactuals in terms of would-counterfactuals known as the *duality thesis* (DT). It is straightforward and highly intuitive because it borrows the notions of necessity and possibility from modal logic and applies them to the setting of counterfactual conditionals in a rough-and-ready manner, *i.e.*:

[DT] $\varphi \Diamond\!\!\rightarrow \psi =_{df} \sim(\varphi \Box\!\!\rightarrow \sim\psi)$[2]

Lewis made the case that this definition of might-counterfactuals respects our ordinary usage of 'might' in counterfactual settings. Suppose I say 'If Lionel Messi had played for Real Madrid this season, Real Madrid might have won La Liga.' What I mean is that it is false that Real Madrid fails to win La Liga this season (that $\sim\psi$) in all possible worlds which (a) are at least as similar to the actual world as a world in which Lionel Messi plays for Real Madrid, and in which (b) Lionel Messi plays for Real Madrid. (Let's call worlds meeting these two criteria *relevant*). In other words, I am claiming that there is at least one relevant world which makes the conditional 'If Lionel Messi had played for Real Madrid this season, Real Madrid would not have won La Liga' false. If you thought I was mistaken, it seems likely that you should contradict me by saying: 'Even if Lionel Messi had played for Real Madrid this season, Real Madrid would not have won La Liga.'

2.2. Stalnaker's epistemic thesis

Evidently, DT is not available in Stalnaker's semantics for counterfactual conditionals. Recall that Stalnaker's truth conditions for conditionals (which I will adopt in my proposed analysis of the relation between might- and would-counterfactuals) are devised by using the *selection function* operator, f, and let $f(\varphi, i)$ be the selection function that picks out the possible world in which φ is true and which otherwise differs minimally from the base world, i. Then, $\varphi \Box\!\!\rightarrow \psi$ is true (/false) in i if ψ is true (/false) in the nearest φ-world $f(\varphi, i)$. It is a feature of Stalnaker's selection function that it operates under this so-called *Uniqueness Assumption*, according to which there is always at most a *single* φ-world at which to evaluate the truth of ψ. (For simplicity's sake, I will make this assumption too in the remainder of this paper. The analysis of might-counterfactuals I will develop is independent of the polemic about the similarity ordering of possible

[2] This definition is insufficient to deal with counterfactuals with impossible antecedents. For purposes of this paper, the analysis of ordinary language counterfactual conditionals, I will ignore the case of such counterfactuals and restrict the discussion to counterfactual conditionals with antecedents that describe possible states of affairs. The same point applies later in the paper to the provided definition of [ST]. Thanks to an anonymous reviewer for pointing this out.

worlds. Those worried about the ubiquity of comparative similarity ties, should help themselves to Stalnaker's 1981 appeal to supervaluations wherever I talk about the single nearest φ-world.) Seeing as for any φ, ψ will either be true or false at $f(\varphi, i)$, disjunctions like $(\varphi \mathrel{\Box\!\!\rightarrow} \psi) \vee (\varphi \mathrel{\Box\!\!\rightarrow} \sim\psi)$ will always be true; *i.e.*, CXM holds.[3] To see how DT and CXM are decidedly incompatible, consider the following argument known as the *might argument against conditional excluded middle* (Lewis 1973):

P1.	$\varphi \mathrel{\Diamond\!\!\rightarrow} \psi = \sim(\varphi \mathrel{\Box\!\!\rightarrow} \sim\psi)$	[DT]
P2.	$(\varphi \mathrel{\Box\!\!\rightarrow} \psi) \vee (\varphi \mathrel{\Box\!\!\rightarrow} \sim\psi)$	[CXM]
P3.	$\sim(\varphi \mathrel{\Box\!\!\rightarrow} \sim\psi) \supset (\varphi \mathrel{\Box\!\!\rightarrow} \psi)$	[from P2 and DF ⊃]
P4.	$\varphi \mathrel{\Box\!\!\rightarrow} \psi \supset \varphi \mathrel{\Diamond\!\!\rightarrow} \psi$	[from DF □→]
P5.	$\varphi \mathrel{\Diamond\!\!\rightarrow} \psi \supset \varphi \mathrel{\Box\!\!\rightarrow} \psi$	[from P1 and P3]
C.	$\varphi \mathrel{\Diamond\!\!\rightarrow} \psi = \varphi \mathrel{\Box\!\!\rightarrow} \psi$	[from P4 and P5]

Evidently, if we hold both DT and CXM, we arrive at the unhappy conclusion that might- and would-counterfactuals have the same truth conditions (which, it should be clear, is not faithful to their ordinary meaning in English). For the purposes of this paper, it being my aim to defend CXM, the most decisive consideration against DT is that, as seen above, it is incompatible with CXM. So, as an alternative to DT, Stalnaker 1981 proposes an epistemic view of might-counterfactuals, which is based on combining the semantics of 'might' outside conditional contexts with his analysis of would-counterfactuals.[4]

[3] Things change when supervaluations are considered (van Fraassen 1966). ψ will either be true or false for any possible valuation of φ. Thus, $(\varphi \mathrel{\Box\!\!\rightarrow} \psi)$ and $(\varphi \mathrel{\Box\!\!\rightarrow} \sim\psi)$ will each either be supertrue, superfalse or indeterminate, but the disjunction $(\varphi \mathrel{\Box\!\!\rightarrow} \psi) \vee (\varphi \mathrel{\Box\!\!\rightarrow} \sim\psi)$ will necessarily be true and CXM will remain valid.

[4] DeRose 1991, 1999, on whom I will focus in the bulk of this paper, painstakingly follows in Stalnaker's footsteps on this. However, while Stalnaker admits some non-epistemic uses of 'might' in counterfactual contexts (see his 1981: 99. 'But *might* sometimes expresses some kind of non-epistemic possibility. *John might have come to the party* could be used to say that it was within John's power to come, or that it was not inevitable that he not come'), DeRose 1999 thinks 'might' is *never* used to indicate non-epistemic possibility. For this reason, and because DeRose's account handles several types of uses of 'might' better than Stalnaker's does, my critique of the epistemic account of might-counterfactuals will focus on DeRose's account.

$$\varphi \Diamond\!\!\rightarrow \psi =_{df} <e> (\varphi \Box\!\!\rightarrow \psi)$$

He claims that 'might' outside conditional contexts indicates possibility and that the kind of possibility it typically expresses is epistemic possibility. In other words, 'It might be the case that ψ' means something like 'What I know does not entail ~ψ' or 'ψ is compatible with what I know'. Putting together the analyses of 'might' and of would-counterfactuals, Stalnaker's thesis (ST) is that a might-counterfactual such as $\varphi \Diamond\!\!\rightarrow \psi$ as uttered by a speaker *S* means that nothing *S* knows obviously entails that ψ is false in $f(\varphi, i)$. Substituting the *definiens* above:

[ST] $\varphi \Diamond\!\!\rightarrow \psi =_{df} \sim K_S (\varphi \Box\!\!\rightarrow \sim\psi)$

ST has a simple and famous rebuttal that Lewis (1973: 80-1) issued. Suppose I do not know what is in my pocket and I say 'If I had looked in my pocket, I might have found a penny.' The fact is that there is no penny in my pocket. This counterfactual is seemingly false, and DT explains why; *i.e.*, because 'If I had looked in my pocket, I would not have found a penny' is true. On ST, however, the counterfactual I uttered is equivalent to saying 'It is compatible with what I know that if I had looked in my pocket, I would have found a penny,' which is true. Thus, ST gives the wrong reading of counterfactuals such as the one in Lewis's classic *Penny Case*.[5]

Hereafter I will dedicate my attention primarily to DeRose's 1991, 1999 more honed and unswervingly epistemic account. The point here was merely to illustrate that the defender of CXM, like Stalnaker, will naturally favor an epistemic account of might-counterfactuals since it can be seamlessly coupled with the logical and semantic groundwork, put forth by Stalnaker in 1968, which upholds CXM. There is a *prima facie* conflict between giving an ontic reading to might-counterfactuals and preserving CXM to which it is the object of this paper to provide a peaceable resolution.

[5] Readers will be reminded of Stalnaker's 1981 solution to the *penny case*. Indeed, Stalnaker provided a quasi-epistemic solution which dealt with this objection and which, in some ways, resembles the ontic account I will put forth.

3. DeRose and the problem of inescapable clashes

Keith DeRose thinks statements like 'It is possible that *P,*' 'It might be the case that *P,*' and, derivatively, might-counterfactuals express epistemic possibilities *all the time*. In DeRose 1991, he weaves through a range of cases in which a speaker for whom it is epistemically possible that *P* may felicitously assert that 'It is possible that *P*,' and he arrives at the following flexible proposal that statements of this kind are true iff:

(1) no member of the relevant community knows that *P* is false, and
(2) there is no relevant way by which members of the relevant community can come to know that *P* is false (1991: 593-4).

Substituting Stalnaker's epistemic possibility operator, $<e>$, with DeRose's analysis of epistemic possibility one gets:

$\varphi \Diamond\!\!\rightarrow \psi$ is true iff
(1) no member of the relevant community knows that $(\varphi \Box\!\!\rightarrow \sim\psi)$, and
(2) there is no relevant way by which members of the relevant community can come to know that $(\varphi \Box\!\!\rightarrow \sim\psi)$.

Or:

[ET] $\varphi \Diamond\!\!\rightarrow \psi =_{df} \sim K_{rc} (\varphi \Box\!\!\rightarrow \sim\psi) \,\&\, \sim\blacklozenge K_{rc} (\varphi \Box\!\!\rightarrow \sim\psi)$[6]

DeRose admits that the 'relevant community' and the 'relevant way' are vague notions but, by means of a variety of examples, shows that any greater specificity in giving truth conditions – of which the prototypical example is ST – too easily generates counterexamples.

3.1. The problem of inescapable clashes

DeRose 1999 has argued that an inescapable problem haunts non-epistemic accounts of might-counterfactuals.[7] While everyday coun-

[6] Think of ♦ as a highly specific modal operator that designates possible worlds which are accessible in the 'relevant ways' that DeRose has in mind.

[7] For different treatments of the problem of inescapable clashes, see also Eagle unpublished, Hawthorne 2005, and Williams 2010.

terfactuals are subject to this problem (see DeRose's baseball example [1999: 385-6]), Hawthorne 2005 has pointed out how quantum theory threatens to falsify any counterfactual conditional grounded in the principles of classical mechanics. Take the following commonsensically true counterfactual:

(W) If I had dropped the plate, it would have fallen to the floor.

In a world governed by quantum mechanics, we must be prepared to accept that:

(M) If I had dropped the plate, it might have flown off sideways.

Notice that (M) is a might-counterfactual, so that to deny (M) would be to insist that it is impossible for the plate to have flown off sideways and to refuse the conclusions of quantum theory. So, if we grant that (M) is true, this will lead us to believe that:

(M') If I had dropped the plate, it might not have fallen to the floor.

It doesn't seem reasonable to agree to (M) while denying (M'). And once you agree to both (W) and (M'), you face the problem of inescapable clashes:

(W+M') If I had dropped the plate, it would have fallen to the floor; nevertheless, the plate might not have fallen to the floor if I had dropped it.[8]

If one holds, as DT does, that (W+M') expresses an inconsistent proposition, one is forced to backtrack and choose between denying (W) – the *skeptical* option – and denying (M') – the *exclusionary* option – two counterfactual claims both of which have 'a good deal of initial plausibility' (DeRose 1999: 387). According to counterfactual skepticism, since (M') is a weaker commitment than is (W), one is encouraged to think of (M') as making (W) false. The counterfactual exclusionary strategy holds that the falsity of most ordinary counterfactuals is too high a price to pay to acknowledge the possibility of quasi-miracles. It is preferable to argue instead from the truth of the

[8] To be sure, conjunctions of the form $\varphi \Box\!\!\rightarrow \sim\psi \;\&\; \varphi \Diamond\!\!\rightarrow \psi$ are also instances of the phenomenon of inescapable clashes.

ordinary counterfactual to the falsity of the corresponding might-counterfactual. Thus, these theories of counterfactuals exclude remarkably low-probability outcomes and say things like 'If I were to roll a die a billion times, it's not the case that it might land tails every time' (Lewis 1979b, Williams 2008).

3.2. DeRose's escapism from pragmatic clashes

DeRose wants to say that both (W) and (M') are true; thus, that there is no semantic contradiction in (W+M') but that, nonetheless, there is pragmatic tension involved in utterances of (W+M') which explains why they are unassertible. He claims, moreover, that his account is singlehandedly capable of accommodating the truth of (W) and (M') while respecting the intuitive proscription against utterances of this kind. DeRose's pragmatic explanation of inescapable clashes is the following: In flat out asserting (W), the speaker represents herself as knowing that (W) while uttering (M') expresses the epistemic possibility for the speaker that (W) is false.

> Thus, what one says in asserting the second conjunct of [(W+M')], while it's perfectly consistent with what one says in asserting the first conjunct, is inconsistent with something one represents as being the case in asserting the first conjunct. This supports our sense that *some* inconsistency is responsible for the clash involved in asserting the conjunction, while, at the same time, happily removing that inconsistency from the realm of what's asserted: The conjunction asserted is itself perfectly consistent, but in trying to assert it, one gets involved in a contradiction between one thing that one asserts, and another thing that one represents as being the case (1999: 389).

DeRose concludes that ET is superior to DT in that it 'provides a way of avoiding the *really* nasty conclusion—that [(W)] is false' (1999: 390). Furthermore, he claims that other, non-epistemic theories are defective insofar as they define might-counterfactuals in a way that renders conjunctions like (W+M') the right thing to say in certain circumstances, and thereby succumb to the problem of inescapable clashes. DeRose discusses Heller's 1995 theory in Section 8 (1999: 395-6), and Lewis's 1986 'ambiguity thesis' in Section 9 (1999: 396-7). Briefly, Heller's theory claims that $\varphi \mathrel{\Diamond\!\!\rightarrow} \sim\psi$ is true iff there is at least one *close enough* φ-world in which $\sim\psi$ is true. Thus, (W+M') is the right thing to say when ψ is true in the closest φ-worlds and false

in some presumably farther yet close enough φ-world. Lewis's theory is that might-counterfactuals are ambiguous between DT and another reading according to which φ $\Diamond\!\!\rightarrow$ ψ is true iff some relevant φ-worlds are worlds where there is a non-zero chance of ψ being false (1986: 63-4). This theory makes (W+M') the right thing to say when ψ is true in the relevant φ-worlds but there is a non-zero chance of ψ being false in some of the relevant φ-worlds. The spirit of the definition of might-counterfactuals I will present in this paper is especially close to that of Heller's, insofar as I embrace and exploit his claim that might-counterfactuals admit 'gratuitous differences'. Contrary to DeRose's allegation that ontic accounts of might-counterfactuals cannot solve the problem of inescapable clashes, I will develop in the following section an account of the phenomenon of inescapable clashes that can be appended to ontic theories of might-counterfactuals in order to solve this problem.

4. The contextual-shift explanation of inescapable clashes

Consider the following discourse, from a high school principal talking to a teacher, featuring two italicized quantificational claims:

> Mr. D'Elia, I looked through the grade reports of your History 101 class this Spring. *Every single student failed the final exam.* You really ought to lower your expectations of undergraduate students. I understand that you want to promote excellence in the student body, but there are more effective ways to go about it. I remember Mr. Shillington. He's one of the best History teachers that ever came through Bumbletown High. He was rigorous but he always made sure the more promising and hard-working students were rewarded. Let me pull up his grade reports for History 101… Here they are! *10% of students passed the final exam.*

If we take the high school principal's quantificational claims out of context and conjoin them, we get:

> **(A+S')** Every single student failed the final exam, but 10% of students passed the final exam.

When we do this, we generate what has the appearance of a contradiction. But, of course, one would hardly say it *is* one since, in order to evaluate the italicized sentences, we must specify by extracting

from the conversational context the (hitherto implicit) range over which quantification takes place and, when we do so, we see clearly that (A+S') amounts to a non-contradictory, and objectively verifiable proposition, which is true just in case

> every single student *in Mr. D'Elia's History 101* class failed the final exam, and 10% of students *in Mr. Shillington's History 101 class* passed the final exam.

The conjunction of these claims, (A+S'), is seemingly contradictory because there is a fairly prevalent pragmatic rule in natural language use, to do with anaphoric reference, according to which the range of quantification remains stable until it is explicitly set to a new range by the conversational context. In this specific instance, by anaphora, the second sentence would inherit the contextually-salient range of quantification over which the first quantificational statement holds. Whatever this class of 'students' turned out to be, the sequence or conjunction of these two quantificational sentences would lend itself to being interpreted as an obvious contradiction because the proposition expressed would be thought to be:

$$\forall x\,(Sx \rightarrow Fx)\ \&\ \exists x\,(Sx \rightarrow \sim Fx)$$

But, as seen, the range of quantification is left unspecified. There is nothing explicit in the sentence form of either quantificational statement to specify the range over which the proposition quantifies and thus nothing in the sentence form to establish definitively that this utterance involves a semantic contradiction. If the speaker of (A+S') were to insist that (A+S') is true on the grounds that every single student in Mr. D'Elia's History 101 class failed the final exam while 10% of students in Mr. Shillington's History 101 class passed the final exam, we would have to suppose there is something seriously wrong about his dominion of conversational pragmatics, but it would be odd to insist, beyond the unsassertability of (A+S'), that (A+S') was false on these grounds.

I believe an analogous kind of pragmatic failure to make explicit a contextual shift takes place amid (W+M')-type conjunctions. A speaker can hold both (W) and (M') and utter them on separate occasions, as long as a contextual shift is adequately established between them, *e.g.*:

> Quantum mechanics, which I believe in, warns us about the possibility of highly erratic physical phenomena. For example, remember the plate I was spinning on my index finger yesterday. *If I had dropped the plate, it might not have fallen to the floor.* It could have flown off sideways instead. Yet, at the same time, commonsense and good ole' Newtonian mechanics tells me that it is oh-so very likely that the plate will shatter on the floor. You can hardly deny that. *If I had dropped the plate, it would have fallen to the floor.* I cannot be certain of it, but I bet it would happen even while I recognize that quantum oddities are possible.

What she cannot do felicitously is utter (W) and (M') in conjunction or in sequence without thereby *almost* invariably generating the impression of an obvious contradiction.[9] This explanation seems to me to satisfy DeRose's two criteria for an adequate solution to the problem of inescapable clashes, *i.e.*, the proposition expressed by (W+M') is not semantically contradictory, yet

> (1) (W+M') is invariably unassertible.

DeRose might raise the same objection here that he raises against the non-DT version of Heller's 1995 view:

> to get a non-DT version of Heller's view, there should be contexts in which the range of [φ-]worlds relevant to the 'might' counterfactuals is different from (no doubt broader than) the range of worlds relevant to the 'would' counterfactuals. In such contexts, [(W+M')-type] conjunctions should be unproblematic. But there are no such contexts; these conjunctions always clash. So any non-DT version of Heller's view will succumb to the problem of inescapable clashes (1999: 396).

Indeed, on both Heller's and my view, $\varphi \mathrel{\square\!\!\rightarrow} \psi$ & $\varphi \mathrel{\Diamond\!\!\rightarrow} \sim\psi$ can express a consistent, counterfactual proposition; namely, by describing an objective, counterfactual state of affairs across multiple φ-worlds. Thus, as DeRose says, (W+M') indeed should be unproblematic in such contexts. However, as I hope my account has explained, without the requisite contextual shift between (W) and (M'), the

[9] I say 'almost' because I think there are circumstances in which these conjunctions are assertible without the overt contextual shift, namely, when the function of discourse is *exploratory* (see the example on p. 29).

sentence form of (W+M')-type conjunctions renders them almost invariably unassertible.

I have here outlined the contextual shift explanation of inescapable clashes which, I believe, provides – by DeRose's own standards – an adequate account of the phenomenon. (1) The thought expressed is not semantically contradictory such that (M') does not contradict (W) and *vice versa*, thereby providing an alternative to counterfactual skepticism and counterfactual exclusion. Nonetheless, (2) said conjunctions invariably clash; they are never (or almost never) the right thing to *say*. Contrary to DeRose's presumption, the ontic camp can provide a solution to the problem of inescapable clashes.

As a sidenote, these conjunctions can, on my view, be the right thing to *think*, and this should be seen as an important advantage over DeRose's solution. As Eagle has pointed out, though DeRose's epistemic view does successfully dodge counterfactual skepticism and exclusion, it is subject to *weak counterfactual skepticism*: 'the thesis that, even if they are true, ordinary 'would' counterfactual claims cannot be known if the corresponding 'might' counterfactuals are known' (unpublished). This seems like a significant downfall for epistemic theories. It is counterintuitive to suppose that speakers cannot at a single time know both (W) and (M'): after all, a mature epistemic agent knows that she would not have won the national lottery had she picked some other number but that, of course, she just might have. On my account of might-counterfactuals, as on Heller's, one avoids this epistemic brand of counterfactual skepticism too.

5. Semantically underdetermined might-counterfactuals

Besides showing that the contextual-shift solution is an *adequate* ontic account, I will now argue that it is a *plausible* one, *i.e.*, that there are independent reasons to think that corresponding might- and would-counterfactuals are evaluated by taking into account different possible worlds (or sets of possible worlds), and thus that the antecedent of a counterfactual conditional, appearances notwithstanding, makes a different semantic contribution in the context of a might-counterfactual than it does in the context of a would-counterfactual. Consider the following three such reasons: first, multiple antecedent-worlds must be relevant to the evaluation of might-counterfactuals; second, semantic underdetermination allows multiple antecedent-

worlds to be relevant to the truth-conditional evaluation of might-counterfactuals; and third, it is consistent with the discursive functions of might- and would-counterfactuals that speakers would use might-counterfactuals, but not would-counterfactuals, to appeal deliberately to the semantic underdetermination in the antecedent.

As I prefaced, might-counterfactuals, unlike would-counterfactuals, apparently must be evaluated by taking into account multiple antecedent-worlds. Stalnaker (1981: 91-5) defended, in relation to would-counterfactuals, that when a speaker asserts $\varphi \mathrel{\square\!\!\rightarrow} \psi$, she purports to represent and describe a 'unique determinate possible world,' namely the φ-world selected by $f(\varphi, i)$, a possible world which other than accommodating the truth of φ differs minimally from the base world, i. This can be seen by considering the following dialog:

> X: President Carter would have appointed a woman to the Supreme Court last year if there had been a vacancy.
> Y: Who do you think he would have appointed?
> X: He wouldn't have appointed any particular woman; he just would have appointed some woman or other (Stalnaker 1981: 94).

X's response seems bad because the fact is that, if there had been a vacancy in the Supreme Court and President Carter had appointed a woman, he must have appointed some *particular* woman. (If X cannot name her, it is due to her epistemic limitations not to any insurmountable metaphysical vagueness.) This is, according to Stalnaker, because would-counterfactuals are evaluated in each case by taking into account the single nearest antecedent-world. Another consequence of the Uniqueness Assumption is that a speaker cannot go on to say that $\varphi \mathrel{\square\!\!\rightarrow} \sim\psi$ without thereby contradicting herself as to what the nearest φ-world is like.

Now consider the Uniqueness Assumption with respect to might-counterfactuals. A speaker can felicitously and truthfully say things like 'If Messi had played for Real Madrid this season, Real Madrid might have won La Liga but Real Madrid might also have not won La Liga,' *i.e.*,

(M+M') $\quad (\varphi \mathrel{\Diamond\!\!\rightarrow} \psi) \,\&\, (\varphi \mathrel{\Diamond\!\!\rightarrow} \sim\psi)$

If, here again, a single possible world were relevant to the truth conditional evaluation of the speaker's statement, it would be hard to see how her statement could be meaningful. But it seems that these statements are meaningful and sometimes true. So, if (i) Stalnaker is right, that would-counterfactuals are evaluated by taking into account only the single nearest antecedent-world, and (ii) seeing as (M+M')-type conjunctions, in which the contradictory consequents of two conjoined (and true) might-counterfactuals, are assertible, then it must be the case that a class of multiple antecedent-worlds is relevant to the truth conditional evaluation of might-counterfactuals. Thus, the antecedent of a counterfactual conditional must make a different semantic contribution in the context of a might-counterfactual than it does in the context of a would-counterfactual.

How is φ capable of diverging from its semantic contribution to a would-counterfactual? Moreover, how are *multiple* possible worlds at once worlds in which a proposition, φ, is true? It'll be easier to answer the questions in reverse order. Take the proposition 'Popes are not young.' Is this proposition true or false? Or indeterminate? If it is true of the actual world, what is it true *in virtue of*? If it is false or indeterminate, what possible state of affairs would render it true? There clearly are obstacles to determining what such state of affairs would be, and this is due to the inherent semantic underdetermination of this sentence. Firstly, the range of Popes over which this claim quantifies – whether it is all the Popes in history, most of them, only those with which the relevant community is acquainted, only those that are salient in the conversational context, *etc.*, – is unspecified. Secondly, there is no sharp boundary between *being young* and *not being young*, so the property that is predicated of Popes is underdetermined. Even while not knowing what it would take for 'Popes are not young' to be true, it seems obvious that any number of possible states of affairs could make it true (see Fine 1975). Even if the proposition were about a single, identifiable person and contained no vague predicates, *e.g.*,

> Jesulin de Ubrique's cape on the night of his professional bullfighting debut was red,

there would be no single state of affairs that this statement could truthfully report. Suppose we grant that this statement is true in the actual world given the particular cape's actual color. It is true too in

those possible worlds in which it is a very slightly different shade of red, a slightly orangish shade of red, a slightly purplish or pinkish shade of red, *etc*. By the same token, virtually any proposition figuring in a counterfactual antecedent, φ, – 'I look in my pocket,' 'The pirates do not threaten the operation,' 'John does not suffer sudden cardiac arrest,' *etc.*, – is semantically underdetermined: it does not contain enough information to select a single φ-world for evaluation. Take 'Lionel Messi plays in Real Madrid this season.' There is a φ_1-world in which Messi plays three games for Real Madrid and gets injured, a φ_2-world in which Messi plays at least sixty minutes out of every league game for Real Madrid, a φ_3-world in which Messi plays for Real Madrid while Iniesta moves back to Albacete to live the simple life, and so on. In other words, in virtue of semantic underdetermination, multiple possible worlds fit the description in φ and all these φ-worlds are truthmakers for the proposition in a counterfactual antecedent, φ.[10] Since, in a would-counterfactual context, the semantic gaps in φ are 'plugged' by the requirement that the φ-world for evaluation be maximally similar to the base world, φ in a might-counterfactual is able to depart from its semantic contribution to a would-counterfactual.

The contention that semantic underdetermination is at work in the utterance and evaluation of might-counterfactuals should garner intuitive appeal by looking at the discursive functions of might- and would-counterfactuals. A speaker who believes with some degree of confidence that

> ψ would have been the case if it were the case that φ, *where <u>only</u> the nearest φ-world is a world in which φ is true*,

[10] I will not in this paper delve into what the sources of semantic underdetermination are, which I suspect is an empirical matter anyhow. However, here are some suggestions: the enrichment of the antecedent with presuppositions ('If Messi played for Real Madrid this season [*and Iniesta moved back to Albacete to live the simple life*], then…), the resolution of the underdetermination in similarity respects and relations ('If Sheffield were more like San Diego, then…', 'If I exercised [*1.5 hrs/day*] more, then…'), vague terms ('If Popes were generally younger, then…), and non-natural predicates ('If Jesulin de Ubrique's bullfighting-debut cape were not red, then…').

seems warranted in asserting the corresponding would-counterfactual 'If it were the case that φ, then it would be the case that ψ' because she thereby commits to it being true that

$f(\varphi, i) \in \psi$.

By contrast, a speaker who (is not confident that *ψ would have been the case if φ*, but rather) believes with some degree of confidence that

> ψ would have been the case if it were the case that φ, *where every member of the class of φ-worlds selected by the range of admissible precisifications of φ is a world in which φ is true*,

is prudent if he couches his claim in terms of a might-counterfactual, 'If it were the case that φ, then it might be the case that ψ' because he thereby makes only the far weaker commitment that

$f(v_1(\varphi), i) \in \psi \vee f(v_2(\varphi), i) \in \psi \vee f(v_3(\varphi), i) \in \psi \vee \ldots f(v_n(\varphi), i) \in \psi$.

It should seem plausible now that multiple antecedent-worlds, which are irrelevant to the evaluation of the corresponding would-counterfactuals, are relevant to the truth-conditional evaluation of might-counterfactuals and semantic underdetermination explains how they are capable of so being: in the context of a might-counterfactual, the antecedent is semantically underdetermined with respect to the class of its truthmaking antecedent-worlds. This in turn enables speakers to use might-counterfactuals to talk, veridically and prolifically, about counterfactual possibilities.

6. The semantics of semantically underdetermined might-counterfactuals

In the previous section, I hinted at my truth conditions for might-counterfactuals, which I will here lay out explicitly. A given might-counterfactual,

$\varphi \mathrel{\Diamond\!\!\rightarrow} \psi$ is true iff for *some* admissible precisification of φ, $v_k(\varphi)$, the nearest $v_k(\varphi)$-world, $f(v_k(\varphi), i)$, is a ψ-world; and

$\varphi \Diamond\!\!\rightarrow \psi$ is false iff for *every* admissible precisification of φ, $v_1(\varphi)$, $v_2(\varphi)$, $v_3(\varphi)$ … $v_n(\varphi)$, the nearest $v_n(\varphi)$-world, $f(v_n(\varphi), i)$, is a $\sim\psi$-world.

The proposed analysis of might-counterfactuals renders (M') consistent with (W) as seen in the contextual-shift solution to inescapable clashes. A counterintuitive consequence of this is that, contrary to what speakers seem prone to do (as seen in the *Lionel Messi Case* and Lewis's *Penny Case*), it is not generally valid to use (W) to falsify (M'), and *vice versa*. To see this, consider again the *Lionel Messi Case*. On Stalnaker's view of would-counterfactuals (which I adopt here), when we are asked to imagine what would have happened if Lionel Messi had played in Real Madrid this season, we conceive the counterfactual state of affairs most similar to the actual world with those minimal modifications made which are necessary to make Lionel Messi a member of Real Madrid's squad; while, on the proposed account of might-counterfactuals, when we are asked to imagine what *might* have happened if Lionel Messi had played in Real Madrid this season, we let our imagination run looser and conceive a range of counterfactual states of affairs with varying degrees of similarity to the actual world all of which about which 'Lionel Messi plays for Real Madrid this season' is true. This difference in the antecedent-worlds relevant for evaluation is precisely why certain possibilities *might* have been realized if such and such were the case that *would not* have been realized, and why the truth of (W) is consistent with the truth of (M').

One worry, which I will here attempt to assuage, easily engendered by this kind of proposal is that it is perhaps unclear what prevents all might-counterfactuals from being true. First of all, invariantly, counterfactual possibilities can be ruled out that presuppose inadmissible precisifications of the antecedent. If only *in*admissible precisifications of the antecedent select possible worlds in which the consequent is true, then the might-counterfactual in question is false. Here are some examples of invariantly false might-counterfactuals:

> If I had looked in my pocket, I might not have looked in my pocket.
> If the pirates had not threatened the operation, 2+2 might have been equal to 5.
> If Lionel Messi had played for Real Madrid this season, *Mus musculus* (the house mouse) would be capable of prolonged levitation.
> If John had not suffered sudden cardiac arrest, Kerry might have won the 2004 presidential election.

These might-counterfactuals are clearly false from the outset. There are no precisifications of what 'The pirates threaten the operation' means according to which, if the state of affairs described by such precisification held, it might be the case that '2+2 equals 5,' and *mutatis mutandis* for the rest of examples.

Much more frequently, might-counterfactuals can be falsified by a constraint on the admissibility of precisifications regulated by the conversational context. (Note that this conversational constraint must be a contingent matter so as to not incite the problem of the inescapable clashes.) Let's begin with an example in which the conversational context *slackens* the admissibility of precisifications. Suppose the function of discourse is *exploratory*, as in the following example:

> If I had practiced the guitar a lot, I would have had a record deal with Sony. In fact, I might have been rich and famous enough to do without the support of a record label if I had practiced the guitar a lot.

In this discourse type, the would-counterfactual doesn't seem to falsify the might-counterfactual. 'I practice the guitar a lot' is semantically underdetermined in the might-counterfactual above such that antecedent-worlds – which are more remote than the maximally similar antecedent-world in which the speaker's gets a record deal with Sony – are relevant to the evaluation of the might-counterfactual. The function of exploratory discourse renders said precisification admissible (N.B. This is, evidently, not to say that any of the antecedent-worlds selected by such admissible precisifications of the antecedent are worlds in which the speaker is rich and famous enough to do without a record label. That is the *next*, and final, step in the truth conditional evaluation of a might-counterfactual).

In other contexts, the aim of counterfactual discourse is, along Stalnaker's lines, to find out what the actual world would have been like if the antecedent had been true. Now consider counterfactual discourse with this, *truth-aiming* purpose. The familiar might-counterfactual

> Had the pirates not threatened the operation, we might have found the vessel

can be falsified by the would-counterfactual

> If the pirates had not threatened the operation, we would not have found the vessel.

In truth-aiming counterfactual deliberation, remote counterfactual possibilities are rendered false by claims that approximate counterfactual truth. In these contexts, the range of admissible precisifications of the antecedent is *constrained* and this has the consequence of falsifying the might-counterfactual under consideration.

Like many elements of conversational score, precisification-admissibility cannot vary wildly: this would render might-counterfactuals unintelligible. Consider again the conjunction 'If Messi had played for Real Madrid this season, Real Madrid might have won La Liga but Real Madrid might also not have won La Liga.' Let's say this conjunction is true in virtue of it being the case that

> if Messi had played for Real Madrid [and Iniesta were frequently injured] this season, then Real Madrid *would* have won la Liga; &
> if Messi had played for Real Madrid this season [only throughout the second leg], then Real Madrid *would not* have la Liga.

There is a stable criterion – let's suppose that in this conversation it's about the ordinary, significantly likely possibilities throughout a football season – that regulates precisification-admissibility in *both* conjuncts. It would be very odd and implausible if the above conjunction were true in virtue of it being the case that

> if Messi had played for Real Madrid [and Iniesta were frequently injured] this season, then Real Madrid *would* have won la Liga; &
> if Messi had played for Real Madrid this season [and Guardiola traveled back in time to sign Pele onto Barça], then Real Madrid *would not* have won la Liga;

where, initially, only significantly likely football happenings are revelant but then, by the second conjunct, the possibilities that time travel offers are all of a sudden relevant. Thus, like conversational score, the criterion for precisification-admissibility in counterfactual discourse typically cannot vary in a wild manner.

This is a mere sketch of how the semantic underdetermination in might-counterfactuals is regulated by the conversational context. A lot more could be said, but I hope to have outlined the main claims (and thereby given a truth conditions for might-counterfactuals):

counterfactual discourse can either be exploratory or truth-aiming and this discourse-type affects the way in which the admissibility of precisifications of φ is regulated by something akin to Lewis's 1979a *conversational score*, *i.e.*, whether it is slackened or constrained.[11]

6.1. Back to CXM

Throughout previous sections, I have developed an ontic account of might-counterfactuals that – I hope to have demonstrated – meets DeRose's challenge. However, the overarching purpose of this paper, which I will now succinctly take on, was to show that this ontic account (unlike DT) is compatible with Stalnaker's CXM-preserving semantics.

I have claimed that $\varphi \Diamond\!\!\rightarrow \psi$ is true iff it is the case that ψ in at least one of the class of φ-worlds selected by the range of admissible precisifications of φ. To be sure, the precisification of φ that figures in $f(\varphi, i)$ is among the admissible precisifications of φ. Therefore,

$$\varphi \Box\!\!\rightarrow \psi \supset \varphi \Diamond\!\!\rightarrow \psi$$
$$\varphi \Box\!\!\rightarrow \sim\psi \supset \varphi \Diamond\!\!\rightarrow \sim\psi$$

And:

$$\sim(\varphi \Diamond\!\!\rightarrow \psi) \supset (\varphi \Box\!\!\rightarrow \sim\psi)$$
$$\sim(\varphi \Diamond\!\!\rightarrow \sim\psi) \supset (\varphi \Box\!\!\rightarrow \psi)$$

Notice in the bottom pair of validities that the bi-directional entailment, which held in DT, doesn't hold. This is a desired feature of my account which both provides an escape from inescapable clashes and renders the 'might' argument against CXM, outlined in Section 2.2, a non-starter.

7. Conclusion

Owing to the notorious problem of inescapable clashes, idealized epistemic accounts of might-counterfactuals, such as DeRose's 1999,

[11] Lewis's examples 2 and 6 on *Permissibility* and *Relative Modalities* respectively may be particularly relevant to the dynamics of counterfactual discourse function and precisification-admissibility.

have recently gained popularity over ontic accounts. In a different vein, the might argument against conditional excluded middle has rendered CXM a contentious principle to incorporate into a logic for conditionals. The aim of this paper has been to rescue both ontic might-counterfactuals and conditional excluded middle from these disparate debates and show how they are indeed compatible.

According to the proposed account of might-counterfactuals, the antecedent of a might-counterfactual is semantically underdetermined with respect to the antecedent-worlds it selects for evaluation. This explains (1a) how might-counterfactuals are able to select multiple antecedent-worlds as they apparently do and (1b) why the utterance of a might-counterfactual confers a weaker alethic commitment on the speaker than does the utterance of a would-counterfactual, as well as (2) provides an ontic solution to the problem of inescapable clashes. I have also briefly sketched how the semantic underdetermination, and consequently the truth conditions, of semantically underdetermined might-counterfactuals are regulated by the conversational context. Namely, a conversational score keeps track of the stringency of precisification-admissibility and thereby determines the truth conditions of any might-counterfactuals under evaluation.

The proposed account should be favored by those who share my intuition that there are counterfactually possible states of affairs which might-counterfactuals serve to describe. Additionally, unlike with epistemic theories which succumb to weak counterfactual skepticism, the proposed ontic theory is able to account for the concurrent knowability of would- and corresponding might-counterfactuals. Alternately, if the assumption that counterfactuals claims are knowable *at all* turned out problematic, this would undermine the central spirit of epistemic theories while leaving the ontic accounts practically intact.

Ivar Hannikainen
Department of Philosophy
University of Sheffield
45 Victoria Street
Sheffield, S3 7QB, UK
ivar.hannikainen@gmail.com

References

DeRose, Keith. 1991. Epistemic possibilities. *The Philosophical Review* 100, 581-605.

DeRose, Keith. 1999. Can it be that it would have been even though it might not have been? *Philosophical Perspectives* 13, 385-413.

Eagle, Antony. 'Might' counterfactuals. Unpublished. Retrieved through the author's website on September 26, 2010 from dl.dropbox.com/u/6362052/might-cfacts.pdf.

Fine, Kit. 1975. Vagueness, truth and logic. *Synthese* 30, 265-300.

Hacking, Ian. 1967. Possibility. *The Philosophical Review* 76, 143-168.

Hawthorne, Jon. 2005. Chance and counterfactuals. *Philosophy and Phenomenological Research* 70, 396-405.

Heller, Mark. 1995. Might-counterfactuals and gratuitous differences. *Australasian Journal of Philosophy* 73, 91-101.

Lewis, David. 1973. Counterfactuals. Oxford: Basil Blackwell.

Lewis, David. 1979a. Scorekeeping in a language game. *Journal of Philosophical Logic* 8, 339-359.

Lewis, David. 1979b. Counterfactual dependence and time's arrow. *Noûs* 13, 455-476.

Lewis, David. 1986. Postscript to 'Counterfactual Dependence and Time's Arrow'. In *Philosophical Papers Vol. II*. Oxford: Oxford University Press.

Stalnaker, Robert. 1968. A theory of conditionals. *Studies in Logical Theory: American Philosophical Quarterly Monograph Series* 2, 98–122.

Stalnaker, Robert. 1981. A defense of conditional excluded middle. In *Ifs: Conditionals, Belief, Decision, Chance, and Time*. Edited by W. Harper, R. Stalnaker and G. Pearce. Dordrecht: D. Reidel.

van Fraasen, Bas. 1966. Singular terms, truth-value gaps, and free logic. *Journal of Philosophy* 63, 481-495.

Williams, J. Robert G. 2008. Chances, counterfactuals and similarity. *Philosophy and Phenomenological Research* 77, 385-420.

Williams, J. Robert G. 2010. Defending conditional excluded middle. *Noûs* 44, 650-668.

On the transcendental deduction in Kant's *Groundwork* III[1]

Marilia Espirito Santo
UFRGS

Abstract
The purpose of the third section of Kant's *Groundwork* is to prove the possibility of the categorical imperative. In the end of the second section, Kant establishes that a proof like this is necessary to show that morality is 'something' and 'not a chimerical idea without any truth' or a 'phantom' (1785: 445). Since the categorical imperative was established as a synthetic *a priori* practical proposition, in order to prove its possibility it is necessary 'to go beyond cognition of objects to a critique of the subject, that is, of pure practical reason' (1785: 440). Kant names this kind of proof a *deduction*. The present paper intends to (1) show the argument whose purpose is to justify the categorical imperative; (2) show that the argument is a transcendental deduction; (3) present the argument as it is reconstructed by Allison, and (4) show that, although it seems compelling, the position of the commentator could not be accepted by Kant himself.

Keywords
Kant, *Groundwork*, transcendental deduction, moral law, categorical imperative.

Introduction

The notion of a deduction plays a central role in Kant's critical project. Nowadays, it points to a meaning quite familiar to us: it refers to the logical procedure by means of which a conclusion is established

[1] The first version of this paper I wrote with a scholarship from CAPES/Brazil during a research period in New York under supervision of Beàtrice Longuenesse, to whom I'm thankful for her extremely helpful criticisms. The final version I wrote with a scholarship from CNPQ/Brazil. Special thanks are owed to Luciano Codato.

through the relationship between some premises. Kant was familiar with this logical usage of the notion 'deduction', but as Henrich (1989: 31) remarks, it was neither the only, nor the most common usage in the academic language in the 18th century.

During the 18th century, 'deduction' was a notion used by jurists to refer to the written claims exposed to the Court in legal proceedings. Considering the argumentative structure of a deduction, one of its peculiar characteristics is that it must refer to an origin. Since the aim of a juridical deduction was to justify the legitimacy of a possession or a usage, that is, the legitimacy concerning an acquired right, it was necessary to explain how this possession or usage came into being. With this, it could be possible to decide who between both parties in the controversial juridical claim was right. The origin of an acquired right should be found in a fact, which must exist before the right in question came into being. That is why the argumentation presented in a deduction should relate the origin to fundamental facts that constitute it. Kant used the term *deduction* having in mind the deduction writings and not the logical procedure (see Kant 1781: A XII; A84/B116; A751/B779; A752/B780).

Kant called *metaphysical deduction* the task of referring to a non empirical origin, or to identify this origin before justifying the legitimacy of a possession or a usage. Once the *a priori* origin is identified it is possible to go on to the task of justification, which Kant called *transcendental deduction*. What he was trying to justify by a deduction was the possibility of synthetic *a priori* judgments. And this is the main ground for calling his project the critical philosophy. According to him, when we come to *a priori* judgments we have to consider their origin in the nature of reason itself and so to justify them and explain their possibility. These tasks of justification and explanation belong to a critique of reason by itself.

In relation to the practical use of reason, Kant is trying to prove that we, human beings, can act morally and, consequently, judge our actions. His starting point is the common rational cognition. He agrees with everyman that moral judgments are bivalent, that is, that we do can say that some actions are right and others wrong. On the other hand, he disagrees with positivists that moral judgments can be verified or justified by appealing to experience. He internalizes the origin of the moral law and defends that human beings *ought* to act morally well because the moral law is a self-imposed one. So, he identifies the origin of the moral law in reason, and this means that

the moral law is an *a priori* one. But since from the mere analysis of the concept of a human being does not follow the concept of acting morally well, or obeying the moral law, the moral law (for human beings) is said to be a synthetic *a priori* principle. And, as we saw above, it has to be justified by a transcendental deduction.

In other words, willing the good action is not necessarily contained in the volition of a human being, endowed with reason and sensibility, that is, it cannot be analytically derived from the volition of such a being. In 1785: 420n, Kant justifies that the categorical imperative is a synthetically practical proposition *a priori*, because it 'does not derive the volition of an action analytically from another volition already presupposed (for we have no such perfect will), but it connects it immediately with the concept of the will of a rational being as something that is not contained in it'.

Before we move on and analyze the deduction itself, it is important to bear in mind that the content and the origin of the moral law is something Kant has expounded in the first two sections of the *Groundwork*. In the first section, from an analysis of common rational cognition, he arrives at the condition for a moral action, which is obedience to a law. In the second section, he presents the content of the moral law by means of some formulations of it and identifies its origin in reason. The formulations are: (1) the formula of universal law, 'act only in accordance with that maxim through which you can at the same time will that it become a universal law' (1785: 421); (2) the formula of universal law of nature, 'act as if the maxim of your action were to become by your will a universal law of nature' (1785: 421); (3) the formula of humanity, 'so act that you use humanity, whether in your person or in the person of any other, always at the same time as an end, never merely as a means'(1785: 429); (4) the formula of autonomy, presented, initially, not as a command, but as 'the idea of the will of every rational being as a will giving universal law' (1785: 431), and (5) the formula of the kingdom of ends, 'that all maxims from one's own lawgiving are to harmonize with a possible kingdom of ends as with a kingdom of nature'(1785: 436). The formulas of universal law, humanity and autonomy are the primary formulations. The formulas of universal law of nature and of the kingdom of ends derived, respectively, from the formulas of universal law and of autonomy, can be called analogical formulas as Almeida suggests (2002).

The usual interpretation of the formula of universal law as ***the*** categorical imperative may lead one to believe that the formula of autonomy adds nothing to the other two primary formulations, that is, to those of universal law and of humanity. However, Kant does not just claim that the principle of autonomy follows from them, but also that the two primary formulations 'were only *assumed* to be categorical because we had to make such an assumption if we wanted to explain the concept of duty' (1785: 431). But those are not the only reason for us to take the formula of autonomy as the one that best expresses the unconditional duty. The principle of autonomy actually brings two ideas: one is the interest in acting on the categorical imperative; the other is the idea of conceiving not only oneself but all rational agents as universal legislators despite their own particular empirical ends. The idea of the moral agent not merely as acting in accordance with 'that maxim that at the same time can become a universal law', but as conceiving herself as a universal legislator and thus as the source of these maxims characterizes the interest in acting on the categorical imperative. Moreover, it states the rational nature not just as an objective end but also as an end that can motivate us. Besides this, by its analogical formula, the principle of autonomy brings the idea of not just oneself but all other rational agents as universal legislators, that is, as agents who despite their particular empirical ends have the same overriding interest in the universal law. It is only with the principle of autonomy that it is possible to understand how a manifold of agents with different empirical interests can accept a universal law. This idea, as Guyer reminds us (1998: 236), is necessary to prove that the categorical imperative is not just intentionally noncontradictory and coherent but also extensionally realizable, a requirement to demonstrate its real possibility.

The formula of universal law and its analogical universal law of nature express just the *form* of the principle of morality, which consists of universality. The isolated reading of this formula accuses Kant's moral theory of empty formalism. The formula of humanity, on the other hand, expresses just the *matter* of the principle of morality, which consists of the rational being, as an end by its nature and hence as an end by itself, as the limiting condition of all merely relative and arbitrary ends. But even with a matter to fulfill the form, it was still missing an interest in acting on the categorical imperative instead of an empirical interest in acting on hypothetical imperatives. Such an interest is introduced just with the formula of autonomy, together

with the understanding of a manifold of agents with diverse empirical particular ends but with the same overriding interest in a universal law, because it is only with the idea of autonomy that we have the complete determination of all moral maxims. This explains why, although Kant claims that the three primary formulas are just expressions of the very same law, they actually complement each other and that we really need the idea of autonomy to have access to the moral law.

Autonomy is the principle behind moral judgments and aims to be the condition for moral action. It expresses the essence of moral law, and it is the principle on which a rational agent would act if reason had full control over passion. That is, although Kant does not make it clear, the principle of autonomy does not necessarily need to take the form of a categorical imperative. Now, once we know which is the principle of morality, the next step is to ask about how it can be justified.

Kant first tries to justify it as a moral law and then as a categorical imperative. But since the autonomy expresses the essence of the moral law, which is the principle on which a rational agent would act if reason had full control over passion, and since it appears for a human being, who sometimes can act under the influence of passion, as a categorical imperative, that is, as a principle on which she *ought* to act, the question about the justification of the principle of autonomy can be expressed as 'how is a categorical imperative possible?'. And this is precisely the question Kant asks in the headline of subsection 4 of the third section of the *Groundwork*, and to which he will answer with a transcendental deduction.

I

In the third section of the *Groundwork*, Kant's argument is given in subsection 4, under the title 'How is a categorical imperative possible?' as follows:

> A rational being counts himself, as intelligence, as belonging to the world of understanding, and only as an **efficient cause** belonging to this does he call his causality a *will*. On the other side he is also conscious of himself as a part of the world of sense, in which his actions are found as mere appearances of that causality; but their possibility from that causality of which we are not cognizant cannot be seen; instead, those ac-

> tions as belonging to the world of sense must be regarded as determined by other appearances, namely desires and inclinations. All my actions as only a member of the world of understanding would therefore conform perfectly with the principle of autonomy of the pure will; as only a part of the world of sense they would have to be taken to conform wholly to the natural law of desires and inclinations, hence the heteronomy of nature. (The former would rest on the supreme principle of morality, the latter on that of happiness). But **because** the world of understanding contains the ground of the world of sense and so too of its laws, and therefore immediately lawgiving with respect to my will (which belongs wholly to the world of understanding) and must accordingly also be thought as such, **it follows** that I shall cognize myself as intelligence, though on the other side as a being belonging to the world of sense, as nevertheless subject to the law of the world of understanding, that is, of reason, which contains in the idea of freedom the law of the world of understanding, and thus cognize myself as subject to the autonomy of the will; **consequently** the laws of the world of understanding must be regarded as imperatives for me, and actions in conformity with these as duties. (1785: 453-4)

II

The core of the proof begins with the adversative conjunction 'but because…' and goes to the end of the paragraph. To analyze the argument, we can break it down in two main parts: the first one is stated by the element 'because' and provides a reason; the second is stated by the expression 'it follows' and provides a conclusion. The premises of the argument can be rewritten as follows:

> P1 – Because the world of understanding contains the ground of the world of sense;
> Corollary of P1 – and so too of its laws [and because the world of understanding contains the ground of the laws of the world of sense];
> P2 – and [because] is therefore immediately lawgiving with respect to my will (which belongs wholly to the world of understanding) and must accordingly also be thought as such;

Its conclusion, that can also be broken down in two parts, as follows:

> C – It follows that I shall cognize myself as intelligence, though on the other side as a being belonging to the world of sense, as nevertheless subject to the law of the world of understanding, that is, of reason,

> which contains in the Idea of freedom the law of the world of understanding, and thus cognize myself as subject to the autonomy of the will; C – consequently the laws of the world of understanding must be regarded as imperatives for me, and actions in conformity with these as duties.

The first premise is not difficult to understand. It states just that the world of understanding contains the ground of the sensible world. The meaning of the corollary that follows from P1 is also compelling. It is necessary, however, to pay attention to the meaning of *ground* in this passage, once Kant uses it with different meanings depending on the context. Sometimes *ground* is used as a synonym of *ratio*, reason, as well as cause, but it can still be found as a synonym of principle. In the context of the theoretical use of reason, specifically in the domain of his thesis about what it is a representation and how knowledge by representation is possible, Kant introduces a distinction between ground and cause (see Zingano 1989: 85). On the other hand, in the *Reflections on Metaphysics,* he presents *ground* as a *first cause* (see Kant 1769: R3972).

In the same manner, when he talks about practical grounds he claims that they are 'grounds of reason [that] provide the rule for actions universally, from principles, without influence from the circumstances of time and place' (1783: §53).

It is not at all our purpose here to list all the occurrences of the term *ground* and how it is used. The ambiguity of the term is solved in the context of its application, but it is necessary to be careful to avoid misinterpretations. In the text under analysis, Kant claims that 'the world of understanding contains (*enthält*) the ground of the world of sense'. *Ground*, here, points to the *efficient cause*, the first cause, the starting point. 'A rational being counts himself, as intelligence, as belonging to the world of understanding, and only as an *efficient cause* belonging to this does he call his causality a *will*'. The world of understanding *contains* the rational being as an efficient cause. That is, the rational being, through her reason, can bring about changes in the world of sense because it is the primary origin of movement. The rational being can be an efficient cause as long as she can be considered from a double standpoint: member of the world of understanding and part of the world of sense. Since she is also part of the world of sense, her property as an efficient cause could not be realized, that is why the necessity of a command, of an imperative.

The term *ground* can only be understood as *efficient cause* in the passage under analysis because Kant claims that the 'world of understanding *contains* the ground of the world of sense'. If his claim were that the world of understanding *is* the ground of the world of sense, ground should be understood as *ratio*, reason and not cause. If the world of understanding were the *cause* of the world of sense, we would be admitting a transcendent use of the principle of causality. A use beyond the *phenomena* that would transgress the limits that Kant himself is trying to establish. So, if we took *ground* as *cause* we would be attributing to Kant dogmatism. But, since the world of understanding contains the rational being as an efficient cause, a rational being who is also part of the world of sense, we preserve Kant within the limits of criticism.

The second premise is a little bit more problematic. One might think that because of P1 and its corollary, that is, because of the relation between the worlds of understanding and of sense established in P1, the world of understanding is also immediately legislative for my will, which belongs entirely to the world of understanding. However, this cannot be true.

To better understand P2 we can rewrite it in the following way:

> P2 – In relation to my will, which belongs wholly to the world of understanding, the world of understanding is directly legislative, and it must also be conceived as containing the ground of actions and laws of the world of sense.

Actually, we can break P2 down in two premises. The sentence which appears between parentheses should be read as an independent premise. So, we have P2 and P3 as follows:

> P2 – In relation to my will, the world of understanding is directly legislative, and it must also be conceived as containing the ground of actions and laws of the world of sense.
> P3 – my will belongs wholly to the world of understanding.

The meaning of the P2 itself is not problematic, what is problematic is the relation between it and P1; better, the problem is relative to the element 'because' in the beginning of P1 and the function of the expression 'and therefore', which introduces P2. That is, one might think that P2 is a conclusion that follows from P1. However, this would be a complete misinterpretation of what Kant is arguing for. It

would make no sense to support that *because* the world of understanding contains the ground of the world of sense and of its laws, it is also directly legislative for the will. It would make no sense because the will belongs entirely to the world of understanding (P3); it does not belong to the world of sense. So, there is no such type of relation between P1 and P2. The simplest way to solve this misinterpretation is to put a 'because', or any other element that indicates a reason, in the beginning of P2. So, it would be properly read as a premise in addition to P1.

Moreover, it is important to stress, that the 'it', in the second part of P2, refers to the world of understanding. Hence, what P2 expresses is that (1) the world of understanding is directly legislative to the will and (2) in relation to my will, the world of understanding 'must also be conceived as containing the ground of actions and laws of the world of sense'. Thus, from P2 and P3, it is possible to say that the will, as part of the world of understanding, contains the ground of the world of sense and its actions and laws. And this is the gist of the deduction: that is, that the pure will as part of the world of understanding contains the moral law as a categorical imperative for this will affected by desires and inclinations, as part of the world of sense.

The conclusion of the argument is a little bit easier to understand, although it is not completely evident. To analyze it, it is possible to break it down in two parts, which are separated by the occurrence of the element 'consequently'.

> **[…] it follows** that I shall cognize myself as intelligence, though on the other side as a being belonging to the world of sense, as nevertheless subject to the law of the world of understanding, that is, of reason, which contains in the idea of freedom the law of the world of understanding, and thus cognize myself as subject to the autonomy of the will;[…]

In this first part, Kant maintains that the human being (*I*), inasmuch as she considers herself as intelligence and *at the same time* as a being that belongs to the world of sense, is subject to the law of the world of understanding and to the autonomy of the will. It is important to bear in mind the conjunction *and at the same time* since the beginning of the

sentence, although it does not appear in this passage[2]. It is not the being only as intelligence that is subject to the law of the world of understanding and to the autonomy of the will; it is the being that is both: intelligence and sensible, that is subject to reason (which contains the law of the world of understanding) and to the autonomy of the will. It would be a misunderstanding to take Kant to be supporting that a rational being *only* as intelligence is subject to the law of the world of understanding and to the autonomy of the will. If it were this, it would be impossible to explain how a human being is subject to the moral law and can take it as a motive to her actions, that is, it would be impossible to prove morality under human conditions. Moreover, the element 'nevertheless' would not be necessary. For a being who is only intelligence, it is not necessary to consider herself in an adversative way subject to the law of the world of understanding and to the autonomy of the will.

In relation to the second part of the conclusion, **'consequently** the laws of the world of understanding must be regarded as imperatives for me, and actions in conformity with these as duties', it is possible to say that what is a law of the world of understanding must appear, or be considered by a human being (*me*), as an imperative to such a being. This is because she is not only intelligence, she is also part of the world of sense, so her will can also be affected and hence motivated to act by a law of this last world. And since an imperative is just 'the formula of a command of reason and is expressed by an 'ought'' (Kant 1785: 413), the actions according to it are called duties.

Thus, almost without realizing, the reader is faced with a complete deduction[3]. This deduction is the answer Kant provides to the

[2] In a number of passages throughout the *Groundwork,* Kant emphasizes the simultaneity of both perspectives (intelligible and sensible) in relation to human beings and the imperative character of the moral law. This simultaneity plays a fundamental role here.

[3] For Liddell, the paragraph we analyzed and took to be the whole deduction is just the second part of Kant's deduction in the *Groundwork*. The author affirms that the deduction begins in the subsection 'Freedom must be presupposed as a property of the will of all rational beings', and its second part is presented in the subsection 'How is a categorical imperative possible?' (1972: 401-2). We cannot agree with Liddell. For us, the argument presented in the subsection 'Freedom must be presupposed (…)' is important for the deduction, but it is just a preparatory argument together with the subsection 'Of the interest attaching to the ideas of

question 'How is a categorical imperative possible?'. The gist of the deduction is that the pure will as part of the world of understanding contains the moral law as a categorical imperative for this will as part of the world of sense. Behind the deduction there is the idea that somehow, the rational perspective of a human being is superior (in the sense of being an efficient cause) to her sensible perspective, and that the pure practical will contains the supreme condition of the will affected by sensible desires.

In the paragraph following the one we analyzed, Kant again answers the question 'how is a categorical imperative possible?' (see 1785: 454), and this answer can be taken as a summary of the deduction just presented, one of the formal characteristics of a good deduction as Henrich (1989: 34) points out. Next, Kant claims, with a concluding remark that, 'the practical use of common human reason confirms the correctness (*Richtigkeit*) of this deduction' (1785: 454)[4].

Now we can conclude that (1) if the purpose of a transcendental deduction for Kant is to justify the legitimacy of a possession or a usage of a synthetic *a priori* judgment or principle; (2) if the autonomy, under human conditions, was proved to be a categorical imperative and hence a synthetic *a priori* principle; (3) if membership in the world of understanding is what justifies that the human being, a finite rational being, is autonomous and hence can act morally well, and (4) if the paragraph we analyzed is the argument that proves that the human being is a member of the world of understanding, therefore this paragraph presents a complete transcendental deduction.

morality'. This seems to be also Kant's idea, since he affirms, at the end of the first subsection 'The concept of freedom is the key to the explanation of the autonomy of the will', that he cannot yet answer the question of how a categorical imperative is possible, because 'some further preparation is required'. So subsections 2 and 3 are the preparatory argument to the answer that will be given in subsection 4, where Kant presents the deduction itself.

[4] In the second *Critique*, Kant again confirms the success of the deduction of the *Groundwork*: 'It [the *Critique of Practical Reason*] presupposes, indeed, the *Groundwork*, but only insofar as this constitutes preliminary acquaintance with the principle of duty and provides and **justifies** a determinate formula of it; (…)' (1788: 8).

III

It is important to note that our interpretation is distinguished from interpretations of well-known Kant scholars. Our present purpose is to analyze Allison's thought in relation to the deduction of the third section of the *Groundwork* and show why we think Kant himself would not concur[5].

According to Allison, the deduction, whose pivotal point is the move from possession of reason to membership in the intelligible world, can be presented in seven steps:

(1) 'Now I assert that every being who cannot act except under the idea of freedom is by this alone – from a practical point of view – really free' (Kant 1785: 448).
(2) 'And I maintain that to every rational being possessed of a will we must also lend the idea of freedom as the only one under which we can act' (Kant 1785: 448).
(3) All laws 'inseparably bound up with freedom' are valid for every being with reason and will.
(4) But the Reciprocity Thesis establishes that the moral law is 'inseparably bound up with freedom'.
(5) Therefore, the moral law is valid for every being with reason and will.
(6) Since beings such as ourselves have reason and will, the moral law is valid for us.
(7) Since we do not necessarily follow the dictates of the law (these dictates being 'objectively necessary' but 'subjectively contingent'), the law for us takes the form of a categorical imperative, that is, we are rationally constrained, although not causally necessitated, to obey it.

Steps 1 and 2 consist of a preparatory argument, as the commentator calls it. Step 7 consists of 'a distinct deduction of the categorical

[5] Allison's work on Kant's moral theory has been criticized for some time now. That is why it is important to point out that, as we understand them and as far as it goes with our reading, the criticism raised here differs completely from those raised by Stephen Engstrom, Andrews Reath, Karl Ameriks and Paul Guyer to whom Allison replies in his 'Kant on freedom: a reply to my critics', 1996.

imperative' (Allison 1995:224). Steps 3 to 6[6] consist, therefore, of a deduction of the moral law. And here it is the first distinction between our and Allison's analysis.

Allison notes that the third section of the *Groundwork* is one of the most enigmatic of Kantian texts. Although it is clear that its main purpose is to justify the supreme principle of morality, articulated in the first two sections, and for that Kant appeals to a deduction; it is not clear whether the deduction is of the moral law, the categorical imperative, freedom, all three, or even whether it can be properly characterized as a deduction at all. Allison's argument is for a deduction of the moral law, and his underlying presupposition is the reciprocity thesis.

As already indicated, our first disagreement with this interpretation is about what Kant is trying to justify by a deduction. We argue for a deduction of the categorical imperative, Allison argues for a deduction of the moral law.

In section two of the *Groundwork,* Kant claims that the principle of autonomy is a categorical imperative 'cannot be proved by mere analysis (…), because it is a synthetic proposition' (1785: 440). For such a proof 'one would have to go beyond cognition of objects to a critique of the subject, that is, of pure practical reason' (1785: 440), a business that 'does not belong in the present section' (1785: 440) says Kant. Moreover, in the end of the same section, Kant emphasizes that the proof that morality is not a chimera to human beings 'follows if the categorical imperative (…) is true and absolutely necessary as an *a priori* principle' and this 'requires a possible *synthetic use of pure practical reason*, which use, however, we cannot venture upon without prefacing it by a *critique* of this rational faculty itself, the main features of which we have to present, sufficiently for our purpose, in the last section' (1785: 445).

In the first part of section three, Kant affirms that 'if (…) freedom of the will is presupposed, **morality** together with its principle **follows** from it **by mere analysis** of its concept. But the **principle of morality** (…) **is** nevertheless always **a synthetic proposition** (…)' (bold added) (1785: 447). Then he goes on to state that free will must be attributed to every rational being (1785: 448). It

[6] According to Allison, steps 1 and 2 constitute the explicit argument Kant provides, steps 3 to 7 constitute the natural extension of the argument he does not make.

would be contradictory to suppose a rational being who could be regarded as not free; that is, if we deny freedom, we necessarily deny reason. So, if we have to attribute free will to every rational being, if from freedom of the will follows morality by mere analysis, we can say that from the concept of a rational being follows morality by mere analysis. We must note that this is true when we think about a pure rational being. To a finite rational being, the principle of morality is 'always a synthetic proposition', because by analysis of its concept does not follow to act morally well. Given its finitude, for such a being it is possible to act different from what morality dictates. That is why to complete the deduction of the categorical imperative Kant has to appeal to a distinction between two worlds (world of understanding and world of sense) as two standpoints from which imperfectly rational beings may regard themselves. Moreover, in the same section, right after the argument that justifies the possibility of the categorical imperative, Kant claims that 'the practical use of common human reason confirms the correctness of this deduction' (1785: 454).

Now, according to the passages we quoted it seems reasonable to regard Kant's attempt to prove the possibility of the categorical imperative, and not of the moral law, by a deduction. Moreover, since a deduction is to prove the transcendental conditions of a possession or a usage of a synthetic *a priori* judgment or principle, and since the moral law follows analytically from the concept of reason, a deduction would not be necessary to justify it. A deduction of the moral law, that is, a deduction of an analytic principle does not have the character of justification, but of demonstration, and once more this is not the task Kant intends to develop here.

But, even if Allison's suggestion was right that the deduction is of the moral law[7] and not of the categorical imperative, he argues for a

[7] What Allison may have in mind (he does not make it clear or explicit) when he affirms that the deduction is of the moral law is that what Kant is trying to justify by a deduction is why **this** one (autonomy, 'the idea of the will of every rational being as a will giving universal law' (Kant 1785: 431)) and not another is **the** moral law. The answer we can provide to him is that **this is the moral law** because of the *complete determination*. It is only because of autonomy, whose principle according to Kant (1785:431) follows from the conjunction of the principles of the universal law and of the humanity, that a rational agent can conceive herself as capable of renunciation of all interest in volition from duty, which is the specific mark of a moral

failure of the deduction due to a fatal ambiguity in two central notions. The first is in that of the intelligible world and the second in that of the will.

In relation to the ambiguity in the notion of an intelligible world, Allison asserts that Kant refers to both a *Verstandeswelt* (world of understanding) and an *intelligibele Welt* (intelligible world) and shifts from the former to the latter without sufficient justification. In doing this, Kant cannot avoid providing a 'non-question-begging deduction of the moral law in *Groundwork* III' (Allison 1995:228).

The *Verstandeswelt* is to be understood negatively as encompassing whatever is nonsensible or 'merely intelligible', that is, whatever is thought to be exempt from the conditions of sensibility (the *noumenon* in the negative sense). The *intelligibele Welt*, on the other hand, is to be understood positively as referring to a supersensible realm governed by moral laws, a 'kingdom of ends' or 'the totality of rational beings as things in themselves' (Kant 1785: 458) (the *noumenon* in the positive sense).

Allison notes that Kant's goal is to show that human beings are members of an *intelligibele Welt* because this would entail that they stand under the moral law. The problem is that the possession of reason only gets us to a *Verstandeswelt*, and since this world is an indeterminate concept, it cannot provide any conclusion about the nature of the rational being as a whole nor about her will.

The second difficulty, related to an ambiguity of the notion of the will, is a corollary of the former. The main point is that given the identification of will and practical reason, the claim that rational beings possess a will can mean (1) merely that reason is practical or (2) that pure reason is practical. The former (practical freedom) is sufficient for us to affirm that we are genuine rational agents rather than automata; but it is the second (transcendental freedom) that is necessary to establish our autonomy.

The problem, again, is that the membership in the *Verstandeswelt* provides support just for practical freedom, but it is transcendental freedom that is necessary and sufficient to establish morality on the basis of a nonmoral premise about our rationality.

will, and understand how herself and all others rational agents in spite of their particular empirical ends have an overriding interest in a universal law.

Thus, Allison concludes for a failure of the deduction, which, according to him, Kant himself may have recognized. Assuming that, the commentator claims that

> we can see why he <Kant> would abandon the attempt to establish the practicability of pure reason on the basis of any nonmoral premise. Thus, instead of beginning with the concept of a rational agent and moving from this first to the presupposition of freedom and then, via the Reciprocity Thesis, to the moral law, Kant there <*Critique of Practical Reason*> moves directly from the consciousness of the moral law as the 'fact of reason' to the practicability of pure reason and the **reality** of transcendental freedom (1995: 228). (Emphasis mine)

IV

Our critique of Allison is based on three points. First, knowing the way in which Kant uses the terms, it is more reasonable to consider him to be using the terms of a *Verstandeswelt* and of an *intelligibele Welt* not in a univocal sense, but, sometimes, interchangeably. Second, that in the *Groundwork*, it is not Kant's purpose to prove the reality (objective validity), but the real possibility[8] of the categorical imperative, and for this the *noumenon,* in the negative sense, is necessary and sufficient. Finally, that which gets us to the intelligible world it is not just the possession of reason but also the *consciousness of the spontaneity of reason*.

The *noumenon* in the negative sense is a being of understanding 'insofar as it is not an object of our sensible intuition' (Kan 1781: B307), but it can be under determination of space and time, that is, it is the

[8] See the passage of the first *Critique* where Kant talks about the work of the jurists (1781: A84/B116), that they distinguish between what is lawful (*quid juris*) and what concerns the fact (*quid facti*) and that they call the first a deduction. In the third section of the *Groundwork* Kant is working with a lawful question, the *quid fact* he will deal with only in the second *Critique*. Where, by the way, he claims that '**the moral law** is given, as it were, as a **fact of pure reason** of which we are *a priori* conscious and which is apodictically certain (...). Hence the **objective reality** of the moral law **cannot be proved by any deduction** (...)'. (1788: 47). And further he claims again that 'the objective reality of a pure will or, what is the same thing, of a pure practical reason is given *a priori* in the moral law, as it were by a fact – for so we may call a determination of the will that is unavoidable even though it does not rest upon empirical principles' (1788: 55).

object that *can appear* (as a *phaenomenon*). Admittedly, it is the *noumenon* in the negative sense that Kant needs to justify the real possibility of the categorical imperative because the categorical imperative is the way the moral law, a law of a being of understanding, appears to a being that is not only a being of understanding, but it is also a being of sense, a being that is under determination of space and time. To prove the possibility of the categorical imperative is to prove that the sensible affected will can give meaning, through its actions, which appear in space and time, to the rules of the pure rational will, and for that, the *noumenon* in the negative sense is necessary and sufficient. If we have just the *noumenon* in the positive sense, that is, 'the object of a non-sensible intuition (…), namely intellectual intuition' (1781: B307), we would have two different worlds and no connection between them. Hence, it would be impossible to justify how a being that is also part of the sensible world could be motivated to act by a law of the intelligible world. It is the *noumenon* in the negative sense that allows us to understand the intelligible world and the sensible world as a double standpoint of the same world, a double standpoint that the human being considers herself and allows to understand why she has to act morally well.

Finally, our last objection to Allison is that what gets us to the intelligible world it is not just the possession of reason but also the *consciousness of the spontaneity of reason*. And this seems to give positive content to our thought of ourselves as members of the intelligible world and, hence, a positive content to the concept of an intelligible world itself. That is, it seems reasonable to support that the consciousness of the spontaneity of reason presupposes a law different from that of nature, and this allows a positive characterization of the intelligible world and, therefore, a characterization of the human being as a *noumenon* in the positive sense.

Notwithstanding, Almeida (2009: 45) notes that the characterization of the human being as a *noumenon* in the positive sense, allowed by the consciousness of the spontaneity of reason, can lead us to another problem. The problem is that although such a rational being can 'transport' herself to the intelligible world by the consciousness of the spontaneity of theoretic reason this is not sufficient to ascribe the same spontaneity to practical reason. For this, an independent moral premise would be necessary. And this, according to him, is the unsolvable problem that made Kant abandon his attempt to prove the

supreme principle of morality by a deduction and appeal to a *fact of reason* in the second *Critique*.

However we will need to address this problem in future work. In the beginning of this paper, we stated as our objectives to show the argument whose purpose is to prove the real possibility of the categorical imperative; to show that the argument is a transcendental deduction; to present the argument as it is reconstructed by Allison, and, finally, to show that, although it seems compelling, the position of the commentator would not be accepted by Kant himself.

Marilia Espirito Santo
Universidade Federal do Rio Grande do Sul (UFRGS)
Programa de Pós-graduacao em Filosofia
Av. Bento Goncalves, 9500 – predio 43311, bloco Al, sala 110
CEP 91501-970 – Porto Alegre, R.S., Brazil
mariliae@yahoo.com.br

References

Allison, Henry. 1995. *Kant's theory of freedom*. Cambridge: Cambridge University Press.

Allison, Henry. 1996. Kant on Freedom: a reply to my critics. In *Idealism and Freedom: essays on Kant's Theoretical and Practical Philosophy*. Cambridge: Cambridge University Press.

Almeida, Guido. 2002. Sobre as "Fórmulas" do Imperativo Categórico. In DOMINGUES, Ivan., PINTO, Roberto Paulo M., DUARTE, Rodrigo. *Ética, Política e Cultura*. Belo Horizonte: ed. UFMG.

Almeida, Guido. 2009. *Fundamentação da Metafísica dos Costumes* / Immanuel Kant. Tradução com introdução e notas por Guido Antônio de Almeida. São Paulo: Discurso Editorial: Barcarolla.

Guyer, Paul. 1998. The Possibility of the Categorical Imperative. In *Kant's Groundwork of the Metaphysics of Morals: critical essays*. Lanham: Rowman & Littlefield Publishers.

Henrich, Dieter. 1989. Kant's notion of deduction and the methodological background of the first Critique. In Foster, E. (ed.) *Kant's Transcendental Deductions. The Three "Critiques" and the "Opus Postumum"*. Stanford, California: Stanford University Press.

Kant, Immanuel. 1769. R3972. In *Kant's gesammelte Schriften*. Berlin: Walter de Grunter & Co, 1926. V 17.

Kant, Immanuel. 1781. *Critique of Pure Reason*. Translated by Paul Guyer and Allen Wood. Cambridge: Cambridge University Press, 1998.

Kant, Immanuel. 1783. *Prolegomena to Any Future Metaphysics*. Translated by Gary Hatfield. Cambridge: Cambridge University Press, 2007.

Kant, Immanuel. 1785. *Groundwork for the Metaphysics of Morals*. Translated by Mary Gregor Cambridge: Cambridge University Press, 2009.

Kant, Immanuel. 1788. *Critique of Practical Reason*. Translated by Mary Gregor Cambridge: Cambridge University Press, 2009.

Liddell, Brendan. 1972. Kant's 'deduction' in the '*Grundlegung*'. In Proceedings of Third International Kant Congress: 401-406, ed. by Beck, Lewis White. Dordrecht-Holland: D. Reidel Publishing Company.

Zingano, Marco Antonio. 1989. *Razão e História em Kant*. São Paulo: editora Brasiliense.

Visual Experience and Demonstrative Thought

Thomas Raleigh
Becario del Programa de Becas Posdoctorales de la UNAM,
Instituto de Investigaciones Filosóficas

Abstract
I raise a problem for common-factor theories of experience concerning the demonstrative thoughts we form on the basis of experience. Building on an insight of Paul Snowdon 1992, I argue that in order to demonstratively refer to an item via conscious awareness of a distinct intermediary the subject must have some understanding that she is aware of a distinct intermediary. This becomes an issue for common-factor theories insofar as it is also widely agreed that the general, pre-philosophical or 'naïve' view of experience does not accept that in normal perceptual cases one is consciously aware of non-environmental (inner, mental) features. I argue then that the standard common-factor view of experience should be committed to attributing quite widespread referential errors or failures amongst the general, non-philosophical populace – which seems an unattractively radical commitment. After clarifying the various assumptions I am making about experience and demonstrative thoughts, I consider a number of possible responses on behalf of the common-factor theorist. I finish by arguing that my argument should apply to any common-factor theory, not just avowedly 'indirect' theories.

Keywords
Visual Experience, Demonstrative Thought, Common-factor, Intentionalism, Paul Snowdon.

0. Overview

There is an ancient debate as to whether what visual consciousness provides – the 'visual field' – is, at least sometimes, access to the mind-independent environment itself, or to some sort of features that can be common to perception and hallucinations/dreams. This latter,

orthodox view, is held by various theories as to the metaphysics of experience – sense-data theories, adverbial theories and most intentional or representational theories, are all *common-factor* theories[1]. I want to raise an issue for common-factor (CF) theories concerning the account they yield of demonstrative thoughts formed on the basis of experience.

According to what I will henceforth call the natural story, when we selectively attend to – or 'single-out' – some feature or element in our visual field, we can thereby form a demonstrative thought about it. Sometimes one can look at an item and attend to it in one's visual field as part of thinking about some other, distinct item. For example: S can attend to a waxwork sculpture of Barack Obama in her visual field and think '*That* is the first black president of the Harvard Law Review'. S's thought here can refer to Obama rather than to the sculpture. However, it seems that in such a case S needs to have some *understanding* that the 'proximal' item in her visual field is distinct from, but related to, the other 'distal' item. If S really would not acknowledge that she is looking at a model/representation of Obama rather than the real Obama, then it seems that her demonstrative thought should count as false. She has mistakenly demonstrated and so referred to something other than the first black president of the Harvard Law Review.

Why should this be a problem for common-factor theories? Well, it is commonly accepted that such theories are *revisionary* with respect to everyday, non-philosophical beliefs. In other words, non-philosophers would *not* in general acknowledge as true the claim that the elements comprising the visual field are distinct from (though related to) any environmental features. (Hence the label: 'naïve realism'.) So they would be regularly mistaking a non-environmental feature for an environmental feature. But given the above line of reasoning, such a mistake should mean that the non-philosophical laity would be quite systematically failing to refer to features in their environment when they form demonstrative thoughts via experience – and so quite systematically forming false demonstrative thoughts via experience (or perhaps failing to form thoughts that refer at all). This

[1] Some versions of intentionalism, e.g. Tye 2009, allow for object-dependent content in perceptual cases, with 'gappy' content in the case of hallucinations, and so are not so clearly 'common-factor' theories.

seems like an unattractively radical consequence for a theory to have, though it is not necessarily a decisive flaw.

The plan for this paper is as follows:

(1) First I'll outline some assumptions contained in what I've called the 'natural story' about demonstrative reference via experience.
(2) Secondly I'll flesh out a little more the central claim, which I draw from Paul Snowdon 1992, that experience of a distinct intermediary can only be a means for demonstratively referring to some other item if the subject understands that it is the intermediary that they are experiencing.
(3) Thirdly, I'll consider some possible complicated cases, which might be thought to be counter-examples to Snowdon's point, where there are competing means or channels that might be determining the reference of a demonstrative thought.
(4) Fourthly, I'll consider some possible responses by CF theorists.
(5) Finally, I'll make a few remarks about the distinction between 'direct' and 'indirect' awareness and about intentionalist theories of experience.

1. Some Assumptions

It has seemed plausible to many philosophers that when we perceive an item (object or feature) we are able to think about it in a certain way. Perceptual experience allows for *demonstrative thought*. Demonstrative thought here contrasts with descriptive thought. A descriptive thought about O involves thinking of O via a descriptive or conceptual mode of presentation, whereas in a demonstrative thought one refers to O without any such descriptive/conceptual mode of presentation. Here is Robin Jeshion articulating this distinction:

> 'I can think that a particular rose is lovely by thinking "*the tallest yellow rose in the garden is lovely*". My thought is about that particular rose because it satisfies, 'fits', the descriptive condition *the tallest yellow rose in the garden*. Alternatively, I can think of these individuals in a way that does not depend essentially on my mode of conceptualising them. I can visually attend to the rose itself and think *that is lovely*, where "that", as it functions in my thought, refers deictically to the object I attend to – that very rose.' (Jeshion 2010: 1)

I will not attempt to defend or motivate this widely accepted distinction between demonstrative and descriptive thoughts – I will simply assume that there is such a distinction. Jeshion's passage contains what I take to be a very natural and plausible story: by selectively attending to some element in one's visual field, one is able to form a demonstrative thought about it and to demonstratively refer to it. On this story, the act of directing one's (cognitive) attention is (partially) determining the reference of the demonstrative component of one's thought. (I will generally speak of 'singling out' for this mental act of focusing/directing cognitive attention at one particular element in one's visual field.)

In his *Reference and Consciousness*, John Campbell 2002 argued for the thesis that:

- Conscious experience of O is necessary for demonstrative reference to O.

Against this thesis it could be claimed that, at least in principle, non-conscious perception of O could allow for demonstrative reference to O. Or indeed it might be claimed that perception of O, whether conscious or not, is unnecessary for demonstrative thought about O. I will *not* attempt to defend or motivate this thesis[2]. The natural story about experience and demonstratives requires only a weaker thesis:

- Normal humans can, and often do in fact, make demonstrative reference to O via their conscious experience of O.

In other words, whether or not it is the only way to demonstratively refer to O, I assume that a subject's conscious experience of O *can* play a role in allowing the subject to demonstratively refer to O.

In order to be able to selectively attend to an element in one's visual field (i.e. to 'single it out') and so to 'tag' it with a demonstrative '*that*…', one need *as yet* know nothing at all about the element. One could simply wonder: 'What's *that*?' without yet having formed any beliefs about the item one sees and attends to. Perhaps one is always *in a position* to form some true belief about an element in one's visual experience – i.e. I don't rule out that some form of infallibility thesis

[2] Smithies 2009 also argues in favour of this thesis, though on somewhat different grounds to Campbell.

is true. But attending to some aspect of one's experience is surely explanatorily prior to forming a belief about it; one does not need any belief about the element *in order* to attentively single the element out. And perhaps one will always know something trivial, such as: *this* is the item I am currently visually aware of. But such knowledge is not explanatory of the successful demonstrative reference – we can imagine a cognitively unsophisticated creature, which lacks any concept of visual experience, yet which still has the capacity for simple demonstrative reference to the items it sees.

These all strike me as very plausible and natural claims and so are fairly minimal assumptions to have made – though, no doubt, there might be ways to challenge them.

Now, the term 'visual field' is ambiguous between the portion of physical environment that lies within a subject's visual range, and the subject's visual experience of that environment from a particular perspective – which, according to a common-factor theorist at least, does not constitutively involve the environmental features. But this ambiguity suits our purpose here, as we need a term that is neutral between relational and common-factor ('direct' and 'indirect') theories of perceptual experience. As I am using it in this paper then, the 'visual field' consists of *whatever* it is that visual consciousness delivers or makes available for the subject. So I am allowing that environmental features and/or mental/inner features *might* be elements within the visual field.

As well as the traditional metaphysical debate about the nature of the visual field, there is a further question concerning what might be called the structure or articulation of the visual field. Imogen Dickie 2010 argues against the idea, which she labels the 'Old Empiricist View', that 'perception delivers only a shifting mosaic of features, which you will call "colour (or texture, or shape) patches" or "sense-data" depending on whether you are prepared to allow that they exist independently of our experience of them'[3](Dickie 2010: 214). So

[3] Actually, the term 'sense-data' was used by G. E. Moore to mean mind-independent but non-environmental features, and by Bermúdez 2000 to mean mind-independent but non-objectual environmental features (surfaces). But this is just terminology.

orthogonal to the traditional metaphysical debate[4] there is also the following question:

> Whether perceptual experience provides 'an array of features laid out around us and developing over time' (Dickie 2010: 214), or whether it delivers a visual field already divided into 'units' or 'visual objects'?

Dickie points to empirical results[5] which suggest the latter answer; the visual field normally comes pre-divided into 'visual objects', units which have such characteristics as spatio-temporal continuity and moving/acting as a whole, not as a mosaic of transient, shifting quality patches. (By pre-divided, Dickie means the 'processing' involved is sub-personal, pre-attentive and pre-conceptual.) These strike me as fascinating results, which are entirely in line with the natural story. But I don't need to take any position on the structure/articulation of the visual field for the purposes of this paper.

2. Snowdon's Distinction

Let's now turn to considering the sort of everyday case that would intuitively count as an instance of indirect awareness – that is, cases in which one might be able to 'see' one environmental item in virtue of *really* seeing some other environmental item. E.g. I see O's shadow, but I do not see O itself. Or, I see a photograph of O but I do not see O itself. These are cases in which I clearly have visual awareness of some item, M, which is distinct from O. (Of course there are also all sorts of difficult, unclear cases – I discuss these below in section 5.) Here the following seems quite obvious: if I have no idea that M bears any sort of relationship to O, then when I mentally single out M in my visual field and think 'What's *that*?', or think '*That* is F', my demonstrative refers to M and not O. E.g. I see (what is in fact) the

[4] E.g. Bermúdez 2000 argues that the visual field *is* constituted by mind-independent, environmental features but is *not* segmented/articulated into objects. Bermúdez claims that we directly experience environmental sense-data, i.e. external/mind-independent quality patches or 'surfaces' that are not standard physical objects.

[5] Dickie recommends, in her second footnote, Pylyshyn 2003, 2007 for a survey of this evidence. See also Campbell 2006 and the empirical work cited therein.

shadow of a rabbit and think 'What's *that*?', or think '*That* is moving quickly'. If I have no grasp or understanding whatsoever that the thing I see is related somehow to the rabbit, but distinct from the rabbit, then there is no way that my demonstrative thought can be referring to the rabbit. I must have referred to the shadow.

Of course, as soon as I have some grasp or understanding of the fact that M and O are distinct (but related) items, then I might single out M in my visual field whilst my thought refers to O. So long as I minimally realise that what I'm looking at is not literally the rabbit, but some distinct thing related to it, then I could, for example, mentally single out the shadow in my visual field whilst my thought refers to the rabbit. E.g. given such understanding[6], my thought '*That* is a rabbit' would be true.

Now let's consider a case in which I *mistake* M, the item I see and attentively single out, for O. E.g. I see a brilliantly life-like trompe l'oeil picture of my rabbit that fools me – I wrongly believe that the thing I'm looking at literally is my rabbit. I fail to understand that the item I see is an entity distinct from my rabbit – I would not acknowledge that there is a picture before my eyes. And now, whilst singling-out the picture in my visual field, I think: '*That* is my rabbit'. What does my demonstrative thought refer to here?

I think it is pretty clear that this thought must count as false, the demonstrative having picked out a picture rather than my rabbit. At the very least we can say that:

[6] When I say that a subject needs some grasp or understanding of the indirect nature of their situation, I do not mean that they must *explicitly think* of it at the time of making their demonstrative judgement. I only mean that they *would* have acknowledged, if asked, that they were really aware of something distinct from O. (Again, I follow Snowdon 1992 here). E.g. whilst watching a film I can become "immersed" in the action so that I lose any explicit, occurent awareness that I am looking at an image on a flat screen. I might then think: '*That* is Bruce Willis'. My thought here does still successfully refer to the real Bruce Willis so long as I *would* acknowledge, if asked, the truth of some such claim as: '*That* is really an image of Bruce Willis'. I don't need to consciously endorse such a claim when I think my original demonstrative thought. But if I really would not have acknowledged this fact, had I been asked, then I am in trouble – for I must believe that Bruce Willis is literally located inside my TV. I would be mistaking a flat, coloured image for a real human being.

> *If* my demonstrative thought has its reference fixed by my attentive singling-out act, then I have referred to M (the picture), not O (the rabbit).

So my thought would be false, but other thoughts might, fortuitously, have been true – e.g. if I had thought '*That* is white' where picture and rabbit are both white. The point is that when I am mistaken about which entity it is that's being attentively selected within the array, then a demonstrative based on this singling-out will not refer as I expect it to.

Snowdon 1992 suggests that the distinction between direct and indirect awareness can be elucidated on this basis. When S has direct visual awareness of O, S does not need any particular knowledge or belief about O in order to be able to demonstratively refer to O (via a visual singling-out act). But when S has indirect visual awareness of O, via visual awareness of some intermediary M, S is only able to demonstratively refer to O (via a visual singling-out) if S has *some grasp* that what she is singling out is something distinct from O and bears some sort of linking relation to O. Snowdon puts this distinction in terms of "dependent" and "non-dependent" demonstrative reference:

> **SNOWDON'S DISTINCTION**: Indirect awareness of O allows only for *dependent* demonstrative reference to O – the success of which *depends* on the subject grasping something about O (its distinctness from M).

Direct awareness of O, in contrast, allows for non-dependent demonstrative reference – for the success of the demonstrative referring to O does not depend on anything other than the awareness itself (and the subject's minimal ability to selectively attend to elements within this awareness and 'tag' them with a demonstrative 'that…' thought).

Once we have allowed a distinction between direct and indirect visual awareness, it seems very hard to deny that Snowdon's distinction will apply to demonstrative thoughts whose reference is fixed *via visual awareness*. The very idea of indirect awareness must surely involve two distinct items: the subject has indirect awareness of one in virtue of having direct awareness of the other. And so it seems clear that: *if reference is being fixed via the visual singling-out act*, and if you are mistaken about which item it is you are singling out, then a demonstrative that refers via this singling out will not refer to the item you take yourself to be referring to. For your demonstrative to

successfully refer to one item (O), whilst singling out another distinct item (M), you must understand which item is which (and that the two are connected in some way) – otherwise you will simply be mistaken as to what it is you are singling out and so what your demonstrative refers to. Indirect awareness of O then is a sort of awareness that can facilitate demonstrative reference to O *only given some understanding* about what one is directly aware of and singling out. Direct awareness of O, in contrast, allows one to demonstrate O in thought without any such understanding – it requires only that one can mentally select a portion of one's visual field and 'tag' it in thought with a '*That*…'.

3. Some Complications

I want to turn now to considering some possible cases where it is unclear whether the singling-out act does fix the reference of a demonstrative thought. Such cases might be thought to be counter-examples to the Snowdon point – for if the singling-out act is not fixing reference, then mistakenly singling out the wrong item, M, need not prevent reference to another item, O. My simple strategy will basically be to put these complicated cases to one side. I only need to show that there are at least some cases, common enough, in which reference *is* fixed via visual experience and singling-out. In these cases Snowdon's point should apply and so the issue for CF theorists will arise. I do not need to adjudicate what, if anything, determines the reference for all instances of demonstrative thought.

In cases where the subject mistakes M for O, one might worry whether the reference of the demonstrative *is* actually fixed by the singling-out act. One might think that if I am mistaking M *for O*, then I'm bound to have an *intention* to refer *to O*. Now I will also presumably have an intention to refer to the item I have singled-out. But someone might try to argue that it is the former intention that takes precedence – i.e. the intention to refer to O is what matters to fixing the reference; the singling-out act (or the intention to refer via such an act) is just an idle accompaniment.

The first thing to say in response is that this would no longer be a case of *demonstrative* thought. If I have already formed an intention to refer to O when I think '*that* is F', and it is this intention that is determining the reference of my thought, then I will be thinking about O via some pre-existent concept, or 'file', I have for O. This

would not be a demonstrative thought. I can allow that there can be cases of this kind. I just need to show that there are also cases of genuinely demonstrative thought, in which the singling-out act, or perhaps the intention to refer via the singling-out act, *is* a determinant of reference.

And there surely will be cases in which the act of singling-out M fixes the reference of the thought despite an intention to refer to O. Recall the case in which I am mistaking a sculpture of Obama for Obama himself and I think 'That is the first black president of the Harvard Law Review.' Suppose now that the real Obama enters stage-left[7], so that I come to realise my mistake – I now grasp that the item I originally singled out in my visual field was not Obama. I might now react in one of two ways:

> I might say: 'Oops, so my previous thought was wrong: *that's* the President, not *this*' – i.e. the reference of my previous thought was fixed by the attentional singling-out act rather than any intention to refer to Obama.

Or I might try to maintain that: 'My previous thought was correct; I was thinking about *Obama* being the first black president of the Harvard Law Review, though I was mistaken about the item I happened to be attending to in my visual field.' – i.e. the reference of my previous thought was fixed by my intention to refer to an item via a pre-existent/non-demonstrative concept rather than the singling-out act.

We must allow for the possibility that a subject reacts in the former way. Such a mistake would not even be possible if an intention to refer to O *always* determined reference. Moreover, if, as surely sometimes happens, the subject has formed a genuinely *demonstrative*

[7] Digression: Dean 1946 suggests that an actor's entrance is less likely to be noticed from stage-right than from stage-left. There is a wealth of interesting evidence that there is 'over-attention' to the left side of space, sometimes termed 'pseudo-neglect', perhaps as a result of the right hemisphere's specialisation for spatial processing. E.g. there is a tendency to bisect lines to the left of centre (Bowers & Heilman 1980), and to bump into objects on our right (Turnbull & McGeorge 1998). Also, paintings are more likely to be given titles referring to left-side objects (Nelson & MacDonald 1971) and to depict the source of illumination on the left (Sun & Perona 1998).

thought based on experience, the former response would seem to be the only plausible one.

So there are cases, common enough, where reference is fixed by the singling-out act. These are cases where visual experience plays a role in fixing the reference of a thought rather than just being the prompt or cause for forming a thought. In such cases Snowdon's point should apply.

There are a number of other possible means or channels that might determine the reference of a demonstrative thought – and so there can be other difficult cases in which two (or more) means or channels are in conflict.

a) There are the other senses. (Notice, of course, that CF theorists are typically CF theorists about all of the senses, not just vision.)

And so there could be cases in which two different senses provide rival determinants of reference for a thought[8]. E.g. I might be visually aware of A and tactilely aware of B, but wrongly take what I am seeing and what I am touching to be the same item. If I then form a thought '*This* is F' it can be quite unclear which item my thought refers to.

b) There is other people's testimony – e.g. I might be listening to someone speaking about an animal and think '*That* is my rabbit'. And there is memory[9] – I might attend to a conscious memory and think: '*That* was my rabbit'.

Again: I could be listening to testimony about A whilst looking at what is in fact B, and wrongly take what I'm looking at to be the same thing as what is being talked about.

[8] Shoemaker 1968 considers this kind of problem case, though he is not concerned to adjudicate which item if any gets referred to in such cases. See also Siegel 2002.

[9] Though perhaps it is not so clear that in the memory case we would normally be employing a *demonstrative* element in thought, where this is something like a bare label or tag. When we recall an object previously encountered we would normally have some kind of richer concept of, or 'file' on, the object. Though we may not have an explicit name for the object, we would not be *thinking* of it using the mental equivalent of a mere re-usable label or pointer.

I have no particular theory to offer as to how reference might be determined in these conflicted cases – or whether there is any referential fact of the matter.

> (c) There are also 'complex' demonstratives involving a noun or noun-phrase; as in 'That rabbit is cute' or 'This shirt is dirty' etc. There is much disagreement on the role of such noun phrases in determining the reference of the demonstrative. There is a question as to whether complex demonstratives are singular terms whose meaning (in a context) is simply the referent, or whether they are really quantificational in something like the way definite descriptions are held to be quantificational. And even assuming that complex demonstratives are still singular terms, there is then the question of what the role of the noun phrase is – does its meaning restrict the possible reference of the complex demonstrative? Is the meaning of the noun phrase part of the content of the overall proposition? I don't want to wade in to these controversies here[10].

Complex demonstratives also raise the prospect of conflicting determinants of reference. E.g. I am looking at a brilliant portrait of the rabbit that I mistake for the real rabbit. I direct attention at the portrait and think '*That* rabbit is furry'. Here I presumably intend to refer to the item in my visual field I am directing attention at, but I also intend to refer to a *rabbit* – I wrongly believe these are intentions to refer to the same one item[11].

To repeat: I take the Snowdon-style point to unproblematically apply only to cases where the subject is clearly demonstrating *via her visual awareness,* rather than via some other channel or using a pre-existent concept. And I think that such cases will still be plentifully common in everyday situations. E.g. when a subject fixes attention on an element in her visual awareness and simply wonders: 'What's *that*?' Here there are no referential intentions to pick out something of a certain kind or something one has previously encountered or

[10] Larson & Segal 1995, Salmon 2002 and Richard 1993 all treat complex demonstratives as singular terms, though they differ over the role played by the noun phrase. Neale 2004 treats complex demonstratives as quantificational.

[11] This is, I take it, what is going on in Donnellan's 1966 well-known examples illustrating 'referential', as opposed to 'attributive', uses of definite descriptions.

referred to and there is no question of the other senses complicating matters.

A final point I should make explicit is that I have been concerned with demonstrative *thought* rather than public language utterances, and with *mental* singling-out acts rather than with the sorts of overt, external factors that might be determinants of reference for publicly uttered demonstratives – e.g. physical finger pointing. Thus, some of the sorts of contextual factors that might be invoked as relevant to the reference of publicly uttered demonstratives are not relevant to the case of demonstrative thoughts with which I am concerned. E.g. Wettstein 1984 argues, very roughly, that the referent of a public '*that*…' utterance, is what a competent listener would take the speaker to be referring to given the context. So it may be that my publicly uttered claim: '*that* is a rabbit', where I am completely fooled by a brilliantly life-like replica of my rabbit, could yet succeed in referring to the real rabbit rather then the replica I'm looking at. For perhaps a competent audience, given the context, would take my utterance to refer to the real rabbit. And perhaps the audience would arrive at this interpretation *even if* the audience knew of my mistaken belief that the replica I was looking at was the real rabbit. Nevertheless, were I to fix attention on the replica in my visual field and simply *think*: '*that* is a rabbit', where I am fooled by the replica, my thought must refer to the replica, not the rabbit. In the case of thought, we cannot invoke an imaginary audience to overrule or overlook my mistake. The interpretation an imaginary audience would give to my demonstrative thought (supposing this even makes sense) is surely irrelevant, for we are considering what *I* am thinking of, not what anyone else might interpret my judgement to be about.

4. Some Possible Responses

Let me briefly recap the problematic issue for CF theories:

If what visual consciousness provides is some common factor distinct from anything environmental then one needs to *grasp* this fact if demonstrative reference to the environment *via visual experience* is to succeed. But it is typically accepted that CF theories are revisionary with respect to pre-philosophical views. That is to say, most non-philosophers do not hold that the visual field is (always) comprised of non-environmental features. So most people would be making very

widespread referential errors in their demonstrative thoughts. Vast swathes of humanity would be failing to refer to the world via attending to elements in their visual fields. And this might well seem a rather unattractive commitment for a theory to have.

The precise range of demonstrative judgements that would then turn out false will depend on the range of properties that the CF theorist allows both the common factor and environmental items to possess. E.g. if both sense-data and roses are allowed to be red, then a thought such as '*That* is red' might turn out to be true despite the fact that the subject fails to realise she is mentally singling out a red sense-datum and not a red rose. But if you think that roses are red but sense-data can only be red*, some distinct inner property of experience caused by environmental redness, then the subject's thought is bound to be false. But whatever the exact range of potentially shared properties, it looks like the CF theorist will end up being committed to attributing quite widespread falsity of demonstrative thoughts as a result of the widespread referential mistakes. Again, an unattractive consequence for a theory to have.

Note: Evans 1982 argues[12] that failing to realise one's visual experience is an hallucination would mean that a demonstrative formed on the basis of this experience would not refer to anything at all – one would have formed a 'pseudo-thought'. Conflating a mental and a physical feature would then be different to conflating two physical items – rather than referring to the wrong item, one would have failed to refer entirely. But this is clearly no help to the CF theorist; it is just as unattractive a consequence that people's demonstratives frequently fail to refer as that they frequently mis-refer.

Faced with this alleged consequence of their theory, the CF theorists might, it seems to me, respond in one of four ways:

(i) They could just accept that it is indeed a consequence that any subjects who do not understand their experiential situation will generally be failing to refer to the environment. Compare: Hume thought that the common man mistakenly attributes a property of their visual experience, colour, to environmental objects that do not possess such a property. But he thought that (as with our mistaken views on morality and causation) such conflation was practically harmless, or even served some pur-

[12] See also Dickie 2010.

pose. So the idea would be that whilst a wide range of everyday demonstrative thoughts are *strictly speaking* either false or fail to refer entirely, they are nonetheless a useful or at least harmless means by which we can guide our behaviour and actions in the world. I think it is fair to say that, dialectically speaking, one would need very compelling philosophical arguments – e.g. from illusion or hallucination – before one was happy to swallow such a picture of widespread referential error or failure.

(ii) The CF theorists might want to deny that the natural pre-philosophical position is one in which people fail to grasp their indirect experiential position. They would claim instead that non-philosophers in general *would* acknowledge, if asked, that the elements making up their visual field are *always* distinct from any elements in the environment. Although I don't personally find this plausible, I think it is perhaps the CF theorist's best option. Firstly, I suppose it is ultimately an empirical question what people would or would not generally acknowledge about the nature of their visual experiences. Secondly, the CF theorist might point to the prevalence of Cartesian-sceptical-style scenarios in popular culture –e.g. 'The Matrix' etc – as evidence that the general population *would* be prepared to acknowledge that what their visual awareness delivers are elements distinct from anything in the physical environment. Thirdly, there is also perhaps anthropological and psychological evidence that humans are natural Cartesians, innately disposed to conceive of a mental realm distinct from the physical realm[13].

Having said all that, I do not think this second response is convincing. Whilst the notion of hallucination – elements in the visual field that are not environmental – is no doubt part of most people's natural understanding of how experience *could potentially be*, the idea that *all* experience (normal perceptual experience) similarly involves the visual presentation of entirely non-environmental features is not, I suggest, something that people would naturally acknowledge, the popularity of 'The Matrix' notwithstanding. If the CF position were

[13] See Boyer (2003) for anthropological evidence that belief in disembodied spirits is universal. See Kuhlmeier et al (2004) for evidence from developmental psychology that children are born dualists as they do not expect people to be subject to physical laws.

the natural view, one might wonder why so many philosophers have felt it worthwhile to provide arguments from hallucination and illusion, in the apparent belief that they are showing something novel and surprising. (Admittedly, it is hardly unknown for philosophers to provide elaborate arguments in support of uncontroversial theses…)

Notice that I am *not* claiming that non-philosophers do generally hold a determinate direct-relational view of the metaphysics of experience. Indeed I am not sure it even makes sense to speak of people having any set 'pre-philosophical' opinion on a philosophical question. (For this reason I dislike the label 'naïve realism'.) I am only claiming that the general population does not typically hold a determinately CF view of experience. This is, of course, ultimately an empirical claim for which I offer no evidence. Still, it strikes me as very plausible and it is something that CF theorists have also traditionally been happy to accept.

I think that cognitively unsophisticated creatures are relevant here. It seems plausible that young children and perhaps various animals are able to entertain simple demonstrative thoughts about the world on the basis of their visual experience – they can wonder: 'What's *that*?' But they lack any concept of 'inner' visual experience distinct from the (apparently) worldly objects of visual awareness – so they could not grasp the CF nature of experience. Attributing widespread referential failures to children is also quite a revisionary and, prima facie, unattractive consequence for a theory.

> (iii) The everyday examples I provided in section 2 to illustrate Snowdon's point involved the subject conflating two environmental features. But a CF theorist might think that failing to acknowledge a *mental* or *sensory* intermediary is somehow different to the parallel mistake involving two normal, environmental objects. What goes for the environmental cases, they might argue, need not hold when, say, sense-data or experiential representations are involved. Even if one is mistaking a common-factor for some distinct environmental feature, an act of singling out the sensory item from the sensory array might still, somehow, allow for successful reference to the environmental item.

One of the few theorists I have encountered who explicitly considers the issue of demonstrative reference via experience is Barry Maund

2003, who endorses this third line of thought. Maund states explicitly that:

> 'the naïve perceiver takes it that there is only one item there… The truth of the matter, however… is that there are two items… [and] the perceiver conflates the two.' (Maund 2003: 84)

Nonetheless, Maund wants to maintain that:

> 'perceptual experiences allow us [even the naïve] to have demonstrative thoughts about such [environmental] objects as cups.' (Maund 2003: 84)

Maund, in the spirit of Hume, insists that the 'conflation of [sensory] sign and thing signified… far from being damaging, or even harmless, is in fact beneficial' (Maund 2003: 85). Why might the naïve confusion of environmental object and mental/sensory feature be less ruinous to demonstrative thought than conflating two environmental items? Maund provides two reasons why conscious awareness of environmental objects via sensory intermediaries allows us to demonstratively refer to the environment *even if we fail to grasp the indirect nature of our situation*.

(1) The experience allows for successful behavioural targeting of the physical object.
(2) The experience is caused (in the right way) by the physical object.

This is not a strong argument. Neither of these factors provides any reason why we should treat the case of conflating a sensory intermediary with some environmental item any differently to a case in which two environmental items are conflated. Both of the factors Maund mentions – enabling successful targeting behaviour and causal connections of the right kind – might equally be present with a physical intermediary.

Consider 'successful behavioural targeting' first: E.g. A cleverly located hologram of a cup may allow me to target my behaviour successfully at the actual cup – by placing the hologram in front of the actual cup I may walk in the right direction, be able to inform others of the actual cup's location, reach towards it etc. None the less it should be clear that, despite such successful behaviour towards the actual cup, if I fail to grasp that what I am visually aware of is really a

hologram rather than the cup itself then when I single out the hologram in visual field, my thought: '*That* is made of bone china' will be false. Successful behaviour notwithstanding, I am *mistaking* a hologram for a cup; my demonstrative is, unbeknownst to me, actually latching onto something that is neither a cup nor made of china. If I were to think a thought like '*That* is on the table', where the hologram is in fact on the table, then my thought would, fortuitously, be true, but my demonstrative would still be referring, unbeknownst to me, to the hologram I have singled out in my visual field and not to any cup.

And we could certainly build into the hologram example that the hologram is causally sensitive to the state of the actual cup and co-varies with it to just the same degree as the proposed co-variance of sense-data and cup. The existence of causal links between a proximal intermediary and a distal object are simply irrelevant. If I have mistaken what it is that has been singled out – I have mistaken an entity distinct from the cup, for the cup itself – then demonstrative reference via that singling out will not refer to what I expect.

Since the two conditions Maund adduces make no difference when the subject is conflating two environmental items, we have no reason to think they should make any difference when the conflation involves a sensory item. Conflating a sense-datum, or qualia, or a mental representation with an environmental feature is still a conflation. Until some better argument than Maund's is provided, we should continue to conclude that CF theorists are committed to attributing a widespread failure to refer to the environment to those who do not accept their theory[14].

(iv) Finally, the CF theorist might seek to deny some part of the 'natural story'. I take it this would also be an unintuitive cost for the CF theorist, who would then need an alternative story as to how experience and attention are involved in fixing the reference of demonstrative thoughts. (Or perhaps a story as to why there is really no such thing as demonstrative thought.)

[14] To be clear: I have not provided an argument ruling out even the possibility that a CF theorist might provide some reason, different to Maund's, why the failure to grasp the presence of a mental intermediary is crucially different to cases where the subject fails to grasp that she is aware of a environmental intermediary. All I have shown is that we have no reason so far to treat these two kinds of cases differently.

5. Some Final Remarks

Before making some final remarks it is worth emphasising: I have not been trying to provide any sort of knock-down refutation of common-factor theories. I have merely sought to highlight what is apparently an unintuitive and revisionary consequence of such theories, one that seems not to have received much mention.

Some philosophers (e.g. Austin 1962: 15-19) have complained that the distinction between direct and indirect awareness is ill-defined or unhelpful. Snowdon's paper was meant to clarify this distinction by invoking demonstrative reference. I have illustrated Snowdon's point by appealing to everyday cases in which it seems quite clear that we have visual awareness of O only in virtue of really/directly seeing something else – seeing O's shadow, seeing O's photograph. However, there remains a range of problematic cases in which it can seem quite unclear where we want to draw the line between seeing directly or indirectly. E.g. seeing O through spectacles, seeing O through a magnifying glass, through binoculars, through a telescope, through a periscope, through left-right inverting goggles, through 'night-vision' goggles, via a digital movie camera in 'real time', via a digital movie camera with a 1-hour delay . . .

Snowdon's point does not, by itself, help to decide such problem cases one way or the other, for it does not tell us how to decide whether something is a 'transparent' or 'invisible' medium through which we gain awareness of O itself, or whether it constitutes a distinct visible intermediary[15]. For instance, if we accept that there is such a visible entity in the

[15] When M is *totally* or *perfectly transparent* then it is simply invisible/unsensible. It is just not possible for the subject to have visual awareness of M at all, and so nor could the subject discriminate M from its surroundings, nor fix attention on M as an element in the visual field and mentally "point" to it. When M is only *imperfectly transparent*, like a slightly dirty pane of glass, then M is potentially visible and so it would be possible to perform two different attentional acts – one could attend to O (through M) or to M (itself). But when M is an *intervening* entity/medium then there is not the possibility of two distinct acts of attention within visual experience – the only way to (indirectly) attend to O *just is* to (directly) attend to M.

visual field as a mirror image of O, something distinct from O, then seeing O in a mirror is seeing it indirectly. Whereas if we think there is no such genuinely distinct entity as an image, then seeing O in a mirror is just an unusual way of seeing (directly) O itself, like seeing O through clear glass or through thin air. (The mirror would then provide a way of 'transparently' looking at O itself, rather than showing a distinct image of O.) I have offered no general method for deciding whether something is just a transparent medium or constitutes an opaque intermediary.

Still, an obvious and compelling reason for thinking that one has visual awareness of an intermediary, rather than O itself, would be if one's visual field could *remain essentially unchanged* despite O changing or vanishing. Which is, of course, just the scenario envisaged in arguments from hallucination. On this basis, any theorist who accepts that there is conscious awareness of a common-factor in both a perception of O and an indistinguishable non-perceptual case would appear to be committed to the presence of an intermediary, distinct from O, in the visual field. In other words, any phenomenological common-factor theory will involve awareness of an intermediary distinct from O, and so Snowdon's point would apply.

Intentional theorists are typically common-factor theorists, indeed this is considered a main virtue of the theory. But they also typically conceive of themselves as direct theorists. McGinn is explicit on this point:

> 'My view is that we see objects 'directly' by representing them in visual experience.' (McGinn 1983: 129)

Barry Maund goes so far as to claim that direct-ness is a commitment of all intentional theories:

> '...theorists who admit the role of representations... are all agreed that the theory is a form of direct realism: representational states are involved in perception but neither they, nor components of them, are said to be 'objects' for the perceiver, nor objects that constitute a veil between perceiver and the world.' (Maund 2003: 7-8)

However, Maund surely goes too far in claiming that *all* intentionalist theorists are agreed on the direct status of such theories. For example, Tim Crane, himself a prominent intentionalist, is equally explicit in his repudiation of this direct status:

> '…critics of intentionalism are right when they say that on the intentionalist view, perception 'falls short' of the world, and in this sense creates what Putnam calls an 'interface' between the mind and the world. The essence of perception – perceptual experience itself – does fall short of the world. But, according to the intentionalist, this is not something which should create any epistemological or metaphysical anxiety; it is simply a consequence of a general aspect of intentionality as traditionally conceived.' (Crane 2006: 141)

I happen to think that Crane is quite right to concede that intentionalism is an indirect view, but nothing essential to my argument turns on the meaning of the terms 'direct' or 'indirect' as applied to theories of conscious experience. I hope this paper will have raised a *referential* anxiety for any intentionalist who accepts that experiences are essentially such that they can be common factors. A common factor must, ex hypothesi, be something distinct from anything environmental; and so the possibility opens up of *mistaking* this non-environmental feature for something environmental. This would presumably be a mistake committed frequently by any of us who have not yet accepted that the elements comprising the visual field are non-environmental elements, elements *common to perceptions and hallucinations*. If the argument of this paper has been correct, the referential difficulties that would result from such a mistake are not only an issue for avowedly indirect sense-data theories, but for any common-factor theory[16].

[16] This paper was presented at the 13th T.I.F. conference in Barcelona, 2011. I am very grateful to Giovanni Merlo for his insightful reply on that occasion. Ancestral versions of this paper were presented to audiences at the 2008 London Graduate Conference, Cornell University and King's College London. I would also like to thank: Karen Bennett, Alex Davies, Gabriel Lakeman, Michael O'Sullivan, Stephen Raleigh, Gabriel Segal, Paul Snowdon, Charles Travis and Sasha Vereker for their comments.

Thomas Raleigh
Becario del Programa de Becas Posdoctorales
Instituto de Investigaciones Filosóficas, U.N.A.M.
Ciudad Universitaria, 04510, Coyoacán
México, D.F.
traleigh@gmail.com

References

Austin, John L. 1962. *Sense and Sensibilia*. Oxford: Oxford University Press.

Bermúdez, Jose L. 2000. Naturalized Sense-Data. *Philosophy and Phenomenal Research* 61: 353-374

Bowers, Dane & Heilman, Kenneth M. 1980. Pseudoneglect: Effects of hemispace on a tactile line bisection task. *Neuropsychologia* 18: 491-498.

Boyer, Pascal. 2003. Religious thought and behaviour as by-products of brain function. *Trends in Cognitive Science* 7(3): 119 - 124

Campbell, John. 2002. *Reference and Consciousness*. Oxford: Clarendon Press.

Campbell, John. 2006. Sortals and the Binding Problem. In *Identity and Modality*, ed. by MacBride, Fraser. Oxford: Oxford University Press.

Crane, Tim. 2006. Is There a Perceptual Relation? In *Perceptual Experience*, ed. by Gendler, Tamar. & Hawthorne, John. Oxford: Oxford University Press.

Dean, Alexander. 1946. *Fundamentals of play directing*. New York.

Dickie, Imogen. 2010. We are Acquainted with Ordinary Things. In *New Essays on Singular Thought*, ed. by Jeshion, Robin. Oxford: Oxford University Press.

Donnellan, Keith S. 1966. Reference and Definite Descriptions, *Philosophical Review* 77: 281-304.

Evans, Gareth. 1982. *The Varieties of Reference*. Oxford: Clarendon Press.

Jeshion, Robin. 2010. Introduction to New Essays on Singular Thought. In *New Essays on Singular Thought*, ed. by Jeshion, Robin. Oxford: Oxford University Press.

Kuhlmeier, Valerie, Wynn, Karen & Bloom, Paul. 2004. Do 5-Month-Old Infants See Humans as Material Objects? *Cognition* 94: 95-103.

Larson, Richard & Segal, Gabriel. 1995. *Knowledge of Meaning*. Cambridge, MA: MIT Press.

Maund, Barry. 2003. *Perception*. Acumen Publishing

Neale, Stephen. 2004. This, That, and the Other. In *Descriptions and Beyond*, ed. by Reimer, Marga & Bezuidenhout, Anne. New York: Oxford University Press.

Nelson, Thomas & MacDonald, Gregory. 1971. Lateral organization, perceived depth and title preference in pictures. *Perceptual and Motor Skills* 33: 983-986.

Pylyshyn, Zenon. 2003. *Seeing and Visualising*. Cambridge MA: MIT Press.

Pylyshyn, Zenon. 2007. *Things and Places: How the Mind Connects with the World.* Cambridge MA: MIT Press.

Richard, Mark. 1993. Articulated Terms. *Philosophical Perspectives* 7: 207-230.

Salmon, Nathan. 2002. Demonstrating and Necessity. *Philosophical Review* 111: 497-537.

Shoemaker, Sidney. 1968. Self-Reference and Self-Awareness. *Journal of Philosophy* 65: 555-67.

Siegel, Susanna. 2002. .The Role of Perception in Demonstrative Reference. *Philosophers' Imprint* 2 (1) (www.philosophersimprint.org/002001).

Smithies, Declan. Forthcoming. What is the Role of Consciousness in Demonstrative Thought? *Journal of Philosophy*, forthcoming.

Snowdon, Paul. 1992. How to Interpret 'Direct Perception'. In *The Contents of Experience*, ed. by Crane, Tim. Cambridge: Cambridge University Press.

Sun, Jennifer & Perona, Pietro. 1998. Where is the sun? *Nature Neuroscience* 1: 183-184.

Turnbull, Oliver & McGeorge, Peter. 1998. Lateral bumping: A normal-subject analog to the behaviour of patients with hemispatial neglect? *Brain and Cognition* 37: 31-33.

Tye, Michael. 2009. *Consciousness Revisited.* Cambridge MA: MIT press

Wettstein, Howard. 1984. How to Bridge the Gap Between Meaning and Reference. *Synthese* 58: 63-84.

Book reviews

The Disordered Mind: An Introduction to Philosophy of Mind and Mental Illness, by George Graham. London and New York: Routledge, 2010, pp. xiv + 288. P/b £18.99.

Perhaps more than any other scientific discipline, psychopathology relies on pre-theoretical intuitions that are unavoidably philosophical. Beneath psychiatric theory and practice lie issues that fall squarely within the philosophy of mind —issues such as the mind-body problem, mental causation, personal identity, subjectivity, consciousness and the emotions— not least, of course, the very concept of rationality. Professor George Graham's book, *The Disordered Mind*, is primarily concerned to show that the understanding and explanation of the mental disorders typically found in psychopathology manuals demands precisely that we apply the concepts and tools distinctive of that branch of philosophy. Unlike those who think about the philosophy of psychiatry as a research field within the philosophy of science, Graham urges us to acknowledge the dependence of psychopathology on the philosophy of mind. At the same time, although to a lesser extent, he draws on the phenomenon of mental illness to introduce and discuss central concepts and positions in that discipline. The book should thus be welcomed as a novel, gripping and doubly unorthodox textbook. It offers a critically acute introduction to the philosophy of psychiatry, while discussing some key themes in the philosophy of mind.

The Disordered Mind fosters realism about mental disorders. This is perhaps the one single feature that best characterizes the book. Mental disorders are real *qua* mental —Graham contends. They should not be treated as neurological illnesses, even though they are brought about by a combination of mental and brute somatic causes. Graham's realism is even in contrast to the view that takes cognitive neuroscience to be the natural niche for categorizing and treating mental illness —a view he, interestingly, labels 'anti-realist' thus assuming, at the risk of alienating its advocates, that it carries with it a reductionist commitment. Neurochemistry notwithstanding, Graham

strongly advises us not to think of the mentally ill as having chemically unbalanced brains, but muddled minds.

To explain the 'mental' in 'mental disorder', and in consonance with his non-reductive approach to mental illnesses, Professor Graham draws upon two cornerstones of the philosophy of mind: intentionality and consciousness. The irreducible mentality of mental disorders is best captured, he claims, by attending to both what the representational states of the mentally ill are about, and the phenomenal character of their conscious experiences. How mentally ill patients represent themselves and the world, and their own experiences of such representations also become essential taxonomic parameters for Graham. Intentionality and consciousness thus turn into the coordinates for Graham's preferred theory of concepts: a theory based on prototypes. The choice of parameters works particularly well when characterizing disorders in which the subject is capable of reflecting upon his own condition. Major depressive disorder is paradigmatic in this respect. The depressed subject typically represents the world as providing no motivation while, at the same time, experiencing himself as helpless or aggrieved. Representational and experiential features thus help us, on Graham's view, both to identify and to understand the onset conditions of a disorder. Yet, while there is much to recommend about this analysis when applied to mental illness of a certain type, intentional parameters are of little use in other, paradigmatic, disorders such as delusion. It would certainly seem meaningless, as Graham himself rightly notes, to talk about e.g. the deluded subject's own experience of his condition contributing to the individuation of the disorder in any way.

Intentional-cum-phenomenological considerations are also paramount to Graham's analysis of 'disorder' in 'mental disorder'. A disorder is presented as an "a-rational gumming up of the rational works" (p. 160): some basic mental capacities are gummed up in the mentally ill, he claims. Not surprisingly, the capacities listed by Graham illustrate the non-reductive and holistic nature of his approach. Among them, we find bodily/spatial self-location, self/world comprehension, care, commitment and emotional engagement (pp. 147-149), to mention just a few. These are all capacities, Graham claims, necessary for conducting a decent life. A disordered mind is thus characterized as one in which the mental capacities involved in living a decent life are gummed up in such a way as to be harmful for the agent —harmful to the point of requiring treatment or assistance.

Graham's philosophy of psychiatry thus comes with a moral psychology programme in tow. His appeal to the idea of a decent life as a regulatory criterion in determining the relevant faculties for the characterization of mental illnesses definitely marks his project as one that is genuinely humanistic and person-centred. As such, *The Disordered Mind* could not be further from the hard-line geneticists and molecular biologists flooding psychiatry journals with their attempts to account for mental illnesses by carving nature ever finer at the joints. These micromanagers of psychiatric nosology will, without a doubt, become suspicious of Graham's strategy, and so will anyone inclined to analyse the intentional in more moderate naturalistic terms. The book, in this sense, seems to be best suited for those who, instead of looking for a causal account of mental illnesses, look rather for a proper understanding of the mentally ill's experiences and their import —the kind of understanding that can, although perhaps not happily, sidestep certain issues that touch on the causal efficacy of the subpersonal.

From a strictly philosophical point of view, and also related to Graham's engagement with the purely intentional, there is something not quite clear about his basic characterization of mental disorder. Graham presents a classic picture of rationality as the smooth working of inferential processes between mental contents, and asks us to view mental illnesses as the breakdown of such processes by a-rational muddles. It would thus seem natural to understand 'a-rational' here as referring to some sort of mechanical, brute, causes. Yet, as already pointed out, Graham warns us against taking the proximate causes of mental illnesses to be purely mechanical. Then again, his reticence to allow for brute causes to become efficient on their own makes it very difficult to draw certain boundaries where, pre-theoretically, we find them. Some archetypal mental illnesses —e.g., schizophrenia— seem to be the outcome of purely mechanical breakdowns —their symptoms just feeding into intentional patterns that reinforce and deepen the disorder. This reticence also makes Graham's analysis of some disorders —such as addiction— slant dangerously towards the self-righteous; for the addict is presented as someone who, in breaking his own promises to restrain from relapsing, is best characterized as someone who lacks responsibility for himself.

Graham does defend his position from these charges, which he takes to be two forms of scepticism about his proposal: moral and metaphysical. On the one hand, the moral sceptic, a Szasz advocate of

sorts, takes the category of mental disorder to be morally ill-conceived inasmuch as it carries with it evaluative judgments about the mentally ill which normal physical diagnoses lack. The Szaszian argument is based on the idea that science is not normative, while psychiatry makes normative judgments; so psychiatry is not scientific and should be abandoned. Graham's reply consists of arguing that general medicine is as value-ridden as psychiatry; that there is not much difference between the normative assumptions guiding the diagnoses of physical and of mental illnesses. Metaphysical sceptics, on the other hand, are those who take Graham's realist approach to mental disorders to be clearly dualistic. *The Disordered Mind* is indeed a good illustration of how to deny that for every instantiated mental property F, there is some physical property G such that F = G. Graham offers instead a particular kind of non-reductive physicalism with clear Davidsonian overtones: "the same condition of a person may be … both a physical condition and a mental disorder" (p. 80). It is, of course, highly unlikely that either the Szaszians or the metaphysical sceptics would feel defeated by Graham's considerations against their views. The Szaszians are likely to argue that there are clearly two notions of normativity at work in the physical and the mental branches of medicine; the metaphysical sceptics will remind us of the well-known weaknesses of token physicalism as a form of physicalism. The discussion of these issues, however, provides the clearest example of the way in which the book also plays the role of an introduction to the philosophy of mind and it does so in a very clear and engaging fashion.

The Disordered Mind is divided into nine chapters and an epilogue. The first six chapters are dedicated to an explanation and defence of the general approach favoured by the author —the main object of this review. The last three chapters examine a few central cases, such as addiction, delusions like paranoia or thought insertion, and multiple personality disorder, among others. Finally, in the epilogue, Graham takes us through some of Kierkegaard's most moving passages, used as a platform for discussing the metaphysics of the self vis-à-vis his view of psychopathology. Here, Graham aspires to legitimize the permeability of boundaries in what he takes to be prototypes of mental disorders, thus eluding some of the anticipated objections regarding the inherent vagueness of his approach; and he does a very good job of showing the prevailing fuzziness of psychopathological taxonomies.

The land of the mentally unsound is poignant territory, which attracts all kinds of scientific and philosophical projects. To understand it involves, in part, identifying the underlying causal patterns that allow for correct classification, assessment, and treatment. From Professor Graham's book we learn that to understand the land of the mentally unsound also involves being able to draw a moral psychological model of human flourishing —one that preserves dignity and self-respect. *The Disordered Mind* will definitely be of interest to philosophy undergraduates and to anyone interested in a philosophical account of the fine balance between sanity and insanity. It is written in an engaging and accessible way for students, yet its contributions will also appeal to psychiatrists, psychologists and mental health practitioners.

Josefa Toribio
Departament de Filosofia
Universitat Autònoma de Barcelona
Facultat de Filosofia i Lletres, Edifici B
08193 Bellaterra, Barcelona
Spain
jtoribio@icrea.cat

LOT 2: The Language of Thought Revisited, by Jerry Fodor, Oxford, Oxford University Press, 2008, 228pp.

In the course of some characteristically wry autobiographical comments at the beginning of this book, Jerry Fodor remarks that when he published *The Language of Thought* in 1975, he thought of himself as reporting an emerging consensus in the study of cognition. His views have inspired much discussion but little outright agreement, and this proclaimed sequel is polemical in nature. Fodor sees himself as in an embattled minority, and here he returns the fire of his critics.

The book is a short one, but covers a great deal of ground. Beginning with some remarks on the history of the development of cognitive science and analytical philosophy, Fodor addresses propositional attitude ascriptions, concept possession, the nature of perceptual representation and the sense/reference distinction, among other things. The pace is brisk, and Fodor's famous wit is again on display. His humorous approach to philosophical writing

often alienates his supporters and amuses his critics, but those who find the tone objectionable risk missing some good jokes.

Fodor is here, as ever, concerned to defend a version of the computational theory of mind, which in his hands is a view both about mental states and about mental processes. He sees mental states as constituted by relations between agents and mental representations. These representations are expressed in a language of thought, sometimes referred to as 'Mentalese', which is semantically prior to natural languages such as English. In particular, it is both syntactically and lexically unambiguous. Ambiguities in natural language are to be explained by positing, for example, that the English word 'bank' corresponds to at least two Mentalese expressions. Further, mental processes are seen as computations on mental representations. For an agent to engage in inference, for example, is for one representation to give rise to another, where the causal powers of the antecedent representation are given by its Mentalese syntax.

Early on in this book, Fodor identifies his enemy as 'pragmatism', 'perhaps the worst idea that philosophy ever had' (p. 9). His characterisation of this enemy is somewhat vague, but what Fodor opposes is the explanation of mental content in terms of abilities or dispositions to act, rather than vice versa. 'Dewey, Wittgenstein, Quine, Ryle, Sellars, Putnam, Rorty, Dummett, Brandom, McDowell' are listed (p. 11) as exponents of pragmatism. Later (p. 194), Kant is added to the list.

This is an astonishingly broad definition of pragmatism, and it yields an astonishingly disparate list of pragmatists. But perhaps we can accept it as a stipulative definition devised for Fodor's particular purposes. It is meant to refer quite generally to views according to which we must explain or characterise in more primitive terms the possession of concepts and beliefs. This explanatory demand is seen as threatening Fodor's extreme realism about the mental, and by extension the necessity of appealing to a mental language underlying our capacities for thought and natural language use.

Unfortunately, the vagueness of Fodor's characterisation of his enemy undermines the effectiveness of his counter-attack. He accuses 'pragmatism' of vicious circularity. The only sort of action, the argument goes, which could possibly be adequate to ground an account of mental content is intentionally directed

action. But ascription of such activity to an agent presupposes a capacity to think on the part of that agent, since the agent must be capable of conceptualising the objective of his or her action.

Fodor's argument here is much too brief, and he makes no attempt to deal with the specific positions of the authors he lists as pragmatists. Quine, for example, is working with a conception of behaviour which is much less rich than the one that Fodor here assumes. In Wittgenstein's case, it is not even clear that he is attempting any sort of reductive explanation of thought, in terms of action or anything else. These authors' projects may well come aground, but Fodor's arguments are not sufficient to show that and where they do.

For the rest of the book, Fodor is concerned not so much with the refutation of 'pragmatism' as with the constructive enterprise of showing how his mentalist computationalism impacts on various problems in the study of cognition, what objections it itself may be open to, and how those objections might be overcome.

The first stop is the sense-reference distinction. Here, Fodor is concerned to argue that Frege's puzzle does not arise for Mentalese expressions. If that is so, he thinks, the semantics of Mentalese can be purely referential. In a way, Fodor's account corresponds to traditional attempts to distinguish the logical from the surface form of natural language expressions, the difference being that Fodor takes logical form to be straightforwardly the correct syntax for an unlearned, unarticulated language, quite distinct from any natural language.

While interesting in its own right, Fodor's discussion is orthogonal to much of the literature about the distinction. That is because his concern is not to argue that the meaning of, say, a proper name in common use in English is given entirely by its reference. Rather, he is concerned to show that the psychology of cognitive processes can carry out causal explanations without appeal to any such entities as 'senses'. It is to expressions of Mentalese that such a psychology will appeal, so that the semantics of English becomes irrelevant.

He goes so far as to speculate (p. 219) that perhaps *only* Mentalese has a semantics, that natural languages do not have any semantics at all. If true, this would indeed be a solution (a very radical one) to Frege's puzzle as conventionally understood, but it is not Fodor's main purpose to argue the point.

Next comes an interesting discussion of what Fodor calls the problem of 'locality'. Here he is pointing to a limitation of the computational theory of mind as currently understood. The problem is that the computational theory does not seem apt to account for certain sorts of cognitive process, specifically those involving inductive inference. A viable account of how creatures like us succeed in reasoning inductively would appear to demand that a nontrivial criterion for the relevance of empirical data to a particular inference be specified, and it is just this that the computational theory has trouble providing. The chapter indicates the sort of work which Fodor thinks philosophers of mind ought to be engaged in.

The remaining three chapters of the book, grouped together under the general heading of 'Minds', might be seen as contributions to the programme, which motivates the computational theory, of naturalising the mind. Generally speaking, Fodor downplays concerns about the importance of normative notions to a proper understanding of mental processes, particularly the processes of concept acquisition and perceptual belief formation.

Fodor's approach is to prioritise causal-explanatory psychological explanation over normative epistemology, a procedure described here (p. 170) as letting 'the epistemological chips falls where they may'. For example, Sellars, Davidson and McDowell have argued that perceptual content cannot be nonconceptual, because if it were then perceptual beliefs could not count as justified. Fodor dismisses the argument. It is, he thinks, an empirical question whether perceptual representations are conceptual. If it turns out that they are not, and perceptual beliefs thereby prove unjustified, then so much the worse for perceptual beliefs.

Fodor draws on some evidence from the cognitive-scientific literature on perceptual illusions to argue that perceptual content is indeed nonconceptual. He does not engage with the arguments of John McDowell (in, for example, 'The content of perceptual experience', *Philosophical Quarterly*, XLIV, 1992, pp. 190-205) to the conclusion that such empirical evidence is consistent with different philosophical views as to the nature of perception. Perhaps more valuably, he provides a clear explanation of how, in his view, perceptual representation works. It is 'iconic', rather than 'discursive', as for example linguistic representation is. In particular, iconic representations, unlike sentences, have no canonical

decomposition. Some parts of a sentence, for example the name 'John' in 'John loves Mary', themselves count as representations, but others (such as the phrase 'John loves') do not. By contrast, each part of a picture represents some part of what the whole picture represents. Perceptual representations, like pictures, lack subject-predicate structure; correlatively, they have no conceptual content. Perceptual beliefs are produced from perceptual representations by way of sub-personal computations on such iconic representations. Whether or how such computations can be seen as preserving epistemic warrant is seen as a secondary consideration.

The discussion of concept acquisition is more frankly speculative. Such accounts are typically threatened by vicious circularity. It is unclear how one can acquire the concept *dog* from experience of dogs without already being able to apply the concept. Fodor again advocates an approach emphasising processes at the subpersonal level, with questions of normativity sidelined. The difference is that the posited process is not only subpersonal but also subcomputational. To simplify greatly, agents are innately disposed, as a matter of their neurology, to form certain concepts provided they have had certain experiences. The concept *dog* could not be inferred merely from experiences of dogs, no matter how many experiences occur and no matter how typical the dogs. It is rather a moot point whether such accounts make all concepts innate: either way, what is innate is the human tendency to produce certain representations given exposure to certain extensionally-defined features of the world.

Fodor makes his proposal in some detail, but without much appeal to evidence. The important point for him is that proposals of that sort, if true, would be adequate to the problem of concept acquisition. The discussion thus serves to illuminate how he sees the problem. The issue for him is how creatures like us acquire our capacity to represent the world, and this is seen as a causal-explanatory question in answering which we are free to appeal to any feature of human nature or experience.

A word, in closing, on Fodor's methodology. His approach to the philosophy of mind is very much constructive rather than critical, and this determines a great deal of the course of his book. The consideration that a particular line of thought presents the best or clearest available explanation of some cognitive phenome-

non tends to trump fundamental objections. Fodor's remarks on the epistemology of perceptual belief, mentioned above, may serve as an example. The approach tends to result in unsatisfyingly glib and superficial responses to other philosophers.

This book works better as a clear exposition of Fodor's current views than as polemic. It contains little detailed engagement with alternative views, and few attempts to provide compelling arguments against them. It clarifies the author's position, but will not convince sceptics.

Michael O'Sullivan
Dept. of Philosophy
King's College London
Strand, London WC2R 2LS
michael.j.o'sullivan@kcl.ac.uk

Narratives and Narrators: A Philosophy of Stories, by Gregory Currie. Oxford: Oxford University Press, 2010, 237 xx + 237 pp.

Gregory Currie's new book, *Narratives & Narrators: A Philosophy of Stories*, discusses a concept which has not received sufficient attention from the community of analytic philosophers, namely, the concept of 'narrative'. How is it possible to characterise such a concept, avoiding the use of unhelpful technicalities or, worse, the dominant ideologies underlying much current literary analysis? Which instruments can the philosopher introduce or exploit in order to clarify the intricate network of concepts around this notion, many of which mesh with the study of fiction? More ambitiously, what is the significance human beings give to their being imaginatively engaged with narratives? What is, finally, the role and function of narrators in the societies to which they belong and about which they narrate? If in this bunch of questions I have tried to summarise some of the most pressing issues dealt with in the book, what remains to be seen is how Currie intends to answer them and whether he succeeds in this demanding task.

The book is articulated into four main directions of investigation: (i) an account of the intentional and representational properties determining what a narrative is; (ii) a pragmatic framework where the nature and presence of an implied author in narratives are discussed and where the differences between narrative texts and forms con-

veyed by different media, be they static pictures as in photography or dynamic ones as in cinema, may be spelt out; (iii) a set of hypotheses concerning those evolutionary aspects that have supposedly determined the emergence of narrative in modern society and the ways human beings have refined their experience of themselves and of their community in relation to the diffusion of narrative forms; (iv) finally, and most important, an account of what Currie considers the most essential feature of a narrative, i.e., its *expressive* power. In what follows, I will mostly focus on (i) and (ii), where I think the view Currie pushes forward is not without problems. I will say something on (iv), the part in which I think Currie obtains the most interesting results. I will leave (iii) aside, given the highly hypothetical nature of those remarks (and, in fact, Currie has opted for leaving them in appendices to some chapters).

As anticipated, the first line of investigation undertaken by Currie is a defence of the claim that narratives are *intentional-communicative artefacts*, that is to say, artefacts whose function is not only that of encoding certain story-like representations, but also to communicate the intentions of their makers as substantiated in a narrative shape. Currie extends this view over the first three chapters, with the important corollary of the last two, in which he defends the psychological notion of 'character' *in* narrative (not to be confused with the more basic notion of characters *of* a narrative).

Chapter I develops the idea of narratives as representational bodies (*corpora* in Lewis's terms) showing some rich internal organization. The stress is, however, put on the activity itself of making a narrative rather than on the content of what is encoded by a given artefact. Against the opinion of some philosophers – most notably, Walton (*Mimesis as Make-believe*, Harvard (1990): Harvard University Press) – the fact that two different artefacts encode the same piece of information does not suffice for both to be successful narratives. In fact, in order for a narrative to be successful, it must enable an audience to know 'the artefactual function of that narrative' (p. 6). To know whether a certain narrative instantiates its artefactual function properly, an audience has to infer, using pragmatic inference, the story content the author intends to communicate. Assessment of truth is therefore at most redundant on Currie's account of narrative, and for that matter of fiction as well (see Currie, *The Nature of Fiction*, New York (1990): Cambridge University Press). What really counts is

whether a statement or a set of statements is part of a narrative, and this can only be assessed through pragmatic inference (see below).

However, it is still legitimate to ask whether an appeal to intentions makes a complete job here. We certainly have narratives (*Odyssey*, *El Canter del Mio Cid*, and others) where it is no longer possible to trace the original author's intentions – if not the author herself –, since these stories were handed down by use and as pieces of folkloristic knowledge (at least in certain phases of their historical transmission, when orality was predominant). Perhaps, it was not in the intention of the creator of a certain body of information to make it function as a narrative, but this would not seem to prevent a future audience from reading it as such. Further, appealing to the author's intentions seems only to overburden the reader's *immediate* experience of a narrative, and the fact that Currie particularly appeals to these to discuss examples where we need to interpret some author's obvious mistakes (see p. 10), seems to confirm my suspicion that his general claim is defended on a thin ground constituted by some innocuous exceptions.

A good way to consider the role of intentions in the understanding of narrative is, perhaps, to concentrate upon *when* a reader's attention to the author's intention is asked for. A reasonable answer seems to me that the reader's attention is required when the relevant question is how *good* a particular narrative is, or, in other words, what strategies the author is responsible for to render that narrative *effectively* successful in communicating her a story. A merit of this book is to persuasively show how a narrative may be successful not only in communicating a story, but also in communicating it *expressively*. We will discuss some application of this point later. On the other hand, Currie insists that intentions have a more central role, that of driving an audience to the explicit content of a story. According to Currie's definition:

> 'P is explicit content when we can find some statement or set of statements in the text, S, which meets two conditions: (I) S is naturally interpretable in such a way as to convey directly, rather than merely to implicate, the thought that P, and (2) an overall best interpretation of the text is one which treats S as reliable' (p.13).

For Currie, whatever does not fulfil these two conditions, may still be considered as part of the story content if it belongs to the class of

propositions entailed by the story. He further hypothesises that the class of propositions entailed by the story corresponds to the class of conversational implicatures. This would seem to offer a solution to the problem of inconsistent stories, stories according to which two propositions, say P and Q, hold, but where Q entails a further statement, say T, implicating not P. Presumably, T should then be treated as a cancellable implicature, avoiding in this way the inconsistency and offering, at the same time, a 'closure condition' for the story content. I am not sure – nor is Currie – how this hypothesis would effectively work though.

Finally, Currie thinks that all there is for his intentionalist account to be successful is the fact that pragmatic inference is omnipresent in narrative (pp. 25-26); for instance, even in the understanding of what critics usually call the 'implied author' of a narrative. The implied author is that imagined or constructed figure, not necessarily the narrator herself, in whose regard a work is (to be) interpreted. Now, Currie sets out the hypothesis of an identity between the implied author's intended meaning with what Levinson, in the context of a pragmatic theory of communication, has called 'achieved meaning', i.e., the meaning which an attentive hearer is able, and also expected, to infer from an utterance. However, even granting that the achieved meaning in a normal communicative setting is nothing but the 'reasonably expected communicative uptake', it seems problematic to apply this view to narrative discourse. First of all, the nature of meaning here at stake is not clear, nor is the nature of the achievement itself: is it what a reasonable reader has been able to infer? That seems to be too weak. Is it what the narrator or the novelist intend the reader to uptake? Or a combination of both? (for an attentive account of how all these possibilities are given see: Richard J. Gerrig, *Experiencing Narrative Worlds*, New Haven and London (1993): Yale University Press).

The structure of narrative is further investigated in Chapter II, where Currie's intent is to offer an account of what is a distinctive feature of most narratives, namely, their particular unity, which is often revealed by the presence of 'sustained temporal-causal relations between particulars, especially characters (p. 28)'. After having scrutinised different notions of causation which all, at the end, seem inconclusive to pin down the multiformity of narrative connections, Currie goes on, opposing Velleman's view according to which the very idea of causal connectedness is not essential to narrative dis-

course. For Velleman, we should get rid of this idea and instead make room for the idea of an 'emotional cadence' governing the structure of a story. To this, Currie replies that if Vellerman's suggestion were on the right track, then it would be possible to count as narrative a mathematical proof or the like, since even for the structure of a deductive system one could experience an 'emotional resolution' of the kind already in effect with narrative. Of course, – Currie adds – we want to say something more exclusive about the representational properties of narrative. On the other hand, we should not be too anxious to ascribe precise borders to the notion of connectedness in narrative, since we are better off with the multifarious range of representations of dependencies that narratives present us with (p. 32).

This last point invites two important considerations as to the evaluation of Currie's overall project. Both are somehow negative theses in relation to the original task of defining what a narrative is. The first point concerns the importance given by Currie to the concept itself of narrative. Notwithstanding the book's title, the notion of narrative is less interesting than one may have initially been led to think, or so does Currie claim. In fact, inspired by a point of Lamarque's, Currie claims that a general definition of narrative as the telling of events, possibly causally related, does not suffice to discriminate works of a rich narrative structure from simple utterances like 'He went to the shop to buy a packet of cigarettes'. Therefore, it would be better to consider the category of *narrativity* as the privileged object of study, given its being a more flexible tool that allows for the differentiation of threshold levels, with the top level occupied by what Currie calls 'exemplary narratives' (p. 35). These are narratives that have as their most significant mark a 'thematic unity', which, according to Currie, may be particular or general. When it is particular, the events are narrated under a specific focus constituted by some common thread or activity. Sometimes, in the absence of a particular thematic unity, there may be a more general one – e.g moral, theoretical, religious, etc. –, which serves to close off the reader's experience of that narrative. Other times, we simply 'import' from our world bits of knowledge which may be important to form our responses to the narrative's having made salient certain possibilities in the story. However, a problem with this otherwise reasonable account is the widespread diffusion of noncanonical forms of narrativity in the post-modern panorama. It is sufficient to think about the

nonlinear narrative of Joyce's *Finnegans Wake*, or the constant plots' interruption which the stories narrated by Calvino in *If On a Winter's Night a Traveller* undergo, or the proliferation of unrelated stories and surplus of information in Wallace's *Infinite Jest*. What about, then, stories where the unity of the novel is invisible to the reader, and therefore is neither particular nor general, in Currie's analysis? I have in mind, in particular, Perec's novel *La Vie Mode de Emploi*, whose narrative structure is determined by an algorithm based on the so-called 'Knight's tour' problem. Besides, these narratives seem to pose a problem to Currie's analysis of narrative in terms of intentional-communicative artefact. The reader will remember that, according to Currie, a narrative is successful when it puts an audience in the condition of knowing its artefactual function. But these cases seem to go in the opposite direction: to block the reader from understanding the exact artefactual function their stories encode, or, said otherwise, they prevent her from having access, or full access, to the story content their authors intend to communicate. Should we say that these novels are unsuccessful? Not at all. Hence these novels pose interesting epistemological problems, and I am not sure Currie's analysis could deal with them. I leave it to the reader to work out a Currian reply, or to find out some alternative analysis.

The second point regards the authors' adoption of frameworks which somehow colour the narration of events and characters. What is a framework and how can we trace it in a story? There is a certain amount of frustration in Currie's wavering between two conceptions of framework, which do not seem to have much in common at first glance:

> (a) a framework as a preferred set of cognitive, evaluative, and emotional responses to the story;
> (b) a framework as a set of ideas an author may convey in her story, asking the reader to adopt it so as to make sense of the story-content.

Let's concentrate on point (b) first. In Chapter VI, a chapter devoted to the discussion of various forms of resistance to the narrative experience, Currie presents two cases that ask for some discussion here. According to him, Kurosawa's *Rashomon* and Proust's *À la recherche du temps perdu* are cases in which the authors would encourage their audiences to either naive or extravagant metaphysical ideas. These works should be blamed because they suggest 'their metaphysical

themes, without going to the trouble of showing how the metaphysics is integrated into the story–something, I suggest, that would be just impossible' (pp. 119-20). In other words, these narratives are the expression of a 'metaphysical anxiety', which is the result of their letting the story-content be obscured by ideas external to it. Currie then proposes a criterion to distinguish between story-content and framework in terms of pragmatic features associated to either of them: while story-content is relatively stable and mandatory, frameworks in the sense indicated by (b) are instead optional and detachable (p. 120), since a reader may understand and enjoy the story without being committed to any further authorial idea about how to interpret it. However, Currie's criterion of detachability does not apply easily to certain cases. Consider, for instance, how Pasolini creates a world of perversion in *Salò: or the 120 Days of Sodoma* by making it the case that the torturers were able thinkers imbued with a very gloomy metaphysics capable of justifying their atrocities. Is the framework expressed by the work detachable from its story-content? How would that be possible? And does Currie offer precise standards for detachability? What makes a framework detachable in one work, while not in another? Besides, as Genette warns us (*Narrative Discourse*, Ithaca, N.Y. and London (1980): Cornell University Press, pp. 159-60), Proust himself was aware that his ideas should not be taken too seriously, but only in the light of 'the purely compositional aspect of the matter' (Proust's words).

The idea of framework also comes in relation to point (a). When this is the case, we enter the realm of expressivity; the expressive features of a narrative work are as relevant to its evaluation as its representational ones (Chapter 3, pp. 51-2). The central chapters of the book are devoted to defending this position, and it must be said that Currie's achievements are notable. Currie first well manages to show how the very pillars of a narrative structure, e.g., time, specificity, causation, can be altered by the *expressed authorial attitude to story content*. This entails that only when we take an *external* perspective on the author's agency, are we able to make sense of those subtle modulations of time, causal idiosyncrasies, lack of specificity that a narrative may present us with. Further, if Currie is right on the previous point then we should also pay attention not to postulate 'internal narrators' in narratives unless necessary; it may be the case that their presence is not required after all, and that the author as a narrator is using some sophisticated pretence to make her presence

visible in the story so as to rhetorically emphasize certain key passages (Chapter 4). Finally, an analysis of various narrative forms also shows how an author systematically exploits expressive techniques, which all seem to be related by the fact that the author's *persona*, the narrator or some of the characters in the story may adopt or reject points of view that are represented in, or simply suggested by, narrative discourse and in so doing certain attitudes to such points of view are made manifest (I am simplifying the point here, since not always does a point of view constitute the target of a relevant attitude).

Chapters 5 and 7 are extremely interesting in this sense, providing both a theoretical framework and several reflections upon narrative strategies to express attitudes. Chapter 5 offers a detailed analysis of how the adoption of a certain framework is crucially related to our ability to acquire certain perspectives on the story. Such perspectives are the expressions of *points of view*, which the narrators or her characters may constantly have or come to adopt rapidly even in the course of a single sentence. Currie offers some valuable insight, both epistemological and psychological, into the significance of points of view in narrative: for instance, that points of view are just the kind of things that arise because of an agent's limitation to access, and act upon, the world. But such sort of limitations, instead of being an obstacle to the narrative experience, becomes a powerful instrument in the service of the narrator. The case of free indirect discourse (FID) is one such notable example in narrative. (FID) is that linguistic device which allows a narrator to report someone's speech or thought, creating an effect of vivid *mimicry* and opening in this way to the expression of a point of view. First, suppose a character X in a story says or thinks this:

> (1) 'Tomorrow is Monday, Monday, the beginning of another school week!'

Now, ask yourself how a narrator could render this utterance in such a way as to convey the same piece of information, but in an expressive manner, exactly the manner typical of X as expressed by (1). The answer is by using (FID), which makes it possible for the narrator to shift the values of some indexicals while keeping others' values unaltered. So, in the narrator's voice (1) becomes

> (2) 'Tomorrow was Monday, Monday, the beginning of another school week!' (quot. from Philippe Schlenker, 'Context of Thought and Context of Utterance: A Note on Free Indirect Discourse and the Historical Present', Mind & Language, 19, 3, 2004, p. 280)

The shift has regarded only the tense, while the indexical 'Tomorrow' has remained unchanged (Schlenker defends the view that we need to allow for two contexts here, e.g., the context of utterance, which is fixed by the rules of grammar that determine the opportune shifts in reports, and the context of thought, which is flexible enough to track the intentions of the thinker). An analysis of this form of reporting in Austen's and James's narratives leads Currie to conclude that FID is the privileged mode of what he calls 'character-focused narration' (p. 143), where narrators try to imitate some of the characters' psychologies, thus professing some critical attitudes toward them. This also helps the reader to feel empathy or some sort of moral contagion for these characters, an experience which Currie considers as being crucial to our cognitive life in general.

I conclude by saying that this book has the indubitable merit of presenting original and stimulating ideas, which comes as no surprise given Currie's ability to master several fields with ease. In this sense, as the introduction states clearly, the book is addressed to a vast audience of readers. It will certainly help a fledgling discipline as philosophy of literature to grow. Hence, I think philosophers of literature will find the book of the most interest and value. Aestheticians as well as scholars and common readers will find Currie's examples taken from literature, cinema and photography highly entertaining and sometimes debatable. As an example, Currie's discussion in Chapter 9 of Hitchcock's *The Birds*: here Currie undertakes the challenge to defend one of the crucial points of the book: an interpretive minimalism that does not demand of an audience of a story more than the understanding, and enjoying, of its narrative structure. Thus the artificial sound of the birds in the movie would seem to serve Hitchcock's purpose of providing an overall ironic narration; ironic in making this artifice, as well as others, salient to the spectator. Although I am attracted by this view, its appeal is not immediate and, in fact, could be criticised as requiring a little of imagination to grasp it.

The book, as the subtitle suggests, is a philosophy of stories. Accordingly, the systematicity in the arguments at times gives way to a more readable approach.

Francesco Gentile
Department of Philosophy
University of Nottingham
apxfpg@nottingham.ac.uk