Centro de Filosofia da Universidade de Lisboa

Cognição e Conteúdo

Cognition and Content

Actas do Seminário de Filosofia Analítica
2003-2004

Proceedings of the Seminar Series in Analytic Philosophy
2003-2004

Organização de João Branquinho

Edited by João Branquinho

2005

Índice

Table of Contents

Prefácio

O presente volume contém uma porção significativa dos ensaios que de algum modo serviram de base às, ou tiveram a sua origem nas, comunicações apresentadas ao Seminário de Filosofia Analítica ao longo do ano académico de 2003-2004.

Os tópicos e problemas filosóficos discutidos no volume são, por conseguinte, de natureza bastante variada: a natureza da complexidade computacional no processamento de uma língua natural; a relação entre o significado linguístico e o sentido Fregeano; as conexões entre a a agência e o poder; o conteúdo semântico da ficção; a explicação dos impasses éticos; a natureza dos argumentos cépticos; as conexões entre as dissociações cognitivas e o carácter modular da mente; a relação entre a referência e o significado. Estes tópicos deixam-se subsumir num tema mais geral, o tema das ligações múltiplas entre a cognição e o conteúdo, mental ou linguístico. O tópico do conteúdo, a questão de determinar como é que muitas das nossas elocuções e muitos dos nossos estados mentais são dotados de conteúdo, representam algo (correcta ou incorrectamente), e o tópico da cognição, a investigação da natureza e dos mecanismos envolvidos na cognição humana, no processamento de informação tipicamente proveniente do exterior, são inegavelmente tópicos centrais da investigação filosófica, presente ou passada.

Dado o carácter inclusivo e pluridisciplinar dos tópicos cobertos, não é surpreendente que os ramos da Filosofia representados no presente volume sejam igualmente diversificados: a Filosofia Moral, a Filosofia da Linguagem, a Teoria do Conhecimento, a Filosofia da Mente, os Fundamentos da Ciência Cognitiva. E também não é surpreendente que nele também estejam representados outros ramos do conhecimento cuja relevância para o estudo do conteúdo e da cognição é conspícua: a Psicologia Cognitiva; a Linguística Computacional e a Inteligência Artificial.

Os autores dos ensaios são de proveniências diversas. De um lado, há um conjunto de especialistas nacionais a trabalhar na tradição analítica em Filosofia, como João Branquinho, ou a trabalhar em algumas áreas filosoficamente importantes da Ciência Cognitiva, como António Branco (Processamento da Língua Natural), José Frederico Marques (Psicologia Cognitiva) e Helder Coelho (Inteligência Artificial). Do outro lado, há um conjunto de especialistas estrangeiros de elevada reputação internacional: Allan Gibbard, da Universidade de Michigan, autor de livros influentes

na área da Ética e Filosofia Moral; Stephen Schiffer, da New York University, uma das principais figuras actuais da Filosofia da Linguagem e da Semântica Geral; Paolo Leonardi, da Universidade de Bolonha, um especialista nas áreas da Filosofia da Linguagem e Filosofia da Comunicação; e Manuel García-Carpintero, da Universidade de Barcelona, um dos grandes vultos da Filosofia Analítica praticada em Espanha e na Europa Continental.

Entre outras coisas, o presente volume é mais um reflexo da actual vitalidade da filosofia praticada à maneira analítica no nosso país.

João Branquinho
Lisboa, 15 de Maio de 2005

1
Cognitive Science and its Computational Foundations: A Natural Language Perspective

António Branco
Universidade de Lisboa
antonio.branco@di.fc.ul.pt

1. Introduction

Cognitive Science has been developed upon the cross-fertilization of results from a set of disciplines that includes Computation Theory and Artificial Intelligence, Psychology, Philosophy of Language and Mind, Neuroscience, and Linguistics. The intertwining of these contributions into a coherent scientific endeavour is pursued under the unifying rationale envisaging the human mind as a system of processes that handle information. Under an influential formulation "The mind is what the brain does; specifically, the brain processes information, and thinking is a kind of computation" (Pinker, 1998, p.21).

This rationale rests on two touchstones contributed by computer science. The *Fundamental Result of Computability* offers the proof that all the independent proposals designed to precisely characterize the intuitive notion of effective computation give rise to the same class of functions. And this result in turn supports the plausibility of the *Church-Turing Thesis* according to which that class of functions coincides with the intuitively circumscribed class of effectively computable functions. Colloquially, the Church-Turing thesis helps supporting the view that the information processing taking place at the human mind is exhaustively characterized by the notion of computability as it has been laid down in computation theory, while the Fundamental Result ensures that computational processes can be transferred across different computing devices: Taken together, they help to cogently support the hypothesis that the brain is one of such computational devices.

A major implication of this research framework is that, in principle, it should be possible to replicate in other computational devices, the human processing of information taking place at the brain. Exploring this possibility has been the prospect of Artificial Intelligence (AI) and the efforts devoted to push it forward have led not only to technological innovation but also to shed light over foundational issues of

Cognitive Science. In particular, many AI results are confluent in pointing towards the fact that the best algorithms that have been designed so far to replicate important human cognitive skills display such a high level of computational complexity that makes them unlike candidates to match the eventual natural algorithms underlying those skills. In this connection, the conclusions that might be drawn concerning the processing of natural language are thus of utmost relevance, given this is a high-level, if not the highest level cognitive capacity, which has been taken ever since as the hallmark of human distinctiveness in the universe of known entities.

But before entering into the subject matter of the present paper, it is worth considering first how this somewhat unsatisfactory outcome has been put into perspective. An immediate appreciation could just invoke the circumstance that AI research has no more than a few decades, and underline the fact that not having found yet a suitable solution to a problem does not necessarily imply that such a solution does not exist or could not be found. This type of position could even be deepened into the more radical observation that no matter how the P≠NP conjecture may seem plausible, it has not received a proof, and that this leaves open the chance that it might turn out to not to hold after all. Besides the tremendous scientific, technological and civilizational revolution this conjecture provably not holding would unleash, a profound implication might also be that such a proof could turn out to unveil the secret of intelligence and that the AI intractable solutions found until now can receive an efficient rendering after all.

Colloquially, this *P≠NP Conjecture* contributed by the Theory of Computation holds that for problems for which the best-known computational solution for every instance of the problem is not practical, no alternative, practical solution can eventually be found. A solution being practical for a given problem here means that when it is applied on a problem instance of size n (i.e. n words, n nodes in a graph, or n digits, etc., depending on the representation of the problem at stake), in the worst case, the returning of the corresponding answer takes a time at most proportional to the time needed to complete around n^k basic computation steps – the problem is then said to be of polynomial complexity. In turn, a solution not being practical means that, in the worst case, the returning of the corresponding answer takes a time proportional to the time needed to complete around k^n basic computation steps, and the problem is thus said to be of exponential complexity.

The mainstream appreciation of the difficulties uncovered by AI research, however, does not live in the hope that the P≠NP conjecture eventually turns out to be proved false. This is so not only because there are good grounds to sustain its plausibility, but also because some sensible additional observations are taken into account. One is that though a solution for a given problem may be shown not to be practical in general, that is, for every problem instance, of any size, it may yet render a good service when put to use for problems of limited size. Another observation is that though the best solution for a problem may provably be not practical, this does not imply that a good enough solution cannot be found that serves practical purposes in most if not all of the relevant cases. These considerations usually go on a par with the suggestion that human abilities have been shaped by evolution to handle with good enough success problems of reasonably limited size in as many as possible situations as they may be encountered in daily life conditions.

Against this background, in this paper we aim at addressing the issue of cognitive science and its computational foundations by focussing on the computational complexity of language processing. Given the above remarks, whatever are the results obtained thus far on natural language complexity, they bear nowadays a less stressing impact on the foundations of cognitive science than they were perhaps envisaged to bear a few decades ago. Still, it appears useful to compile an updated overview of the main, stable results in this area in as much as they provide the ground to critically assess how they have been programmatically interpreted and how they are guiding current research frameworks and shaping the results to be obtained in the foreseeable future.

2. Processing problems

Differently from fully specified problems and corresponding computational solutions, and in particular from artificial languages and associated processing algorithms, the determination of the complexity of natural language processing comes close to a chicken-egg issue, where a step forward in our understanding of one of the terms of the opposition opens the chance to improve what we know about the other term, and vice-versa: Human language is an entity of the natural world and to figure out how its processing takes place, it helps to know within which boundaries lies its computational complexity, and vice-versa.

Surely, there is empirical evidence of different sorts upon which to draw hypothesis about the processing of natural language. It ranges from latency times obtained in experimental settings from a population of subjects to linguistic judgments based on individual introspection, including quantitative data collected from corpora or images and registers of neurological activity in the brain, among others. In the current state of our scientific knowledge about natural language, the empirical data uncovered thus far have lend themselves to be accounted for under different hypothesis concerning the processing of natural language. Accordingly, to a very large extent, the strength of the conclusions about the computational complexity of natural language processing drew in the scope of these competing models hung dependent upon the corresponding model-internal assumptions and primitives.

Besides, it is worth noting that the processing of natural language is unlikely to constitute a single monolithic procedure. Taking into account, for instance, just the perception side, that permits the mapping of a linguistic form into the linguistic meaning it conveys, several procedures are likely to be involved, e.g. the detection of the different phonemes, their grouping into individual lexemes, the grouping of lexemes into phrases, the compositional calculation of their meaning from the meaning of their parts, etc. All such different dimensions and sub-problems of language processing do not have necessarily to be addressed by a single computational method or solution, or by different solutions of the same level of computational complexity.

Given this scenario, the chances to find firm results on the complexity of language are as much higher as it is simpler the sub-procedure under consideration, and as the empirical evidence is more elementary and less controversial, i.e. less prone to model-driven interpretation or accommodation. Guided by these concerns, important progress has been obtained when the issue of complexity is addressed by studying what is known as the *recognition problem*: given a string *s* of lexical forms of a natural language *L*, how complex it turns out to be the procedure to determine whether *s* is or is not a sentence of *L*? Not only this is a very simple sub-problem in terms of natural language processing when put into perspective with respect to the vast intricacies of human language, as the empirical evidence needed for the investigation of its complexity are not more than strings of lexemes forming sentences. Still, the results its investigation help to obtain may be of utmost

importance for the progress of the science and technology of human language, as the discussion below will help to make evident.

3. Complexity levels

For the sake of perspicuity, the recognition problem is typically rendered as a set membership problem. When for methodological purposes, the empirical evidence to be taken into account is confined to strings of lexemes, a language *L* lends itself to be regarded the set S_L whose elements are precisely those strings of lexemes that are its sentences. Seeking a computational solution for the problem whether a string of lexemes *s* is recognized as being a sentence of language *L* is thus seeking a solution for the question whether the string *s* is a member of the set S_L.

This would be a problem with an immediate solution if a human language could be extensionally presented as a listing with all and only its member sentences: it would just take to exhaustively scan that list seeking for the input string. But that is not the case, and the set of sentences of a language has rather to be presented under an intensional definition. Such a definition relies upon a number of criteria specifying the conditions of membership that allows their formalization into a finite set of properly defined rules. This set of rules constitutes a grammar for the language.

Accordingly, a solution for the membership problem turns out to consist into designing a parser that takes as input a string *s* and a grammar G_L for the language *L* and after a finite number of steps delivers the answer *yes* in case *s* belongs to the set S_L defined by G_L, and the answer *no* otherwise. Under the methodological constraints pointed out above, determining the computational complexity of the processing of a language boils down to determining the complexity of the least possible complex parser for a grammar of that language.

It turns out that a given language can be defined by infinitely many different grammars. Interestingly, when considering all possible languages thus definable, the corresponding grammars can be grouped into classes according to the level of complexity of the least possible complex parser required to solve the membership problem. There can be different such groupings depending of the classificatory granularity adopted. In view of our discussion here, a relevant grouping can be one that sets apart those grammars whose best parsers are practical solutions for the membership problem from those grammars for which no practical solutions could be found.

In this connection, it has been common practice to use a threefold complexity hierarchy proposed by Chomsky that groups grammars into regular, context-free and context-sensitive types. All regular grammars are context-free grammars, and the set of all languages defined by the former are properly included in the set of all languages defined by the latter. Similar considerations hold with respect to context-free and context-sensitive languages, respectively. While no practical parser could be found for every context-sensitive grammar, the best parsers for any regular or context-free grammar are always practical solutions for the membership problem, with the best parser for regular grammars being a highly practical one.[1]

This complexity hierarchy has been a useful yardstick to help determining the complexity of natural language processing in as much as this is methodologically circumscribed to the determination of the solution for the recognition problem. As this problem can be formulated in terms of the set membership problem, assessing its level of complexity turns out to consist in empirically clarifying what least as possible complex type of grammar is suited to define natural languages.

4. Grammar types

The claim that natural languages are not regular was originally put forward by Chomsky (1956), and valid arguments based on the English language can be found in (Gazdar and Pullum, 1987, p.394), or in the more accessible textbook (Partee *et al.*, 1993, p.477). An argument based on Portuguese can be deployed as follows.

Consider the following sequence of Portuguese example sentences built by successively embedding direct object relative clauses modifying subjects into each other:

O gato morreu.
The cat died.
O gato que o cão mordeu morreu.
The cat the dog bit died.
O gato que o cão que o elefante pisou mordeu morreu.
The cat the dog the elephant stepped over bit died.
O gato que o cão que o elefante que o rato assustou pisou mordeu morreu.

1 In more concrete terms, the best parser for any context-free grammars has polynomial, cubic complexity, while parsers for regular grammars have linear complexity, with time for obtaining a solution for a problem of size n proportional to n in the worst case (Hopcroft et al., 2001).

The cat the dog the elephant the mouse frightened stepped over bit died.

...

Based on these examples, and letting

A = {*que o cão, que o elefante, que o rato, que o periquito,...*}
B = {*pisou, mordeu, assustou, perseguiu,...*}

be finite sets of simple noun phrases preceded by the relative pronoun and transitive verbs, respectively, the following infinite subset of Portuguese can be defined

P' = {*o gato* A^n B^n *morreu* | $n \geq 0$}

where A^n and B^n are finite sequences of size *n* of members of A and B, respectively.

Notice that *P'* is the intersection of the set *P*, containing all sentences of Portuguese, with the set of the regular language

R = {*o gato A* B* morreu*}

where *A** and *B** are finite sequences of any size of members of *A* and *B*, respectively. Given that *P'* is not regular (Pumping Lemma[2]), and regular sets are closed under intersection, hence set *P* is not regular.

While it is not feasible to check this result for every one of the almost 7 000 existing languages in the world (Grimes, 2000), the position that natural languages are not regular on the basis of this kind of argument has not be challenged in an much as it can be easily replicated with other type of syntactic constructions besides centre-embedded relative clauses above, and for natural languages other than English or Portuguese (Gazdar and Pullum, 1987, p.395; Partee *et al.*, 1993, p.478).[3]

2 The Pumping Lemma for Regular Languages: Let L be a regular language. Then there exists a constant n (which depends on L) such that for every string w in L of length l≥n, we can break w into three strings w=xyz, such that y is not the empty string, the length of xy is less than n+1, and for all k≥0, the string xykz is also in L. (Hopcroft et al., 2001, p.126)

However the members, of any length, of P', are broken, no string can be found that matches the pattern xykz.

In order to find the place of natural languages in the complexity hierarchy above, this result implies that the question to be asked next is whether they are context-free. And for more almost three decades, different attempts were made to support the claim that they are not, based on data from English comparatives (Chomsky, 1963), Mohawk noun-stem incorporation (Postal, 1964), *respectively* constructions (Bar-Hillel and Shamir, 1964; Langendoen, 1977), Dutch embedded verb phrases (Huybregts, 1976; Bresnan *et al.*, 1982), number Pi (Elster, 1978), English *such that* clauses (Higginbotham, 1984), or English sluicing clauses (Langendoen and Postal, 1985). But the ones that were to be retained as empirically and formally valid arguments until present day are based on reduplication in noun formation in Bambara (Culy, 1985), and German Swiss embedded infinitival verb phrases (Shieber, 1985).[4]

The argument based on German Swiss data runs like this. Consider the following sequence of example sentences built by successively embedding verb phrases in subordinate clauses:

Jan säit das mer em Hans es huus haend wele hälfe aastriche.
Jan said that we the Hans-DAT the house-ACC have wanted help paint
Jan said that we have wanted to help Hans paint the house.
Jan säit das mer d'chind em Hans es huus haend wele laa hälfe aastriche.
Jan said that we the children-ACC the Hans-DAT the house-ACC have wanted let help paint
Jan said that we have wanted to let the children help Hans paint the house.
...

Based on these examples, and letting

A = {*d'chind*,...}
B = {*em Hans*,...}
C = {*laa*,...}
D = {*hälfe*,...}

3 Recently, Fitch and Hauser (2004) proposed that the divide between regular and non-regular computational process is the key to tell the difference between human and non human cognitive capacities. This claim was based on arguments of the sort just described. The validity of the argument given the experimentally elicited data obtained to sustain it was strongly challenged however (Liberman, 2004; Pinker and Jackendoff, 2005).

4 For overviews and critical assessment, see (Pullum and Gazdar, 1982; Pullum, 1984; Partee et al., 1993).

be finite sets of accusative noun phrases, dative noun phrases, accusative object taking transitive verbs, dative object taking transitive verbs, respectively, the following subset of German Swiss can be defined

G' = {*Jan säit das mer* A^n B^m *es huus haend wele* C^n D^m *aastriiche* | $n, m \geq 0$}

Notice that *G'* is the intersection of the set *G* with all sentences of Swiss German with the set of the regular language

R = {*Jan säit das mer A* B* es huus haend wele C* D* aastriiche*}

Given that *G'* is not context-free (Pumping Lemma[5]), and context-free sets are closed under intersection with regular sets, hence the set *G* is not regular.

5. Research programs

For the methodological purpose of gaining stepwise insight into the computational complexity of natural language processing, the focus of inquiry reported above was restricted just to the complexity of recognizing a string of lexemes as a sentence, i.e. to what was termed as the recognition problem. This methodological circumscription would have turn out to be of little help in case its outcome had been that for every natural language, this problem has highly practical computational solutions. Also of immediate interpretation would be an uncontroversial result at the other extreme of the range, viz. the one holding that for every natural language, this recognition problem has no computational solution of whatever level of complexity: In this event, it would be the foundations of cognitive science that would be under non negligible threat.

Interestingly, the outcome does not seem to lie at any of these extremes: The research summarized above is methodologically productive as it helps to uncover a constraint of utmost significance concerning the nature and processing of natural languages. Given this constraint is a landmark that no theorization can ignore, the way

5 The Pumping Lemma for Regular Languages: Let L be a context-free language. Then there exists a constant n (which depends on L) such that if z is any string in L such that the its length is at least n, the we can write z=uvwxy, subject to the following conditions: (i) the length of vwx is at most n; (i) vx is not the empty string; (iii) for all i>=0, uviwxiy is in L (Hopcroft, et al., 2001, p.275).

However the members, of any length, of G', are broken, no string can be found that matches the pattern uviwxiy.

it has been addressed and accounted for has been a key factor on how different grammatical research frameworks for natural language have been shaped in recent years.

5.1 Matching the complexity of the recognition problem

Given the results above, one possible research path is to study and design formalisms for natural language grammars that match this constraint with a lowest price as possible in terms of computational complexity. This implies going minimally beyond context-freeness, just as far as to the extent needed for the recognition problem of all sentences to receive a solution. The aim is thus to ensure both complete empirical coverage and parsing solutions at a practical level of complexity, if possible.

This goal has been pursued by exploring the fact that not all context-sensitive languages beyond context-freeness require a grammar whose parser is of non practical complexity.[6] The grammar formalisms for languages of this type have then been used to develop natural language computational grammars able to handle the known natural language constructions beyond the power of context-free grammars (Wahlster, 2000; Uszkoreit, 2005), thus providing a constructive argument that such constructions do not necessarily push the processing of natural language to computationally unpractical solutions. This is the line of research pursued most notably by GPSG[7] (Gazdar *et al.*, 1985), and by its successor HPSG[8] framework (Pollard and Sag, 1987, 1994), among others.

5.2 Approximating the complexity of the recognition problem

Another research path is based on a different position with respect to the complexity results presented in the previous section. This position grounds its rationale on a few facts deemed to be worth putting into perspective (Van Noord, 1998).

First, the only empirical evidence known to push natural language complexity beyond context-freeness complexity are those two results (Culy, 1987; Shieber, 1985) reported above in the previous section. Not only it took almost three decades of

6 For a critical overview, see (Gazdar and Pullum, 1985; Partee et al., 1993, Chap. 21).
7 Generalized Phrase Structure Grammar.
8 Head-driven Phrase Structure Grammar.

continued research effort to get at this first evidence, involving Bambara and German Swiss, as no other constructions or other languages were identified since then as replicating the same sort of implication in terms of complexity. Moreover, the cross-serial dependencies between verb phrases and their complements in German Swiss get harder to be recognized by native speakers beyond triple embedding (Shieber, 1985, p.329). These circumstances have been invoked to support the view that natural languages are in its essence within the context-freeness level of complexity.

Second, the centre-embedding constructions pushing natural language complexity beyond regular grammar are easy to replicate in different languages with different kinds of constructions. Nevertheless, also in this case, human speakers find themselves at odds to recognize sentences with more than a few embeddings. In this respect, it is interesting to note the contrast between the increasing difficulty of processing sentences in the sequence of centre embeddings used to argue for the strict context-freeness of natural languages

O gato morreu.
The cat died.
O gato [*que o cão mordeu*] *morreu.*
The cat the dog bit died.
O gato [*que o cão* [*que o elefante pisou mordeu*]] *morreu.*
The cat the dog the elephant step over bit died.
O gato [*que o cão* [*que o elefante* [*que o rato assustou*] *pisou*] *mordeu*] *morreu.*
The cat the dog the elephant the mouse frightened step over bit died.
...

with the much lower level of difficulty in processing a syntactically similar sequence but now with peripheral left-embedding

O gato morreu.
The cat died.
O gato [*que mordeu o cão*] *morreu.*
The cat [that bit the dog] died.
O gato [*que mordeu o cão* [*que pisou o elefante*]] *morreu.*
The cat [that bit the dog [that stepped over the elephant]] died.

O gato [*que mordeu o cão* [*que pisou o elefante* [*que assustou o rato*]]] *morreu.*
The cat [that bit the dog [that stepped over the elephant [that frightened the mouse]]] died.
...

This contrast has been used to support the view that there might be a finite upper bound for centre embedding in natural languages, in which case the latter could be described by a regular grammar. This view is further reinforced by the fact that peripheral embedding, though not centre-embedding, can be account for by regular grammars (Langendoen, 1975; Gazdar and Pullum, 1987; Van Noord, 1998).

These two points, together with the observation that humans process language very efficiently taking time that is a linear function of the length of the sentences, support the claim that regular grammars can provide at least good approximations to the description of natural languages. This is the line of research pursued since late eighties at different sites (Roche and Schabes, 1997; van Noord, 1998).

While the first research line, underlying HPSG, assume that the complexity results uncovered with respect to the recognition problem are to be matched, this second research path has rather tried to envisage them as an upper bound to be approximated. Although they are different, these two perspectives are not necessarily in conflict. Its complementarity has actually been explored under the rationale that for the sake of maximize efficiency, less complex solutions should be used as much as possible until the point where resorting to more complex solutions turns out to be unavoidable in face of the eventual sub-problems to be solved. Typically, regular methods have been applied to shallow linguistic processing, whose outcome feeds augmented context-free grammars in charge of deep linguistic processing, responsible for yielding fully-fledged grammatical representations (Crysmann *et al.*, 2002).

Accordingly, when it comes to the accommodation of the impact of the complexity results presented in the previous section, the largest divide is not that much between these two research programs but rather between them and a third one, to be described below.

5.3 The complexity of the recognition problem in a trade off

Rather than putting into perspective the empirical data supporting the complexity results, a third line of research calls instead for an examination of the

complexity metric used. In particular, it is noted that the distinction polynomial vs. exponential is a very coarse-grained measure of complexity, aimed at abstracting away as much as possible from the varying details of the atomic operations of different computing devices. Consequently, this distinction is a reliable indicator of the actual superior efficiency of algorithms, in general, and parsers, in particular, for problems beyond a sufficiently large size, such that a polynomial growth of the time needed to complete its operation will never be outperformed by an exponential growth.

These considerations are put together with the observation that sentences are made at most of a few dozens of words on average, that is, that the recognition problem for a natural language parser has a very limited size. Under such circumstances, for the actual time required to find a solution to a problem of this length, it is the grammar – with its considerable size in terms of number of rules, internal data structure to encode them, etc., - rather than the (exponential) complexity of parser that turns out to be responsible for the largest share. Moreover, moving from weaker (and more efficient) to more powerful (and least efficient) grammars types permits that a given language may be described more succinctly. Consequently, grammars well beyond context-freeness, even if requiring exponential parsers, may help processing actual sentences much faster than context-free grammars as they turn out to be more compact.

The point here is thus that given the very small size of the recognition problem in natural languages, the key issue for matching observed human efficiency is not to find the best general parser but the best trade-off between the complexity brought to the procedure, on the one hand, by the parser, and on the other hand, by other factors that are relevant given the dimension of the problem at stake, in particular by the size and shape of the grammar. Accordingly, natural language grammar is very likely to be of context-sensitive type (and its parser of exponential complexity).

This position is fully articulated in (Berwick and Weinberg, 1982), and LFG[9] (Kaplan and Bresnan, 1982) is a research program that lends itself to be classified as a grammar framework admitting context-sensitive grammars for natural languages (Kaplan and Bresnan, 1982; Berwick, 1982).[10]

9 Lexical Functional Grammar.

10 The GB (Government amd Binding) research line and its successor MP (Minimalist Program) (Chomsky, 1981, 1995) are deemed to embrace this position, though it turns out that these research traditions do not use any properly defined grammar formalism let alone to be able to support the development

6. Final remarks

The three programs of research on natural language grammar just described present three ways to accommodate the empirical results on the computational complexity of the recognition problem. Given the Chomsky hierarchy for computable solutions, they fill the whole spectrum of key hypothesis, ranging from the assumption that natural languages are strictly regular, to the postulation that they are strictly context-sensitive, including the claim that they are strictly context-free.

Restricting the focus of inquiry to the recognition problem was a productive methodological move that permitted to progress and gain new insights into the processing of natural language. Yet, as noted at the outset, this is just one of the sub-procedures involved in the whole task of natural language processing. As more empirical data, from more sources of evidence (e.g. behavioural records, brain imagiology, neurological findings, etc.), become available, one should expect that the number of working hypotheses about its complexity can be narrowed down. Such a convergence on the more suited research program is expected also to be fostered by extensive comparative testing of computational grammars of comparable linguistic coverage at the laboratory bench.

of any computational grammar for which complexity issues can be properly determined (Johnson and Lappin, 1997, 1999; Lappin et al., 2000). Besides, its empirical adequacy and explanaratory power has been also seriously challenged: for an overview see (Pinker and Jackendoff, 2005).

References

Bar-Hillel, Yehoshua and E. Shamir, 1964, "Finite State Languages: Formal Representations and Adequacy Problems", In Y. Bar-Hillel (ed.), *Language and Information*, Reading, Addsion-Wesley, pp.87-98.

Berwick, Robert and Amy Weinberg, 1982, "Parsing Efficiency, Computational Complexity, and the Evaluation of Grammatical Theories". *Linguistic Inquiry*, 13, pp.165-191.

Berwick, Robert, 1982, "Computational Complexity and Lexical Functional Grammar". *American Journal of Computational Linguistics*, 8, pp.97-109.

Bresnan, Joan, Ronald Kaplan, Stanley Peters and Annie Zaenen, 1982, "Cross-serial Dependencies in Dutch", *Linguistic Inquiry*, 13, pp.613-635.

Chomsky, Noam, 1956, , The Hague, Mouton.

Chomsky, Noam, 1963, "Formal Properties of Grammars", In R. Luce, R. Bush and E. Galanter (eds.), *Handbook of Mathematical Psychology*, vil. II, oo.323-418.

Chomsky, Noam, 1981, Lectures on Government and Binding, Dordrecht, Foris.

Chomsky, Noam, 1995, The Minimalist Program. Cambridge, The MIT Press.

Crysmann; Berthold, Anette Frank; Kiefer Bernd; Stefan Mueller; Guenter Neumann; Jakub Piskorski; Ulrich Schaefer; Melanie Siegel; Hans Uszkoreit; Feiyu Xu; Markus Becker; and Hans-Ulrich Krieger, 2002, "An Integrated Archictecture for Shallow and Deep Processing", In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp.441-448.

Culy, Christopher, 1985, "The Complexity of the Vocabulary of Bambara", *Linguistics and Philosophy*, 8, pp.345-351.

Elster, J., 1978, *Logic and Society: Contradictions and Possible Worlds*, New York, Willey.

Gazdar, Gerald and Geoffrey Pullum, 1987, "Computationally Relevant Properties of Natural Languages and their Grammars", In W. Savitch et al., 1987, pp.387-437; Reprinted from *New Generation Computing*, 1985, 273-306.

Grimes, Barbara (ed.), 2000), *Ethnologue, Volume 1: Languages of the World*,, SIL International, 14th ed.

Fitch, Tecumseh and Marc Hauser, 2004, "Computational Constraints on Syntactic Processing in a Nonhuman Primate", *Science*, 303, pp.377-380.

Higginbotham, James, 1984, “English is not a Context-free Language”, *Linguistic Inquiry*, 15, pp.225-234.

Hopcroft, John, Rajeev Motwani and Jeffrey Ullman, 2001, *Introduction to Automata Theory, Languages, and Computation*. New York, Addision-Wesley..

Huybregts, M., 1976; “Overlapping Dependencies in Dutch”, *Utrecht Working Papers in Linguistics*, 1, pp.24-65.

Johnson, David and Shalom Lappin,.1997, “A Critique of the Minimalist Program”, Linguistics and Philosophy, 20, 273-333.

Johnson, David and Shalom Lappin, 1999, Local Constraints vs. Economy. Stanford, CA: CSLI Publications.

Kaplan, Ronald and Joan Bresnan, 1982, ” Lexical-Functional Grammar: A formal system for grammatical representation”. In Joan Bresnan (ed.), *The Mental Representation of Grammatical Relations*, Cambridge, The MIT Press, pp.173--281.

Langendoen, Terence, 1975, “Finite-state Parsing of the Phrase-structure Languages and the Status of Readjustment Rules in Grammar”, *Linguistic Inquiry*, 5, pp.533-554.

Langendoen, Terence, 1977, “On the Inadequacy of Type-2 and Type-3 Grammars for Huiman Languages”, In P. Hopper (ed.), *Studies in Descriptive and Historical Linguistics*, Amsterdam, John Benjamin.

Langendoen, Terence and Paul Postal, 1985, “English and the Class of Context-free Languages”, *Computational Linguistics*, 10, pp.177-181.

Lappin, Shalom, Robert Levine and David Johnson, 2000, “The Structure of Unscientific Revolutions”. Natural Language and Linguistic Theory, 18, pp.665-771

Liberman, Marc, 2004, “Humans context-free, monkeys finite-state? Apparently not”, *Language Log*,
http://itre.cis.upenn.edu/%7Emyl/languagelog/archives/001399.html

Partee, Barbara Alice ter Meulen and Robert Wall, 1993, *Mathematical Methods in Linguistics*, Dordrecht, Kluwer.

Pinker, Steven and Ray Jackendoff, 2005, “The Faculty of Language: What's Special about it?”, Cognition, 95, pp.201-236.

Pinker, Steven, 1998, *How the Mind Works*, The Penguin Press.

Pollard, Carl and Ivan Sag, 1987, *Information-based Syntax and Semantics*, Stanford, CSLI Publication.

Pollard, Carl and Ivan Sag, 1994, *Head-Driven Phrase Structure Grammar*, Cicago, The University of Chicago.

Postal, Paul, 1964, “Limitations of Phrase Structure”, J. Fodor and J. Katz (eds.), *The Structure of Language: Readings in the Philosophy of Language*, Englewood Cliffs, Prentice-Hall.

Pullum, Geoffrey and Gerald Gazdar, 1982, “Natural Languages and Context-free Languages”, *Linguistics and Philosophy*, 4, pp.471-504.

Pullum, Geoffrey, 1984, “On Two Recent Attempts to Show that English is not a CFL”, *Computational Linguistics*, 10, pp.182-188.

Roche, Emmanuel and Yves Schabes, 1997, *Finite-State Language Processing*, Cambridge, The MIT Press.

Savitch, Walter, Emmon Bach William Marsh and Gila Safran-Naveh (eds.), 1987, *The Formal Complexity of Natural Language*, Dordrecht, D. Reidel.

Shieber, Stuart, 1985, “Evidence against the Context-freeness of Natural Language”, *Linguistics and Philosophy*, 8, pp.333-343.

Uszkoreit, Hans, 2005, *Deep Linguistic Processing with HPSG*, http://www.delph-in.net/index.php.

Van Noord, 1998, *Algorithms for Linguistic Processing*, Groningen, Alfa-informatica, http://odur.let.rug.nl/%7Evannoord/alp/proposal/pion.html

Wahlster, Wolfgang (ed.), 2000; *Verbmobil: Foundations of Speech-to-Speech Translation*, Berlin, Springer.

2
Sense and Meaning

João Branquinho
Universidade de Lisboa
jbranquinho@netcabo.pt

Consider the following familiar and seemingly cogent anti-Fregean argument.[1] Take a pair of strict synonyms in English, such as (presumably) 'ketchup' and 'catsup'. As these terms have the same meaning in all respects, it seems indubitable that they have the same propositional value - or semantic content - with respect to every possible context of use. Now consider a speaker, Sasha, whose mother tongue is not English and who learns the meanings of 'ketchup' and 'catsup' by means of ostensive definitions in the following way, not being told at the outset that they are straightforward synonyms. Sasha acquires the words by reading the labels on the bottles in which ketchup (or catsup) is served during meals. It happens that the same condiment is regularly served to him in bottles labelled 'catsup' at breakfast, when it is eaten with eggs and hash browns, and in bottles labelled 'ketchup' at lunch, when it is eaten with hamburgers. And such a situation induces Sasha to think that he is consuming a different condiment in each case (though one which is similar in taste, colour and consistency). Therefore, whereas 'Ketchup is ketchup' is uninformative to Sasha, 'Ketchup is catsup' would be quite informative to him: his knowledge would be substantially extended if he came to know that the condiment is one and the same in both cases. Hence, by the sort of strategy labelled by Nathan Salmon "the generalized Frege's Puzzle",[2] one would come to the conclusion that the information value of 'ketchup' (whatever it is) differs from the information value of

1 See Salmon 1990, 220-3.

'catsup' (whatever it is), which clearly contradicts the obvious principle that synonymy preserves information value.

I shall now discuss three possible rejoinders to the above sort of anti-Fregean argument. To begin with, an indirect counter-argument could be adduced to the effect that the argument in turn contradicts the following equally obvious principle:[3]

> **(E)** Necessarily, if a speaker **x** understands two expressions **E** and **E'** in a language **L**, and **E** and **E'** are (strict) synonyms in **L**, then **x** knows that **E** and **E'** are synonyms in **L**.

Principle **(E)** seems to be quite plausible: having grasped the meanings of **E** and **E'**, and given that **E** and **E'** have the same meaning, one is bound to be aware of this fact. And such a principle is of course violated in the anti-Fregean argument. On the one hand, Sasha is credited with an understanding of the words 'ketchup' and 'catsup' (he is supposed to have learnt the meanings of the words). On the other, the words in question are taken to be strict synonyms in English. Yet, Sasha does not know that they are synonyms. Therefore, one should apparently conclude that either principle **(E)** is false or the Millian argument is wrong.

Nonetheless, it should be noted that principle **(E)** is not unchallengeable.[4] Consider the following parallel principle:

> **(E*)** Necessarily, if a speaker **x** understands two expressions **E** and **E'** in a language **L**, and **E** and **E'** are *not* (strict) synonyms in **L**, then **x** knows that they are not synonyms in **L**.

Now **(E*)** turns out to be false. For instance, competent speakers of English will claim that words such as 'stop' and 'finish', or 'accident' and 'mistake', are synonymous, until

---

2 See Salmon 1986, 73.

3 This principle is subscribed to by Michael Dummett; see e.g. Dummett 1981, 323-4.

they are presented with examples which make clear the non-synonymy of the words as those speakers themselves use them. And such a sort of result about **(E*)** might be exploited to cast some doubt upon **(E)**. Thus, concerning a synonymous pair **E** and **E'**, it might be claimed that a speaker who understands both **E** and **E'** might be inclined to count them as synonymous, but withhold belief in synonymy because her experience of counter-examples to **(E*)** makes her suspect that she is wrong.

Of course, this could hardly be taken as evidence that principle **(E)** is false. And if the above sort of dilemma were inescapable one would be naturally inclined to take the latter horn of it; indeed, principle **(E)** is intuitively compelling and should not be given up on *that* basis. However, as we shall see, there is just no need to argue from the truth of principle **(E)** to the unsoundness of the anti-Fregean argument, and hence our dilemma turns out to be clearly escapable. I would regard the foregoing reflection about principle **(E*)** as at least showing that the indirect counter-argument from principle **(E)** is not as persuasive as one might think, in the sense that the intuitive strength of **(E)** may be after all insufficient to yield a convincing refutation of the Millian argument.

A second sort of reply to the anti-Fregean argument, which in a way complements the one just outlined, consists in what we might call the *objection from partial* (or *imperfect*) *understanding*. It might be argued that the 'ketchup'/'catsup' story does not satisfy a requirement which turns out to be crucial to Frege's original argument about informativeness. The requirement in question is that the speaker fully understand both sentences **S** and **S'** and therefore the singular terms out of which these sentences are composed. And it is alleged that it is doubtful whether the anti-Fregean argument meets this kind of demand since, on the one hand, Sasha is not a native or fully competent speaker of English, and, on the other, his peculiar way of learning the use of the words

4 I am very much indebted to Tim Williamson for this point.

'ketchup' and 'catsup' might be regarded as revealing that he has only a partial (or imperfect) grasp of the meanings of these words; and a full mastery is indeed required in the Fregean argument.

Now I have doubts about the effectiveness of such a line of attack. Indeed, it seems to be vulnerable to the following sort of intuitively powerful objection. Suppose that Sasha had learned the meaning of 'ketchup' in the peculiar way described before, but without the word 'catsup' coming into the story. This would normally be quite adequate for understanding. On the other hand, also learning something about 'catsup' should not undermine that. Hence, one may say that Sasha understands 'ketchup'; and, by a parallel argument, one would say that he also understands 'catsup'. Of course, there is no reason to think that such an objection would be decisive; maybe some reasonable reply could be framed against it. And one might even be inclined to think that the issue whether or not a speaker like Sasha should be credited with an adequate understanding of the words 'ketchup' and 'catsup', is a moot issue; or that it is unlikely that anything like an appeal to our ordinary intuitions about understanding would enable us to settle the dispute. Anyway, I guess that we are at least entitled to conclude that, given its relative weakness and lack of intuitive support, the objection from partial understanding is far from representing a good move against the Millian argument.

Finally, let me sketch a third sort of argumentative strategy one might pursue in dealing with the 'ketchup'/'catsup' story and similar cases from a Fregean perspective. Let us begin by taking for granted the premiss about understanding employed in the Millian argument. And let us recall that the argument is intended as a *reductio*, the allegedly absurd conclusion of which is the following claim:

(*) 'ketchup' and 'catsup' have different propositional values (with respect to Sasha's story);

Incidentally, some Millian theorists (especially Salmon) would take the *reductio* hypothesis to be the claim that 'Ketchup is catsup' is genuinely informative to Sasha. And the crucial premisses in the argument are these:

> **(@)** If expressions **E** and **E'** are synonymous (in a language **L**) then **E** and **E'** have the same propositional value (with respect to every possible context of use).
>
> **($)** 'Ketchup' and 'catsup' are synonymous (in English).

(*) is deemed implausible because, given **($)**, it comes out as inconsistent with **(@)**, and **(@)** and **($)** are both supposed to be obviously true. Now a Fregean reply could proceed in either of the following two directions.

On the one hand, one could just reject premiss **($)**, while keeping **(@)** and endorsing **(*)**. As a result, **(*)** could no longer be taken as a *reductio* of anything at all. But how could **($)** be reasonably challenged? Well, one might begin by maintaining that the notion of synonymy has no clear application to the case of proper names; indeed, ordinary proper names have no linguistic meanings, in the sense that definitional clauses like those one may find in a dictionary are not, in general, available for them. Then one might claim that words like 'ketchup' and 'catsup' may be thought of as having a semantic status which is very similar to that of proper names: they are names of substances or names of kinds of stuff. One could then apparently conclude that, strictly speaking, words of that sort have no linguistic meanings either; hence, the notion of synonymy has no straightforward application to them. However, I do not think that such an approach is convincing. First, and less important, it turns out that some authorized English dictionaries[5] actually count the words 'ketchup' and 'catsup' as being strict synonyms, the latter being - along with 'catchup' - just a spelling variant of the former (a

5 E.g. Collins English Dictionary and The Oxford Encyclopedic English Dictionary.

variant used mainly in the U.S.A.). Second, and more important, even if one happens to be reluctant to apply the notion of synonymy to names of artificial kinds, it turns out that an argument can be mounted which parallels the 'ketchup'/'catsup' argument and yet involves only colour words; and the objection from the inapplicability of the notion of synonymy would hardly make sense with respect to colour words. Thus, in Portuguese there are two different words for red, viz. '*vermelho*' and '*encarnado*', which have literally the same meaning; I am pretty sure that every native (or fully competent) speaker of Portuguese would promptly acknowledge such words as being strictly synonymous. Now suppose that Ronald, a monolingual speaker of English, is taught Portuguese by the direct method and learns '*vermelho*' and '*encarnado*' under the following sort of circumstances. First, he learns the meaning of '*vermelho*' by being presented with samples of a particular shade of red. Then he comes to learn '*encarnado*' by being presented with samples of what is in fact the very same shade of red. It just happens that, on the later occasion, Ronald does not remember the particular shade of red he saw when he learned '*vermelho*'; so, when he acquires the word '*encarnado*', he does not even entertain the question whether '*vermelho*' holds of the samples then seen. Let us agree that one is entitled by ordinary standards to credit Ronald with an adequate understanding of the Portuguese predicates '*vermelho*' and '*encarnado*'. Then it would be possible to draw from the above case conclusions which parallel those drawn from the 'ketchup'/'catsup' story, a significant difference between the two arguments being that in the '*vermelho*'/'*encarnado*' argument the premiss about synonymy seems to be incontrovertible. In particular, it would not be difficult to imagine a set of circumstances under which the Portuguese sentence '*Vermelho é* (is) *encarnado*' (as uttered on the later occasion) would carry non-trivial or informative information to Ronald (whereas '*Vermelho é* (is) *vermelho*' would be clearly uninformative to him). The objection might

be raised that as soon as Ronald considered the matter, he would realize that the words in question are synonymous. Yet, a possible reply might be given as follows: Ronald may realize that '*vermelho*' and '*encarnado*' have similar meanings, but feel unable to rule out the possibility that he will one day see a shade that will strike him as *vermelho*, but not as *encarnado*.

Alternatively, and this is the kind of move I would be inclined to favour, one could just reject premiss **(@)**, while accepting premiss **($)** and fully endorsing claim **(*)**. Again, it would follow that **(*)** could no longer be taken as a *reductio* of anything at all. But how could one reasonably reject **(@)**? Well, it turns out that from a Fregean standpoint, a standpoint in which information values are (at least partially) senses or modes of presentation, claim **(@)** is by no means compulsory. Indeed, it seems to me that a Fregean theorist might, plausibly and fruitfully, hold the view that sameness of linguistic meaning does not entail sameness of sense.

Notice that the connection holding between the notions of linguistic meaning and Fregean sense is a very loose one, at least according to the general conception of sense with which we are willing to work. The linguistic meaning conventionally correlated with a given singular term, e.g. an indexical expression, is certainly an objective feature of the term; it is something which remains necessarily constant across speakers and across occasions of use. By contrast, the Fregean senses associated with singular terms are, in many cases, non-conventional and subjective; it is always possible for singular modes of presentation to vary from speaker to speaker and/or from occasions of use to occasion of use. Thus, different speakers may be in a position to attach distinct particular senses to a given singular term token **t** (at a given time), or to tokens **t** and **t'** of the same type (at the same or at different times), even when **t** and **t'** are co-referential (in given contexts of use); i.e., they may entertain different particular ways of thinking of the

object referred to. And the same speaker may be in a position to attach distinct particular senses to singular term tokens **t** and **t'** of the same type (at different times), even when **t** and **t'** are co-referential (in given contexts); i.e., she may entertain on distinct occasions different particular ways of thinking of the object referred to. However, in all such cases, it is obvious that the linguistic meaning of the singular term tokens - which is conferred upon them by the types of which they are tokens - is necessarily the same. On the other hand, for any tokens **t** and **t'** of different types which are co-referential in contexts **c** and **c'**, it is obviously not the case that if **t** and **t'** express the same particular sense in **c** and **c'** relative to a given speaker, then **t** and **t'** are synonymous (or belong to synonymous types); according to some neo-Fregean accounts, certain uses of indexicals such as 'here' and 'there', or demonstratives such as 'this' and 'that', illustrate this point.

Moreover, one may even introduce cases in which singular term tokens **t** and **t'** which are co-referential (in given contexts of use) and which belong to different but *synonymous* types are nevertheless to be seen, in the light of certain brands of Fregeanism, as having *different* senses with respect to a given subject. Thus, one may safely assume that the expression-types 'yesterday' and 'the day (just) before today' have exactly the same linguistic meaning (dictionaries usually give the latter as the meaning of the former). But consider tokens of such types as uttered by a speaker, say Jones, under the following sort of circumstances. At 11:58 pm on a day **d** Jones asserts 'Yesterday was mild', having thus a belief about **d-1**; and one hour later, looking at his watch, he comes to assert 'The day before today was not mild', apparently having thus a belief about **d**. Yet, Jones happens to be unaware that Summer Time ends precisely at midnight on **d** and that then clocks go back one hour, so that the time of his later assertion is in fact 11:58 pm on **d** and the associated (putative) belief a belief about **d-1**. Now if one thinks of the modes of presentation correlated with temporal indexicals as

consisting in ways of tracking a time - or re-identifying it - throughout a period of time, then it will not be the case that Jones entertains on both occasions (or, rather, at what is conventionally the same time) the same singular sense.[6]

The preceding considerations motivate a picture of the relationship between linguistic meaning and information value on which there is a considerable gap between the two notions and on which claim **(@)** is not, in general, true. Claim **(@)** is simply taken for granted in the anti-Fregean argument; and this is so because, considered in its application to ordinary proper names and to names of (natural or artificial) kinds, it comes out as trivially true under a strict Millian account. In effect, the object or the kind referred to by any syntactically simple singular term of the above sort (in a given context) is regarded on such a view as playing a double semantic role: it is (or at least it determines) the linguistic meaning of the term; and it is also the propositional value assigned to the term (in the context). But it seems to be somehow unfair to invoke this doctrine - as a means of validating claim **(@)** - in the course of assessing an argument whose aim is to show that such a doctrine is wrong. And once one drops the Millian conception of the information values of simple sentences as being singular propositions, which are by definition psychologically insensitive, in favour of a conception of such information values as being Fregean thoughts, which are by definition psychologically sensitive, claim **(@)** ceases to be compelling.

I am therefore prepared to endorse the claim that, in general, it is possible for expressions which are strict synonymous (in a given language) to have different senses in a speaker's idiolect. Concerning the 'ketchup'/'catsup' story, I would say that Sasha employs different ways of thinking of the same condiment, the 'ketchup'-way of thinking

6 This is a very rough description of the case under consideration. I examine the notion of indexical sense employed in my Oxford D. Phil. Thesis Direct Reference, Cognitive Significance and Fregean Sense.

and the 'catsup'-way of thinking. He is obviously not aware that he is being presented with a single kind of stuff at breakfast and at lunch; no wonder then that the thought that ketchup is catsup is informative to him. Given their analogy with ordinary proper names, names of natural or artificial kinds are - to use Evans's terminology[7] - *information-invoking* singular terms. Accordingly, one could sketchily represent Sasha's distinct modes of presentation of ketchup as consisting in different chains of information, or in separate mental files titled 'ketchup' and 'catsup', formed on the basis of his disparate cognitive encounters with the condiment at breakfast and at lunch. And a parallel treatment might be provided to the '*vermelho*'/'*encarnado*' case, the difference being that even a Millian theorist would acknowledge that predicates are to be assigned something very akin to Fregean senses as their propositional values in possible contexts of use. Indeed, on Salmon's theory of predicative reference, in contradistinction to the case of syntactically simple singular terms, syntactically simple predicates are thought of as having two sorts of semantic value: their information values, which are taken to be certain intensional entities like n-ary attributes; and their references, which are taken to be certain extensional entities like functions from n-tuples of objects to truth-values. But Salmon would presumably treat synonymous predicates like '*vermelho*' and '*encarnado*' as invariably contributing one and the same unary attribute to the information contents of sentences in which they might occur. And this would not enable us to accommodate possible differences in cognitive significance which, *pace* Salmon, we wish to take as basic data in need of explanation, such as the potential difference in informative value - relative to Ronald and to his story - between a thought expressed with the help of '*vermelho*' and a thought expressed with the help of '*encarnado*'. Thus, I would say that Ronald employs in thought different ways of thinking of redness; or, if one prefers, he

7 See Evans 1982, 384-5.

employs different ways of thinking of that function or Fregean concept which yields, for any red surface as argument, the True as value. And Ronald's case seems to motivate a *De Re* view of the kind of senses expressed by colour terms, i.e. a view on which such senses are to be seen as being (partially) dependent upon certain perceptual relations holding between a thinker and colour samples in her environment; in effect, it is the presence of this sort of non-conceptual factors which ultimately explains why redness is presented to Ronald under distinct modes of presentation.

A consequence of the above way of countering the anti-Fregean argument is that principle **(E)** should be, after all, given up. We are committed to the result that e.g., though '*vermelho*' and '*encarnado*' are synonyms (in Portuguese), Ronald does not know that they are synonyms. If Ronald knew this then he would know that '*vermelho*' and '*encarnado*' are co-extensional predicates and thus that one and the same colour is presented to him on both occasions; but then a sentence such as '*Vermelho é encarnado*' would not express a thought which would be informative to him. Therefore, since we take as intuitively sound the claim about informativeness, and since we take the objection from imperfect understanding as intuitively dubious, we are forced to reject principle **(E)**. Now I think that there is nothing essentially wrong in pursuing this train of thought. Underlying principle **(E)** is a certain form of cartesianism about meaning, in the sense that our knowledge about sameness of meaning is taken to be infallible. But one may have good reasons, in this and in other areas of philosophical inquiry, to be suspicious about such cartesian principles; it is very likely that linguistic meaning is not as transparent as it is claimed, and that even fully competent and reflective speakers may be mistaken about synonymy.

References

Dummett, Michael 1981 The Interpretation of Frege's Philosophy, London, Duckworth.

Evans, Gareth **1982** *The Varieties of Reference*, edited by John McDowell, Oxford, Clarendon Press and New York, Oxford University Press

Salmon, Nathan **1986** *Frege's Puzzle*, Cambridge, Massachusetts and London, England, The MIT Press, Bradford Books

Salmon, Nathan **1990** 'A Millian Heir Rejects the Wages of Sinn' in C.A. Anderson and J. Owens (eds), *Propositional Attitudes: The Role of Content in Logic, Language and Mind*, Stanford, CSLI.

3
Faces do Poder de um Agente

Helder Coelho
Universidade de Lisboa
hcoelho@di.fc.ul.pt

Resumo**:** A questão do poder de um agente é actual nos dias de hoje pois a simulação das actividades económicas ou das actividades políticas interessa às organizações em geral, e é um meio de melhorar a tomada de decisão em ocasiões de grande complexidade. O poder tem sido abordado em muitas disciplinas, com destaque para a Filosofia e para a Psicologia, e no presente artigo vamos nos restringir ao poder individual o qual é indispensável para caracterizarmos um modelo e uma arquitectura de um agente artificial, inteligente e autónomo, e capaz de acção directa em ambientes não triviais. Por detrás deste poder encontramos a vontade, a qual está relacionada com o comportamento associado à capacidade de fazer escolhas e ao sentimento que se tem dessas escolhas quanto às acções que se tomam e às consequências que daí decorrem.

1. Introdução

> **"**Não sendo uma potência, a Filosofia não pode conduzir uma batalha contra as potências. Pelo contrário, é preferível que ela se envolva numa guerra sem batalha, numa guerrilha contra elas." Gilles Deleuze, 2003.

Quando se estuda a concepção de agentes artificiais, com particular destaque para a natureza das suas mentes, através da proposta de novos modelos e arquitecturas computacionais, somos imediatamente confrontados com vários mistérios actuais ainda por resolver, tais como: Como as mentes trabalham e se fabricam comportamentos? Como decidimos, e depois, passamos animadamente à acção? Como escolhemos a melhor coisa que devemos fazer? Como descobrimos as consequências positivas ou negativas que podem surgir se realizarmos esta (boa/má) acção? Como enfrentamos o desconhecimento sobre o que acontecerá a seguir a um evento? Que impulsos governam o nosso dia a dia? As respostas para estas questões dizem respeito aos modos como o poder de um agente (natural ou artificial) se envolve com a sua vontade, a autonomia e a tomada de decisão, e podem ser encontradas por detrás de arquitecturas computacionais que misturem as capacidades de reacção e de deliberação (o modelo Crenças-Desejos-Intenções ou BDI, de "Beliefs-Desires-Intentions", de Bratman é a principal proposta que importa pôr em causa e discutir).

Tais dúvidas em redor dos modelos e das arquitecturas dos agentes podem ser enquadradas num imenso campo multidisciplinar, onde se destacam trabalhos sobre vários temas e com intervenção de numerosos cientistas, a saber:

Filosofia Política: Alma, Multitude, Poder (Aristóteles, Espinosa, Locke, Hume, Negri).

Filosofia da Mente: Agência, Intenção, Acção (Grice, Pörn, Bratman, Cohen, Levesque , Dennett).

Psicologia Cognitiva: Mentalidade, Decisão (James, Fodor, Putnam, McCarthy, Minsky, Devlin).

Psicologia Social: Dependência, Poder (Gasser, Epstein, Tuomela, Castelfranchi).

Inteligência Artificial Distribuída: Agência, Decisão, Autonomia (Georgeff, Shoham, Sloman, Jennings, Wooldridge).

Um ser humano tem uma noção do que deve fazer sobre um certo curso de acções (espaço do livre arbítrio), e do mesmo modo um agente artificial deve também ter essa noção, a qual é puramente individual. Há que distinguir o que o agente deve realizar das acções que ele tem a intenção de executar para atingir aquele resultado. Por um lado, ele necessita de considerar as consequências das acções alternativas, e por outro lado ele pode ou não ter a obrigação de agir (vontade) para alcançar e obter esse mesmo resultado. Perante cada ponto de ramificação (estrutura da escolha) o agente deve ponderar a evitabilidade dos acontecimentos, distinguindo entre acções e não acções (possibilidades). De facto, uma acção que é determinada pelas características dessa situação não envolve livre arbítrio (McCarthy, 2000). A maioria das acções humanas (e as dos animais) ocorre como reacção directa (sem deliberação) a uma certa situação e não implica a antecipação das consequências das acções alternativas, mas exige uma vontade (instinto de sobrevivência), um impulso, em vencer e ultrapassar a incomodidade dessas situações.

O desenho dos comportamentos dos agentes é condicionado pelos mecanismos que seleccionamos para incluir nas suas mentes, e durante algum tempo o modelo BDI foi o padrão a seguir. Mas entre agir logo e organizar (ter a intenção de) a execução de uma acção há um espaço de oportunidades que condicionam o tipo de agente que queremos criar para um certo contexto situacional (no teatro, um encenador escolhe os actores para os papeis de uma peça, obedecendo às exigências e ao perfil psicológico dos candidatos). Teremos ocasião, ao longo do artigo, de

reflectirmos sobre esta problemática e também sobre as alternativas para conceber criaturas mais empenhadas (agressivas) em intervir.

Comecemos então por descrever o modo como, do ponto de vista experimental, se articulam aquelas cinco disciplinas. Esta clarificação é indispensável para se compreender a consiliência entre as ciências humanas e as engenharias. A metodologia de trabalho adoptada na Inteligência Artificial (IA) articula-se em cinco passos, 1) Colocar hipóteses ("A mente é constituída por uma sociedade de agentes"), 2) Fazer conjecturas ("Existe uma tabela periódica dos estados mentais"); 3) Arrumar teorias ("O comportamento de um agente emerge a partir da interacção entre dois espaços, o mental e o arquitectural"); 4) Construir instalações (Agile, INTERSECTIONS); e, 5) Executar simulações (Observar a inovação criada pelas empresas na geografia de um parque de ciência e tecnologia). Esta metodologia, como se ilustra na figura 1, permite, hoje em dia, atacar problemas e situações complexas de uma forma eficaz, e, no caso da geração de inovação, estabelecer uma estratégia quanto à selecção dos novos ocupantes desse parque, assim como da sua melhor localização para potenciar sinergias.

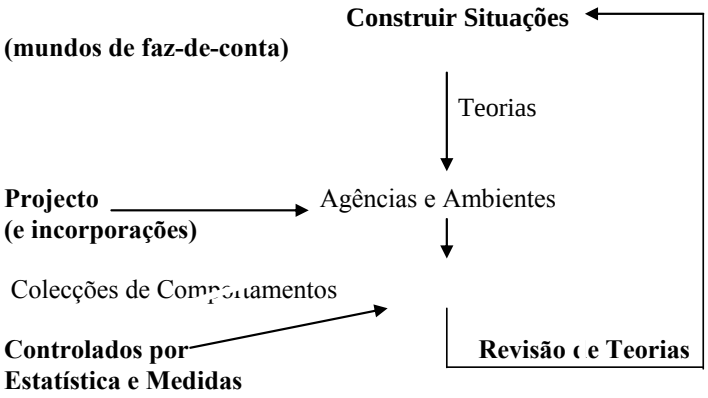

Figura 1: Campo de intervenção das ciências da complexidade

A simulação de mundos de faz-de-conta (simplificações da realidade) é acompanhada de medidas rigorosas capazes de garantir a correcção das experiências e de apoiar a emergência de conjecturas, as quais são transformadas depois em novas perguntas do tipo "o que-se". O ambiente envolvente, a instalação e as ferramentas,

para a sua construção e acompanhamento, devem ser cuidadosamente aferidas antes de se projectarem os ensaios definitivos. Especial atenção tem de ser dedicada à agência (organização dos agentes) e aos agentes per si. As suas arquitecturas são muito importantes para garantirmos a validade das conclusões a retirar de toda a experimentação. Sabemos que essas configurações possuem uma enorme diversidade de mecanismos, sendo o seu desenho ajustado ao tipo de situações e de ambiências que as enfrentam.

2. Pontos de vista

As influências sobre este tipo de investigação vêm de inúmeras disciplinas. Por exemplo, na Inteligência Artificial (IA), (John McCarthy, 1979) procurou uma linguagem apoiada em estados mentais, inspirada na Psicologia Popular (vulgo, "Folk Psychology"): "Atribuir crenças, livre arbítrio, intenções, consciência, habilidades ou querer a uma máquina é legítimo quando uma tal atribuição expressa a mesma informação acerca da máquina daquela que é expressa sobre uma pessoa". "É útil quando a atribuição ajuda-nos a compreender a estrutura da máquina, o seu comportamento passado ou futuro, ou como repará-la ou melhorá-la". (Baum, 2004) defendeu que "A mente é complexa, porque é o resultado da evolução. Para se compreender a mente, precisamos de entender os processos de pensamento, e o processo evolucionário que o produziu, no nível computacional", e também que "A mente é um programa de computador." Por outro lado, (Minsky, 1986) propôs uma teoria sobre a Sociedade da Mente e afirmou que "A mente é uma imensa colecção de agentes que realizam um grande leque de funções, tais como esperar, prever, reparar, lembrar, rever, actuar, depurar, comparar, generalizar, exemplificar por analogia, simplificar, e muitas outras tarefas cognitivas". A teoria da Sociedade da Mente aborda como os grupos de agentes se podem organizar em comunidades (via colaboração) com mais capacidades de que um único agente poderia alguma vez ter. Na Filosofia, Espinosa (séc. XVII) avançou que é importante "fazer do corpo uma potência, que não se reduz ao organismo, e fazer do pensamento uma potência que não se reduz à consciência". Finalmente, no terreno da Biologia e das Neurociências, (Damásio, 2003) defendeu que "sem reacções emocionais, não saberíamos o que fazer", e que "a Lógica diz-nos o que se segue após diferentes acções, mas são as emoções que nos ajudam a enfrentar as consequências". De facto, as emoções de um agente estão associadas com a sua autonomia e a escolha, e por sua vez a autonomia

com a sua vontade e poder individual, enquanto a escolha está ligada intimamente com a tomada de decisão. Damásio estabeleceu uma cadeia [Pensamentos, Emoções, Sentimentos], onde as emoções são reacções públicas no teatro do corpo (comportamentos) e os sentimentos são imagens mentais privadas no teatro da mente (percepções de emoções). As emoções surgem assim como acções ou movimentos, e, por conseguinte, são facetas que importa reunir quando se caracteriza a potência de um agente. No presente artigo, por razões de espaço, não abordaremos estas facetas emocionais, focando toda a nossa atenção sobre o triângulo mental (BDI) da cognição e os processos para o subverter.

3. Teorias da Mente

> "A volição é o exercício actual do poder que a mente tem em ordenar a consideração de uma ideia, ou em se abster de a levar em conta." John Locke, 1690.

Quando fazemos uma panorâmica sobre o que sabemos acerca da natureza da mente somos confrontados com a seguinte questão sobre o que somos: Seremos meros agentes que planeiam num mundo social, obsecados só com o futuro? A resposta é simples: preparamos regularmente planos de acção, antecipadamente, e suportamos formas complexas de organização para interactuar com o ambiente que nos envolve e face ao estado futuro do mundo. No entanto, não nos restringimos a simples planeadores, pois no quotidiano presente agimos de forma instintiva e graças às nossas intuições. Após aprender os antecedentes causais de certos padrões de comportamento, um agente gera uma nova oportunidade de escolha. Para intervir, precisa de vontade, de força, pois agir não é só fazer! Muito do que executamos, e somos, é expresso e explicado em termos do que sentimos, acreditamos, desejamos, tencionamos, esperamos, receamos, etc. A compreensão do nosso comportamento é feita, assim, em função dos nossos estados mentais, ou seja das explicações das nossas acções (Corrêa e Coelho, 2005).

Na Psicologia Popular recorremos aos estados mentais, tais como os objectivos (estados que o agente alcança), as crenças (o que o agente imagina ser o estado do mundo a partir das informações e dos conhecimentos), os desejos (as preferências do agente, as motivação), as intenções (os objectivos ou os desejos com

os quais o agente se comprometeu trabalhar), as expectactivas (as situações em que se espera a ocorrência de algo), e as emoções (acções, modos de pensamento). E, por isso, as criaturas artificiais são modeladas sobre tais estados mentais.

Ao longo da história da Filosofia os desejos foram consideradas por Hume como causas de acções e a vontade como um sub-produto dos processos causais subjacentes, enquanto para Kant o papel dos agentes, ou o poder e os princípios universais das suas acção, estavam por detrás dos desejos (como bases potenciais da acção). Para Espinosa, a vontade podia existir sem o desejo, mas não o desejo sem a vontade que estava sempre por detrás do desejo. James defendia que a vontade era energia, movimento, esforço de atenção, e as acções podiam ser rápidas ou lentas, conforme o grau de esforço e os impulsos. Estas ideias levaram-nos a re-considerar o modelo BDI, popularizado pela indústria de jogos de computador, e hoje em dia quase uma norma dos agentes artificiais. E, a crítica avançada mais à frente foi motivada por novas exigências de aplicações organizacionais muito complexas, onde a gestão dos recursos impõe o domínio da inteligência emocional. Como pode um agente controlar a razão sob tensão quando enfrenta outros agentes numa empresa?

Antes de agir um agente autónomo pensa e ganha energia e força para enfrentar o ambiente que o cerca. Tudo o que se passar em seguida, depende dos seus sentimentos, da sua alma (impulso, anima)! Por isso, faz sentido perguntar: Onde colocamos a alma na arquitectura da mente (interioridade) de um agente artificial? E, onde encontra um tal agente a energia (potência) para intervir e para se auto-motivar? Donde surgem as motivações e os recursos da sua potência? O que suporta os seus impulsos?

Quando pretendemos explicar como uma mente funciona e estabelecer uma escala da cognição recorremos ao diagrama da figura 2 (Devlin, 1991).

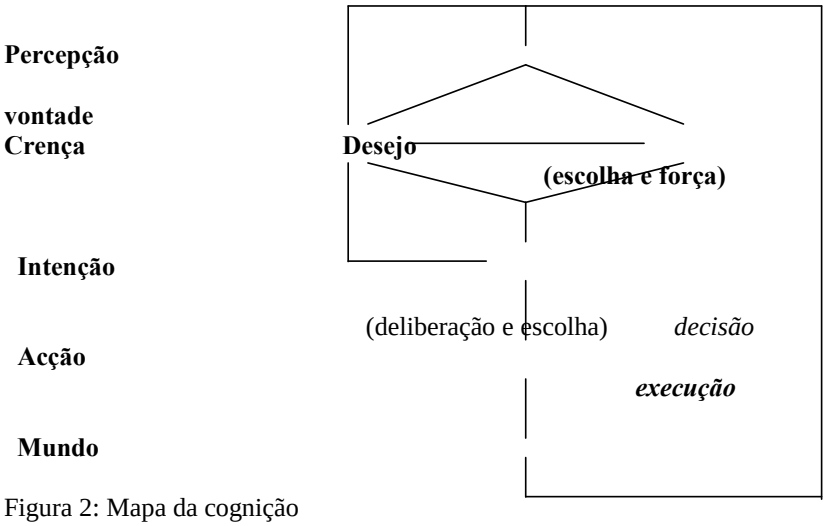

Figura 2: Mapa da cognição

Na figura 2 optámos por situar a vontade entre as crenças e os desejos, seguindo a sugestão de Espinosa, e por assinalar os diversos lugares da escolha e da tomada de decisão, os quais foram quase ignorados pelos defensores do modelo BDI.

A ideia da aplicabilidade da noção de vontade pode ser compreendida olhando para a teoria da Agitação de (Debord, 1957) e para a construção de situações (experiências, exercícios, simulações), de ambiências colectivas, dos conjuntos de impressões que determinam a qualidade do momento, onde a intervenção não é só execução. Vejamos um exemplo de acção directa (de agi-prop), onde é imposto a organização do lugar, a escolha dos participantes, e a provocação dos acontecimentos, com vista a obter a ambiência desejada e a comunicação das mensagens. A teoria da Construção de Situações, elaborada pelo grupo dos Situacionistas de Debord, é tão boa para o cinema, como para analisar o futebol sintético construído com agentes artificiais e explorado pelos jogos de computador! Em ambos os casos, procura-se um valor mais que distinga um grupo do outro, e, em nossa opinião, a vontade faz a diferença.

De facto, os agentes (com poder individual) aparecem em diversas situações, tais como:

Futebol: organiza uma jogada, envolvendo outros (trabalho colectivo/equipa) para marcar golo.

Agitação: capta a atenção de um grupo, dos indivíduos, convence com argumentos (agita) e mobiliza-os para irem para o lugar X.

Emergência: após um desastre, toma o comando das operações (liderança), emitindo ordens para todos os outros, coordenando, realizando acções e dando exemplos de intervenção e ajuda.

Nestes exemplos simples surgem situações de dependência (aguarda ordens dos outros; ajuda/prejudica), de independência (faz acções, sem esperar pelo comando; intervém, transforma), e, de autonomia (assume iniciativas, em função da análise das situações; lidera, contesta, controla, ou coordena), onde a ideia de poder está sempre presente.

A noção de poder surge também por detrás de uma equipa que tem de realizar uma sequência de acções visando atingir um objectivo ou cumprir uma missão, como durante uma partida de RoboSoccer: durante o desenrolar de um jogo de futebol, em que a sua equipa está a perder, um jogador arrasta dois companheiros para virar o resultado, constroi um lance colectivo, conjugando e articulando esforços para lançar um contra-ataque demolidor (jogada de laboratório). De novo, o que faz a diferença é a bondade manifestada por aquele jogador.

No entanto, há que distinguir entre a potência (poder individual), organizado em torno do mobilizar, do agir e do fazer (Coelho e Coelho, 2003), onde se destacam as questões de mudança e re-organização, alteração do status-quo para enfrentar o novo, e o poder social, organizado em torno da dependência e da influência (redes de dependência), com particular destaque para o normativo, o disciplinador e o institucional (David, Sichman e Coelho, 2001). Neste último caso surgem problemas como a resolução de conflitos e a feitura de coligações ou de alianças. No primeiro caso, o poder individual tem a ver com a acção directa, e no segundo caso, o poder social, com o domínio. Também, a noção de poder individual (capacidade de concretização de objectivos, virada para os efeitos das acções sobre o mundo) está ligada à autonomia (capacidade de gestão do espaço mental, virada para os processos interiores), e esta articulação é importante de reter quando se projecta uma criatura artificial.

Do ponto de vista da tipologia, a emergência do poder começa de forma ascendente a partir do poder-de (poder individual de realizar uma acção directa, objectivo, resultado) e sobe para o poder social de influenciar. Entre estes dois estratos surge o poder-sobre, onde temos de entrar em linha de conta com a

dependência (Castelfranchi, 2000). Ao longo deste artigo, iremos concentrarmo-nos apenas sobre o poder-de.

Temos ainda de distinguir entre a acção individual e a acção colectiva. No primeiro caso, não se espera que os indivíduos actuem consistentemente e que adoptem os objectivos do grupo. Pelo contrário, a maioria aproveita-se ("free-riders"), a seu favor, das contribuições dos outros (incerteza dos comportamentos). No segundo caso, os objectivos a serem alcançados conjuntamente pelos membros de um grupo são um bem público, que pode ser usofruído sem rivalidades, e de cuja fruição ninguém pode ser afastado.

No contraponto entre individual e colectivo, a ideia de multitude, importada de Espinosa, adquire uma enorme relavância, face à noção de povo defendida por Hobbes no seu Contrato Social. Por multitude, entende-se a rede de indivíduos, o conjunto de singularidades, o elemento orgânico do trabalho colaborativo, a multiplicidade, o movimento de elementos, sem convergir para um Um, que se compõem e constroem sem se evaporarem de forma centrípeta (em direcção ao centro), ou ainda a ideia de constituinte, com uma multiplicidade produtiva e auto-organizada. O povo ilude a individualidade, reclama a presença do chefe ou líder, e exige a centralização do poder como se o todo fosse apenas a soma anónima das partes.

O Um confronta o Múltiplo (Plural), ou ainda o poder-de individual joga contra o poder colectivo (combinação de poderes individuais através da cooperação, onde o todo não é a soma das partes). Importa compreender como agrupar sem necessidade de unidade (Um), sem subordinar as suas singularidades ou negar as suas diferenças, numa forma de organização, em rede, onde não exista uma hierarquia rígida e um chefe que imponha uma vontade sem contestação (facilita o contra-poder contra a forma centralizada e sob direcção responsável). De realçar que a multitude é fundada sobre a comunicação e a colaboração, isto é através da produção de linguagem, de imagens e de afectos. Por detrás há um bem público (conhecimentos, informações e relações), que interessa defender.

Quando se passa para a construção de uma organização há um movimento em direcção à complexidade. Um indivíduo não é anónimo: o Um compõe o Múltiplo, mas não se desfaz nele, pois o movimento dos Uns é centrífugo e não centrípeto (centralismo)! Logo, a cooperação é viável. O Um tem de ter qualidades, personalidade, potência para se afirmar e fazer afirmar o múltiplo (através da

colaboração). Muitas vezes, nas organizações, esquecem-se estes aspectos devido à forma hierárquica como são construídas. Que agente democrático é este? Como se constroi este poder colectivo, sem domínio central, ou como devemos estudar a eterna oposição entre centralização e descentralização? Que democracias estão por aparecer no futuro (Hardt e Negri, 2000)?

4. Idealizações e Construções

O programa Massive (Multiple Agent Simulation In Virtual Environments) de Stephen Regelous, explorado na trilogia fílmica de O Senhor dos Aneis de Peter Jackson, inspirou-se nas teorias de Karl Sims (1994). Foi concebido como vida artificial e sistema multiagente, e não apenas como sistema de animação de multidões, e as acções de cada agente são uma função da percepção do seu ambiente, dos seus conhecimentos e das suas particularidades. Não há decisão centralizada e cada agente pode tomar 24 decisões por segundo. Existe a possibilidade de se gerarem milhares de agentes, autónomos e todos diferentes, o que levou Hollywood a interessar-se por esta tecnologia para regressar às produções à Cecil B. DeMille dos anos 60.

Cada um dos agentes reconhece o ambiente externo, com quatro sentidos (visão, ouvido, tacto, cheiro) e é sujeito a forças físicas (sons, luzes, impactes). Emitem gritos e ruídos para baixar a moral dos seus adversários, e a sua percepção é indispensável para se orientarem nos campos de batalha e para acharem os seus inimigos. Na movimentação dos agentes o choque é evitado (observação sobre como as pessoas não se chocam nas ruas), assim como outros obstáculos.

A personalidade dos guerreiros de O Senhor dos Aneis tem várias dimensões, e a raça e a morfologia podem variar a seu belo prazer. O seu carácter inclui tonalidades de agressividade, de medo e de força. O seu corpo adquire várias formas e características, ajudando a concepção de anões, de agentes com boa vista ou com pele escura. As suas acções (lista de 350) incluem espada para cima, espada para baixo, um passo em frente, um passo para trás (1 seg. cada), e são determinadas pelo cérebro escolhido (100 a 8.000 nós lógicos comportamentais). São também governadas por colecções de regras capazes de perceber, de interpretar, de responder, de decidir, de agir, de comunicar, de controlar a agressão, o estilo de luta e o movimento através de um terreno variado.

O programa Massive permite gerar com facilidade multidões, e misturar o caos e as acções com fins bem definidos: agentes com escolhas autónomas e imprevisíveis (cérebros digitais), em vez da dinâmica baseada em simples partículas (tecnologia anterior adoptada pela indústria do cinema de animação). Tal possibilidade suporta as mudanças de comportamento: os agentes reagem ao terreno que pisam, andando de forma diferente quando sobem ou descem uma colina. Não há duplicação (cópia) das acções mecânicas, e os agentes individuais são todos diferentes e com poder de agir próprio: combatem mesmo a sério! Mas, neste cinema, não há ainda poder individual e colectivo.

Consideremos um segundo exemplo de inteligência social (Henke e Bassler, 2004), importado da Biologia Molecular, onde foram estudados dois mecanismos das bactérias Vibrio harveyi e Vibrio fischeri: 1) Libertar e detectar moléculas do tipo hormona (autoindutores): quando se atinge o limiar de uma concentração de autoindutores (sensibilidade ao quorum) inicia-se uma cascata de sinais e produz-se luz (evitam que o sistema imunitário do hospedeiro seja activado cedo demais, antes de atingirem a massa crítica); 2) A V. haveyi produz os autoindutores AI-1 e AI-2, os quais são detectados por uma proteína sensora particular: o circuito AI-1 é usado para a comunicação intra-espécies e o AI-2 para a comunicação interespécies (para fazer actos de cooperação).

A comunicação entre bactérias pode ser entendida se olharmos para a linguagem, constituída por sinais microbianos (mensagens químicas) e por um léxico (vocabulário) de sinais para comunicar produtos químicos capazes de gerarem sensibilidade ao quorum. O resultado consiste em as bactérias emitirem luz (azul suave), quando em grupo, em resposta ao aumento da densidade da população celular. E, o objectivo é atrair para a reprodução, distinguir amigo de inimigo, construir exércitos, organizar a divisão do trabalho, ou ainda cometer suicídio em massa para o bem da comunidade.

Os circuitos da bactéria Vibrio harveyi permitem modificar os seus comportamentos, sob formas subtis, dependendo destas estarem em minoria ou em maioria na comunidade das espécies. As bactérias são criaturas sociais, sendo o mecanismo de sensibilidade ao quorum, usado para a cooperação (coordenar os elementos de um organismo ou criar superorganismos). Mas, existem comportamentos subversivos, pois a bactéria Bacillus subtilis produz uma molécula que modifica os autoindutores usados por muitas outras bactérias, arruinando a sua

eficácia (através do empastelamento da comunicação). As duas estratégias de defesa, 1) formação mediada por sensibilização ao quorum de biofilmes, e 2) produção de metabolitos específicos para inibir o predador, são essenciais para a construção da inteligência social das bactérias e do seu poder colectivo.

A sensibilidade ao quorum das bactérias (a comunicação microbial assegura a cooperação entre os micróbios) é usada para romper com os sistemas de comunicação dos inimigos (sabotagem) e para fabricar contra-medidas de subversão. Um tal mecanismo é interessante para a indústria farmacéutica, e por isso estes novos conhecimentos apoiam actualmente a concepção de novas drogas antibacterianas (diferentes dos antibióticos correntes): não matam os micróbios individuais, mas destroem a sua capacidade em comunicar entre si e em cooperar, e subvertem ainda o trabalho em equipa dos micróbios sociais.

Uma assembleia só atinge autoridade quando o número dos seus membros (bactérias, virus) atinge um certo quorum: então pode-se votar e tomar decisões (expressão da vontade), quase de forma idêntica ao que se passa num parlamento! Estamos no seio do comportamento animal: os indivíduos juntam-se em colónias, comunicam entre si e decidem o que fazer em seguida: constroem biofilmes (filmes de bactérias, comunidades de micróbios com centenas de espécies distintas) os quais permitem às bactérias defenderem-se contra os predadores e gerarem factores de malignidade.

O terceiro exemplo foi elaborado por (Urbano, 2004) e trata da formação descentralizada de consensos em sociedades de agentes artificiais sem nenhuma estrutura organizacional (sequência de escolhas consensuais diferentes, de consenso em consenso, como um passeio colectivo aleatório). A transição entre escolhas colectivas diferentes deve obedecer a critérios de desempenho, pois não é desejável que a sociedade oscile entre várias decisões possíveis: a escolha de uma delas deve ser tão rápida quanto possível. A natureza das escolhas colectivas e a duração dos consensos são completamente aleatórias.

O consenso é visto como uma forma de decisão colectiva comum que pode desempenhar um papel importante na coordenação das escolhas dos agentes, por exemplo, na selecção de regras que regulem os conflitos inter-individuais. A tendência progressiva para atribuir mais autonomia aos agentes dá sentido à investigação dos mecanismos de emergência endógena de convenções. Seguindo o princípio descentralizado do projecto de sistemas, a formação de consensos vai ser o

resultado espontâneo das interacções entre os diversos agentes, onde não existe a figura de um líder, nem de um centro coordenador (confrontar com a noção de multitude de Espinosa, atrás introduzida). Assim, torna-se imperioso compreender a emergência dos comportamentos colectivos (leis sociais, léxico partilhado), a selecção de uma escolha comum e o modo como ela pode evoluir ao longo do tempo. Para esse efeito foi introduzido um comportamento individual muito simples, adaptado de um modelo de auto-organização sobre a formação de uma ordem de domínio do mundo animal, e a experimentação foi conseguida com grupos de micro-pintores capazes de criarem padrões artísticos aleatórios ao longo de uma sucessão de escolhas consensuais.

Estes três exemplos chamam a nossa atenção para a diversidade dos mecanismos de controle individual (reactividade), muitas vezes desprezáveis face ao forte poder de deliberação (cognição: selecção de acções) dos agentes artificiais. Mas é urgente reflectir sobre o jogo entre a racionalidade e a autonomia/controle, ou entre a deliberação e o controle, pois as intenções (Bratman, 1990) dependem das acções escolhidas:

Intenções=AcçõesSeleccionadas(Vontade,Crenças)

Se a mente é composta de estados mentais, nada melhor que procurar os blocos de construção primitivos e analisar as suas propriedades (comportamentos) e as relações que se estabelecem entre eles para definir os comportamentos compostos de molde a construir diversos tipos de agentes: realistas, altruístas, egoístas, timídos, audaciosos, prudentes e interventores. Será possível a desconstrução da potência (oscilações entre a vontade e a intenção) de um agente?

Na investigação realizada pelo grupo de IA da FCUL abordou-se em primeiro lugar a possibilidade de desenhar agentes auto-motivados em redor do modelo BDI (Bratman, Israel e Pollack, 1987), com um destaque especial para se compreenderem os estados de alma. A ideia extensiva de mentalidade foi desenvolvida sobre o modelo BDIE (Corrêa,1994) com quatro estados (crenças, desejos, intenções e expectactivas) e uma nova arquitectura computacional inspirada nas propostas de (Genesereth e Nilsson, 1987). Mais tarde, e após numerosas experiências, foi proposta uma teoria da Sociedade de Estados Mentais (Corrêa & Coelho, 1998) capaz de agregar os conhecimentos sobre os atributos, as relações e os controles dos estados mentais numa única tábua, onde os comportamentos mentais são compostos como átomos dessa mesma tábua. Cada estado mental (individual ou colectivo) é definido por uma lista

de atributos (insatisfação, urgência, certeza, intensidade, insistência, importância), um conjunto de 10 leis de causalidade, e por 11 regras de controle (filtros, accionadores), e a aplicabilidade desta teoria na Psicologia Clínica é objecto de investigação em progresso (Corrêa e Coelho, 2005).

Em seguida, estudou-se a decomposição da cadeia da tomada de decisão (Antunes, 2001), recorrendo-se ao modelo BVG ("Beliefs, Values, Goals"): agentes com qualidades (valores) e o cálculo da importância. Aprendeu-se a operar na faixa "Desejo, Intenção, Acção" (veja-se a figura 2) e identificaram-se os lugares da escolha, assim como se pode controlar a autonomia através dos valores e das emoções. Actualmente, aborda-se o desenho de agentes interventores e a construção da vontade ou potência de agir (Coelho e Coelho, 2003), ou seja a construção de agentes agitadores, contestatários, interventores ou transformadores. E, para isso investiga-se como operar entre "crenças e desejos", isto é a descobrir qual é o lugar do campo da vontade. Trata-se de saber como controlar o agir, graças à potência do agente, para no futuro se pesquisar a identificação do lugar dos afectos e dos sentimentos, alargando o espaço dos estados mentais úteis.

A construção da acção (segundo o modelo BDI) passa pela montagem dos processos do raciocínio prático (RP), dirigido à acção, em contraponto com o raciocínio teórico (RT), dirigido unicamente às crenças. No raciocínio prático existem dois processos computacionais: 1) Decidir que estado queremos atingir (deliberação), com destaque para o resultado que é a escolha; e, 2) Decidir como queremos atingir esse estado (via raciocínio meios-fins), com destaque para o resultado que é o plano. A maquinaria computacional (Bratman, Israel e Pollack, 1987), avançada na instalação IRMA sobre o mundo dos tabuleiros de xadrez (TileWorld) é composta por dois ciclos, o do raciocínio, desde a percepção ao raciocinador de meios-fins, passando pela filtragem das opções e a deliberação das soluções potenciais, e o da actuação, entre a deliberação e a geração de estruturas de intenção, onde as razões desejos-crenças aparecem como considerações para serem avaliadas e pesadas (com valores) na deliberação final entre as opções relevantes e as admissíveis.

A evolução do algoritmo básico (Wooldridge, 2001), para construir uma arquitectura computacional para o modelo BDI, permite compreendermos como o desenho de um agente poderá incluir, além de outros dispositivos, os mecanismos da vontade. Começaremos pela proposta mais simples de uma arquitectura BDI:

1. while for verdade
2. observe o mundo;
3. actualize o modelo interno;
4. delibere sobre qual a intenção que deve alcançar a seguir;
5. use o raciocínio meios-fins para obter um plano para essa intenção;
6. execute o plano
7. end-while

O segundo algoritmo pode ser um pouco mais refinado:

1. B:=Bo; /* Bo são crenças iniciais*/
2. while verdade do
3. get next percepção P;
4. B:=brf(B,P);
5. I:=delibere(B);
6. Pi:=plano(B,I);
7. execute(Pi)
8. end-while

onde se identificam, de forma explícita, três componentes do ciclo de controle, a saber:

Função de revisão de crenças (brf): f(B) x Per → f(B)
Processo de deliberação (delibere): f(B) → f(I)
Raciocínio meios-fins (plano): f(B) x f(I) → plano

O terceiro algoritmo, com uma maior capacidade de deliberação, inclui mais duas componentes, além das três anteriores, as opções e o filtro:

brf: f(B) x Per → f(B)
delibere: f(B) → f(I)
plano: f(B) x f(I) → plano
opções: f(B) x f(I) → f(D)
filtro: f(B) x f(D) x f(I) → f(I)

```
1.    B:=Bo; /* Bo são crenças iniciais*/
2.    I:=Io; /* Io são intenções iniciais */
3.    while verdade do
4.    get next percepção P;
5.    B:=brf(B,P);
6.    D:=opções(B,I);
7.    I:=filtro(B,D,I);
      Pi:=plano(B,I);
      execute(Pi)
10.   end-while
```

O modelo BDI tem aspectos descritivos e normativos. Tenta capturar a estrutura básica da concepção da mente ligada à compreensão comum das intenções e da acção, e articula ainda uma concepção da racionalidade prática. Outras alternativas, como o modelo BVG proposto por (Antunes, 2001), concentram a sua atenção apenas sobre as etapas da tomada de decisão, recorrendo à inclusão de valores para melhorarem a filtragem e a ponderação (deliberação), etapas anteriores à actuação do agente. O desaparecimento dos desejos é táctico, como medida simplificadora, enquanto as intenções são substituídas pelos objectivos.

As intenções e os planos desempenham um papel de coordenação na definição da vontade. No âmbito das liberdades e responsabilidades de um agente ele faz escolhas com valores e propósitos para determinar os cursos das suas acções e decidir o que fazer. As intenções estão ancoradas no processo de deliberação, e a vontade está aqui implícita na própria intenção! Há que desconstruir a intenção para que a vontade possa aparecer claramente.

O raciocínio prático pode ser defindo pela equação:

Acção = Crenças + Desejos

a qual deve ser confrontada com a equação clássica de (Cohen e Levesque, 1990),

Intenção = Escolha + Compromisso

onde por detrás do compromisso temos um mecanismo para determinar quando e como se podem deixar cair as intenções. Portanto, temos aqui duas dimensões agregadas e confundidas: a vontade (como relação entre intenção e acção) e o

raciocínio (dirigido à intenção futura). Grice adoptou duas equações diferentes, querendo tornar claro que é necessário explicitar a vontade:

Intenção = Vontade + Crença

Intenção = Volição

É preciso incluir, na arquitectura de um agente, um mecanismo para distinguir um agente activo (interventor) de um passivo (espectador): a ideia da potência! Não basta querer, é preciso agir! As contribuições de Espinosa, Hume e Kant são essenciais para se entender o lugar da vontade entre as crenças e os desejos.

Assim, ao desconstruirmos o significado de intenção

Causar A

(visando ou dirigindo o comportamento)

Chegar perto de A,

Esforçar-se por A.

distinguimos três peças de comportamento, a saber:

1) Alguém tem intenção de fazer A.

2) Alguém esforça-se por fazer A.

3) Alguém age (faz A) intencionalmente.

Um tal exercício é fundamental para entendermos o que é um agente com potência, ficando ainda várias perguntas no ar, tais como donde vem esse esforço? E, como se gera esta força? Será através das sequintes equações?

Acção = Crenças + Vontade +Desejos

Intenção^{+} = Escolha + Compromisso + Acção

A resposta parece ser negativa, pois os mecanismos da vontade não são simples. Considerando um comportamento activo de um agente, aperecem quatro aspectos que devem ser objecto de meditação:

1) Alguém quer fazer A por causa do ambiente que o rodeia e das suas próprias crenças.

2) Ganha força para desfazer-se de alguns desejos e para escolher os compromissos certos.

3) Pondera as alternativas e escolhe a sua preferência para intervir.

4) Age, realizando A, intencionalmente.

Vejamos então como poderemos incluir estes aspectos num primeiro algoritmo com vontade, onde a escolha dos objectivos não é focalizada, e preste-se atenção à re-definição da intenção através das acções seleccionadas, referida anteriormente:

```
1.  D:=Do; /* Do são desejos iniciais*/
2.  Pi:=Pio; /* Pio são planos iniciais */
3.  while verdade do
4.  get next percepção P;
5.  B:=brf(P);
6.  G:=filtro_vontade(B,D);
7.  I:=máquina_racional(G,B,Pi);
8.  Acção_selec:=filtro_valores(B,I);
9.  execute(Acção_selec)
10. end-while
```

O que faz falta para um agente enfrentar um ambiente complexo em tempo real (exploração espacial, combate a incêndios, simulação organizacional)? Em primeiro lugar, ser capaz de lidar com várias propriedades de um ambiente: acessibilidade vs. inacessibilidade, determinismo vs. não determinismo, episódico vs. não episódico, estático vs. dinâmico, discreto vs. contínuo, e imprevisível, ou seja em mudança constante. E, para atingir uma tal aptidão, requer não só autonomia e decisão, mas também vontade.

Como construir a volição de uma forma efectiva, atendendo a que um agente tem pressa de agir? Como garantir respostas responsáveis, onde o pensar ocorrerá sempre antes de reagir? A solução em que temos estado empenhados passa por 1) regenerar o modelo BDI, ao longo da cadeia do raciocínio prático, ou então 2) de subvertê-lo. A primeira alternativa consiste em recorrer à arquitectura PRS (Procedural Reasoning System) do modelo BDI (Ingrand, Geogeff e Rao, 1992), o qual mistura o raciocínio dirigido por objectivos com o comportamento reactivo. De facto, a arquitectura PRS adequa-se melhor a ambientes incertos, dinâmicos e operando em tempo real, onde a vontade é indispensável. De um ponto de vista esquemático, a organização de um agente é montada em redor de um interpretador (raciocinador), com particular destaque para os módulos da base de dados (crenças), da biblioteca de conhecimento (planos), para os objectivos e para a estrutura de intenções. O algoritmo PRS é apresentado em seguida:

```
1.  B:=crenças_iniciais ( )
2.  I:=intenções_iniciais ( )
```

3. while verdade do
4. B:=revisão_de_crenças(B,estímulos);
5. D:=possibilidades(B, I);
6. I:=escolha(B,D,I);
7. T:=tarefas(B,I,acções);
8. while not (vazio(T) ou realizado(I,B) ou impossível(I,B))
9. t:=próxima_tarefa(T);
10. execute(t);
11. T:=restantes_tarefas(T);
12. B:=revisão_de_crenças(B,estímulos);
13. if reconsiderar(I,B) then D:=deliberação(B,I) I:=escolha(B,D,I);
    else adequado(T,I,B) then T:=tarefas(B,I,acções)
14. end-if
15. end-while
16. end-while

Este algoritmo apresenta dois ciclos computacionais, onde o ciclo dedicado à execução da computação está encastrado num sub-ciclo. A computação do compromisso (veja-se a equação de Cohen e Levesque) está colocada no grande ciclo da escolha da acção, o que torna o agente mais lento a agir. Convém referir que a estrutura em dois ciclos não proporciona bons agentes locais, fundamentais para operarem em ambientes hostis ou incertos pois o cálculo é mais complexo. O que é necessário então para enfrentar esses ambientes complexos? A resposta é trivial: um agente com aptidões locais e globais. O algoritmo LGS (Local Global System) de (Coelho e Coelho, 2005) é a nossa resposta a este desafio, pois mistura as capacidades de deliberação e de reacção:

1. bB:=crenças_iniciais( )
2. bD:=desejos_iniciais( )
3. bP:=planos_iniciais( )
4. bS:=resolvedores_iniciais( )
5. while verdade do
6. bB:=revisão_de_crenças(bB,estímulos);
7. <aB,aD,aP,s>:=*vontade*(bB);

8. C:=possibilidades(aB,aP); {can do}
9. N:=necessidades(aD,aP); {need do}
10. I:=escolha(s,C,N);
11. execute(I)
12. end-while

(onde aB= crenças activas; s= estratégia de escolha; L= local, G= global)

Agora a computação do compromisso não é feita no grande ciclo (como no PRS), mas no interior da vontade. Todo o ciclo da escolha da acção, no PRS, tem de ser modificado. E, o poder individual (potência) do agente depende do custo computacional do processamento global da intenção, e não está relacionado com o processo usado para executar um desejo.

O que está verdadeiramente em causa na concepção destes algoritmos com vontade? Atingir um pensamento intuitivo, isto é permitir que os agentes possam saltar logo para as conclusões e pensar rápido (graças à intuição), ou seja agirem em directo. De facto, demasiado tempo, e muita informação, confunde-nos e cega-nos, e também aos agentes artificiais! As experiências associadas ao último algoritmo estão a decorrer, sendo ainda cedo para podermos entender completamente como se domina a intuição, e como se articula convenientemente reacção e deliberação.

5. Conclusões

> "Não basta saber, é preciso também aplicar. Não basta querer é preciso agir."
> Goethe (1742-1832).

O fascínio do pensar o sistema de oposições no tipo de agentes (passivo/activo, obediente/intempestivo, racional/irracional) leva-nos a conceber novos modelos e arquitecturas computacionais, e a procurar ingredientes capazes de inspirarem novos mecanismos. Há um vai e vém entre o interesse e o desinteresse de um agente e isso marca a sua própria potência em agir. Na busca de novas personalidades, há uma necessidade de introduzir agentes morais com carácter e qualidades, que se orientem cada vez mais por valores, e menos por interesses, e também neste caso a vontade é um facto relevante. No caso dos agentes que actuam

nas organizações somos obrigados a conceber papeis mais complexos, possuídos de certas aptidões individuais e colectivas, como a liderança ou o trabalho em equipa, e de novo a vontade parece ser uma faceta distintiva a ter em conta se quisermos simular a dinâmica empresarial.

A contemplação do trabalho em equipa leva-nos imediatamente a recorrer à ideia de multitude, complexo em que o poder resulta da colaboração e cooperação entre agentes individuais e democráticos. A clarificação do lugar da vontade na personalidade de um agente é apenas um contributo para clarificarmos como se podem erigir essas multitudes. O alvo determinante é conhecermos como as agências podem ser organizadas, e descobrirmos que importância podem ter as formas de as estruturar para atingirmos depois objectivos sociais. As sociedades electrónicas e artificiais são hoje relevantes do ponto de vista económico, sobretudo na Internet, e a autonomia dos agentes envolvidos não é completamente clara, causando algumas apreensões por causa do livre arbítrio que possa daí surgir. Por isso, o estudo em torno do poder individual (e da vontade) dos agentes é um passo para esclarecermos que espaço difuso inclui as consequências da automatização. A existência de agentes responsáveis, com características próximas dos seres humanos, poderá sossegar-nos quanto aos perigos da tecnologia, e também quanto aos benefícios que dela podemos retirar. A questão pertinente em aberto continuará a ser a definição do quadro das responsabilidades, e por isso esta linha de pesquisa poderá dar algumas contribuições significativas.

Referências

Antunes, L. – Agentes com Decisão Baseada em Valores, Dissertação de Doutoramento, Universidade de Lisboa, 2001.

Antunes, L., Faria, J. e Coelho, H. – Enhancing Autonomy with Value-Based Goal Adoption, in Gmytrasiewicz, P. e Parsons, S. (eds.), Proceedings of the AAAI Spring Symposium on Game and Decision Theories on Agent Design, AAAI Press, 2001.

Baum, E. – What is Thought?, The MIT Press, 2004.

Aristóteles – Da Alma, Edições 70, Textos Filosóficos, 2001.

Bratman, M. – Intention, Plans, and Practical Reason, Harvard University Press, 1987.

Bratman, M. – What is Intention? in Intentions in Communication, P. R. Cohen, J.L. Morgan and M. Pollack (eds.), The MIT Press, 1990.

Bratman, M., Israel, D. e Pollack, M. – Towards an Architecture for Resource-Bounded Agents, Technical Report CSL 87-104, SRI, Stanford University, August, 1987.

Castelfranchi, C. – The Micro-Macro Constitution of Power, Proceedings of the 2nd SARA Workshop, Lisboa, 2000.

Coelho, F. e Coelho, H. – Towards Individual Power Design, in Progress in Artificial Intelligence, Pires, F. M. e Abreu, S. (eds.), Lecture Notes on AI, Volume 2902, pp. 366-378, Springer-Verlag, 2003.

Coelho, F. e Coelho, H. – Mighty Agents for Hard Worlds, submitted in 2005.

Cohen, P. e Levesque, H. – Intention is Choice with Commitment, Artificial Intelligence Journal, 42:213-261, 1990.

Corrêa, M. – A Arquitectura de Diálogos entre Agentes Cognitivos Distribuídos, Dissertação de Doutoramento, Universidade Federal do Rio de Janeiro, 1994.

Corrêa, M. e Coelho, H. – From Mental States and Architectures to Agents' Programming, Proceedings of the 7th Congress Iberomerican on Artificial Intelligence (IBERAMIA98), Lisbon, October 6-9, Springer-Verlag, Lecture Notes in AI 1484, pp. 64-75, 1998.

Corrêa, M. e Coelho, H. – Collective Mental States in Extended Mental States Framework, in Proceedings of the International Conference on Collective Intentionality IV, Siena, Certosa di Pontignano (Italy), October 13-15, 2004.

Corrêa, M. e Coelho, H. – Deconstructing the Mind, Paper Draft, 2005.

Costa, A. R. e Dimuro, G. P. – Drives and the Functional Foudation of Agent Autonomy, Working Report, UC de Pelotas, 2003.

Damásio, A. – Looking for Spinoza, Joy, Sorrow and the Feeling Brain, William Heinemann, 2003.

David, N., Sichman, J. e Coelho, H. – Models and Experiments with Dependence Based Structures in Social Reasoning, Proceedings of the Second Workshop on Multi Agent Based Simulation (MABS2000), Boston, 9 July, 2000, S. Moss e P. Davidsson (eds.), Lecture Notes in AI 1979, Springer-Verlag, 2001.

Debord, G. – A Sociedade do Espectáculo, Mobilis in mobile, 1991.

Deleuze, G. – Pourparlers 1972-1990, Les Éditions de Minuit, 2003.

Deleuze, G. e Parnet, C. – Dialogues, Flammarion, 1996.

Dennett, D. – Freedom Evolves, Penguin Books, 2003.

Devlin, K. – Logic and Information, Cambridge University Press, 1991.

Espinosa, B. – Oeuvres I, Court Traité, Garnier-Flamarion, 1964.

Ferreira, M. L. – Uma Suprema alegria: Escritos sobre Espinosa, Quarteto, 2003.

Genesereth, M. e Nilsson, N. – Logical Foundations of Artificial Intelligence, Morgan Kaufmann, 1987.

Grice, H. – Meaning, Philosophical Review, 66, pp. 377-388, 1957.

Hardt, M. e Negri, A. – Empire, Harvard University Press, 2000.

Henke, J. M. e Bassler, B. L. – Three Parallel Quorum Sensing Systems Regulate Gene Expression in Vibrio harveyi, Journal of Bacteriology, 186, pp. 6902-6914, 2004.

Hobbes, T. – The Leviathan, 1660
http://oregonstate.edu/instruct/phl302/texts/hobbes/leviathan-contents.html.

Hume, D. – An Enquiry Concerning Human Understanding, 1748, http://www.hti.umich.edu/cgi/t/text/pageviewer idx?sid=2ef962d8ec90d6a117157c32072abaf1&idno= aje6344 0004.001&c=moa&seq=00000013; Investigação sobre o Entendimento Humano, Edições 70, Textos Filosóficos, 1989.

Ingrand, F., Georgeff, M. e Rao, A. – An Architecture for Real-time Reasoning and System Control, IEEE Intelligent Systems Journal, Vol.7, No. 6, December, pp. 34-44, 1992.

James, W. – Psychology, The Briefer Course, Henry Holt and Co., 1892; University of Notre Dame, 1985.

Kant, I. – Crítica da Razão Prática, Edições 70, Textos Filosóficos, 1986.

Locke, J. – An Essay Concerning Human Understanding, 1690, http://socserv2.socsi.mcmaster.ca/~econ/ugcm/3113/locke/Essay.htm.

McCarthy, J. – Ascribing Mental Qualities to Machines, Report STAN-CS-79-725, Computer Science Department, Stanford University, 1979.

McCarthy, J. – Free Will, Even for Robots, Working Report, Stanford University, 2000.

McCarthy, J. – Deterministic Free Will, Working Report, Stanford University, 2003.

Minsky, M. – The Society of Mind, Simon and Schuster, 1986.

Pörn, I. – The Logic of Power, Basil Blackwell, 1970.

Sims, K. – Evolving Virtual Creatures, Computer Graphics (Siggraph '94 Proceedings), pp.15-22, July 1994.

Urbano, P. – Jogos Descentralizados de Consenso ou de Consenso em Consenso, Dissertação de Doutoramento, Universidade de Lisboa, 2004.

Wooldridge, M. – Reasoning about Rational Agents, The MIT Press, 2001.

Wooldridge, M. – An Introduction to MultiAgent Systems, John Wiley, 2002.

Apêndice: Tipos de Agentes

Agente Reactivo

```
1.
2.      B:=Bo; /* Bo são crenças iniciais*/
3.      I:=Io; /* Io são intenções iniciais */
4.      while verdade do
5.              get next percepção P;
6.              B:=brf(B,P);
7.              D:=opções(B,I);
8.              I:=filtro(B,D,I);
9.              Pi:=plano(B,I);
10.             while not vazio(Pi) do
11.                     Alfa:=cabeça(Pi);
12.                     execute(Alfa);
13.                     Pi:=cauda(Pi);
14.                     get next percepção P;
15.                     B:=brf(B,P);
16.                     if not correcto(Pi,I,B) then
17.                             Pi:=plano(B,I)
18.                     end-if
19.             end-while
20.     end-while
```

Agente que Deixa Cair Intenções

```
1.
2.      B:=Bo; /* Bo são crenças iniciais*/
3.      I:=Io; /* Io são intenções iniciais */
4.      while verdade do
5.              get next percepção P;
6.              B:=brf(B,P);
7.              D:=opções(B,I);
8.              I:=filtro(B,D,I);
9.              Pi:=plano(B,I);
10.             while not (vazio(Pi) or sucedido(I,B) or impossivel(I,B)) do
11.                     Alfa:=cabeça(Pi);
12.                     execute(Alfa);
13.                     Pi:=cauda(Pi);
14.                     get next percepção P;
15.                     B:=brf(B,P);
16.                     if not correcto(Pi,I,B) then
17.                             Pi:=plano(B,I)
18.                     end-if
19.             end-while
20.     end-while
```

Agente Prudente

```
1.
2.      B:=Bo; /* Bo são crenças iniciais*/
3.      I:=Io; /* Io são intenções iniciais */
```

```
4.      while verdade do
5.              get next percepção P;
6.              B:=brf(B,P);
7.              D:=opções(B,I);
8.              I:=filtro(B,D,I);
9.              Pi:=plano(B,I);
10.             while not (vazio(Pi) or sucedido(I,B) or impossivel(I,B)) do
11.                     Alfa:=cabeça(Pi);
12.                     execute(Alfa);
13.                     Pi:=cauda(Pi);
14.                     get next percepção P;
15.                     B:=brf(B,P);
16.                     D:=opções(B,I);
17.                     I:=filtro(B,D,I);
18.                     if not correcto(Pi,I,B) then
19.                             Pi:=plano(B,I)
20.                     end-if
21.             end-while
end-while
```

Agente entre Audacioso e Prudente

```
1.
2.      B:=Bo; /* Bo são crenças iniciais*/
3.      I:=Io; /* Io são intenções iniciais */
4.      while verdade do
5.              get next percepção P;
6.              B:=brf(B,P);
7.              D:=opções(B,I);
8.              I:=filtro(B,D,I);
9.              Pi:=plano(B,I);
10.             while not (vazio(Pi) or sucedido(I,B) or impossivel(I,B)) do
11.                     alfa:=cabeça(Pi);
12.                     execute(alfa);
13.                     Pi:=cauda(Pi);
14.                     get next percepção P;
15.                     B:=brf(B,P);
16.                     if  reconsidere(I,B) then
17.                             D:=opções(B,I);
18.                             I:=filtro(B,D,I);
19.                     end-if
20.                     if not correcto(Pi,I,B) then
21.                             Pi:=plano(B,I)
22.                     end-if
23.             end-while
24.     end-while
```

4
Fiction-Making as a Gricean Illocutionary Type[1]

Manuel Garcia-Carpintero
Universidat de Barcelona
m.garciacarpintero@~~mat.~~ub.edu~~s~~

1. Preamble

There are propositions constituting the content of fictions – sometimes ~~the~~ of the outmost importance~~t~~ ~~ones~~to understand them – which are not explicitly presented ~~in them~~, but must somehow be inferred. This paper deals with what these inferences tell us about the nature of fiction. I will criticize three well-known proposals in the literature, by Lewis (1978/83), Currie (1990) and Walton (1990). I will advocate a proposal of my own, which I will claim improves on theirs. Most important for my purposes, I will argue on this basis, against Walton's objections, for an illocutionary-act account of fiction, inspired in part by some of Lewis' and Currie's suggestions, but (perhaps paradoxically) above all by Walton's deservedly influential views.

I start with a quotation from a story by the Argentinian writer Julio Cortázar, short enough to be given in full; it will provide a crucial example to help present~~ing~~ my argument:

> He had begun to read the novel a few days before. He had put it down because of some urgent business conferences, opened it again on his way back to the estate by train; he permitted himself a slowly growing interest in the plot, in the characterization. That afternoon, after writing a letter giving his power of attorney and discussing a matter of joint ownership with the manager of his estate, he returned to the book in the tranquility of his study which looked out upon the park with its oaks. Sprawled on his favorite armchair, its back toward the door – even the possibility of an intrusion would have irritated him, had he thought of it – he let his left hand caress repeatedly the green velvet upholstery

1 This work, as part of the European Science Foundation EUROCORES Programme OMLL, was supported by funds from the Spanish Government's grant DGI BFF2002-10164 and the EC Sixth Framework Programme under Contract no. ERAS-CT-2003-980409, from DGI HUM2004-05609-C02-01, DURSI, Generalitat de Catalunya, SGR01-0018, and a Distinció de Recerca de la Generalitat, Investigadors Reconeguts 2002-2008. José Díez and Esther Romero suggested revisions on a previous version ~~which~~ that led to improvements. The paper was presented at the Seminário de Filosofia Analítica, Universidade de Lisboa, ~~and~~ at the II Barcelona-Milano Philosophy Meeting held in Barcelona and at the Facoltá di Lettere e Filosofia, Universitá del Piemonte Orientale a Vercelli. I thank the audiences there for helpful criticism, in particular to Sandro Zucchi, who was commentator at the second occasion. I am also indebted to Michael Maudsley for his grammatical revision.

and set to reading the final chapters. He remembered effortlessly the names and his mental image of the characters; the novel spread its glamour over him almost at once. He tasted the almost perverse pleasure of disengaging himself line by line from the things around him, and at the same time feeling his head rest comfortably on the green velvet of the chair with its high back, sensing that the cigarettes rested within reach of his hands, that beyond the great windows the air of afternoon danced under the oak trees in the park. Word by word, licked up by the sordid dilemma of the hero and heroine, letting himself be absorbed to the point where the images settled down and took on color and movement, he was witness to the final encounter in the mountain cabin. The woman arrived first, apprehensive; now the lover came, his face cut by the backlash of a branch. Admirably, she stanched the blood with her kisses, but he rebuffed her caresses, he had not come to perform again the ceremonies of a secret passion, protected by a world of dry leaves and furtive paths through the forest. The dagger warmed itself against his chest, and underneath liberty pounded, hidden close. A lustful, panting dialogue raced down the pages like a rivulet of snakes, and one felt it had all been decided from eternity. Even to those caresses which writhed about the lover's body, as though wishing to keep him there, to dissuade him from it; they sketched abominably the frame of that other body it was necessary to destroy. Nothing had been forgotten: alibis, unforeseen hazards, possible mistakes. From this hour on, each instant had its use minutely assigned. The cold-blooded, twice-gone-over reexamination of the details was barely broken off so that a hand could caress a cheek. It was beginning to get dark.

Not looking at one another now, rigidly fixed upon the task which awaited them, they separated at the cabin door. She was to follow the trail that lead north. On the path leading in the opposite direction, he turned for a moment to watch her running, her hair loosened and flying. He ran in turn, crouching among the trees and hedges until, in the yellowish fog of dusk, he could distinguish the avenue of trees which led up to the house. The dogs were not supposed to bark, they did not bark. The estate manager would not be there at this hour, and he was not there. The woman's words reached him over the thudding of blood in his ears: first a blue chamber, then a hall, then a carpeted stairway. At the top, two doors. No one in the first room, no one in the second.

> The door of the salon, and then, the knife in hand, the light from the great windows, the high back of an armchair covered in green velvet, the head of the man in the chair reading a novel.
>
> JULIO CORTÁZAR, A Continuity of Parks, *The End of the Game* (1956)

Consider an utterance of (1) below by Cortázar, as part of the longer utterance by him of the full discourse that, with a measure of idealization, we can think constitutes "A Continuity of Parks" (*ACP*, for short, henceforth). (This is of course itself part of the idealization; we should really be speaking of an utterance of the Spanish sentence 'había empezado a leer la novela unos días antes', the actual part of the story created by Cortázar and published by him in his 1956 collection *Final del juego*.)

~~(1)~~(1) He had begun to read the novel a few days before.

Formatadas: Marcas e numeração

(1) is in the declarative mood, which by default expresses ~~in English~~ assertion in English. Nonetheless, most accounts of fiction would follow Plantinga's (1974, 161) view that the author of a fiction "does not assert the propositions that form his stock in trade", and hence would not count such an utterance as assertoric in illocutionary force at all: the context in which it occurs overrides the default interpretation for (1)'s mood.[2]

Just for the sake of having a specific ~~account~~ proposal about assertion in mind, let us assume what I in any case take to be a plausible account of what such an interpretation amounts to, based on Williamson's (1996/2000).[3] According to this view, the following norm or rule (the *knowledge-transmission rule*) is constitutive of assertion, and individuates it:

> (KTR) One must ((assert that *p*) only if one puts thereby one's intended audience in a position to know that *p*)

By default, the declarative mood indicates that utterances of sentences in that mood are subject to that norm (which, of course, does not mean that they fulfill the obligation that it imposes). By uttering (1) in the context in which he did, Cortázar makes it clear that he is not doing something that commits him to KTR. Let us use the verb 'to fictionalize' to refer to what he is alternatively doing. A fictionalizing context overrides the default relative to which the declarative mood is interpreted, and

2 Exceptions include followers of the views by Goodman to be mentioned later.

3 García-Carpintero (2003) defends this version of Williamson's account.

therefore Cortázar is not committed to KTR. Presumably he lacks knowledge of any proposition that an utterance of (1) in an otherwise normal context would express, and is therefore unable to put anybody in a position to acquire such knowledge; but he is not thereby violating a norm.

This paper is about what *fictionalizing* is, about what Cortázar is alternatively doing in uttering the discourse of which (1) is part. I want to defend the view – rejected, among others, by Walton (1990) – that fiction ~~-mak~~alizing is a type of speech-act, like promising or voting, a ~~n~~n illocutionary type, ~~like promising or voting~~ understood along Gricean lines. Once again for the sake of having a specific ~~account~~ proposal in mind, I will adopt Currie's (1990) ~~Gricean illocutionary type~~ analysis of fictionalizing, or fiction-making; on this view, to fiction-make a proposition by uttering something (or by painting, or by having people acting on a stage, etc) is to so utter with the communicative intention to put an intended audience in a position to make believe (imagine) that proposition. ~~More in~~ In more detail, his proposal (1990, 33) is this: *U*'s utterance of *S* is fictive iff there is a Φ and there is a χ such that *U* utters *S* intending that anyone who has χ would: (1) recognize that *S* has Φ;[4] (2) recognize that *S* is intended by *U* to have Φ; (3) recognize that *U* intends them (those ~~possessors~~ who ~~of~~have χ) to make believe that *P*, for some proposition *P*; (4) make believe that *P*; (5) take (2) as their reason for (3); (6) take (3) as their reason for (4).

The main objection to such accounts is that they incur ~~in~~ some form of the "intentional fallacy", which, ~~;~~ in their famous manifesto purportedly exposing it, Beardsley & Wimsatt (1954, 4) characterized as the view that "in order to judge the poet's performance, we must know what he intended," as against which they argue that "the design or intention of the author is neither available nor desirable as a standard for judging the success of a work of literary art" (*ibid.*, 3). ~~–~~I will ~~say something about~~come back to this at the end. The core of the paper is ~~to elaborate~~ an argument that ~~our~~ nonnegotiable intuitions about what propositions ~~obviously~~ constitute the content of a given fiction ~~can only be knowledge if~~are best accommodated if fiction ~~-mak~~alizing ~~is~~ is an illocutionary type.

2. The Content of Fictions

4 The variables χ and Φ are intended to pick up, respectively, features charactering the intended audience (so that it has whatever is required to understand the speaker's intentions) and features of the utterance accessible to such an audience, such as its having a certain conventional meaning.

~~To~~ I will start developing the argument~~, let us~~ by considering a different speech-act that one could make in uttering ~~it~~(1), related to Cortázar's story. One who is familiar with the story could utter ~~(1)~~it in the context of telling someone else, or otherwise discussing, the content of the story, its plot, what goes on in it, for instance by uttering (1) after saying 'the story is about someone who reads a novel'. In such a context, the utterance does constitute an assertion, one moreover that appears to satisfy KTR and is therefore true. For this to be so, it must express a proposition; and there is an obvious problem here: what is the contribution of the referential expressions in the utterance, 'he', 'the novel', the implicit indexical governed by 'before'? Neo-Meinongian views reply that the referents of those expressions are fictional characters, like the ones that are explicitly referred to in utterances like (2):

~~(2)~~(2) The man who reads a novel in ACP is the sort of character with which any reader immediately identifies.

Formatadas: Marcas e numeração

Van Inwagen (1977) argued that an acceptable semantic account of the content of assertions like (2) requires an ontology of "creatures of fiction", fictional characters that can be referred to by singular terms like the definite description in it. Given that one accepts his arguments, a similar neo-Meinongian treatment is available for the sort of assertion of (1) one makes in stating the content of the story~~the sort of assertion of (1) we were considering before, and this is what neo-Meinongian accounts suggest~~s.

There are well-known problems with this proposal.[5] Consider uses of (3) and (4) intended to make assertions about the story's plot, analogous to the one discussed before for (1), in a similar context introduced by 'the story is about someone who reads a novel'.

~~(3)~~(3) He was born in Patagonia.

~~(4)~~(4) He wasn't born in Patagonia.

Formatadas: Marcas e numeração

Both assertions appear to be false, because Cortázar has not given us in the story any indications one way or the other. However, to the extent that, as neo-Meinongian ~~these~~ proposals require, (3) and (4) have the logical forms that they apparently have – a property is predicated of an object in (3), and the same property is denied of the same object in (4) – this appears to violate a logical law, $\forall x(\mathrm{P}(x) \vee \neg \mathrm{P}(x))$.

5 Thomasson (1999), 100-5 is a fuller short discussion.

There are ~~different~~ several ~~suggest~~ways~~ions~~ to deal with this problem~~, on behalf of~~according to neo-Meinongian accounts. One~~; a~~ possibility, advanced by van Inwagen (1977), is to deny that the copula expresses predication in (3) and (4), contending instead that it expresses a relation in which a fictional character stands to a property when the property is ascribed to the character in a certain fiction. But this suggestion ~~makes~~ closely approaches the proposal ~~all but a convoluted variant of~~to a simpler well-known view, developed among others by David Lewis (1978/83).[6] According to this proposal, in the logical form of the relevant assertions of (1), (3) and (4) there is an implicit operator, 'ACP makes it fictional that …', which behaves in closely similar ways to operators that have been ~~very much~~ studied in depth in contemporary semantics, like 'S believes that …'. To the extent that we can invoke a semantic account of the significance of referential expressions like 'he' when they occur in contexts governed by those operators on which they do not necessarily contribute their ordinary referents outside them, we avoid any problems caused by their lacking those referents. (We could still grant van Inwagen's view that assertions like (2) do have their apparent logical forms, their singular terms genuinely referring to creatures of fiction.) And there is no problem with both assertions (3) and (4) being false, when understood as suggested: like most belief-systems concerning many propositions and their negations, ACP is noncommittal on the matter of the reader's origins in Patagonia.

Let us henceforth use '$\mathbf{F}_{\mathbf{acp}}(p)$' as an abbreviation of 'ACP makes it fictional that *p*', and consider now an assertion of the following sentences, uttered again in a context discussing ACP's plot introduced by 'the story is about someone who reads a novel':

(5) $\mathbf{F}_{\mathbf{acp}}$(he is killed)

~~(5)~~**(6)** $\mathbf{F}_{\mathbf{acp}}$(~~a man~~he reads a novel about what is in fact the scheme of his wife and her lover to kill him, whose denouement is about to happen ~~while~~ as he reads ~~unsuspecting~~ about it unsuspectingly)

The proposition that ~~(5)~~(6) claims to be fictional in Cortázar's story is not just actually fictional there; it is *the* fictional truth in the story, the main one. (For there is an order of importance among the ~~classes of~~ propositions that are fictional in a given fiction, to which we are sensitive when summing up the plot of a film or a novel.) A ~~baffled~~

Formatadas: Marcas e numeração

6 I am sympathetic to Zalta's (2000) suggestion that his preferred neo-Meinongian account, and the sort of proposal I will be endorsing, can be seen as mere notational variants.

puzzled child who misses ~~(5)~~(6) has not understood the story. We will consider presently direct evidence for (5) and (6) in the text; ~~T~~a~~h~~s indirect evidence for the truth of (6)~~us~~, there is firstly the fact that~~the truth of (5)~~ it allows us to make sense of the story's title: it is the parks in the novel (the reader reads that the lover-killer "could distinguish the avenue of trees which led up to the house") that are ~~continuations~~continuous ~~of~~with~~ous~~, ~~and~~ indeed identical to, ~~with~~ the parks with oaks upon which the reader's study looks out. ~~And~~ Secondly, it allows us to appreciate a point that Cortázar might well be trying to convey by telling us this story:

~~(6)~~(7) There might be fictions such that all ~~propositions~~ propositions that are fictional ~~propositions~~ in~~side~~ them are actually true

Formatadas: Marcas e numeração

There is nothing fictional about ~~(6)~~(7); ~~(6)~~(7) conveys a plain assertion, in fact a philosophical claim about fictions that would be rejected by ~~many~~ some philosophers ~~that~~ who have dealt with these matters – including, as we will presently see, David Lewis~~, including David Lewis~~. ~~I will come back to this presently.~~ For now ~~on~~, I will focus on trying to state in virtue of ~~which~~ what ~~is~~ ~~(5)~~(6) is true. I will argue, mostly against Walton (1990), that only an illocutionary-type account of fictionalizing can properly supply an answer.

Together with many other writers, Walton (1990) distinguishes facts about what is fictional in a given fiction that are somehow explicitly there, from others that are merely implicitly there. The proposition that ~~(7)~~(8) correctly claims to be fictional in ACP is not put in so many words in the story; it is something we infer from what we are explicitly told.[7]

~~(7)~~**(8)** $\mathbf{F_{acp}}$(the hero and heroine in the novel plan to kill the heroine's husband)

Formatadas: Marcas e numeração

I will follow Walton in characterizing the distinction as a contrast between the directly and the indirectly generated (or implied) facts; the latter are generated indirectly in that the propositions they correctly state to be fictional in the given work are derived in part on the basis of others, previously determined to be fictional in that work. Those facts about what is fictional in a given fiction that contribute to determine others, without being themselves determined on the basis of others, are the directly generated ones.[8] A fundamental problem in giving the truth-conditions of

7 Our evidence: their dilemma is "sordid", their passion "secret"; the heroine's caresses "sketched abominably" for the hero "the frame of that other body it was necessary to destroy"; she is familiar with the house, which she has described in detail to her lover.

8 As ~~it~~ will become clear later, a given ~~fact~~ claim about what is fictional in a given fiction ~~is~~ only defeasibly~~y~~ states a directly generated ~~in this sense~~fact; considerations dependent on other propositions being fictional in the given fiction can defeat any claim that a proposition is fictional in it.

claims like ~~(5)~~(6) is to characterize what Walton calls 'principles of generation', the principles relative to which indirectly generated facts are determined.

3. Lewis' Account: Preliminaries

Lewis (1978/83) gives the truth conditions of claims like (5), (6) and (8) inside his well-known possible-worlds semantic framework~~Lewis (1978/83) resorts to the possible-worlds semantic framework to give the truth conditions of claims like 8(5) and (7)~~. As a first approximation, he contemplates to consider "exactly those worlds where the plot of the fiction is enacted, where a course of events takes place that matches the story. What is true in the Sherlock Holmes stories would then be what is true at all of those possible worlds" (*op. cit.*, 264). However, he rejects this suggestion, particularly because of the following problem. Let us assume that Conan Doyle wrote the Sherlock Holmes stories as pure fiction; in particular he had no knowledge of anyone who did the deeds he ascribed to Holmes. Still, it might be that the actual world is one of the worlds where the plot of the Conan Doyle stories is enacted. ("Improbable, incredible, but surely possible!" (*op. cit.*, 265).) Then we have that it is false in the actual world that the name 'Sherlock Holmes', as used in the stories, refers to someone; however, it is true in the stories that the name 'Sherlock Holmes' refer to someone; so there is something that is true in the stories, but false in one of the worlds where their plot is enacted.

~~For reasons of his own that I lack the space to discuss here,~~To deal with this problem, Lewis, like Searle (1974/9), adopts a *pretence theory* of the act of fictionalizing: "Storytelling is pretence. The storyteller purports to be telling the truth about matters whereof he has knowledge. He purports to be talking about characters who are known to him, and whom he refers to, typically, by means of their ordinary proper names. But if his story is fiction, he is not really doing these things" (*op. cit.*, 266). Given this, to use the possible worlds frameworks to analyze the truth-conditions of claims like ~~(5)~~(6), the worlds to consider "are the worlds where the fiction is told, but as known fact rather than fiction. The act of storytelling occurs, just as it does here at our world; but there it *is* what here it falsely purports to be: truth-telling about matters whereof the teller has knowledge" (*ibid.*)

This is, by the way, why Lewis is, as I said before, one of those philosophers who would reject~~It is relative to these assumptions that Lewis~~ ~~rejects~~ the claim ~~(6)~~(7) ~~that~~ which I said Cortázar might be intending to convey with ACP: "Our world cannot

be such a world; for if it is really a fiction that we are dealing with, then the act of storytelling at our world was not what it purported to be" (*ibid.*). Any act of storytelling occurring in the actual world purports to be a (protracted) assertion, complying with something like what I earlier took to be its constitutive norm, KTR; there are possible worlds in which the very same act (or an appropriate epistemic counterpart of it) is such a thing, but the actual world (Lewis submits) cannot be one of them, because there the act is a mere pretence of a normatively correct assertion. On this view, claims like (9) below are always true statements about the contents of stories. (I have replaced Lewis' *telling as known fact* by, simply, *asserting*, given the previous explication of that speech act.)

(9) $\mathbf{F}_{acp}$(its utterer asserts that someone had begun to read a novel)

On Lewis' account, the embedded propositions in claims like (9) are true in all the relevant worlds we need to properly characterize the content of fictions, and therefore are part of their contents; and this aspect of their content is, according to him, always false of the actual world. This is how he intends to deal with the problem posed by the possibility that the plot of the Holmes stories is enacted in actuality. It has the consequence that (7) is false, on Lewis' view; at least propositions like the one embedded in (9) are a false part of the content of every fiction. And it has also the consequence that, on Lewis' view, Cortázar's story is one of those modernist fictions with an impossible content. Because, if I interpret it correctly, the main point of the story, as captured by the truth of (5) and (6), requires that the very world in which the main character is reading a novel is one of those worlds where the full content of that novel is enacted; and, on Lewis' philosophical account of fiction, that can never be the case. (5) and (6) become on this view questionable; they pose, in a very sharp form, the problem that possible-worlds accounts have with conceptually incoherent fictions.

Let me emphasize that this is not the real problem posed by the derivation of (6) for Lewis' account that I will be mostly concerned with. Lewis could maintain the features of the view he is committed to that we have considered so far, rejecting (7), and still appeal to some of the procedures he discusses to deal with conceptually impossible fictions so as to account for the truth of (5) and (6). However, let me briefly depart from the main course of my argument to say why I think this that this aspect of his view is an unpersuasivenot convincing argument, because its discussion

will help later to appreciate the real difficulties for Lewis' (and ~~Currie's~~others') proposals to account for the generation of ~~(5)~~(6).

Even if the pretence theory of fictionalizing is correct (later I will reject it as a fully satisfactory account, but I will grant that there is something to it), someone who pretends to correctly assert that *p* may still be correctly asserting that *p*; indeed, he might be correctly asserting that *p by pretending* to assert that *p*. We can, I think, coherently imagine that the story that Cortázar tells us actually obtains. Perhaps the reader's estate manager caught pieces of conversations between the reader's wife and her lover, and surmised their conspiracy; uncertain about the response that direct exposure to his suspicions would provoke in his employer, he wrote a novel under a pseudonym and made sure that h~~is employer~~e read it, hoping (to no avail) that the details given in it would lead ~~him the latter~~him to recognize the author's assertoric communicative intentions and its implications. Improbable, incredible, but surely possible! ~~However, if Lewis is right, given that his is a conceptual claim, this should be impossible. Against~~If this is so, the proposition that ~~(5)~~(6) state~~ay~~s is fictional in ACP is possible; it is possible that the world according to ACP, in which the reader reads the novel, is one of the worlds in which the content of that novel obtains, so as to make (5~~2~~) true.[9] Cortázar might well be correct in his implied claim ~~(6)~~(7) – if that is what it is. This disposes of accounts of fictionalizing like Goodman's, much more heavily committed – given their philosophical ambitions – than Lewis' to the view that some ~~at least~~ of the declarative sentences giving a fiction's content should be untrue, literally taken. But I think that nothing of substance would change in Lewis' views is we modified them in response to the previous objection.[10]

4. The Problem with Lewis' Account

As I said, however, this is not the main problem I want to discuss; as far as the core claims in the paper are concerned, we could assume that my criticism is misguided.[11] Let us go back to Lewis' account of the truth conditions of claims like

9 In fact, one could~~d~~ take the possibility of the story to be ~~implied~~ implicitly conceded by what Lewis (1978/83, 278-9) grants in postscript C to the originally published version of the paper, "Fiction in the service of truth."

10 To deal with the ~~Kripke objection~~problem that we saw worries him (the incredible possibility that the plot of the Holmes' stories obtains in the actual world) ~~(op. cit., 265-7)~~, Lewis ~~merely~~ really only needs that ~~in some cases (including the Holmes' case that Kripke apparently raised), not necessarily in all, in the actual world the teller of the story merely pretends to assert and to refe~~claims like (9) state facts (directly or indirectly generated) about what is fictional in stories~~r~~.

11 Currie's account of fictionalizing includes as a necessary condition that fictions are not non-accidentally true, to deal with what he takes to be counterexamples to an account without it (op. cit., 42-49). If this is correct, the novel that the main character in Cortázar's story reads can at most be an apparent novel, in the counterexample that I have described to criticize Lewis. The counterexample can be modified to deal with this; perhaps the state

(5), ~~(5)~~(6) and ~~(7)~~(8). A first stab ~~towards~~ at a possible-worlds account already suggested by the pretence theory of fictionalizing is this: A sentence of the form 'In the fiction F, ϕ ' is true if and only if ϕ is true at every world where F is told as known fact rather than fiction. But this proposal would only allow for the fictionality of what is explicitly stated in fictions. It is compatible with Cortázar's story being told as known fact rather than fiction that it is not the heroine's husband that the hero in the novel ~~that~~ being read by the main character ~~reads~~ purports to kill, but anybody whose house the woman can correctly describe. To allow for indirectly generated fictional truths, Lewis offers two different analyses; the two capture some of the principles guiding our inferences towards what is implicit in fictions. I will only consider the second one, which ~~of the two is~~, even if both are ultimately similarly unsuccessful, is ~~the best~~better suited ~~candidate~~ to dealing with the problem I will raise:

> (L) A sentence of the form 'In the fiction F, ϕ ' is true if and only if, for any collective belief world w of the community of origin of F, there is some world where F is told as known fact and ϕ is true which is more similar to w than any world where F is told as known fact and ϕ is false.

Assuming Lewis' well-known analysis of counterfactuals, (L), more simply put, has it that 'In the fiction F, ϕ ' is true if and only if ϕ would be true if F were told as known fact and the beliefs constituting common knowledge in the community where F originated were also known fact.[12] We can fairly assume that it is commonly believed in the community where ACP originated that, if a love affair between a man and a woman is kept secret, if it is felt to be sordid, if jealous thoughts of another man are evoked in ~~the ma~~him~~n~~ by the woman's caresses, if all this leads to a murderous conspiracy for which the woman provides crucial information, the third person ~~at stake~~involved is (typically, at least) the woman's husband. On these grounds, we can grant that (L) accounts for the generation of the implicit fictional truth ~~(7)~~(8), that it is the heroine's husband that the hero in the fictional novel purports to kill.[13]

---

manager is paranoid, did not have any evidence and made up the conspiracy, did have the intention to warn his boss by writing the novel, and came upon the truth by accident.

12 'Common knowledge' is used for the concept introduced by Schiffer, Lewis and others so as to provide Gricean analyses of meaning, convention and related notions.

13 As a matter of fact, this is not so clear, as Phillips (199~~9~~, 279-81) correctly points out. If we were told as known fact a story like the one in the novel in ACP, we would infer at most that it is probable that the victim is the husband, while when we are told of it in the fiction, we are certain ~~about~~ of this. But let us grant it, for the sake of illustrating how (L) is supposed to work. Later I will show how my proposal avoids the difficulty.

Lewis' analysis (L) appeals to the beliefs that are common knowledge in the community of origin of the fiction to account for inferences to what is implicit in it from what is explicit. His alternative analysis, which I have not given, appeals instead to what obtains in fact in the actual world. It should be obvious that ~~none of these~~neither suggestion~~s~~ can account for the generation of (5) and ~~(5)~~(6), as I said ~~before,~~ the main fact about what is fictional in ACP. I have argued before, against Lewis, that the proposition that ~~(5)~~(6) claims to be fictional in ACP is possible, that one of the worlds in which ~~the~~ all propositions fictional in the novel ~~that is~~ being read in ACP obtain might in fact be the world in which~~, as we are told,~~ the novel is read. However, if we were told ACP as known fact, and accepted it as such, no appeal to what we take to be common knowledge (now, or in Cortázar's time), even less to what is in fact the case in the actual world, would lead us to infer that the actual world is one of the worlds of the novel which we are told about. We would take it as a rather insipid narrative about someone enjoying a novel, sat in a green velvet armchair with its back toward the door of a room with ~~great~~ large windows, ending abruptly when, in the novel, a fictional character also sat in a green velvet armchair with ~~its~~ his back toward the door seems about to be murdered. Neither the coincidence in the upholstery and relative position of the armchair, nor in other~~s~~ details (the parks, the estate manager), would suffice – if the narrative is given as true assertion, and not as fiction – given what we take to be mutually known, to outbalance the enormous amount of implausibility required to identify the reader ~~with~~ as the victim in the novel that he is reading~~s~~.

So, how is it that as experienced readers we effortlessly infer ~~(5)~~(6) when, exhilarated with the increased narrative speed, we ~~assist astonished to~~come to the revelations in the final sentence of the story? Intuitively speaking, it is a matter of Gricean *relevance*, relative to the aims we ascribe to the teller of a story. The main piece of evidence has to do with that ending, when put in the context of the piece of discourse in which it is supposed to belong. This discourse is supposed to be a piece of narrative, to tell us a meaningful story; and stories have a peculiar kind of explanatory structure (the story typically highlights an event, the denouement, and disposes others to account for it in appropriate ways) which would be missing, unless we generate (5) and (6).[14] ~~Roughly speaking, it is a matter of *relevance*, relative to the aims we ascribe to the teller of a story. Firstly, o~~Only the generation of (5) and ~~(5)~~(6)

makes ~~literary~~ narrative sense, and, of course, the coincidences just mentioned~~, rounded off in that final sentence~~support them. Then there are the indirect evidences we already mentioned~~;~~, ~~secondly, it~~ the account~~s~~ ~~for~~ of the title; and, last but not least, that it allows for ~~(6)~~(7) to be, so to say, the moral of the story.[15]

Walton (1990, 161-183) notices ~~this~~ the importance of this sort of relevance on the basis of several related examples: "We know what creators of representations are up to, that a large part of their job is to make propositions fictional. When an artist has arranged for a work to generate fictional truths that in one way or another call attention to some further proposition, it is often apparent that his reason for doing so was to make this proposition fictional as well. There is likely to be an understanding approximately to the effect that when this appears to have been the artist's objective, the salient proposition is fictional, its fictionality being implied by the fictional truths that call attention to it" (*ibid*., 166). But I feel that he is not sensitive to the full implications of this. For, what does it tell us about the truth conditions of claims like ~~(5)~~(6), and ultimately about the nature of fictionalizing?

It tells us, I suggest, that we take fictionalizing to be a type of speech act, an illocutionary type understood along Gricean lines. Relevance figures prominently in Grice's (1975) well-known account of conversational implicature, as the only maxim in his third ~~sub~~ category~~maxim~~. The account depends on Grice's theory of speaker's meaning, developed in a series of influential papers. Speaker's meaning, according to Grice, is meaning resulting from a type of rational activity guided by a *communicative intention*; roughly, the (indexical self-reflexive) intention of rationally inducing an audience to ~~rationally~~ form specific mental states, on the basis of the recognition of that very intention.[16] Variations on the types of mental states to be formed, and on the expected rational procedures for the audience to be guided into them, account for differences in type of illocutionary force. Granted a number of (philosophically substantive) presuppositions, some of them concerning the very interpretation of Grice's proposals, the account of assertion based on KTR suggested ~~before~~ above will count as a Gricean one, for a fundamental type of illocutionary force.[17]

14 See Velleman (2003) for an interesting account of the peculiar explanatory force uniting narratives, and related references.

15 It is this very same sort of relevance that is required for a full account of the generation of the implicit fictional truth considered in the previous but last footnote, that it is not just probable but the (fictional) case that it is the heroine's husband that the hero in the fictional novel purports to kill. This ~~will~~ would be explained by the elaboration of this idea to be provided in what follows, in terms of an illocutionary-act theory of fictionalizing.

16 Bach (1987) advocates such an indexical self-referential analysis of communicative intentions. For an illuminating recent account of the nature of those states, applied both to communicative intentions and to common knowledge, see Peacocke (forthcoming).

17 García-Carpintero (2001) and (2003) elaborates on some of those presuppositions.

The Gricean account of speaker's meaning features a sign, the meaning-vehicle, which is not just a particular token, but consists of the instantiation of some recognizable properties; and it involves not just any non-descript audience, but audiences with specific properties.[18] In literal communication, speakers convey meaning by producing signs instantiating types that conventionally have certain meanings, for the sake of audiences that share their knowledge with the speaker. In conversational implicature, speaker's meaning is typically conveyed by ~~producing~~ uttering signs with literal meanings such that, given certain assumptions taken to be common knowledge with the intended audience, the utterance would blatantly violate the conversational maxims~~, because~~ ; this ~~will~~ leads the audience to what is meant, as the only sensible way of ~~restoring the expectation of respect for~~having the utterer conforming to the maxims.

Grice's explicit theory of implicature does not cover the full scope of the phenomenon. The maxims or sub-maxims that he specifies are adequate for speech-acts that can be properly evaluated as true or false, of which assertion is the fundamental one. The account works well for Grice's (1975) examples, where an utterance which, if made literally, would assert that *p*, conversationally implicates something else. However, there are many cases (including indirect speech acts) for which the general Gricean framework supplies a correct account, but which do not fit this schema; thus, for instance, when, by uttering 'Could you pass the salt?' the speaker does not express the literal question he is asking (~~whose~~ the answer to which he knows very well), but a request.[19] The sub-maxims of quantity and quality do not apply in those cases, because if the utterance is taken literally no amount of information is supposed to be given, nor is there any knowledge that the speaker intends to convey. The ~~sub~~ maxim~~s~~ of relation ~~and manner~~ appears to work better with these cases; as Levinson (2000, 17) argues, this is because Gricean relevance~~, say,~~ is a determinable principle of attending to interlocutors' goals or plans that should be further determined relative to the types of speech act setting up the specific goals at stake. What counts as relevance when it is an act governed by KTR that ~~would be~~is made, were the utterance taken literally, differs from what counts as such

18 Currie (1990, 30-5) gives references and provides some of these reasons, which I cannot but illustrate here, for the particular case of a Gricean account of fictionalizing.

19 Searle's (1975) account is in the spirit of Grice's. The relation between plain assertions like (~~7~~6), and the fiction-making ~~(Cortázar's utterance of ACP, in this case)~~ giving rise to them (Cortázar's utterance of ACP, in this case) is in my view essentially of this very kind, a ~~form~~ kind of conversational implicature properly extended beyond Grice's (1975) proposal so as to cover indirect speech-acts in general. Close (1972) uses facts like this, about the way fictions manage to indirectly convey assertions and other speech acts, to argue for a speech-act account of fiction-making.

when it is a request ~~is made~~, or a question ~~asked~~. A similar point applies to the sub-maxims of manner.

Any form of activity guided by communicative intentions, not just assertion, is thus governed by Grice's (1975/89, 26) *Cooperative Principle*, "Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged", together with more specific ~~sub-~~maxims that follow from this given the specific "purpose or direction of the talk" constituted by the intended illocutionary types. The indirect generation of truths like (5) and ~~(5)~~(6) depends on specific sub-maxims derived from the Cooperative Principle for the specific case of fictionalizing, a particular illocutionary type. The distinctive nature of fiction-making does not therefore lie in the truth-properties of the ~~conveyed~~ fictional propositions conveyed, but in the norms distinguishing such an illocutionary type from others, assertion in particular.

The pretence theory of fictionalizing is a step in this direction,[20] but, as other writers have shown, not a fully satisfactory one.[21] Pretended assertions (or any other speech-acts) sometimes have nothing to do with fictionalizing: they could just be a parody of somebody's speech. And fictionalizing sometimes has no use for pretended speech-acts (as in silent movies or pictures), even though, in the case of literary fiction, the pretence of assertion (and other speech-acts) does play an important role. Pretended speech-acts are not, in general, at the core of what fictionalizing is~~t~~; pretence itself, or make-believe, is, as Walton (1990) has persuasively argued. Walton (*ibid.*, 85-9), however, extends his criticisms of the pretence theory of fictionalizing to any illocutionary-type account – to any account on which fictionalizing essentially involves communicative intentions, or any specific kind of intention on the part of the fiction-maker. Others have convincingly replied to his criticisms.[22] My argument here for illocutionary-type theories lies in that they give proper elaboration to the intuition that a principle relevance guides us to infer ~~(5)~~(6), subscribed by Walton in a text I quoted earlier, and the best account of the truth-conditions of claims like ~~(5)~~(6).

At the outset I presented Currie's (1990, 33) illocutionary-type analysis of fictionalizing, as adequate for my purposes to illustrate the kind of analysis of fiction-

20 This is why, ~~although~~ as I said in footnote 9, ~~we saw before Lewis rejects (6),~~ once Lewis ~~he~~ has adopted an illocutionary type theory he does not need to rejects (7) ~~do so,~~ – in contrast ~~with~~ to theorists like Goodman, who analyze fictionalizing in terms of the truth-properties of the conveyed propositions.

21 See the criticisms by Currie (1990, 12-18), Walton (1990, 81-5) and Lamarque & Olsen (1994, 62-9).

22 See Currie (1990, 35-42) and Lamarque & Olsen (1994, 46-9).

making which ~~I think examples like (5) supp~~I want to defend~~ort~~. However, although Currie's (1990, 8~~01~~) own analysis of the truth-conditions of claims like ~~(5)~~(6) and ~~(7)~~(8), (C), ~~is~~ does better ~~off~~ than Lewis', it still fails to account for ~~(5)~~(6):

> ~~(C)~~(6) A sentence of the form 'In the fiction F, ϕ ' is true if and only if, it is reasonable for an informed reader of the text to infer that the fictional author of F believes that ϕ.

Formatadas: Marcas e numeração

According to Currie, "what is true in the fiction is what the teller believes. But it is important to realize that the teller is himself a fictional construct, not the real live author of the work" (*op. cit.*, 75); "as readers, our make-believe is that we are reading a narrative written by a reliable, historically situated agent (the fictional author) who wants to impart certain information" (*ibid.*, 80). There are obvious problems with this proposal. It is unclear how to apply it to fictions that do not consist~~s~~ of (pretended) assertions, for instance a poem consisting only of questions. There are narratives with explicit, but unreliable narrators; ~~it is not fictional~~ what they purport to believe in these fictions ~~is not fictional; in fact it is what they purport to believe,~~ usually the opposite that it is fictional~~/?please check the opposite instead~~. To deal with the latter problem~~is~~, Currie contends that an explicit narrator is never his fictional teller; the latter tells us about the former (*op. cit.*, 124). There are fictions that preclude the possibility that they could be told as known fact, fictions about situations without intelligent life, for instance. To deal with this, he contends that th~~o~~ese narratives have contradictory contents (*ibid.*, 125-6). These replies are problematic.[23] Be this as it may, the ~~previous~~ above discussion shows that Currie's proposal cannot at any rate account for the generation of ~~(5)~~(6). If we took ACP to be a narrative written by a reliable agent who wants to impart information, and tried~~, wanting~~ to ~~infer transmit~~infer his beliefs, we would never ascribe to him the proposition embedded in ~~(5)~~(6). Currie's proposal appears to be motivated by the fear ~~to~~ of incurring ~~in~~ the intentional fallacy (*op. cit.*, 109-16); presently I will appeal to the distinction between communicative and non-communicative intentions in order to evade it.

5. A Gricean Alternative

My own proposal appeals directly to the fiction-making intentions of the author, instead of the beliefs of Currie's fictional teller. The creator of a fiction wants

23 See Phillips' (199~~8~~, 282-6) discussion.

us to imagine propositions shaping a story, so as to entertain us, to lead us to reflect on the consequences of the possibilities he thereby depicts, and so on. There is an established practice of using certain means for that goal. One of them is to pretend to use sentences of natural languages in their agreed ways, although this is by no means the only one; there are others, like having people pretending to act in certain ways on a stage, arranging colors on a canvas, and so on.

When assertoric sentences are used for this purpose, it is common knowledge that this is going to serve it by leading us to imagine that they are used in their agreed ways, to imagine what typically goes on when they are correctly used in their agreed ways, and so on. However, it is common knowledge that there is only the pretence of assertion, ultimately for the sake of giving rise to interesting make believe; the creator of the fiction is by no means obliged to observe the rules governing serious assertion. In producing ACP, Cortázar is merely pretending to know about a reader reading a novel with the features he represents for his fiction-making purposes. Were we to assume that it is knowledge that he is trying to convey, it would not cross our mind that the reader he is telling us about is the victim in the novel he is reading, in spite of the coincidences (the upholstery of the armchair, the large windows); if the story is told to express knowledge about the actual world, it makes much more sense to think that these are just accidental. We only derive (6) when we take into consideration that Cortázar is merely pretending to assert, and that his real purpose is to entertain us with an amusing story, perhaps one in addition with interesting consequences to reflect upon, like (7). Thus, it is something like (G-C) that captures in a general way the basic assumption that we use in determining the content of a fiction:

(G-C) A sentence of the form 'In the fiction F, ϕ ' is true if and only if it is part of the communicative intentions of F's creator, as expressed by recognizable features Φ of F, to put audiences with intended features χ in a position to make believe ϕ.

(G-C) accounts for the generation of (6); for, as I said before, the three pieces of evidence mentioned in the intuitive justification offered above for its truth are in

fact considerations of Gricean *relevance* relative to the communicative goals constituting fiction-making.

(G-C) is not in opposition to the two sorts of principles that Lewis appeals to in his two analyses, (L) and the alternative one appealing to the way the actual world is instead of what is common knowledge. I take both of them, in those cases in which each is intuitively applicable, to be entailed by (G-C). Reliance on shared beliefs (or on not so commonly known actual facts, for instance autobiographical ones facts in poetry) to go beyond what is explicit in the fiction is just one of the most usual make-believe devices presupposed by the communicative intentions of authors, one of the ways authors lead sensible reasonable audiences to imagine propositions. What are "sensiblereasonable" /"sensitive"/? audiences? The most that can be said here is that there is a practice of criticism (in which, of course, not just official critics, but most people fond of fictions engage), according to which not every proposition is fictional in a given fiction, on whose standards depend the nature of the illocutionary-type at stake.[24] In general, I share Walton's (*op. cit.*, 183-7) skepticism about the prospects for working out an exhaustive catalogue of the resources successfully used to generate fictional contents. To put it bluntly, given such a catalogue, a clever author would produce a fiction relying on a procedure not in the catalogue to generate content for its fiction, which most informed appraisers would agree in counting as fictional in the work. (G-C) is of course vaguely general, but I think that the philosopher's task finishes with such a proposal; beyond that, it is the critic's work to characterize specific principles of generation.

(G-C) could thus be put in terms closer to Lewis' possible worlds account, by including a restriction on the relevant worlds additional to Lewis' shared belief restriction.[25]

($G\text{-}C_{pw}$) A sentence of the form 'In the fiction F, ϕ ' is true if and only if ϕ would be true if, to the extent that this is compatible with the realization of the fiction-making intentions of the author as ascertainable from F by intended audiences, F were told as known fact, and the beliefs constituting

24 Lamarque & Olsen (1994, 29-52) usefully discuss the nature of this practice, in terms that I find congenial. They reject accounts that, like the present one, analyze fiction-making as a type of illocutionary force, because "there is little to be gained" from them (ibid., 72); this paper tries to articulate part of what is there is to be gained.

25 Put in those terms it can be more easily compared to Bonomi & Zucchi's (2003, 117-9) related proposal; their appeal to "conventions for the fiction" can be understood, along the lines here advocated, as deriving from a speech-act account of fiction-making. Phillips' proposal (op. cit., 287) is also close: "A sentence of the form 'In the fiction F, ϕ' is true if and only if, it is reasonable for an informed reader to infer from the text that, under ideal conditions, the author of F would agree that ϕ is a part of F." By suggesting resources for elaborating on what an author should agree on under ideal conditions, implicit in the appeal to an illocutionary-type account of fiction-making, mine improves on this.

common knowledge in the community where F originated were also known fact.

How useful the appeal to possible worlds will be depends in general on how adequate the possible worlds account of the content of fictions is, for instance to deal with the crucial issue that referential expressions like those in (1), (3) and (4) lack their ordinary referents, or with intentionally impossible fictions. My own view is that it is useful, if we take into consideration only the "primary intensions" of two-dimensional semantics on a certain neo-Fregean interpretation of them;[26] but this should be properly elaborated elsewhere.

6. The Intentional Fallacy

Most criticisms of illocutionary-type accounts of fictionalizing accuse them of incurring ~~in~~ the alleged "intentional fallacy," which, as previously indicated, Beardsley & Wimsatt (1954) put as the view that "in order to judge the poet's performance, we must know what he intended."~~.~~ In evaluating these criticisms, it is very important to keep in mind that on the present account not any intention of the fiction-maker is relevant~~:~~, only *communicative* intentions are. Let me provide an illustration of this distinction, taken from recent disputes regarding the nature and role of demonstrations in the interpretation of demonstratives. In his earlier work, Kaplan took demonstrations to be something like visual presentations of objects discriminated by pointing gestures (Kaplan 1989a, p. 490), or the pointing gestures themselves. Later, Kaplan (1989b, pp. 582-584) proposes a revision of that theory, according to which demonstrations are to be considered sets of "directing intentions". To justify the revision, he mentions a famous example that he (1978/90, 30-1) had ~~had given~~gi~~a~~ven earlier: the speaker points at a picture of Agnew while wrongly believing to be pointing at a picture of Carnap that used to hang in the place on the wall where he is pointing at, while uttering 'that is a picture of one of the greatest philosophers of this century'. He says that he had adopted the earlier view ~~having~~ with this example in mind, thinking that to take demonstrations to be directing intentions instead of pointing gestures "seemed to confound what Donnellan might call the *referential* and the *attributive* uses" (Kaplan 1989b, p. 583), while now he has decided to disregard this example "as a rather complex, atypical case" (*ibid*., p. 582, fn).

26 García-Carpintero, M. ~~& Macià, J.~~ (forthcoming) outline~~s~~ this interpretation.

This has suggested to some readers that a motivation for the revised theory is the belief that – against what he himself (1978/90, 30-1) said earlier in presenting the example – a correct theory should entail that it is Carnap's picture, and not Agnew's ~~picture~~, that the demonstrative refers to. Reimer (1991) criticizes Kaplan's revised theory on this assumption, endorsing the earlier account. I think she is right that Carnap's picture is merely the speaker's referent, not the demonstrative's semantic referent~~, which is Agnew's picture~~. In defense of the revised theory, however, Bach (1992) points out that the speaker in the story *also intends* to refer to Agnew's picture. It is just that what he ~~*immediately*~~ *ancillary intends to refer to* (the demonstrated entity, Agnew's picture) does not coincide with what he *ultimately intends to refer to* (Carnap's picture) by enacting the first, ~~immediate~~ ancillary referential intention. Thus, says Bach, Reimer's intuitions about the example are still borne out by Kaplan's new theory of demonstration, to the extent that we assume that when a pointing gesture takes place, it is this gesture that gives the primary indication of the speaker's directing intentions when determining the semantic referent.

Putting aside Kaplan's exegesis, the important point is that Bach is surely correct that the two intentions are present, even if the one the speaker focuses on is his ultimate intention – the other being a merely ancillary one. ~~What is more~~More ~~relevant~~importantly, Bach is also right that there is good reason, inside a Gricean framework, to take ~~only~~ the speaker's ancillary intention expressed by his pointing gesture, as opposed to his ultimate intention to refer to the picture which used to be hanging there, as most relevant to determine the semantic referent; namely, that only the former can be ~~sensibly~~ reasonably taken to be a *communicative* referential intention, an intention that ~~can be~~is expected to succeed by its being recognized.

When using demonstratives, speakers intend the propositions they express to be about individuals made salient to their audience when they utter them in agreed ways, particularly by the use of accompanying pointing gestures. Typically, they also intend for those propositions to be about individuals having further recognizable features (in the example, the features represented in the picture of Carnap that the speaker believes to be behind him). It is the latter intention that they have fully in mind, because it involves more useful properties than that of being made salient in a particular act of demonstration for the cognitive handling (in inference, memory and so on) of the propositions they want to convey to their audiences. Ancillary and ultimate referential intentions usually pick up the same individuals; otherwise,

demonstratives would not be serviceable, and there would not be conventions establishing their use. ~~But, w~~When they do not, the demonstrative's literal referent is certainly not determined by the ultimate referential intentions ~~the ancillary referential intentions~~, because it is the conflicting ancillary ~~they~~ ones that are *communicative* – capable of producing their intended effects by their being recognized in the required way. Whether the most sensible treatment of the cases have the ancillary referential intentions determining the semantic referent, or whether they should be treated as reference-failures, is up for grabs; intuitions certainly waver, depending on the case. Perhaps the most that can be said is that different cases deserve different diagnoses.[27]

These points apply *mutatis mutandis* to acts of fictionalizing. Firstly, not ~~any~~ every intention to put their audiences in a position to ~~make~~ make believe (imagine) propositions that authors might have in mind determines the contents of their fictions, no matter how important those intentions are for them~~:~~; only those that it is sensible to count as recognizable ~~for~~ by the intended audiences on the basis of features of their acts ~~do~~. Thus, even if Cortázar very much wanted us to imagine that the novel in ACP was written by alien beings, this is not a fictional truth in the story.[28] Secondly, just as the speaker in Kaplan's example ~~has~~ may have both unintentionally (relative to his ultimate intentions) ~~but still~~and intentionally ~~nonetheless~~ (relative to his communicative intentions) conveyed a proposition about Agnew's picture,[29] an author might well unintentionally (relative to his ultimate intentions) but intentionally (relative to his communicative intentions, ascertainable by sensible audiences in his story) lead ~~to~~ his audience to correctly imagine certain propositions. Thus, ~~(7)~~(8) ~~is~~ may still be true, even if Cortázar declares not to have thought of the ~~hero~~ victim of the conspiracy in the fictional novel of his story as the heroine's husband. For he knows very well our critical practices in interpreting fictions, which he exploits for his fiction-making purposes in his writings. He should thus agree that he intended ~~his~~ that propositions that can be derived from what is explicit in his story on the basis of shared beliefs ~~/OK? to b~~are part of ~~the~~ its content ~~of his story if they can be derived from what is explicitly put in there on the basis of shared beliefs~~; and this, together with the pieces of evidence we gave before in justifying ~~(7)~~(8), entails that he intended, even if only in those general terms, the relevant proposition to belong to ACP's content.

27 Wettstein's (1984) discussion of several cases shows this.

28 Currie (op. cit., 109) has a similar example.

It may be thought that the demonstrative analogy is misleading, in that the criteria for proper manifestations of communicative referential intentions are in that case clear, while they are not in the case of fictions. But the situation is not so different. Pointing gestures are acknowledged ways of conveying referential communicative intentions for demonstratives, but they are not the only ones; and there is an open-ended list of other equally serviceable criteria that speakers make use of. In some cases, there might well even be as a result two or more equally acceptable competing candidates, given by two or more equally acceptable criteria. The same applies to the interpretation of fictions.[30]

~~Let me~~It could be useful to ~~hammer home the last point by~~ discuss~~ing~~ ~~another~~ an example.[31] ~~I am an author, and I am dumb~~but ~~a~~A ~~bad~~n unskilled author ~~one~~. ~~I~~ write~~s~~ a mystery. He~~I~~ want~~s~~ to write an open-ended story, one at the end of which the reader is supposed to ~~make~~ make believe that the case is not solved. But ~~I am~~he is just no~~t~~ good at it. Without realizing it, he~~I~~ construct~~s~~ the plot in such a way that it implies that the killer is the night porter.

So, (10~~5~~) is true:

~~(8)~~(10) In the mystery, the killer is ~~likely to be~~ the night porter.

Formatadas: Marcas e numeração

But this is ~~totally~~ unintended by the writer ~~by me~~, ~~I~~ he just did not figure out that th~~os~~e clues, once one puts them together, ~~strongly suggest~~establish that the killer is the night porter. So, there is no intention on ~~my~~ his part that the audience ~~make~~ believe that the killer is ~~likely to be~~ the night porter. Thus, there is no communicative intention of this sort either.

~~There are two possibilities to deal with this kind of objection, depending on how much weight is put on the dumbness of the author's lack of skill. If too much weight is put, the case is similar to those objections by Walton I mentioned earlier, the b~~B~~ible or Greek myths as fiction, the wonderful literary piece accidentally produced by monkeys, and so on. In that case, the sort of reply that, as I mentioned earlier, Currie and others have suggested, is adequate. The only clear thing about these cases is that we *treat them* as fiction, not that they are fiction. It is acceptable for an otherwise theoretically well-motivated account to insist that they *are not* fiction; to explain why we treat them as if they were should not be difficult. If not too much~~

29 Whether a proposition has been properly expressed about Agnew's picture depends of course on whether or not the case is treated as reference-failure.

30 Currie (op. cit., 66) provides as an ~~interesting~~ example~~,~~ the well-known rival interpretations of James' "Turn of the Screw".

31 Together with other similarly interesting ~~ones~~examples, this one ~~it~~ was proposed by Alessandro Zucchi, to whom I am grateful for discussion of this and related points.

~~weight is put on the dumbness of the author's lack of skill, then the previous line works also here. The author knows about the critical practice of determining the content of mysteries; indeed, in fact he intentionally exploits it in constructing his mystery. It is part of his communicative intention to rely on this practice to fill out the explicit content of his mystery. When the clues determining the night porter as the likely murderer are pointed out to him, he should acknowledge that he intended this to be part of the content of the mystery, although only in general terms, not having himself carefully worked out all the (intended, in those general terms) consequences of his assumptions.~~

Commonplace criticisms of the *intentional fallacy* thus have a point, which is why they are so popular; but the point they have is compatible with the truth of our intentionalist proposal, and thus with the alleged intentional fallacy, as characterized by Beardsley and Wimsatt, not being a fallacy at all. Firstly, an author can unsuccessfully intend to make part of the content of his fiction a proposition that he considers in its full specificity. Secondly, an author can wrongly deny that a proposition also considered in its full propositional specificity is part of his fiction's content, in that he does have general intentions that make the proposition part of the content.

Objections like this should be answered in two ways, depending on how much weight is put on the author's lack of skill. If too much weight is put, the case is similar to objections by Walton to illocutionary-type accounts, based on the fact that we take the Bible or Greek myths as fiction, or on examples like the wonderful literary piece accidentally produced by monkeys, and so on.[32] In that case, the sort of reply that Currie and others have suggested is also adequate here. The only clear thing about these cases is that we *treat them* as fiction, not that they are fiction. It is acceptable for an otherwise theoretically well-motivated account to insist that they *are not* fiction; to explain why we treat them as if they were should not be difficult. In the example, it is clear that we treat the mystery as one for which (10) is true, but it is not clear that it is one such. If not too much weight is put on the author's lack of skill, then the previous line works also here. The author knows about the critical practice of determining the content of mysteries; indeed he intentionally exploits it in constructing his mystery. It is part of his communicative intention to rely on this practice to fill out the explicit content of his mystery. When the clues determining the night porter as the likely

murderer are pointed out to him, he should acknowledge that he intended this to be part of the content of the mystery, although only in general terms, not having himself carefully worked out all the (intended, in those general terms) consequences of his assumptions.

7. Conclusion

(6) is but a particularly clear-cut illustration of a type of inference to indirectly generated fictional truths that is part of our appreciation of the content of most fictions. (G-C) handles other cases similarly, which Lewis himself saw as problematic for his account, and which critics have dwelt upon since in discussions of his proposal, such as: the inference that the singer of the ballad of Mack the Knife in Brecht's *Threepenny Opera* is a treacherous fellow (Lewis, *op. cit.*, 274); unreliable narrators, like the one in "the puzzle of the flash stockman" (*ibid.*, 279-80); fictions with intended contradictory contents (*ibid.*, 274-5, 277-8); Walton's silly questions, as whether Othello spoke a wonderfully nuanced English for a military man.[33] I think that, given the failures of the competing views, all these cases provide decisive evidence in favor of Gricean illocutionary-type accounts of fictionalizing.

32 Those writings mentioned in fn. 16 provide convincing replies to these criticisms.

33 For further discussion of these problems, see Currie (1990, 62-70, 83-9); Lamarque & Olsen (1994, 90-5), and Phillips (1999~~98~~), 277-281.

References

Bach, Kent (1987): "On communicative intentions: a reply to Recanati." *Mind and Language* 2: 141–154.

Bach, Kent (1992): "Paving the Road to Reference," *Philosophical Studies* 67, pp. 295-300.

Beardsley, M. & Wimsatt, W. (1946/54): "The Intentional Fallacy," in W. Wimsatt & M. Beardsley, *The Verbal Icon*, Lexington: University of Kentucky.

Bonomi, A. & Zucchi, S. (2003): "A Pragmatic Framework for Truth in Fiction," *Dialectica* 57, 103-120.

Close (1972): "*Don Quixote* and the 'Intentionalist Fallacy'," *British Journal of Aesthetics* 12, 19-39.

Currie, Gregory (1990): *The Nature of Fiction*, Cambridge: Cambridge University Press.

García-Carpintero, Manuel (2001): 'Gricean Rational Reconstructions and the Semantics/Pragmatics Distinction', *Synthese* (USA), 128, 93-131.

García-Carpintero, Manuel (2003): "Assertion and the Semantics of Force-Markers", in C. Bianchi (ed.), *The Semantics/Pragmatics Distinction*, CSLI Lecture Notes, Chicago: The University of Chicago Press.

García-Carpintero, M~~. & Macià, J.~~anuel (forthcoming): "~~Introduction~~A neo-Fregean interpretation of Two-dimensionalism", in ~~in~~ J. Macià, & M. García-Carpintero ~~*ibid.*~~ (eds.), *Two-Dimensional Semantics*, Oxford: Oxford University Press.

Grice, H. P. (1975/89): "Logic and Conversation", in P. Cole and J. Morgan (eds.), *Syntax and Semantics*, vol. 3, New York: Academic Press. Also in Grice, H.P., *Studies in The Ways of Words*, Cambridge, Mass.: Harvard U.P., 1989, pp. 22-40.

Kaplan, David (1978/90): "Dthat", in P. Cole (ed.), *Syntax and Semantics*, New York: Academic Press, 221-43. Reprinted in P. Yourgrau, *Demonstratives*, 11-32, Oxford: Oxford University Press, 1990.

Kaplan, David (1989a): "Demonstratives", in J. Almog, J. Perry and H. Wettstein (eds.), *Themes from Kaplan*, Oxford: Oxford University Press, pp. 481-563.

Kaplan, David (1989b): "Afterthoughts", in J. Almog, J. Perry and H. Wettstein (eds.), *Themes from Kaplan*, Oxford: Oxford University Press, pp. 565-614.

Lamarque, P. & Olsen, S.H. (1994): *Truth, Fiction and Literature*, Oxford: Clarendon Press.

Levinson, Stephen (2000): *Presumptive Meanings*, Cambridge, Mass.: MIT Press.

Lewis, David (1978/83): "Truth in Fiction," *American Philosophical Quarterly* 15, 37-46. Reprinted with postscripts in D. Lewis, *Philosophical Papers, vol. 1*, pp. 261-280, Oxford: Oxford University Press, 1983.

Peacocke, Christopher (forthcoming): "Joint Attention: Its Nature, Reflexivity, and Relation to Common Knowledge", in *Joint Attention: Communication and Other Minds*, N. Eilan, C. Hoerl, T. McCormack and J. Roessler (ed.), Oxford: Oxford University Press.

Phillips, John F. (1999): "Truth and Inference in Fiction", *Philosophical Studies* 94, 273-293.

Plantinga, Alvin (1974): *The Nature of Necessity*, Oxford: Clarendon Press.

Reimer, Marga (1991): "Demonstratives, Demonstrations and Demonstrata". *Philosophical Studies* 63, pp. 187-202.

Searle, John (1974/79): "The Logical Status of Fictional Discourse", *New Literary History*, 6; also in his *Expression and Meaning*, Cambridge: Cambridge Univ. Press, 1979.

Searle, John (1975): "Indirect Speech Acts", in P. Cole and J. Morgan (eds.), *Syntax and Semantics*, vol. 3, New York: Academic Press, 59-82.

Thomasson, Amie (1999): *Fiction and Metaphysics*, Cambridge: Cambridge U.P.

van Inwagen, Peter (1977) "Creatures of Fiction", *American Philosophical Quarterly* 14, 299-308.

Velleman, David (2003): "Narrative Explanation", *Philosophical Review* 112, 1-25.

Walton, Kendall (1990): *Mimesis and Make-Believe*. Cambridge, Mass.: Harvard U.P.

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

Williamson, Timothy (1996/2000): "Knowing and Asserting", *Philosophical Review* 105, 1996, 489-523; included with some revisions as chapter 11 of his *Knowledge and Its Limits*, New York: Oxford U.P., 2000.

Zalta, Edward (2000): "The Road Between Pretense Theory and Abstract Object Theory," in A. Everett & T. Hofweber, *Empty Names, Fiction and the Puzzles of Non-existence*, Stanford: CSLI, 117-147.

5
Ethical Impasse: Two Accounts

Allan Gibbard
University of Michigan, Ann Arbor
gibbard@umich.edu

I one time settled into my seat for a trans-Atlantic flight, and as soon as I started chatting with the passenger beside me, he announced that he was president of the Dutch spiritualist association. The whole flight was very uncomfortable for me. At the start, we continued chatting, and he told me, in a quiet an amiable tone of voice, a sequence of tales I considered I had no reason to believe, saying "These things happen, you know." Eventually, I was able to bury myself in reading interspersed with drowsing, but I certainly wished I were sitting with someone else. Now, it's puzzling why my predicament was so uncomfortable, though some things it's easy enough to say about it. Trying to convert him to rational ways of thinking would have been Quixotic. Still, I didn't want to rudely ignore him or politely nod in agreement. On the other hand, I didn't want to spend my energies contradicting everything he said. The normal rewards of conversation just weren't to be had; on the contrary!

Conversation is something we ordinarily crave, and it figures crucially in my own views of normative thought and talk—but I'll touch on that later. First I want to bring on stage three current kinds of metaethical theories. By metaethical theories, we traditionally mean theories of the nature of ethical concepts and the meanings of ethical claims. I mean what I say also to extend, though, beyond ethics, to normative concepts in general. By *normative* concepts I mean ones laden with some kind of "oughtiness", concepts that are somehow", in the phrase attributed to Wilfrid Sellars, "fraught with ought, in a way that cries out for philosophical explication. I'll be talking not so much about moral obligation specifically, but of what a person "ought" to do in a sense that is very thin. It's what the person has most reason to do all told, taking account of all considerations, whether moral, prudential, or otherwise. I mean my discussion also to take in the concepts of normative epistemology, concepts of justified belief, of what we ought to believe in a sense that looks to the evidence and not directly to the practical advantages of believing one thing over another. I'll be

talking of what a person has reason to do and reason to believe, and how the reasons stack up to settle what she ought to do and what she ought to believe in light of her evidence. The word I'll use to capture all these things will be just plain 'ought'. For what one "ought" to do, understand what it makes most sense to do, what one has most reason to do.

Start first, then, with one kind of hard-line normative realism, the non-naturalistic, intuitionistic kind: On this view, what we ought to do or believe is just a special kind of fact, which we are equipped to apprehend, more or less. It's never a purely natural or empirical fact, and so normative claims never mean exactly the same as empirical claims; they can't be established by purely empirical, scientific investigation. Our special power to apprehend such truths is called intuition. We should note that in describing this position, I've been using the term 'fact' in a specially broad sense. The term originally finds its place in a contrast between matters of fact and matters of law, and in newspapers we distinguish matters of fact in proper new stories from matters of opinion on the editorial page. The point, though, is that non-naturalistic intuitionists think there's nothing much to this contrast apart from what's empirically accessible and what's accessible only in some other way.

Non-naturalism is respectable again these days, with such philosophers as Derek Parfit, Thomas M. Scanlon, Thomas Nagel, Ronald Dworkin, and Donald Regan taking positions that could perhaps be described as non-naturalistic and intuitionistic. I won't be evaluating non-naturalism today, though, as an analysis of normative claims. I'll rather be discussing two views that take it that non-naturalism isn't a credible account of normative concepts. I can mention now that the difficulties I'll be discussing have counterparts in the epistemology of non-naturalism—but that isn't a claim I'll try to make good on. I'll treat non-naturalistic views just as a part of intellectual milieu from which the views I'll be discussing emerge.

Non-naturalists arrive at their claims by refuting various forms of naturalism, the view that ethical concepts aren't sharply different from concepts in empirical sciences like psychology, sociology, and anthropology. Naturalists in turn, some of them, have become very ingenious in refining their analyses. One broad form of naturalism is subjectivism, the view that normative, "oughty" claims are claims about human attitudes of some kind. Crude forms of subjectivism are clearly non-starters; elementary ethics classes quickly dispatch such claims as that 'X is good' means I like X. But more sophisticated versions of naturalistic subjectivism look to Adam

Smith and aspects of Hume, flowering in the work of Roderick Firth and Michael Smith, among others. It's Michael Smith's version I'll be discussing—though I'm not entirely sure that his view fits the classification I'm putting it in. Smith sometimes uses the term 'desirable' in his analyses, but I'll stick to the term 'ought' to couch his theory. To believe that I ought to perform an act is, according to Smith, is to think "that I would want myself so to act if I had a desire set that was purged of all cognitive limitations and rational failings." And this "is for one's desire set to be maximally informed and coherent and unified."[1]

What does it mean for a desire set to be "maximally informed and coherent and unified"? Coherence is partly a matter of satisfying the laws of logic and using terms in accordance with their meanings, meeting whatever requirements you more or less have to meet if you are to count as deploying the concepts you do. But "coherence and unity" in Smith's sense must go beyond this. One person might conceivably think, say, that accomplishment is desirable in itself; another might think that though accomplishment is often desirable, it's never desirable just in itself, for its own sake—and both might satisfy all requirements of sheer logic and meaning. The history of philosophy, I take it, is replete with failures to show that logic and meaning by themselves can settle our preferences, or even our straight beliefs. On strictly formal grounds, after all, we can't even rule it out that the dead speak to people, or that the universe started 6000 years ago with fossils and DNA all in place—or six minutes ago, for that matter. Thinking what to seek and why calls for judgment. We may proceed from cases we find clear and apply the principles that seem to be at work in them, rejecting distinctions we find arbitrary. And these procedures may bring us to agreement—it hasn't happened yet, but I myself cling to hopes that it might. Conceivably, though, what I find arbitrary you might not. We may just have different reactions to matters. Or, given a choice between unity and simplicity in one's desires, on the one hand, and closer fit with our spontaneous desires, on the other, I might conceivably find one choice compelling and you another. The formal part of coherence, as Smith uses the term, is just a part. As he insists, what constitutes coherence is hotly contested, and the disputes may not be resolvable by methods of formal analysis alone.

To give us a quick way of drawing on this distinction, I'll confine the term 'coherent' to the formal aspects, and use 'cogent' for the aspects of reason that go

1 Smith (2002)., p. 311. See also Smith (1994).

beyond strict, formal coherence. To be coherent and cogent is to be reasonable, and we can now put Smith's view this way: that one ought to do *X* means that, if rendered ideally reasonable, one would want one's actual self to do *X*. Railton[2] offers a neat way of expressing this view: Let's call the agent who perhaps ought to do something *A*, and let *A*+ be *A* as she would be if she were fully reasonable, fully informed, coherent, and cogent. Then that *A* ought to do some thing *X* means that *A*+ would want *A* to do *X*. What a person ought to do, then, is what her ideal self *A*+ would want her non-ideal self to do. I'll call this view Sophisticated Naturalistic Subjectivism, or SNS for short.

I'll return to this theory shortly, but for now, let me mention the worry that Moore might stress with such a definition. Suppose the standards of what constitutes full cogency were fully and precisely spelled out, in naturalistic terms. Probably we'll never reach that point, but suppose we did. Then we'd have a fully naturalistic definition of what it is for an act to be advisable (or for a preference or other attitude to be well-founded). Label these conditions *N*. But some person might give standards these standards N that we have delineated no role in forming her desires and other attitudes. Instead, she might guide her desires and the like by alternative standards. Intuitively, then, as we might put it, this person thinks that standards I've called *N* aren't really the standards that constitute ideal cogency. Doesn't she, in that case, think that what one ought to do isn't a matter of being what one would want oneself to do in these conditions *N*, but rather in conditions of ideal coherence that are different from *N*?

That brings me to my own, somewhat different view of the meaning of 'ought'. What am I doing when I give advice and evaluate preferences as well-founded or not? I'm putting myself in the other person's shoes, and thinking what to want and what to do in the circumstance. You ask for advice for whether to sacrifice other things to achieve fame after your death. Ought you to work away at a book beyond the point of enjoyment or later reward, or go to a dinner party that you would enjoy? Isn't what you want, in asking me for advice, just this: for me to face, on your behalf, the same problem you yourself are facing? You want me, in effect, to try out doing the thinking that you yourself are faced with doing. You want me to do this in hopes that my sharing your deliberation will help you with your own thinking on the matter. So to come up with sincere advice, I imagine, hypothetically, that I'm in your

2 Railton (1986). The theory I couch with Railton's device is not Railton's own, but (I think) Smith's.

shoes, with your likes and ambitions and reactions to things. I then think, on this hypothesis, what to do and what to pursue. Suppose I conclude against giving any intrinsic weight to fame after death, if in your shoes. I'll say, "Fame after death, fame that will never affect how you experience things, isn't anything worth seeking in itself." In my view, you ought not to want posthumous fame for its own sake. I mean by this, "Don't give intrinsic weight to gaining fame after death." When I conclude for joining the dinner, I'll say, "You ought to let up on work and join the dinner." So, what kind of imperative is this, this imperative of advice? I'm placing myself, hypothetically, in your full circumstance, and then settling, hypothetically, what to do, letting you in on the verdict. That's my own theory of advice.

I have labeled this theory a form of "expressivism". An expressivist like Ayer or Blackburn or me, or like Stevenson late in his career, doesn't offer a straight definition of a term like 'ought'.[3] He's convinced by arguments like G.E. Moore's that no straight definition in naturalistic terms is going to work. There may be a correct account, in naturalistic terms, of what we ought to do, but it won't give the very meaning of the term, the thing meant even by the person with a basically wrong view of what we ought to do in life. Instead of giving a straight definition, the expressivist for a term defines it obliquely, by saying what state of mind the term expresses. Here, then, is an oblique, expressivist definition of 'ought'. As I state it, it won't exactly be my own view of what 'ought' means; rather, I'll put it in a form that makes it pretty directly comparable to Smith's sophisticated naturalistic subjectivism. We define what it is for me as an observer to believe that you ought to do some act *X*. To believe this is to want, for the hypothetical case of being you in your exact circumstance with all of your characteristics, to do that act *X*.

One distinction here is extremely important. Of course if I were in your shoes, in the expansive sense of the phrase that I intend, sharing your every characteristic, I'd just do whatever in fact you are going to do. I'm talking in this analysis not about what I would do if I were in your shoes. Rather, as an advisor, I face a hypothetical decision of what to do if in your shoes. My question is, if I'm forthwith to become you with all your characteristics and surroundings, what then to do. To answer this, I form actual preferences for a hypothetical situation. Hare calls these "actual-for-hypothetical" preferences; they're like "now-for-then" preferences that I can form for

3 Ayer (1936, Chap. 6); Stevenson (1964, Essay 11, esp. pp. 210–214); Blackburn (1993). My own theory of normative concepts is developed in Gibbard (1990) and Gibbard (2003).

a time when I'll likely give in to temptation.[4] So if you, being the way you are, will in fact give up pleasure for fame, that doesn't guarantee that I now want to give up pleasure for pain in case I'm you.

Suppose now I tell you that you ought to go to the dinner party. You, of course, can accept what I say or not. You agree with what I say if you opt for joining the dinner. Why does it matter to you, though, what I myself conclude? Perhaps it won't, or perhaps you'll treat my concluding what I do as just as a biographical fact about me. You don't have to want my opinion on anything, and it remains my opinion nevertheless. But perhaps you'll give a stronger kind of heed to what I conclude. You wanted advice because you were having a hard time settling whether to sacrifice enjoyments and the like for the sake of fame after death, and joining the dinner or not hinges on that. You wanted me to join in your efforts to think the matter through. Two heads are better than one, sometimes, and you wanted a second head to join you in your thinking.

My account, though, is meant to apply even if you find my advice worthless. It's one matter what I'm saying; it's another matter whether you believe what I say. I found the Dutch spiritualist's judgments on whether the dead communicate with us not worth heeding, but that doesn't bear on the content of what he was saying. Just so, when it comes to whether, if in your shoes, to seek fame after death for its own sake, you may not find my conclusions worth any heed. Still, according to my theory, when I say, "Fame after death isn't anything to seek for its own sake," then I'm expressing, as it were, an aspect of a contingency plan for in case I'm in your shoes. I'm telling myself and you, "If in those shoes, don't give intrinsic weight to fame after death."

Still, perhaps an objection remains. If, in your judgment, my conclusions on what to do in your shoes aren't ones to heed, there's at least a certain awkwardness in the conversation. Perhaps, in that case, I can't present my conclusions on what to do as ones for you to heed. And perhaps that means that I'm not really offering advice, or not telling you what I think you really ought to do. Perhaps I need to withdraw my claims when they aren't ones that you can fall in with. And this may lead us back to Michael Smith's sophisticated subjectivism in ethics. That's the possibility I want to explore.

4 Hare (1981), 101–106.

Return to Subjectivism

Cogency is that part of reason that isn't a matter of strict, formal coherence. Judgments of cogency require a sense of plausibility, which includes a sense of what's arbitrary or ad hoc. All this raises the possibility, I said, that different people have plausibility reactions and arbitrariness reactions that differ, in fundamental ways, that some of us just have fundamentally different senses of cogency. Of course in any actual conversation, it's very hard to tell whether the differences are fundamental and whether they would persist if we each achieved full formal coherence—as with the Dutch spiritualist. Perhaps one of us is just failing to focus on crucial aspects of the matter. We can't rule out on a priori grounds, though, the possibility of what I'll call a preference impasse, where each of two people is fully informed on the natural facts and fully coherent, but still disagree in what they prefer to do if in some circumstance.

Here, then, is a worry about my account: On my view, when I ask for advice on what to do and you answer sincerely, you put yourself hypothetically in my shoes, decide hypothetically what to do, and express your decision to me. Suppose, though, that your thoughts about what to do if in my shoes are at odds with my own, and are at odds with anything I'd come to from my own starting point, even if I had full information on all the natural facts that bear on my decision. Are you then really giving advice? Are you really telling me what it makes sense for me to do? Are you telling me what I myself have most reason to do? Aren't you, rather, just planning for yourself as me? How is this in any way giving advice? You have your plan and I have mine, plans for the same case of being exactly like me. Aren't advice and reasons and oughtiness a matter of starting with something that means something to me, not of speaking from you own concerns?

There are answers I can give. Your plans for being in my shoes, to be good plans, must take deep account of what being in my shoes is like, the ideals it involves adhering to and finding compelling, the ambitions it involves having, the kinds of things it involves finding fulfilling. So I'm not saying that well-founded advice can ride roughshod over who you most profoundly are. Then too, as I have said, there's a difference between advice being sincere and advice being correct. I can think that you are really telling me what you think I ought to do, but that you are wrong.

But suppose that we find the objections convincing. Then we might want to modify or abandon the account I have given. When I ask you for advice, I'm indeed asking for help with my thinking. But it must be help with starting from my own

premises and concerns and convictions and propensities to find things plausible. You must think for me in a way that would be accessible to me if only I were doing it better. After all, when you present considerations to me as bearing of what I have most reason to do, you present them as somehow compelling. And if they compel me, they must start from where I am. So if you give me advice, trying out doing some of my thinking for me, you must present your conclusions as relevant to my own thinking. You must present yourself as a qualified proxy for my own deliberating.

What, then, would make you a qualified proxy? Let's try saying this: You must come out the same way I would if I got matters right. I may be short of information; I may be short of time to think; I may be having trouble keeping track of everything I can recognize as bearing on the question of what to do. If you don't have these shortcomings, you may be able to think the matter through for me better than I can think it through myself. You are then thinking as I would if I lacked certain shortcomings. But if you and I would think differently once we lacked these shortcomings, we're now saying, advice to me pertains to how I myself would think. So in giving me advice—to the degree that you stand behind it as good advice—you present yourself to me as an improved version of the way I myself might think matters through.

This leads us back to Smith's sophisticated subjectivism. To model me turned fully informed and fully reasonable is just a way of predicting what I would want if I were fully informed and fully reasonable. And what I would want my actual self to do if I were fully informed and reasonable is just what I ought to do, according to Smith's Sophisticated Naturalistic Subjectivism.

Expressivistic Subjectivism

I now, and in the rest of the paper, propose to argue two things: First, we can't get the sophisticated subjectivism to work and keep it fully naturalistic. Sophisticated subjectivism itself needs to be given an expressivist twist. And in this version, it takes on the same burdens as my own expressivistic theory; it ceases to have the advantage of being a straight, naturalistic analysis. Second, I'll examine whether the resulting expressivistic subjectivism has any advantage over my straight expressivism.

I have been imagining that we have a set of conditions that are in some sense ideal for forming conclusions of what to do. These conditions might be couched in purely naturalistic, ought-free terms, and then we'll have an analysis that is fully

naturalistic, a filling out of what it is to be fully informed, coherent, and cogent. Various philosophers have attempted to formulate such conditions: Brandt talked of having undergone "cognitive psychotherapy", which is vivid and repeated awareness of everything that would impact one's desires. Firth said his ideal observer is, among other things, omnipercipient, and dispassionate, and in the respects he doesn't list, normal. Railton elaborated an account of "objectivized" desires. In speaking of conditions that are ideal for judgment, I have supposed that we had such an account. What, though, is at stake in regarding states of mind that meet these conditions ideal? Two theorists might set up different sets of conditions. Would they then just be using the term 'ought' in two different senses of the word, so that when one says that Jill "ought" to go up the hill and the other says that she "ought not" to do so, they aren't contradicting each other? Aren't they, rather, somehow at odds over which set of conditions really is ideal, which one really does constitute being a reliable gauge of normative matters? It seems so—but what could this mean? What's involved in thinking conditions for judgment to be ideal?

To think conditions ideal is to defer to judgments made in those conditions. That seems to get to the heart of the matter. Suppose we have a set of conditions *N*, specified in fully naturalistic terms. I regard these conditions as making one an ideal planner if, on learning what I would plan if I were in conditions *N*, I then plan that way. I treat judgments made in conditions N as reliable, in that I rely on the judgments I know I'd make if I were in conditions *N*.

We now can preserve the form of our sophisticated subjectivism SS, but give up on trying to put it in strictly naturalistic, ought-free terms. I'll call this analysis expressivistic subjectivism, or ES, and it reads at first much like a naturalistic version of sophisticated subjectivism SS.

> ES: A ought in *C* to do *X* $=_{\text{def}}$ Let *A*+ be *A* as he would be if in ideal conditions for forming wants. Then *A*+ would want *A* in *C* to do *X*.

Now, though, we explain the concept of "ideal conditions" for forming wants. No straight, naturalistic meaning-equivalent can be had, we now suppose; what conditions are ideal for forming wants is a fundamentally contestable matter. Instead, we explain the state of mind of regarding certain conditions as ideal. For *O* to think that conditions *N* are *ideal* for forming wants is for *O* to stand ready to want whatever *O* learns that he would want if he were in conditions *N*.

In calling conditions *N* ideal, then, observer *O* expresses this plan to defer to whatever judgments he would make if in conditions *N*.

I don't know for sure that this hybrid view ES isn't the kind of thing that Michael Smith himself has in mind. He stresses that the conditions of reason are highly contestable, and so his theory needs to be filled in with an account of how contestable concepts work. I take it now that purely naturalistic versions of a sophisticated subjectivism are off the table as accounts of meaning; they don't permit essential contestability. One naturalistic version, it must be, is right in substance; it correctly tells us what the conditions are that make a judgment reliable. No such version, though, gives the very meaning of 'reliable', and none of them gives the very meaning of claims that a person ought to do something.

So we are left with two accounts of the meaning of 'ought' on the table: my straight expressivism, which I'll call SE, and hybrid view expressivistic subjectivism, which I'm calling ES. But aren't we missing the most promising alternative? What about reflective equilibrium? Start where the thinker-planner is, refine her beliefs, preferences, and standards of reason in tandem, and see where she ends up. Well, in a sense, I agree, that's all we can do. And if we get matters fully right, we'll be in an equilibrium, where further considerations won't change our minds about anything. The idea of equilibrium is clear enough: One question we confront is what conditions are ideal for judgment. If we say that conditions N are the ones ideal for judgment, that account can be self-endorsing or self-undermining. The account is self-endorsing just in case, in conditions N, one would accept the account. A self-undermining view has a kind of epistemic incoherence, and so we won't have a reasonable ensemble of views until we have an account of ideal conditions for judgment that is self-endorsing.

But more than one view might be self-endorsing; there might be more than one possible reflective equilibrium. If so, which one of them yields what one really ought to do? The one we'd got to by a process of refinement of our present views, we might try saying—but we could refine them in different ways, ironing out formal incoherencies in our views in different orders. Who knows but that different courses of refinement might not end us in different equilibria?

It's clear enough what our goal is: it's to get the right conclusion. We want to end up thinking that Jill ought to go up the hill if and only if, indeed, she ought to go

up the hill. But we haven't found how to express this goal in naturalistic terms. So far, it's a goal couched only in normative terms.

Two Kinds of Impasse

How might two people both be ideally coherent formally, but still disagree. In many ways, presumably, but we can distinguish two extreme types. Perhaps they agree on what conditions are ideal for judgment, but their basic temperaments or mental constitutions are different, so that they judge differently and prefer different things in these conditions. Jack and Jill have taken their tumble, imagine, and they now must decide whether to go back up the hill. Conditions *N* are ideal for forming wants, they both agree, but in these conditions, Jill would want herself to go up the hill, whereas Jack would want, for the case of being Jill in her exact circumstance, not to go up the hill. This extreme type I call a constitutional impasse. Or alternatively, the two might have exactly the same basic mental constitution, so that in conditions N_L they would both agree on going up the hill if in Jill's shoes, and conditions N_K they would both agree on staying safely at the bottom with an empty pail. In either of N_K and N_L, imagine, Jack and Jill would be in reflective equilibrium, with a self-endorsing set of views. But as it happens, Jack's set of conditions is N_K whereas Jill's is N_L, so that Jill wants herself to go up the hill whereas Jack wants, for the case of being in her shoes, to stay at the bottom. This I call a multi-equilibrium impasse.

The hybrid view Expressivistic Subjectivism says that one such kind of impasse, the constitutional kind, isn't a true normative impasse—if by a true normative impasse, we mean a case where two parties disagree in questions of ought although both are fully informed on non-normative matters and both are fully coherent. If Jack and Jill agree on what conditions are ideal for judgment, ES says, then they will agree on what Jill ought to do: it's whatever Jill would want her non-ideal self to do were she in ideal conditions for judgment. As for what Jack ideally wants to do for the case of being in Jill's shoes, that, according to view ES, has nothing to do with what he thinks Jill ought to do. View ES, though, does allow a multi-equilibrium impasse as a true normative impasse, where Jack and Jill would form the same hypothetical wants in the same circumstances, but coherently each regard a different set of circumstances as ideal for forming judgments.

We can see all this from exploring the case. Both Jack and Jill are fully informed and coherent, and they each have a view of what conditions are ideal for

judgment. Each of their views on this is self-endorsing; Jack and Jill are both in reflective equilibrium. Jill wants to go up the hill; she thinks that if she were in the conditions she regards as ideal for judgment, she would still think those conditions ideal for judgment, and she would still want herself to go up the hill. Jack, for Jill's complete circumstance, wants not to go up the hill. Like Jill, he thinks that if he were in the conditions he regards as ideal for judgment, he would think those conditions ideal for judgment. He also thinks that in those conditions, he wouldn't want to go up the hill for the case of being in Jill's shoes. In the straight expressivist's sense, then, they disagree on what Jill ought to do: Jill thinks that, in this sense, she ought to go up the hill, whereas Jack thinks she ought not to do so. Jack, after all, wants, for Jill's exact circumstance, to stay put at the bottom. They are thus at a true normative impasse if my straight expressivism is right.

What, then, of expressivistic subjectivism, the hybrid view? In the case of a constitutional impasse, the parties agree on what the conditions ideal for judgment are, and so they agree on what Jill would want herself to do if she were in conditions ideal for judgment. They thus agree that Jill ought to go up the hill, in the sense given by expressivistic subjectivism. If this is the right view of matters, there's no true normative impasse in this case; they agree on what Jill ought to do. The case of multi-equilibrium impasse is different, though. In this case, Jack and Jill disagree about what conditions are ideal for judgment. In the conditions that Jill considers ideal for judgment, she would want herself to go up the hill, whereas in the conditions that Jack considers ideal for judgment, she would want herself to stay safely put. They both know this. Thus even in the hybrid, expressivistic subjectivist sense, Jill thinks that she ought to go up the hill, whereas Jack disagrees. Even if expressivistic subjectivism gives the right analysis of 'ought', Jack and Jill still disagree about what she ought to do, and the case is a true normative impasse.

The two candidate analyses differ, then, not on whether a true normative impasse is conceivable—they both agree that it is. They disagree only on whether a particular kind of true impasse is possible, a constitutional impasse. The two positions do still disagree, then, on this question: whether even with full information, full coherence, and full cogency, a person can be mistaken on what she ought to do. Straight expressivism says that she can be, whereas the hybrid view expressivistic subjectivism says that she can't. In the case of a constitutional impasse, after all, Jack thinks Jill to be fully informed, coherent, and cogent. Still, in straight expressivist

sense, he thinks that she ought not to go up the hill, whereas she thinks she ought to do so. She is thus, in Jack's view, fully informed, coherent, and cogent, but mistaken as to what she ought to do. She just has a defective basic mental constitution, thinks Jack. Jack regards Jill, because of her basic mental constitution, as prone to a kind of completely innocent mistake. Jack's view is coherent, and so it's not ruled out on sheer conceptual grounds. If the hybrid view expressivistic subjectivism is the right analysis of 'ought', in contrast, no such informed but innocent mistake is conceivable. If Jill is fully informed, coherent, and cogent, then she'll be right about what she ought in this sense to do.

Conclusions

We could just conclude at this point that we have laid out two coherent possibilities for what 'ought' means, possibilities that each allow for basic, fully informed and fully coherent disagreement as to what a person ought to do, that allow oughts to be essentially contestable. Both are concepts we can now go on to deploy, now that we have distinguished them. Both accounts are expressivistic; neither gives a purely naturalistic rendition of the meanings of normative terms. Expressivism of some form we're stuck with, we've seen if I'm right, if we are to match the phenomena of disagreement with our account and we don't retreat to a mysterious form of non-naturalism in our account of normative concepts. But this last is another topic. I do insist that even the most sophisticated subjectivism can't credibly be couched in fully naturalistic terms; it needs an element of expressivism.

Are there any further arguments, though, in favor of one as opposed to the other? We can't avoid the possibility of a true normative impasse with either account, we have seen. But perhaps a constitutional impasse is the only kind we really have to rule out—and the hybrid view Expressivistic Subjectivism does accomplish this. We could argue for the hybrid view, then, by arguing that a constitutional impasse is especially unacceptable. In a constitutional impasse, after all, Jill would be incoherent if she accepted all of Jack's views of (i) what she ought to do and (ii) what she ought to believe that she ought to do. A coherent set of views, we might argue, must consist in a package that the audience could accept without incoherence. Views as to what a person ought to do in the straight expressivist's sense sometimes just couldn't have this feature, even when one's other beliefs and preferences are in good order. Perhaps that shows that straight expressivism just doesn't define an eligible

concept. We are left with the hybrid view expressivistic subjectivism, which doesn't fail this test.

But this argument against straight expressivism seems too strong. Think of what we might call an "evidential impass"; such a thing is clearly possible, I'll argue, and may have a parallel feature. An evidential impasse I'll define as a case in which parties disagree and both are fully coherent, and both are fully informed as to available evidence. Do the spirits of the dead talk to the living? Consider my Dutch spiritualist seating companion, whom I'll dub Jan. Conceivably both Jan and I could have track of all available evidence and both be formally coherent, but differ basically in what we found credible and what not—a difference in our basic mental constitutions. He, perhaps, would be incoherent if he accepted all of my views of (i) whether spirits talk to us and (ii) whether he ought to believe that spirits talk to us. Still, the claim"Spirits talk to us" doesn't mean "If you were fully cogent, you would believe that spirits talk to us." And I can coherently believe that spirits don't talk to us, but that Jan, in light of his mental constitution, ought to believe that they do, given the evidence.

We can't, then, impose the requirement that rules out my straight expressivism without ruling out what clearly has to be the case with plain factual beliefs. My straight expressivism remains in contention, and parallels beliefs of fact in a way that expressivistic subjectivism does not.

Of course a situation in which we knew the two senses of 'ought' I've been discussing to diverge would be fantastic. When could anyone possibly know enough and have his plans for acting and wanting well enough worked out to have one view of what to do if in Jill's shoes and a distinct view of what to want to do if in Jill's shoes? Now of course if Jill has a firm view on some matter and isn't open to reconsidering it, then she isn't ready to take on opinions that rest on denying what she takes to be true. But that goes for anything that Jack might assert, no matter how prosaically factual. Our ordinary condition as human beings, though, is to have confused fragments of views, with some islands of coherence and clarity. It's still useful to run thought experiments about fully informed, coherent, and decided states of mind; they tell us something about the logic of our concepts. We have to remember, though, that such a state isn't one we ever attain.

That brings me to one of the arguments that draws me to prefer straight expressivism to a hybrid Expressivistic Subjectivism. Imagine that Jill is and

ordinary person like the rest of us, and she asks Jack for advice; she asks him whether she ought to go up the hill. We are studying the workings of the concept ought, and so we now inquire what kind of answer Jack can give if he is fully informed and coherent. Has Jill asked his thoughts on what to do in her circumstance? Then Jack will think about being in her circumstance, and think what to do in that case. He thinks about the same problem as Jill is confronting. Or is Jill asking for his thoughts on what she would want herself to do if she were ideal as a judge of the matter? That amounts to asking what, in his view, she ought to want herself to do. Now it seems to me that she may want his thoughts on the problem she is addressing, namely, what to do—which for Jack is the question of what to do if in her complete circumstance. She may want this more than she wants his thoughts on what to want herself to do. This last, after all, is a more indirect question. To get an answer to it, Jack must think about how to form wants and preferences. Jill might distrust his answer to this, just as she might distrust his answer on what to do. In case the two answers diverge, why should she care more about his thoughts on the indirect question than on his thoughts on the direct question of what to do? It isn't as if she has a coherent view on how to form her wants and preferences, and just needs help applying it.

Jack has a clearer, more coherent view of matters than she has; he has thought out what to do if in her shoes, and also what to want to do if in her shoes. Jill wants to try out his views for size. She can take on either of these views of his, but she can't coherently take on both at once. Jack can coherently hold both views, but Jill can't: if she held both views, she'd have a policy of wanting what she didn't in fact want, she'd violate her own policy for forming wants. She must reject one or the other of Jack's views, on pain of incoherence. It isn't clear, though, that cogency requires her to give more heed to Jack's epistemic views, his views on how to come to a view on what to do and why, than his ground-level views on what to do and why. Maybe what she most wants Jack to think through for her circumstance is what to do, not how to settle what to do. And so maybe this is what we'd best dignify as "advice", as Jack's asserting his view of what Jill ought to do.

We have, then, two eligible analyses of the term 'ought'. Both are expressivistic at one point or another; they need to be if they are to handle phenomena of possible normative disagreement. My own conclusion, though, is for my more straightforward expressivistic account of normative concepts. When we make normative assertions, I myself conclude, we assert oughts in the sense that straight

expressivism elucidates. As with assertions of plain fact, we make our assertions, ordinarily, in hopes that no impasse will arise. Still, we don't keep hedging our assertions with a proviso that it won't. I conclude, then, that the more direct expressivistic is the better one.

References

Ayer, A.J. (1936) *Language, Truth and Logic* (London: Victor Gollancz).

Gibbard, Allan (1990). *Wise Choices, Apt Feelings: A Theory of Normative Judgment* (Oxford: Oxford University Press).

———, (2003). *Thinking How to Live* (Cambridge, MA: Harvard University Press).

Hare, R.M. (1981), *Moral Thinking: Its Levels, Method, and Point* (Oxford: Clarendon Press).

Railton, Peter (1986). "Moral Realism". *Philosophical Review* **95**, 163–207.

Smith, Michael, (1994). *The Moral Problem* (Oxford: Basil Blackwell).

———, (2002). 'Evaluation, Uncertainty and Motivation'. *Ethical Theory and Moral Practice* **5**:3. 305–320.

Stevenson, Charles (1963). *Facts and Values* (New Haven: Yale University Press).

6
One More Argument

Paolo Leonardi
Universitá di Bologna
paolo.leonardi@unibo.it

> There is no … vantage point.
> W.v.O. Quine "Natural Kinds", 1967, p.127
>
> …most people are not aware that this round-about progress through all things is the only way in which the mind can attain truth and wisdom.
> Plato *Parmenides* 136e

Philosophy searches for knowledge, and its way is articulating ideas and investigating them by argument. Scepticism questions any knowledge, asking for articulating ideas more and for more argument. The first's search is under no constraint; the second's quest for rigor accepts no accommodation. The' sceptic looks a philosopher's complement, because he uses the same tools, and seems to pursue the same aims, being critical rather than assertive.

There are many strategies for dealing with scepticism. The most popular strategy is a "don't care" one, as if sceptical objections were riddles. But some philosophers take up the challenge. A few accept that at the core of the knowledge enterprise there are some dogmatic assumptions. What we know, we do directly or because we conclude to it by argument; we cannot know everything by argument because of regress; hence, something we know directly.[1] What is then directly known? Of course, there is no argument to it.[2] Yet, some consider a *proper* dogmatic exit in the debate, though evaluating the opportunity of a dogmatic assumption is still arguing.[3] Someone else deems that if by the above argument we were out of scepticism, we would be nowhere, and try an ultimate line of reasoning against scepticism.[4] A few of them, try to develop a different form of warrant, usually called 'entitlement', which

1 Cf. Moore 1939, p. 150. But see also his 1956, chapters 5 and 6.

2 Moore had clear the structure of the problem, and tried nonetheless to tell what is so known – common sense truths. He was well aware of that there weren't an argument for his claim.

For criticisms and defence of Moore, see Coliva 2004, Lycan 2001, Pryor 2004a, Wright 1985 and 2002.

3 See Pryor 2004b.

4 Wright 1991 suggests that the sceptic wins even by a draw. I disagree, but the nowhere feeling matches that suggestion.

Sextus Empiricus considers the problem, but instead of concluding by accepting some primitive, presents this as one of two modes of epochē. (In Mates 1996, p. 112).

might be by default, contextual or general.[5] Others try to show a direct access to self-evidences, indicating one or two of them. Or, they would claim that there is a non-vicious regress. What is maintained to be self-evident may be knowledge of an external object, as my left hand, or of a supposedly internal something, like a sense-datum, say, a color patch, or a formal argument.[6] The more internal the something, the most self-evident it is taken to be (identical to its own evidence), the least problematic is held to be our access to it:[7] perhaps, there is no blue square painted on the book cover, and I am suffering an illusion, but, anyway, I have a blue square sense-datum; it does not matter to a mathematical proof, as Descartes himself argues,[8] whether we dream of it or not. Having a sense-datum is identical to having access to it; figuring a proof is identical to having one.[9] Finally, some others look for finishing off scepticism, say, as paradoxical or as a disease affecting our abilities to detect mistakes in our conceptions.[10] The strategy of claim and query I'm going to sketch resembles the default and challenge strategy Michael Williams has developed out of an idea by Robert Brandom, but provides for no entitlement.

The problems with the sceptic start already with what is at stake and with describing his stand. What is at stake: throughout I'll not retreat from knowledge to warranted belief,[11] because this wouldn't warrant any truth, and what the sceptic questions is that *we have evidence* of *what there is* and of *how it is*.[12] The sceptic's

5 Cf., for instance, Williams 2001 (who builds on an idea from Brandom 1994), and Wright 2004.

6 Moore could have sided with either alternatives; Descartes and, in our days, Wright 1991 have, with some important proviso, opted for the second alternative. Sextus Empiricus accepted ephemeral appearances (to the individual subject) as evident.

7 The internal is taken to be 'luminous'. Cf. Williamson 2000, ch. 4.

8 Descartes 1641, Meditation 1.

> For whether I am awake or asleep, two and three together always form five, and the square can never have more than four sides, and it does not seem possible that truths so clear and apparent can be suspected of any falsity [or uncertainty].

9 But Nagel 2000 tries to counter this idea even for characteristically phenomenical things such as pains.

10 The phrase 'finish off' is from Johnston 1999; the two positions hinted at are, respectively, Wright 1991 and Williamson 2004.

11 Some forms of scepticism, Pyrrhonism for instance, do make the shift.
Wright makes the shifts.
Gettier 1963 shows that justified true belief isn't knowledge, and a sceptic could exploit it to surmise that a warranted belief doesn't therefore cope with her objections.
Williamson 2000 argues against the shift from knowledge to warranted belief.

12 A cue to it is precisely that the most famous modern and contemporary sceptical conjectures, if not all, take an objective perspective. (Classical scepticism, instead, seems to have aimed at a technique of self-control facing the discomfort of the unattainment of knowledge.) If an evil genius deceives me, and takes care that I am mistaken anytime I count, consider a geometrical shape or develop an argument, I could count 2 plus 3 equal 6, and find out that a triangle has 4 sides. If I am a brain in a vat and there is a smartest scientist operating on me, I can be deluded to be a brain in the head of a person, sitting in an armchair and reading today newspaper, with comments on what happened yesterday in my country and far away, taking electric impulses cleverly distributed by the scientist for traces of paper with printed on it pictures and sentences in a common language. (Malebranche is an ancestor of Putnam' brain-in-a-vat case. See 1696, I § 5.) The conditions express an objective circumstance – the evil genius, the smartest scientist and the brain in a vat – and the issue is evaluated from an objective point of view – I don't trace neither the evil genius nor the smartest scientist and don't realize to be respectively a duped individual or a brain in a vat. Being more careful and claiming agnosticism doesn't affect the point that the understanding of the sceptical quandary requires an objective point of view. A second, and more substantial, aspect is that our emotional and cognitive attitudes are connected with other places and times, and if it weren't so, they wouldn't be the attitudes they

stand. The sceptic takes no commitment; he is not maintaining that since there might be an evil genius or a smartest and as evil scientist, we do not know anything. He questions the philosopher's contentions, and those are consequences she draws in coping with his questions. Ideally, as Pyrrho suggested, the sceptic uncommittally submits a counterargument or a counterthesis, drawing on one of his opponents against another of them. Modern and contemporary sceptical conjectures are original forms of this strategy. If there is any impossibility or anything paradoxical, it is in the dogmatic replies of the philosopher.[13] Here, I'll not disqualify the sceptic, but build on his anti-dogmatism and, by a fairness, or parity, principle, constrain his own search. I'll give form to the sceptic uncommittal attitude, by imagining the sceptic to limit himself to questioning; I'll give form too to his anti-dogmatic stance, by distinguishing between holding a claim and holding it dogmatically, and between questioning a claim and questioning it dogmatically. The strategy allows for evaluating at each moment who is winning on points, but defuses the hope for a final victory. The strategy is partially and freely inspired by Pyrrho, the most classical form of scepticism.[14]

1

Many a reply to the sceptic use multiple strategies. George E. Moore, for instance, keeps all along to a dogmatic reply – there is some posits; there is some (common sense) truths we know. But he considers also that some common sense beliefs are more likely true than the sceptical doubts.[15] Already in "Hume's Philosophy", in 1909, Moore writes:

> There is no reason why we should not, in this respect, make our philosophical opinions agree with what we necessarily believe at other times. There is no reason why I should not confidently assert that I do really *know* some external facts, although I cannot prove the assertion except by simply assuming that I do.

---

are. We remember the past; we classify the moment as of the same kind as some other displaced one(s), etc. Thinking is connecting, and if my thought were limited to what appears to me now, it would me miserable if not null. (Even perception and attention are differential, and in a stable and homogenous environment we would notice and thereby perceive nothing.) That does not make the subjective aspects minor, however, because knowledge involves subjective evidence, i.e. evidence available to the individual subject, of what there is and of how it is.

13 The revival of Pyrrhonian studies goes back to Naess 1968. In my understanding of Pyrrhonism I have profited of Mates 1996 and of Johnsen 2001. My last claim in the text, according to which a restraint to the internal is in the dogmatic reply of the philosopher, misdescribes the facts. The Pyrrhonists subscribe only to what appears them now, and can be interpreted as "internal". Cf. Mates 1996, pp. 17-21.

14 The strategy revises the sceptical stand too. For instance, the sceptic is told to be happy with appearances by Sextus Empiricus, who is our main source concerning Pyrrhonism, and I think that a Sceptic is not obliged even to endorse appearances.

15 See Moore 1941, p. 226.

Besides the texts already quoted, the main Moorean texts on the subject are Moore 1925, 1939, 1941, 1940-44.

> I am, in fact, as certain of this as of anything; and as reasonably certain of it. (p. 163)

Throughout his main lever is the appeal to common sense, which is to be understood as our ordinary processes of reasoning and understanding.[16] The main strategy is restated in closing "Proof of an External World":

> [Kant] implies that so long as we have no proof of the existence of external things, their existence must be accepted merely on *faith.* He means to say, I think, that if I cannot prove that there is a hand here, I must accept it merely as a matter of faith - I cannot know it. Such a view, though it has been very common among philosophers, can, I think, be shown to be wrong - though shown only by the use of premisses which are not known to be true, unless we do know of the existence of external things. I can know things, which I cannot prove; and among things which I certainly did know, even if (as I think) I could not prove them, were the premisses of my two proofs. I should say, therefore, that those, if any, who are dissatisfied with these proofs merely on the ground that I did not know their premisses, have no good reason for their dissatisfaction. (1939, p. 150)

An equilibrated choice of grounds for a search might be acceptable, but dogmatism it is. Moore's specific twist is that he carefully chooses his core moves and (often) does not argue for them. Notwithstanding the title – "Proof of an External World" – what we are given a proof of is not the core truth – 'Here is one hand' – rather, what is proved, on that ground, is that there is an external world. Making it clear what is claimed rather than arguing for it is keeping to the dogmatic stance and halting the debate.[17] Choosing carefully what is claimed makes more problematic denying it. Besides, Moore's solution satisfies the essential requirement of combining subjective and objective elements – common sense, as sense, has a subjective dimension, and because common, has an intersubjective dimension.[18]

16 As Lycan remarks that does not make Moore claim common sense irrefutable.

17 Common sense does not supply an argument but a standard.

18 Moore doesn't explicitly analyze common sense. He is very careful in choosing some statements as instances of common sense statement, but the reasons of the choice are all off-record.

Refusing the split between a philosophical reason and everyday reason, Moore cuts out what later has been called the Humean solution to the skeptical quandary. (Cf. Kripke 1982, p. 66.) If I doubt that this is my left hand, and that there is a bottle here on the table, there is no thing I can do. (As Ludwig Wittgenstein has insisted on, later.)

More recently, Crispin Wright has offered a most thorough analysis and reply to the sceptic. In 1991, Wright has tried to show the sceptic paradoxical, providing «a properly detailed diagnosis and exposé of its power to seduce». (p.89) In 2004, he has suggested a new form of warrant, entitlement to accept, distinct from justification, thereby opting for another strategy, «essentially recognisable by means of traditionally internalist resources». (p. 209) Apart from the many strategy he pursues, Wright shows, I think, that internalist resources aren't sufficient to overcome the sceptical criticism.

The most interesting paradox, in the 1991 paper, concerns the case of the warrant we are not maundering (maundering stands to thinking as dreaming stands to perceiving – if I am dreaming I am not perceiving, and if I am maundering I am not thinking; maundering is taken to be phenomenologically indistinguishable from thinking). Informally, the argument is the following: assuming the sceptical premises, according to which a subject is not warranted not to be maundering and it is warranted that if a subject is thinking he is not maundering; assuming iteration and transmission of warrantedness; from the assumption that it is warranted that a subject is thinking, it is concluded that it is not warranted that if it is warranted that a subject is thinking then he is not maundering, i.e. the denial of the second sceptical premises.[19] The argument relies on the ambiguity between thinking and maundering, dependent on their being phenomenologically indistinguishable. Hence, if I were thinking, it would prove the second sceptical premise false. However, the assumption of phenomenological

19 More precisely the argument is the following (omitting the reference to time):

(P1**) a subject is not warranted not to be maundering;

(P2**) it is warranted that, if one is thinking, one is not maundering;

then, by assumption,

(i) it is warranted that a subject is thinking;

by an iteration rule,

(ii) it is warranted that it is warranted that a subject is thinking;

by transmission from (ii) and (P2),

(iii) it is warranted that a subject is not maundering;

from (P1) and (P2),

(iv) it is not warranted that a subject is thinking.

Besides, there is the possibility of applying the conclusion that it is not warranted that a subject is thinking to the basic idea that (2**) if it is warranted that a subject is thinking then he is not maundering, inferring that

it is not warranted that if it is warranted that a subject is thinking then he is not maundering

which denies (P2**). Hence, one of the sceptical premises themselves is supposedly rebutted.

smoothness, i.e. of the phenomenological indistinguishablility between thinking and maundering, is ungrounded. No *phenomenological* comparison can be made between a maundering and a thinking event can be made. If I were maundering, I could mistake my situation for one of thinking even if it were *not* phenomenologically indistinguishable from it. Besides, I have specifically no warrant to be thinking when evaluating the argument, and hence I have no warrant ever to tell that the argument, if I were thinking, would prove scepticism paradoxical. If I were maundering, the argument just discussed would be a reason for maundering false the maundering counterpart of the sceptical premise according to which it is warranted that if a subject is thinking he is not maundering, rather than a reason for thinking that sceptical premise false. The maundered argument couldn't be told to be an argument, and hence it would prove nothing. In other words, the argument proves thinking not to be paradoxical by assuming thinking. Moreover, if I were maundering, the argument for sure would prove nothing to me, who wouldn't be thinking – it wouldn't prove my «right to claim knowledge». It would prove to a third thinking party that it is not paradoxical to assume that he is thinking.[20]

In 2004, Wright pursues an entitlement strategy. Independently of assessing the success of the pursuit, notwithstanding the claim to rely on internalist resources, externalist ones sneak in. Epistemic entitlements are, according to Wright, essentially recognisable by means of traditionally internalist resources. … What [the sceptic] put[s] in doubt, adds he, is rather our right to *claim* knowledge and justified belief. It is this which the project of making out entitlements tries to address and which, on what seems to me to be a correct assumption, externalism is impotent to address. (2004, p. 210)

Consider, for instance, the conditions licensing, according to Wright, an absolute strategic entitlement:

> A thinker X is absolutely strategically entitled to accept P just in case
>
> (i) X has no sufficient reason to believe that P is untrue; and
>
> (ii), in all contexts, it is a dominant strategy for X to act exactly as if he had justified belief that P. (Wright 2004, p. 183)

20 Incidentally, the supposed ambiguity is one way of understanding a sceptic stance – a way closer to the Academy's kind of scepticism, which the Pyrrhonian believed to be dogmatic. A Pyrrhonian would rather stop at what appears to him at the moment, positing no thinking and no maundering.
Two more remarks. First, deriving a paradox from a sceptic's claim is impossible, let me remind you, because there is no such claim. Secondly, Wright tries to counter also the idea that the sceptic wins also by a draw, discussing second order scepticism, interestingly requiring for a warranted doubt.

In relation to 'having no sufficient reason to believe that P is untrue' it's difficult to imagine that here sufficiency is a question of a priori reflection, i.e. that it can be brought down to logical inconsistency. Any other option might require more than a priori elements and reflection. The idea of a dominant strategy in all *contexts*, quantifying on a position restricted by this term so connected with the idea of what is peculiar to an experience, seems even less just a question of a priori reflection rather than critical reflection on experience.

I would like to express in the most general terms the problem with internalism.[21] Wright speaks of «our right to *claim* knowledge and justified belief»: the problem is that besides 'claim', also 'knowledge' and 'justified belief" would have to be stressed.[22] The quandary is precisely not to give up the double constraint we face: the *subjective claim* to *objective knowledge*, or, in a slightly different wording, the fact that only a *subject knows*. A purely internalist strategy would fail, I suspect, the second constraint; and a purely externalist strategy, such as the reliabilist's, would fail the first one.[23]

2

In order to sketch my different strategy for coping with the double constraint, and as a preliminary to it, I'll straighten some things up.

> I was struck by how Diogenes Laertius reports Pyrrho's stand:
>
> the Sceptics persevered in overthrowing all the dogmas of every sect, while they themselves asserted nothing dogmatically; and contented themselves with expressing the opinions of others, without affirming anything themselves, not even that they did affirm nothing; so that even discarded all positive denial; for to say, "We affirm nothing," was to affirm something. "But we," said they, "enunciate the doctrines of others, to prove our own perfect indifference; it is just as if we were to express the same thing by a simple sign." (300-350 A.D., lib. IX, 74)[24]

21 Above I have preferred the pair subjective/objective to internal/external, though they are rather different. For instance, a priori knowledge, if there is any, is internal and objective.

22 Wright acknowledges much to the externalist, but the core of his warranting strategy is internalist – though with more hedges and provisos in 2004 than in 1991.

23 The double constraint prevents there to be an absolute warrant – which is an aim Wright shares with the Academic sceptic – but leaves room for a best strategy.

24 Cf. Sextus Empiricus I, 27 (in Mates 1996, p. 117).

There are two points I want to emphasize. The first concerns the sceptic's null endorsement. The second is the description of how a sceptic argues his stand. The technique is to contend against the dogmatist by rehearsing, and improving, other dogmatists' arguments, i.e. arguments by people subscribing to a different dogma from that of the occasional opponent of the sceptic. The method is theatrical, and perhaps, because of that, very effective. Sextus Empiricus too insists on null endorsment, and, concerning argument characterizes the Sceptic's practice as one «of opposing to each statement an equal statement». (*Outlines of Pyrrhonism* I, 6)

My proposal is more austere but aims to preserve either feature. If the sceptic opted for questioning rather than claiming, and if he asked his opponent why she does not think as another of his opponent does, illustrating this alternative way of thinking, he would impeccably achieve in one swoop either of what he cares for – showing that there are other alternative views, possibly as good, and being committed to any.

Questioning is uncommittal. The Pyrrhonist was trying to achieve *ataraxia*, and hence was concerned with the subject's point of view, and spoke of no endorsement, rather than no commitment, but here what is relevant is what the sceptic publicly endorses, i.e. what he is committed to. The questioner takes no commitment and defers them to the answerer.[25] [26] His role is to push the philosopher, and he can act his role almost wholly parasitically on his opponent and some third party, though he can contribute something original to the play – it is up to the philosopher to overcome the problem she is faced with. If he keeps to his interrogative stand, the sceptic isn't required even to claim appearances (though Sextus Empiricus maintains he does). Besides, if we reframe the dispute between the sceptic and his opponent as a game of question and answer, the questioner always asking for one more argument, we describe the sceptic as trying to show that what was claimed so far requires some backings, and hence that (up to now) it isn't well grounded.[27] Putting questions is onversationally consequential. Asking is classically a first move, which gives a lead in

25 See Leonardi 1984, p. 83.

26 With a proviso, though. The conditions that have to be satisfied to ask anything might make it impossible to endorse anything. See, for instance, Johnsen, p. 531 ff.

With a somehow Kantian move, Grice suggests that the sceptical view seems already given up implicitly by the assumptions made for setting the sceptical doubt. See, for instance, Grice 1975 [1991], p. 138, and Grice 1986 [1991], pp. 103ff.

27 «The Sceptic Way is called Zetetic [ζήτησις, "questioning"] from its activity in questioning and inquiring», writes Sextus Empiricus (in Mates 1996, p.89).

Actually, Sextus writes that the sceptic «are simply reporting, like a chronicler, what now appears to [them] to be the case». (Mates 1996, p. 89) But I think he has not to tell even that.

the talk exchange, and project a later pulling together of the different threads. The sceptic refrains from the last task, thereby stressing that there is no matter established through the search.

The sceptic – a thing that was much clearer to the Greek philosopher than later – has a specific target, the dogmatist. The Dogmatist is characterized by Sextus Empiricus as the person saying «to have found the truth», as «the followers of Aristotle and Epicurus, the Stoics, and certain others.» (*Outlines of Pyrrhonism*, I, 1) The dogmatist makes some claims, which the sceptic finds need backing, and which the dogmatist denies need it. A revised characterization of the dogmatist might be the following one: the dogmatist is a person saying something to be true but not arguing for its truth. A dialogical characterization of the dogmatist might be as a person who, in the debate, replies to a question with the same answer she has given to a previous one, thereby repeating what her own point of view is. Specifically and somehow more neatly: the dogmatist asserts her thesis; is asked for backing the claim; restates the thesis. However, if we accept the characterization, we can see that the sceptic himself could go dogmatically, by asking twice the same question. Then, he becomes a dogmatic questioner. Being dogmatic would then amount to assert or to question something arguing it by the assertion or the question itself. Of course, the picture is hypersimplified – there can be indefinitely many good reasons for repeating the same reply, or the same question at that – starting from problems in hearing what has been said. Besides, there can be indefinitely many tricks to claim to have said, or asked for, something different. A stylistic variant might make opaque that the same thing has been said. The words 'warrant' and 'justification' have, for instance, a different meaning in Wright's papers, but often are used just as linguistic variants. Hence, it may be difficult to assess a specific case. In ideal conditions, anyway, all these practical inconveniences aren't there, and a repetition would immediately indicate a dogmatic move. Participants to a dialogue are non-dogmatic, as it is clear now, when in such a circumstance they do not repeat the same move. If we value being non-dogmatic, as the sceptic does, we can then suggest, by fairness, as constraint on anyone, sceptic or not, ever to repeat the same move – specifically, the sceptic cannot limit himself to reiterate the question 'What warrant (justification) do you have for that?', or a similar one. As he wants his opponent to reformulate her answer, when the turn is back to him, he has to reframe his

question. Ideally, the sceptic has to oppose a view by asking why not to develop an alternative view, apparently as good in accounting for what his opponent calls data. The sceptic develops alternatives not because of relativism, because looking at the matter from one angle we would judge one way, and from another another way. He develops alternatives to argue that we cannot tell, to emphasize that either views may be as good – and neither is good as yet.[28]

The sceptic seems to oppose plainly holding something true – the dogmatist says «to have found the truth», writes blamefully Sextus Empiricus, as we have seen. But, already for Sextus, the problem is with holding true what we have not evidence for, i.e. with claims we cannot back – his keeping to appearances depends on appearances being evident. The idea of a non dogmatic questioning goes well with Pyrrho's suggestion, which we listened above, of opposing a dogmatic position by arguing for another (without committing oneself to it).

Summing up. If the sceptic limits himself to putting questions, and he is committed to no claim, as he wanted – the issue of what he claims cannot even be raised. A claim is held dogmatically, if upon request of why holding it rather any of many alternatives, she who holds it doesn't back the claim but restate it by the same words. A question is put dogmatically, if upon having been replied, he who puts it does not suggest another problem, but puts it again by the same words.

If the point of the dispute is dogmatism, we can imagine a common standard to assess, step by step, who has the lead between the sceptic and the non-sceptic, which is what the second strives for, without giving up anything of what the first cares for. In the dispute, at each stage, if not repetitious, the last speaker has the lead, and is winning on points. The forthcoming of a further reply or a further question can never be excluded; and hence there is no way to adjudicate a final victory.

There might seem to be a main problem in the issue of repetition, one which I sidestepped: a speaker can seem non repetitious simply by changing topic, i.e. by not abiding by relevance. However delicate may be to establish whether or not persons

28 The proposal respects one remark Diogenes Laertius attributes to the Sceptics. Writes he, in the book of the life of Phyrro:

> [T]he Sceptics reply that they only employ reason as an instrument, because it is impossible to overturn the authority of reason, without employing reason; just as if we assert that there is no such thing as space, we must employ the word 'space,' but that not dogmatically, but demonstratively; and if we assert that nothing exists according to necessity, it is unavoidable that we must use the word 'necessity.' (300-350 a.C. lib. IX, 77)

repeat themselves, there is a clear model: to produce an identical turn. The case of relevance hasn't as neat a standard. The problem is there in any discourse, and isn't specific to the dispute between the sceptic and the non-sceptic. As in any other case, it is up to the participants in the talk exchange to contest each other the relevance of their contributions, as part of the debate itself, and there are some limited standards to claim relevance. Each party can withdraw from the exchange, protesting that the other party is flouting relevance. Anyway, the charge might be a pretext, and as a consequence we might consider a draw the case in which one party withdraws blaming the other party not to abide by relevance.

3

A question-answer sequence as the one outlined would be a quest for *knowledge*. It would keep to the connection between knowledge and justification, without patching the definition of knowledge as justified true belief.[29] It would be grounded on *evidence* available to a subject. And it would balance the subjective claim to evidence with the objective claim to knowledge.

> In "Other Minds" John L. Austin offers the following picture of knowledge:
>
> But now, when I say 'I promise', a new plunge is taken: I have not merely announced my intention, but, by using this formula (performing this ritual), I have bound myself to others, and staked my reputation, in a new way. Similarly, saying 'I know' is taking a new plunge. But it is *not* saying 'I have performed a specially striking feat of cognition, superior, in the same scale as believing and being sure, even to being merely quite sure': for there *is* nothing in that scale superior to being quite sure. Just as promising is not something superior, in the same scale as hoping and intending, even to merely fully intending: for there *is* nothing in that scale superior to fully intending. When I say 'I know', I *give others my word:* I *give others my authority for saying* that – 'S is P'. (1946 [1979[3]], p. 99)

Two aspects are important. First, the remark seems to apply to *saying* 'I know' and not to knowing. 'Know' would be an illocutionary verb, with a commissive color to it (though it is classified as an expositive). Yet, if saying to know voices a commitment,

29 Cf. Gettier 1963.

knowing is suggested to imply one. Perhaps, we can repair the problem, by maintaining that knowledge is in principle public, and that if I am unable to express my supposed knowledge, though I have a normal general competence (motion, perception, language, reasoning, etc.) and the circumstances are normal and there is the opportunity (I can show my knowledge by doing something, possibly by expressing it in words), I cannot claim knowledge. Secondly, the commitment Austin proposes – if I say I know that 'S is P', «I *give others my authority for saying* that 'S is P'» - is auxiliary, and I would rather suggest a different one, in two stages. Both stages depend on the subjective happiness conditions of illocutionary acts:

> (*Γ*.1) Where, as often, the procedure is designed for use by persons having certain thoughts or feelings, or for the inauguration of certain consequential conduct on the part of any participant, then a person participating in and so invoking the procedure must in fact have those thoughts or feelings, and the participants must intend so to conduct themselves, and further
>
> (*Γ*.2) must actually so conduct themselves subsequently. (1962, p. 15)

(a), in harmony with (*Γ*.1), I would rather say that by claiming knowledge I am committed to back my statement if it is doubted. The idea that I give my authority is only derivative, because in the particular case the authority depends on my ability in arguing the claims. (b), in harmony with (*Γ*.2), this *commitment* has to yield a fact, i.e. if we know we succeed in arguing for it in front of any query. (Which is not equivalent to 'in front of any query in any circumstance', but to 'in front of any query in any actual circumstance'.) (Of course, there is the further condition that what we argue for successfully is true. But this is something we *judge* and not something we *state* – basically the procedure I am hinting at is the one we follow to claim to *know* something, and hence the one through which we judge something to be true.)

Then, knowledge is making a true claim, being thereafter under the commitment to back the claim against criticisms and alternatives, and succeeding in so backing it. Knowledge is ascribed to any claim which isn't challenged, or which stands up to challenge.[30]

The double constraint we face, the *subjective claim* to *objective knowledge*, is satisfied in the present view because it is the subject which provides other arguments for

30 Occasionally this allows equating believing true with knowing. Cf. Sartwell 1991.

her claim, and this has to be true. Notice, however, that the double constraint is somehow taken care immediately in that evidence is required to be evidence provided by an individual to another, hence supposedly good for either, for the first person point of view of the individual providing it and for a third person point of view of the individual provided with it, who in principle can be any individual whatsoever.

4

My aim has been to sketch a strategy, or the preliminary of one, to deal with the sceptic, respectful of his stance that takes no commitment and accepts no dogmatism.[31] Moreover, I wanted to show that a challengeable claim is still a claim to knowledge. I have rather engaged other strategies, to state or suggest them as unsuccessful, than I have engaged the sceptic – actually, I haven't attempted to refute any sceptical query, like, say, "Are you sure you are perceiving rather than dreaming?" In closing, I want still to dwell on a framework aspect of a reply to the sceptic, to which I have repeatedly hinted at, and which is my second best strategy – dogmatism or, as it would better be called, 'controlled dogmatism'.

A controlled dogmatism may distinguish different turns at which stop arguing. There are posits of "objects" which are a condition for knowing – I posit my own existence, that of the Earth in the last five minutes, that of other things I am experiencing or I have experienced recently – or of more sophisticated "objects", whose existence is entrenched in our experience but cannot be proved but by assumption – I posit the existence of other minds, etc. Descartes posits his own existence. These posits are starting points, which fix a frame of reference, within which our theories develops. Moreover, there are conditions for knowing that I know (knowing what I know), second order conditions. These further reflections do not posit things, but assume that some forms of intellection are correct, or incorrect, that is they establish some form of cognizing or of thinking as a standard – perception, some concepts, some core logic, etc. Assumption of this kind seems what the Dreaming and the Maundering Argument attack. Controlled dogmatism allows for us being occasionally wrong in what we posit and in our forms of intellection being occasionally refutable. I can judge wrong my

31 I have also aimed to defend the possibility for a person to be a systematic sceptic. If we distinguished between a methodic and a systematic (or radical) sceptic, my sceptic would be rather of the second than of the first kind. His aim is to impair the non-sceptic, and he is neither auxiliary nor subservient, but oppositional.

perception, by looking better; I can downgrade my argument by reflecting on it. Specific arguments require mastering specific concepts. The Dreaming Argument – according to which if we are dreaming, we are not perceiving –, for instance, requires an understanding of "perceiving" and of "dreaming", besides some core logic. But the idea is, often, that in order to deem wrong these forms of intellection we have anyway to rely of them.

Though controlled dogmatism is very tempting, and though one can be dogmatic, *à la* Moore, without giving reasons for one's own choices, controlled dogmatism is an oximoron – dogmatism because a series of posits and assumptions and controlled because somehow deemed nonetheless appropriate. Only within a system we have the tools – referring to what there is, referring to values, using a logic and any other theory – to evaluate the system itself and any other system too, at least up to slightly different ones. Hence, also whether there are invariant posits and assumptions, is a fact that can be argued for only from within one system. This is not to deny any choice a person can make were she to opt for a dogmatic strategy, but to deny that this strategy achieves what it claims – some invariants which are independent of further theoretical and practical choices the person would make. If all the further choices turned out to be a failure, and we were to keep to these supposed invariants, our choices could not be wrong but would have *no reason.*[32]

I owe the use of the somehow Cartesian labels 'methodic' and 'systematic' to a suggestion of Eva Picardi.

32 Earlier drafts of this paper have been presented at Lisboa, at Venice Summer School in Analytic Philosophy on Realism 2004, at Rijeka, at Bologna and at Prague 5th Colloquium on Interpretation. I have profited from remarks and criticisms of Elvio Baccarini, Andrea Bianchi, Joaõ Branquinho, Annalisa Coliva, Adriana Silva Graça, Elisabetta Lalumera, Eva Picardi, Barry Smith and Crispin Wright. My research was supported by the MIUR Prin 2003 research grant on "Rappresentazione e ragionamento", coordinated at the national level by Carlo Penco, Università di Genova, and at the local level, Bologna, by myself.

References

J.L. Austin 1946 "Other Minds" (*Procedings of the Aristotelian Society*, Supplementary Volume XX, 148-87; repr. in J.L. Austin *Philosophical Papers* Oxford at the UP 19793, pp.76-116).

R. B. Brandom 1994 *Making it Explicit* (Cambridge MA Harvard UP).

A. Coliva 2004 "Moore's *Proof*: Transmission-failure, Question-beggingness, Dialectical Ineffectiveness, Failure to Settle a Question, or Just Mere Irrelevance?" (unpublished).

R. Descartes 1641 *Meditationes De Prima Philosophia* (Parigi).

E. Gettier 1963 "Is Justified True Belief Knowledge?" (*Analysis* 23¸ 121-3).

P. Grice 1975 "Method in Philosophical Psychology (from the Banal to the Bizarre)" (*Proceedings and Addresses of the American Philosophical Association* 48, 23-53; repr. in P. Grice *The Conception of Value* Oxford Clarendon Press 1991, pp. 121-61).

- 1986 "Reply to Richards" (in *Philosophical Grounds of Rationality* Oxford Clarendon Press 1991, pp. 45-106; the final section of the paper has been reprinted in P. Grice *The Conception of Value* Oxford Clarendon Press 1991, pp. 93-120).

B.C. Johnsen 2001 "On the Coherence of Pyrrhonian Skepticism" (*The Philosophical Review* 110, 521-61).

P. Leonardi 1984 "On Conventions, rules, and speech acts" (*Journal of Pragmatics* 8, 71-86).

W.G. Lycan 2001 "Moore Against the New Skeptics" (*Philosophical Studies*, 103, 35-53).

N. Malebranche 1696 *Entretiens sur la Métaphisique e sur la Religion* (repri. in *Oeuvres Complàtes* Tome s XII-XIII Paris Vrin 1976).

B. Mates 1996 *The Skeptic Way* (Oxford at the UP).

G.E. Moore 1909 "Hume's Philosophy" (in *The New Quarterly*, repr. in *Philosophical Studies* London Routledge and Kegan Paul 1922, 147-67).

- 1925 "A Defence of Common Sense" (in *Contemporary British Philosophy*, a cura di J.H. Muirhead, Londra George Allen & Unwin, repr. in *Philosophical Papers* London George Allen & Unwin 1959, pp. 125-50).

- 1939 "Proof of an External World" (in *Proceedings of the British Academy* 25, 273-200; repr. in *Philosophical Papers* London George Allen & Unwin 1959, pp. 32-59).
- 1941 "Certainty" (in *Philosophical Papers* London George Allen & Unwin 1959, pp. 226-51).
- 1940-44 "Four Forms of Scepticism" (in *Philosophical Papers* London George Allen & Unwin 1959, pp. 196-225).
- 1939 "Proof of an External World" (in *Proceedings of the British Academy* 25 1939, repr. in *Philosophical Papers* London George Allen & Unwin 1959, pp. 127-50).
- 1953 *Some Main Problems of Philosophy* (London George Allen & Unwin 1959, pp. 32-59).
A. Naess 1968 *Scepticism* (London Routledge & Kegan Paul).
J. Pryor 2000 "The Skeptic and the Dogmatist" (*Nous* 34, 517-49).
- 2004a "What is Wrong with Moore's Argument" (*Philosophical Issues* 14, 349-78).
- 2004b "The Skeptic and the Dogmatist" (*Nous* 34, 517-49).
C. Sartwell 1991 "Knowledge is Merely True Belief" (American Philosophical Quarterly 28, 157-65).
Sextus Empiricus 2nd-3rd Centuries A.C. *Outlines of Pyrrhonism* (in Mates 1996).
M. Williams 2001 *Problems of Knowledge* (Oxford at the UP)
T. Williamson 2000 *Knowledge and its Limits* (Oxford at the UP).
C. Wright 1985 "Facts and Certainties" (*Proceedings of the British Academy* 71, 429-72).
- 1991 "Scepticism and Dreaming: Imploding the Demon" (Mind C, 87-116).
- 2002 "(Anti-)sceptics simple and subtle: Moore and McDowell" (Philosophy and Phenomenological Research 65, 330–48).
- 2004 "Warrant for Nothing (and Foundations for Free)?" (Aristotelian Society Supplement 78, 167-212).

7
Investigando a Organização da Mente: Dissociações e Modularidade em Ciência Cognitiva

J. Frederico Marques
Universidade de Lisboa
jfredmarq@fpce.ul.pt

Desde a antiguidade, a mente tem sido concebida como um sistema composto por uma variedade de processos relacionados entre si mas ao mesmo tempo desempenhando funções distintas e separadas (Dunn & Kirsner, 1988). Já em diversos pensadores gregos como Platão ou Aristóteles podemos encontrar descrições de diversas funções ou faculdades (Terry, 1998/2003). Tendo em conta estas raízes profundas da reflexão sobre a organização da mente, não é assim de estranhar que uma teoria componencial da mente sirva de base estruturante à ciência cognitiva actual no seu objectivo de compreender o comportamento humano. Isso mesmo é patente nas mais diversas definições de ciência cognitiva (ex. (Dunn & Kirsner, 2003; Lyons, 2001; Medler, Dawson & Kingstone, in press; Shallice, 1988) em que, o seu objectivo geral é caracterizado como o de descrever a arquitectura funcional da mente humana que se assume divisível em vários componentes ou módulos articulados entre si. Uma outra forma de dizer teoria componencial da mente é assim afirmar que o nosso sistema de processamento de informação é modular (Lyons, 2001, 2003; Shallice, 1988; Van Orden, Pennington, & Stone, 2001).

O tema da modularidade tem sido amplamente debatido em filosofia e também em ciência cognitiva, onde no entanto a definição do que é que constitui um módulo aparece como menos estrita em relação à muito debatida visão de Fodor (1983). Na sua obra clássica "*The modularity of mind*", Fodor propõe vários critérios que um sistema deverá respeitar para ser considerado um módulo dos quais se destacam: serem encapsulados em termos informacionais, ou seja têm pouco o nenhum acesso aos objectivos e crenças do organismo no seu todo; serem cognitivamente impenetráveis ou opacos à introspecção, no sentido em que as suas funções são automáticas e os seus resultados não podem ser modificados por outros processos cognitivos; e serem sistemas de input no sentido em que recebem informação do exterior (Fodor, 1983).

Em ciência cognitiva, a perspectiva geral que enforma a visão componencial da mente corresponde a sistemas funcionalmente distintos considerados encapsulados mas apenas no sentido em que um módulo pode desempenhar as suas funções sem necessariamente aceder a outros módulos (Lyons, 2001). No entanto, é deixada em aberto a possibilidade de que um módulo possa ser penetrado em determinadas condições (ex. a atenção pode influenciar o input sensorial), para além de se considerar que os módulos possam não corresponder a sistemas de input. Provavelmente por esta razão e pelas implicações de natureza filosófica que o tema da modularidade tem, o termo módulo raramente é utilizado em ciência cognitiva, falando-se antes em sistemas e processos mas que correspondem a uma perspectiva de modularidade geral assumida a priori (Plaut, 2003; Van Orden et al., 2001).

A ciência cognitiva em geral, e psicologia cognitiva e a neuropsicologia cognitiva em particular, têm recolhido ampla evidência em favor de uma arquitectura funcional da mente em termos de sistemas ou módulos através das chamadas 'dissociações de desempenho' ou simplesmente 'dissociações'. Na verdade, embora a descrição dos sistemas/processos mentais seja o objectivo da ciência cognitiva, estes não são directamente observáveis (Dunn & Kirsner, 2003). Em vez disso a sua existência e função têm que ser inferidas a partir do desempenho observável (i.e. mensurável) em uma ou mais tarefas e da comparação entre diferentes condições definidas em termos de variáveis independentes associadas às tarefas e seus materiais (ex. dificuldade da tarefa, modalidade sensorial de apresentação dos materiais) e/ou variáveis independentes associadas aos sujeitos que realizam as tarefas (ex. idade, nível escolar, tipo de patologia). Finalmente, é o registo de 'dissociações' de desempenho em função destas comparações que tem sido utilizado como principal ferramenta metodológica para inferir a existência de sistemas e/ou processos mentais separados. O que constitui então uma 'dissociação'?

Na sua forma mais simples uma dissociação corresponde à situação observável em que uma variável ou factor afecta selectivamente o desempenho numa tarefa A mas não numa tarefa B (McCloskey, 2003; Shallice, 1988). A diferença observável pode ser fruto de uma manipulação relativa à tarefa (ex. número de letras afectando o tempo de leitura mas não o tempo de compreensão de uma palavra) ou relativa aos sujeitos (ex. paciente com uma lesão temporal superior afectando a nomeação de animais mas não a nomeação de objectos e grupo controlo com desempenho normal nas duas tarefas), sendo inferido que as tarefas (ex. leitura vs. compreensão;

nomeação de animais vs. nomeação de objectos) envolvem sistemas de processamento diferentes.

A este tipo de dissociação designado de 'dissociação simples' tem sido contraposto um segundo tipo designado de 'dupla dissociação' que corresponde em termos gerais à conjunção de duas dissociações simples complementares. Assim, temos uma primeira variável ou factor que afecta selectivamente o desempenho numa tarefa A mas não numa tarefa B, ao qual acresce uma segunda variável que afecta selectivamente o desempenho numa tarefa B mas não numa tarefa A. Estendendo os exemplos já apresentados teríamos por exemplo no primeiro caso:

> *Número de letras da palavra* afectando o seu tempo de leitura mas não o seu tempo de compreensão;
>
> *Familiaridade com a palavra* afectando o seu tempo de compreensão mas não o seu tempo de leitura.

E, no segundo caso:

> *Paciente com lesão temporal superior* e nomeação de objectos superior à nomeação de animais;
>
> *Paciente com lesão temporal média* e nomeação de animais superior à nomeação de objectos.

Tal como nas dissociações simples a conclusão é uma vez mais de que as duas tarefas terão subjacentes sistemas de processamento diferentes. Em ambos os casos a lógica é aparentemente clara mas na realidade o seu valor de evidência para a inferência de processos, sistemas ou módulos separados tem sido cada vez mais contestado e debatido. Neste âmbito, várias subcategorias têm ainda sido consideradas dentro destes dois tipos de dissociações (Dunn & Kirsner, 1988; Shallice, 1988; Sternberg, 2003), mas não são consensuais entre a generalidade dos autores. Por esta razão, restringimos a presente discussão às dissociações simples e duplas.

Começando pelas dissociações simples, é reconhecido consensualmente que o seu valor de prova para múltiplos sistemas não é decisivo. No entanto, apesar deste reconhecimento abundam na literatura exemplos da utilização das dissociações simples justamente nestes termos. Em muitos casos, o erro de inferência é mesmo agravado pela inferência automática de dois sistemas de processamento definidos um pela negativa do outro, o que Bedford (1997) designa pela falácia do não-sistema. Muitos exemplos podem ser chamados para ilustrar este erro de lógica onde uma descoberta de um sistema é transformada em dois sistemas. O primeiro sistema (ex.

memória explícita) é descrito em termos de atributos positivos, por aquilo que é, enquanto o segundo sistema é descrito por oposição ao primeiro, ou seja, é descrito por aquilo que não é (ex. memória implícita).

Este tipo de lógica está muitas vezes associado à descoberta de défices de funcionamento selectivos. Na verdade, se a investigação identifica um défice selectivo ou uma variável que associa a um sistema, então este será candidato a constituir-se como um processo/sistema independente e significativo. No entanto, tal não sugere que tudo o resto que o sujeito é capaz de fazer seja também um processo/sistema independente e significativo. O exemplo dado por Bedford (1997) a este respeito é particularmente elucidativo. O facto de ao remover-se o fígado de um rato, ele morrer devido à acumulação de toxinas no sangue, parece conduzir logicamente à inferência de que este sistema independente é responsável por esta função. No entanto, tal não permite a inferência de que tudo o resto que o rato consegue fazer (antes de morrer) seja da responsabilidade de um outro órgão complementar ao primeiro! O mesmo se passa com a inferência de sistemas ou processos. O facto de identificarmos um sistema de memória explícita não nos permite definir por oposição um outro implícito que pode ou não existir ou corresponder a mais do que um sistema. Este argumento não deve ser confundido como contra a existência de categorias heterogéneas ou fraccionáveis mas antes contra a ideia de que uma categoria possa ser estabelecida por negação de uma outra e de que a conjunção de ambas constitua um todo significativo.
Mas as críticas a esta lógica de investigação também se estendem ao próprio estabelecimento do sistema/processo positivo. Na verdade, o facto de que após um défice selectivo em X determinada função se perde, apenas nos permite dizer que X é candidato a constituir-se com um sistema/processo significativo. No entanto, tal é necessariamente verdade pois uma dissociação simples pode ser explicada e em muitos casos corresponde a um único sistema de processamento subjacente. Veja-se a este respeito o exemplo de Chater (2003). Suponham um tempo na antiguidade em que nada sabíamos sobre a fisiologia do sistema digestivo e encontrávamos um indivíduo que era capaz de digerir tudo menos camarões, segundo esta lógica poderíamos inferir um sistema digestivo especializado para este alimento que estaria deficitário neste indivíduo. Obviamente que sabemos que esta explicação não só não decorre necessariamente (diferentes alimentos podem simplesmente ser de mais difícil digestão) como é falsa. O mesmo se passa ao nível dos processos/sistemas cognitivos.

Voltando aos exemplos dados, o facto simples de uma lesão temporal superior afectar a nomeação de animais mas não de objectos não significa necessariamente que teremos um sistema de processamento distinto para cada uma das categorias. Na verdade, as categorias podem ser processados por um mesmo sistema, sendo a dissociação explicada pelo facto de os animais serem simplesmente mais difíceis de identificar e nomear do que os objectos. Assim, registando-se uma lesão do sistema de nomeação que não é total, os itens à partida mais difíceis serão mais afectados do que os itens mais fáceis.

Para além disto, a inferência positiva parte também do pressuposto que tudo funciona à excepção de um sistema que foi afectado. Em muitos casos não podemos realmente saber ou será mesmo pouco provável que um determinado défice corresponda apenas a um sistema. Um défice de desempenho observado poderá realmente corresponder a dois ou mais sistemas independentes afectados ou até à quebra de ligação entre eles. Finalmente, o facto de que sem X algo não funciona e então X deverá ter essa função acaba por constituir uma lógica circular que não nos fornece evidência que X funciona por si só enquanto sistema independente (Bedford, 1997). Para isso, e voltamos à modularidade, temos que partir do pressuposto que quando o sistema global se quebra isso acontece ao nível de sistemas significativos e não simplesmente a nível de componentes de sistemas. Isto pode parecer evidente ao nível de órgãos do corpo humano mas poderá não o ser ao nível de sistemas cognitivos (Bedford, 1997; Lyons, 2001; Plaut, 2003).

A isto tudo acresce o facto de que na maioria dos casos as dissociações simples observadas não são tudo ou nada ou correspondem a casos puros o que, apesar de tudo, consistiria uma evidência mais forte. Na verdade, na maioria dos casos o que se observa é um padrão de resultados em que existe um desempenho muito melhor numa situação do que em outra mas que no entanto não corresponde a um desempenho nulo (ex. casos de lesão temporal superior associada ao défice de nomeação de animais não significam incapacidade para nomear todos os animais).

Desde dos anos cinquenta, em especial com Teuber (1955), foi reconhecido que a validade da inferência não estava garantida nestes casos. Independentemente do facto de que uma interpretação em termos de múltiplos processos ou sistemas possa ser a explicação mais plausível de uma determinada dissociação simples, este padrão de resultados não implica a rejeição absoluta de um modelo alternativo em termos de um processo ou sistema único. Foi o mesmo Teuber (1955) que introduz o conceito de

dupla dissociação como forma de ultrapassar as limitações apontadas à dissociação simples no sentido de permitir o controlo de condições presentes em alguns estudos mas ausentes em outros. De forma particular, as duplas dissociações eliminam interpretações em termos de um único sistema associadas a diferenças de processamento de materiais. Na verdade, se um determinado défice selectivo faz com que a função X desapareça mas não a função Y será tentador concluir que X e Y são desempenhados por sistemas de processamento independentes. No entanto tal como Teuber (1955) o assinalou, X e Y podem ser desempenhadas por um mesmo sistema de forma que se X desaparece e não Y com um défice é apenas porque X é simplesmente mais difícil/complexo/exigente do que Y. Com uma dupla dissociação esta explicação alternativa em termos de dificuldade de processamento cai por terra, pois ao encontrarmos a situação inversa em que Y desaparece mas não X, teríamos que tirar a conclusão contrária de que Y é mais difícil/complexo/exigente do que X. Não podemos concluir ao mesmo tempo que os animais são mais difíceis de nomear que os objectos e que estes últimos são mais difíceis de nomear que os animais.

Se é verdade que as duplas dissociações permitem eliminar este eventual artefacto, elas não asseguram porém que estejamos em presença de dois sistemas de processamento independentes. O próprio Teuber (1955) menciona o problema de equacionar sintomas com funções ou seja a reificação de uma dissociação entre tarefas como uma dissociação entre funções. Voltando uma vez mais ao exemplo do sistema digestivo de Chater (2003), imaginem que para além do nosso indivíduo que é incapaz de digerir camarões mas não apresenta qualquer problema para ostras encontramos um outro que digere os camarões mas não as ostras – poderemos então concluir pela existência de dois sistemas digestivos especializados na digestão de cada um destes alimentos? Sabemos que não e o mesmo se passa ao nível dos sistemas cognitivos em que uma dupla dissociação pode simplesmente reflectir outras dimensões subjacentes que a "olho nu" serão responsáveis pela separação de desempenhos.

Um exemplo disso mesmo é o modelo proposto por Farah e McClelland (1991) para explicar a aparente dupla dissociação entre animais (e mais geralmente seres vivos) e objectos (e mais geralmente artefactos). A sua proposta parte da ideia de Warrington e Shallice (1984) de que a informação no sistema semântico estaria organizada segundo atributos sensoriais e atributos funcionais, os primeiros determinantes para a identificação de seres vivos e os segundos determinantes para a

identificação de objectos. Os atributos sensoriais estariam representados numa proporção muito mais elevada para os seres vivos e também corresponderiam ao maior número de atributos para as duas categorias. Esta composição e rácio determinaria o papel dos dois atributos na identificação dos elementos das duas categorias e seria assim responsável pelos défices observados.

Este modelo que apenas chamo aqui a título exemplificativo é também uma resposta ao facto de que as duplas dissociações registadas no domínio da nomeação não serem absolutas nem totalmente complementares o que, tal como no caso das dissociações simples as tornaria uma evidência mais forte. Ao assumir-se esta dimensão subjacente e pensarmos que nem todos os exemplares de animais e objectos serão igualmente típicos das suas categorias em termos de atributos subjacentes permite explicar padrões de resultados relativos em que o que se regista é simplesmente um desempenho muito melhor numa categoria em relação a outra (Marques, 2000). Na verdade, esta limitação apontada às dissociações simples está também presente nas duplas dissociações. Quando X é dissociado de Y, e Y dissociado de X, os dois parecem complementares mas na realidade só raramente se apresentam como exactamente complementares. E se isto não acontece estamos então na presença de duas dissociações simples, não se eliminando mas antes duplicando a falácia do não-sistema (Bedford, 1997). A este facto acresce que, a não ser que previamente saibamos que o espaço que queremos caracterizar consiste apenas de dois itens X e Y – que é o que habitualmente queremos provar – não podemos concluir nada sobre a coerência de não-X (ou Y) como processo a partir do isolamento de X (Baddeley, 2003; Bedford, 1997).

Como já vimos uma dupla dissociação implica um planeamento experimental de dois indivíduos ou grupos por duas condições (ou mais geralmente de duas variáveis independentes com dois níveis), se o nosso modelo corresponder a três sistemas precisamos então de uma tripla dissociação e um planeamento três por três, o que em casos em que não temos nenhum modelo teórico consistente à partida nos leva a colocar a questão de qual o planeamento escolher, se um destes ou ainda um mais complexo (Baddeley, 2003).

O princípio da parcimónia é aqui habitualmente invocado para guiar as escolhas mas deve estar também presente na interpretação das duplas dissociações quando consideramos uma explicação em termos de um sistema ou processo único. No entanto, em muitos casos os autores parecem optar mais rapidamente pela

explicação em termos de múltiplos sistemas/processos. A este respeito, vários autores têm fornecido estratégias concretas para colocar as duas inferências em pé de igualdade.

Em várias situações será mais plausível pensar que os mecanismos subjacentes ás tarefas dissociadas são praticamente coincidentes com a excepção de um pequeno componente específico de cada tarefa ou categoria de informação, o que parece ser claramente o caso do nosso exemplo digestivo (Chater, 2003). Em outras situações poderá dar-se o caso do sistema subjacente ás tarefas dissociadas ser o mesmo e que o seu défice geral ou num seu determinado componente afecte selectivamente uma das tarefas, simplesmente porque no caso de uma delas existe um outro sistema que permite realizá-la. Imaginem as tarefas de escrita no quadro e de apanhar objectos no meu caso que sou destro. Se partir o braço torno-me incapaz de realizar a primeira mas consigo realizar a segunda pois tenho a mão esquerda como sistema alternativo. Por outro lado se partir uma perna torno-me incapaz de realizar a segunda pois não consigo deslocar-me mas continuo a realizar a primeira e assim esta dupla dissociação observada não corresponde na realidade a sistemas motores independentes mas a diferenças de sistemas alternativos de apoio às tarefas (Chater, 2003).

Na neuropsicologia cognitiva a interpretação das duplas dissociações tem também de ter em conta o facto da investigação contrastar casos de pacientes singulares e assim a variabilidade individual ser um potencial factor explicativo. Por exemplo, existe ampla evidência de que a profissão e os interesses dos indivíduos que sofrem uma lesão cerebral pode enviesar a sua recuperação de funções de forma consistente com essa experiência anterior (Andrewes, 2003). Assim, a identificação de uma dupla dissociação a partir de dois casos singulares deverá também investigar até que ponto os resultados não poderão estar associados em um ou nos dois casos à experiência pré-lesão (ou mesmo pós-lesão) dos pacientes (Andrewes, 2003). Um exemplo mais geral e claro tem a ver com as diferenças entre sexos. Na verdade, sabe-se claramente que comparações entre casos singulares de indivíduos de sexos diferentes podem estar a apontar para diferenças relativas a dimorfismo cerebral em vez de diferenças gerais de processamento ou de sistemas.

Os modelos conexionistas em que é possível simular o efeito de lesões também mostram outra limitação do ênfase em casos singulares relativamente à variabilidade individual (Juola, 2000; Juola & Plunkett, 2000). Isto acontece porque vários investigadores apressam-se a inferir múltiplos processos/sistemas a partir de

um só caso ou de dois casos na literatura. Os estudos com modelos conexionistas com um único sistema ou mecanismo atractor mostram que a sua lesão em termos aleatórios pode resultar algumas vezes em nada de relevante (correspondendo à generalidade dos casos que não são relatados na literatura) e outras vezes, mais raras, pode resultar em défices de desempenho complementares. Deste modo, o registo de uma dupla dissociação, particularmente a partir de casos patológicos muito raros, não constitui por si só evidência suficiente para a conclusão de sistemas ou processos separados.

Estas limitações, entre outras levam alguns autores a contestar o carácter heurístico de todas dissociações e mesmo o pressuposto de modularidade que lhes está subjacente (Medler et al., in press; Plaut, 2003; Van Orden et al., 2001; Van Orden & Kloos, 2003). Na verdade, ao inferirmos das dissociações sistemas/módulos subjacentes temos que também assumir a priori que as dissociações identificadas correspondem em primeiro lugar a um e um só módulo subjacente. Isto é os casos singulares deverão ser casos puros. Para além disso, as tarefas em que se verificam as dissociações deverão também ser processualmente puras para podermos realizar com toda a certeza a inferência pretendida. E o problema destas condições é que não há forma independente de as verificar pois dependem da teoria que propõe esses mesmos módulos, incorrendo num problema de circularidade (Van Orden et al., 2001). No limite, este problema torna possível a defesa perpétua das teorias quer argumentando que os novos casos isolados não são puros, quer incorporando as novas dissociações na teoria e postulando sub-módulos cada vez mais específicos. Como podemos considerar que duas tarefas nunca recrutam exactamente as mesmas funções subjacentes as dissociações serão assim triviais pondo em causa a sua utilidade já que acabaríamos por necessitar do mesmo número de funções mentais quantas as tarefas existentes para realizar (Van Orden et al., 2001).

Consideradas as limitações e contestações apresentadas tanto às dissociações quanto aos seus pressupostos subjacentes analisemos então as soluções que têm sido contrapostas. São várias as soluções e podemos organizá-las em dois grandes grupos tendo em conta aqueles que não negam o pressuposto de um organização modular e os que consideram que esta organização não deve ser um pressuposto de base.

No grupo “modular”, encontramos em primeiro lugar aqueles que consideram que as objecções às duplas dissociações são em parte falsas porque partem de um pressuposto geral de que os cientistas as utilizam no sentido de justificar ou obrigar à

existência de módulos separados (ex. Coltheart & Davies, 2003; Sternberg, 2003). Pelo contrário, estes autores consideram que a ideia dos cientistas relativamente a esta categoria de evidência como para outras é de que a sua observação apoia mas não determina a teoria, e que o valor desse apoio, depende entre outras coisas, da plausibilidade de outras teorias consistentes com os dados. Chamam também à atenção que será errado considerar que a generalidade dos autores advogam a existência de casos puros ou, que muitos deles (embora hajam aqui mais excepções) advogam tarefas processualmente puras (Coltheart & Davies, 2003). Pelo contrário, cada vez mais se assiste à procura de índices que permitam isolar a contribuição de vários processos numa mesma tarefa (ex. Jacoby, 1991, 1998).

Outros autores enfatizam também o facto da contestação se fazer a um nível abstracto que não está em causa na prática da investigação (ex. Chater, 2003; McCloskey, 2003). Ou seja, concordam que as dissociações enquanto princípio abstracto não permitem revelar a estrutura cognitiva mas consideram também que elas não podem ser avaliadas enquanto tal, ou seja, a evidência destas e de outras categorias como as correlações ou associações não podem ser avaliadas no vácuo (Chater, 2003; McCloskey, 2003). Eliminar a priori uma determinada categoria de evidência que permite ligar resultados e teoria é que não será possível nem desejável (McCloskey, 2003).

É assim que dentro do grupo "modular" a generalidade dos autores advoga a continuação do uso das dissociações enquanto categoria de evidência mas atribuindo-lhes um peso decididamente inferior ao que pelo menos alguns autores lhes atribuíam. Neste sentido, alguns (ex. Lyons, 2003) defendem a ideia de que as dissociações permitem inferir dois sistemas disjuntos deve ser substituída pela inferência de dois sistemas distintos, embora podendo partilhar vários componentes. Passa-se assim para uma ideia de que a independência de sistemas é relativa em vez de absoluta e para a questão de qual o possível grau de independência entre vários sistemas. No mesmo sentido vão outras propostas de que as dissociações não permitem inferir a existência de funções mentais separadas mas simplesmente permitem saber mais sobre as propriedades de funções mentais que tenham antes sido estabelecidas a partir de outros dados teóricos ou empíricos (Dunn, 2003). Outros colocam esta mesma solução a um nível mais geral, dizendo apenas que as dissociações como outro tipo de evidências colocam constrangimentos à teorização e assim são relevantes pois

eliminarão todas as classes de modelos que são incapazes de as reproduzir/explicar (Baddeley, 2003; Chater, 2003).

Deste modo, muitos consideram que as dissociações não devem ter um estatuto de evidência especial mas não devem ser eliminadas enquanto categoria de evidência. Antes cada relação proposta entre dissociações e arquitectura funcional deverá ser avaliada em si mesma e à luz de outra evidência experimental, neuropsicológica e computacional (Chater, 2003). Mais ainda, este tipo de evidência deve ser analizada tendo em conta as questões teóricas particulares em avaliação, as diferentes respostas alternativas às questões que foram propostas e o tipo de evidência específica observada. As inferências para um único sistema vs. mais do que um sistema ou processo subjacente devem assim ser consideradas simplesmente como competindo para a inferência à melhor explicação num conjunto de evidências que deverão múltiplas (Coltheart & Davies, 2003).

O aspecto da procura de evidências múltiplas e convergentes é também enfatizado por vários autores do grupo "não modular" (Juola, 2000; Juola & Plunkett, 2000) que chamam à atenção que as duplas dissociações de poucos casos em si não são significativas a não ser que se controle a possibilidade de tais diferenças serem simplesmente aleatórias. Para isso sugerem que em primeiro lugar se considerem múltiplos casos para análise e se comparem expectativas de desempenho com o desempenho observado e distribuições de resultados obtidos. Assim, mesmo para este segundo grupo as dissociações continuam a ser consideradas como relevantes, embora considerando-se que elas não permitem por si só a opção por uma outra perspectiva. No âmbito de uma perspectiva não modular o sistema cognitivo poderá não ser constituído por componentes discretos mas antes corresponder a um sistema distribuído. Finalmente, neste sistema as funções atribuídas a componentes individuais num sistema modular estão distribuídas por múltiplos grupos ou unidades ou mesmo por todo o sistema. Assim será possível pensar que, embora todo o sistema participe no processamento de cada estímulo, diferentes partes do sistema tenham contribuições únicas ou sejam diferencialmente importantes para aspectos particulares do desempenho de uma tarefa, havendo assim uma especialização funcional do sistema (Medler et al., in press; Plaut, 2003). Esta especialização de determinadas partes do sistema pode conduzir a défices relativamente selectivos após lesão dessas mesmas partes. Assim, embora se conteste o valor de dissociações baseadas na variabilidade ou variância das lesões que podem simplesmente traduzir efeitos

idiossincráticos, lesões que traduzem efeitos médios nos sistemas resultando da análise de múltiplos casos podem ser informativas sobre uma especialização funcional, mesmo que esta não corresponda à estrutura do sistema de uma forma transparente como as teorias modulares tipicamente assumem (Medler et al., in press; Plaut, 2003).

Neste contexto de perspectiva não modular (mas também com aplicação a uma perspectiva modular) alguns autores (Lyons, 2003) têm sugerido que se olhe mais para os desempenhos que permanecem intactos em face da lesão do sistema em vez dos défices que daí decorrem. Na verdade, e, com maior certeza, olhar para o que permanece em face da lesão do sistema permite ir isolando partes que não são necessárias para determinada tarefa e assim função (ou funções) que lhes são consideradas subjacentes. No caso de uma rede conexionista, embora seja difícil convencer que no caso de uma lesão provocar um défice tal significa um sistema independente (cf. Juola & Plunket, 2000), o contrário, o caso de apesar de uma lesão o desempenho continuar normal, será difícil convencer que tal parte da rede não corresponderá a um sistema distinto. Deste modo, mesmo que a hipótese de modularidade seja falsa, a ciência cognitiva continua a ser possível e a evidência das dissociações continua a ser relevante.

Em conclusão, penso que as diferentes críticas apresentadas mostram claramente que a evidência das dissociações de desempenho, embora intuitivamente apelativa e de interpretação aparentemente directa e simples, esconde uma teia de possíveis interpretações que tem que ser devidamente apreciada tendo em conta as características específicas da evidência recolhida e procurando outra evidência corroborante.

Assim, independentemente da nossa perspectiva mais geral sobre a arquitectura funcional da mente humana, as dissociações de desempenho, embora não devam ser vistas como evidência mágica à qual atribuímos um estatuto especial, também não devem ser descontadas como irrelevantes.

A história da ciência cognitiva está cheia de exemplos de descobertas importantes identificadas a partir de dissociações (tanto simples, quanto duplas). Tenho a certeza de que se tivermos em conta as diferentes limitações assinaladas, evitando assim conclusões erradas, este tipo de evidência continuará a trazer à ciência cognitiva evidência substantiva para o conhecimento da mente humana.

Referências Bibliográficas

Andrewes, D. (2003). Double dissociation and the benefit of experience. *Cortex, 39*, 158-160.

Baddeley, A. (2003). Double dissociations: Not magic but still useful. *Cortex, 39*, 129-131.

Bedford, F. L. (1997). False categories in cognition: The not-the-liver fallacy. *Cognition, 64*, 231-248.

Chater, N. (2003). How much can we learn from double dissociations. *Cortex, 39*, 167-169.

Coltheart, M., & Davies, M. (2003). Inference and explanation in cognitive neuropsychology. *Cortex, 39*, 188-191.

Dunn, J. C. (2003). The elusive dissociation. *Cortex, 39*, 177-179.

Dunn, J. C., & Kirsner, K. (1988). Discovering functionally independent mental processes: The principle of reversed association. *Psychological Review, 95*, 91-101.

Dunn, J. C., & Kirsner, K. (1988).What can we infer from double dissociations? *Cortex, 39*, 1-7.

Farah, M. J., & McClelland, J. L. (1991). A computational model of semantic memory impairment: Modality specificity and emergent category specificity. *Journal of Experimental Psychology: General, 120*, 339-357.

Fodor, J. (1983). *The modularity of mind*. Cambridge MA: MIT Press.

Jacoby, L. J. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30*, 513-541.

Jacoby, L. J. (1998). Invariance in automatic influences of memory: Toward a user's guide for the process-dissociation procedure. *Journal of Experimental Psychology: Learning, Memory and Cognition, 24*, 3-26.

Juola, P. (2000). Double dissociations and neurophysiological expectations. *Brain and Cognition, 43*, 257-262.

Juola, P., & Plunkett, K. (2000). Why double dissociations don't mean much. In G. Cohen, R. A. Johnson, & K. Plunkett (Eds.), *Exploring cognition: Damaged brains and neural networks* (pp. 319-327).

Lyons, J. C. (2001). Carving the mind at its (not necessarily modular) joints. *British Journal of Philosophy of Science, 52*, 277-302.

Lyons, J. C. (2003). Lesion studies, spared performance and cognitive systems. *Cortex, 39,* 145-147.

Marques, J. F. (2000). The "living things" impairment and the nature of semantic memory organization: An experimental study using PI-release and semantic cues. *Cognitive Neuropsychology, 17,* 683-707.

McCloskey, M. (2003). Beyond task dissociation logic: A richer conception of cognitive neuropsychology. *Cortex, 39,* 196-2o2.

Medler, D. A., Dawson, M. R. W. Dawson, & Kingstone, A. (in press). Functional localization and double dissociations: The relationship between internal structure and behavior. *Brain and Cognition*.

Plaut, D. C. (2003). Interpreting double dissociations in connectionist networks. *Cortex, 39,* 138-141.

Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge: Cambridge University Press.

Sternberg, S. (2003). Process decomposition from double dissociation of subprocesses. *Cortex, 39,* 180-182.

Terry, A. (2003). *História concisa da filosofia ocidental* (D. Murcho trad.). Lisboa: Temas e Debates (obra original publicada em 1998).

Teuber, H. L. (1955). Physiological psychology. *Annual Review of Psychology, 6,* 267-296.

Van Orden, G. C., & Kloos, H. (2003). The module mistake. *Cortex, 39,* 164-166.

Van Orden, G. C., Pennington, B. F., & Stone, G. O. (2001). What do double dissociations prove? *Cognitive Science, 25*, 111-172.

Warrington, E. K., & Shallice, T. (1984). Category-specific impairment. *Brain, 107*, 829-853.

8
What Reference Has to Tell us about Meaning[1]

Stephen Schiffer
University of New York
stephen.schiffer@nyu.edu

The first thing reference has to tell us about meaning is that many of the propositions expressed by utterances containing referring expressions appear to have what I call *the relativity feature*. A proposition has the relativity feature provided it's an *x*-dependent proposition the entertainment of which requires different people, or the same person at different times or places, to think of *x* in different ways. The fact that some propositions appear to have the relativity feature creates a puzzle, for having that feature is incompatible with all currently dominant theories of propositions. The puzzle deepens when we appreciate that these theories can't even explain away the appearance that some propositions have the relativity feature; but then it gets resolved when we see that there is an independently motivated theory of propositions—one I call the theory of pleonastic propositions—which does accommodate the relativity feature. In the end, what reference has to tell us about meaning is that pleonastic propositions are the propositions we say and believe.

I. Starting Points

I begin with what I take to be some obvious truths (some perhaps more obvious than others).

(1) You and I both believe that I'm old enough to vote.

(2) Similarly, if I utter the sentence 'I'm old enough to vote' in response to a question about my age, then in uttering 'I'm old enough to vote' I'm saying that I'm old enough to vote.

(3) The fact that you and I believe that I'm old enough to vote entails that there's something which we both believe—to wit, that I'm old enough to vote;

1 An earlier version of this material was presented in a seminar on reference that Stephen Neale and I gave at NYU in the spring semester of 2004. I'm indebted to the discussion of this material in the seminar, especially to the remarks of Ray Buchanan, Paul Elbourne, Paul Horwich, Stephen Neale, and Anna Szabolcsi. An earlier version of this paper was given as a talk at the University of Lisbon in July 2004, and this version benefited from the discussion on that occasion, especially from the remarks of João Branquinho, Adriana Silva Graça, and Teresa Marques.

and the fact that I said that I'm old enough to vote entails that there is something that I said—to wit, that I'm old enough to vote.

(4) This thing, *that I'm old enough to vote*, which you and I believe and I said, is such that:

- it's *abstract*: it has no mass or energy or anything else that would make it a physical object;
- it's *mind- and language-independent*: it wasn't created by anything anyone said or thought, and while it can be expressed in just about any language, it itself belongs to no language;
- it has a *truth condition*: *that I'm old enough to vote* is true iff I'm old enough to vote;
- it has its truth condition *essentially*: it's a necessary truth that *that I'm old enough to vote* is true iff I'm old enough to vote; and
- it has its truth condition *absolutely*, without relativization to anything else.

From all this we may conclude, by a straightforward generalization, that the things we believe and say are what philosophers nowadays call *propositions*: abstract, mind- and language-independent entities that have truth conditions, and have those truth conditions both essentially and absolutely. Whether or not it's obvious, I'll take it as a working hypothesis of this paper.[2] A related working hypothesis, a challenge to which I will consider later, is that propositional-attitude and propositional-speech-act reports of the form '*A V*s that *S*' are true just in case the referent of '*A*' stands in the relation expressed by '*V*' to the proposition to which 'that *S*' refers. For example, my utterance of 'I believe that I'm old enough to vote' is true just in case I stand in the belief relation to the proposition that I'm old enough to vote, and my utterance of 'I said that I'm old enough to vote' is true just in case I stood in the saying relation to the proposition that I'm old enough to vote (I'm being relaxed about the representation of tense in logical form).

Given the assumption that saying is a relation to propositions, it will help in what follows to have before us what we may call the *semantic content* of an utterance. Suppose *A* says to *B*, 'It's going to rain', and means thereby both that it's going to rain in Central Park that evening and that the outdoor concert they were hoping to attend will be postponed. Then the proposition that it's going to rain in Central Park

2 I argue for this hypothesis in Schiffer (2003).

that evening, but not the proposition that the outdoor concert they were hoping to attend will be postponed, stands in a certain relation of "fit" to the meaning of the sentence type 'It's going to rain' in a way that makes it appropriate to say that it's *the semantic content* of *A*'s utterance. The idea, rendered more explicitly, is that the meaning of an indicative sentence type constrains what a speaker must mean in uttering the sentence, if she's to be speaking literally—that is, in conformity with the sentence's meaning. For example, an utterance of 'It's raining' is literal only if, roughly speaking, for some place π to which the speaker implicitly refers in producing her utterance, she means that it's raining at π. At a certain level of abstraction, it's harmless to think of the meaning that carries this constraint as a propositional form or template. If a propositional form *P* is the meaning of the sentence type σ, then an unembedded utterance of σ is literal just in case, for some proposition *q* of form *P*, the speaker means *q* in her utterance of σ.[3] In the event, I'll say that *q* is the *semantic content* of the utterance of σ. If *p* is the semantic content of *x*'s unembedded utterance of σ, then *x* said *p* in uttering σ; but it doesn't follow from its being true that *x* said *p* in uttering σ, that *p* is the semantic content of *x*'s utterance of σ. For example, Al, who has no idea who invented Velcro, utters 'The inventor of Velcro was a genius', and says thereby that the inventor of Velcro was a genius. You and Beth, however, know that George de Mestral was the man who in fact invented Velcro, and in appropriate circumstances you may be counted as speaking truly in saying to Beth, 'Al said that George de Mestral was a genius'. It would then be true that in uttering 'The inventor of Velcro was a genius', Al said both that the inventor of Velcro was a genius and that George de Mestral was a genius, but only the former proposition would be the semantic content of his utterance.

II. The Relativity Feature

Many of the propositions we believe and say are *object-dependent propositions*. More exactly, an *x*-dependent proposition is a proposition that's partly individuated with respect to *x* and that wouldn't exist if *x* didn't exist. For example, the proposition that I'm old enough to vote is a Schiffer-dependent proposition that's true in an arbitrary possible world just in case I'm old enough to vote in that world. But there's more to the proposition than that.

3 An utterance of σ is unembedded when the uttered token of σ isn't a component of a larger sentence token.

The further feature to which I allude is one I call *the relativity feature*. A proposition has this feature provided it's an *x*-dependent proposition the entertainment of which requires different people, or the same person at different times or places, to think of *x* in different ways. The proposition that I'm old enough to vote arguably has the relativity feature:

- In order for me to believe what I say when I say 'I'm old enough to vote', it's not enough for me to believe of some person who happens to be me that he's old enough to vote. I must also think of that person *as myself*; I must think of myself in a way that only I can think of myself—namely, in the self-conscious way of thinking of oneself associated with the first-person singular pronoun 'I'.
- You also believe what I say when I say 'I'm old enough to vote', but you aren't required to think of me under the self-conscious mode of presentation, which, of course, is a good thing, since it's impossible for you to think of anyone other than yourself in that way. (Roughly speaking, in our present context you must think of me under a demonstrative mode of presentation which enables you to say to me, 'You're old enough to vote.')

In other words, we have this plausible argument to show that the proposition that I'm old enough to vote has the relativity feature:

(1) You and I both believe the proposition that I'm old enough to vote.

(2) A necessary condition of my believing that proposition is that I think of me in the self-conscious way, but it's not a necessary condition of your believing that proposition that you think of me (*per impossibile*) in the self-conscious way.

(3) ∴ The proposition that I'm old enough to vote has the relativity feature.

This little inference is valid, given the definition of the relativity feature, and, I submit, each of its two premises is plausible.

The proposition *that you are F* also seems to have the relativity feature. In order for you to understand what I say to you when I utter 'You're delirious', you must evidently think of yourself in the self-conscious way, so that you can correctly report 'Schiffer said that I'm delirious'. But of course that's not how I must think of you when I say that you're delirious.

In the case of the propositions that I am *F* and that you are *F*, the saliently operative clause in the definition of the relativity feature is the one pertaining to different people, but there are other propositions which appear to have the relativity feature where the phrase 'or the same person at different times or places' is the operative clause:

- One class of examples involves adverbial temporal demonstratives such as 'now', 'today', and 'yesterday'. Today Jane utters 'It's sunny today' and thereby says that it's sunny today. The token of 'today' in Jane's utterance refers to a certain day *d*, and those who are party to the utterance on *d* must think of *d* in a certain way if they're to understand or believe what she said. This is the especially direct way of thinking of a day when one thinks of it as, so to say, "this very day, the day it is now." But someone can also believe or understand what Jane said in uttering 'It's sunny today' on a day other than the day on which her utterance occurred. As Gareth Evans remarked, a man reading yesterday's newspaper can believe and understand what was said when he reads 'The Prime Minister is holding a cabinet meeting today'.[4] If one knows that the day to which Jane referred was the day before one's current day, then one must think of that day in the way one thinks of a day when one thinks of it as, so to say, "the day before this very day."
- A similar class of examples involves adverbial spatial demonstratives such as 'here' and 'there'. For example, your friend calls you in New York from Lisbon and says 'The weather is lovely here'. In order for your friend to know what she's saying, she must think of the place to which she's referring in the uniquely direct egocentric way in which one thinks of a place when one thinks of it as, so to say, "this place where I am now." But that is not a way you can think of the place to which she's referring, even though you know that she's saying the proposition to which she would refer in saying 'I said that the weather is lovely here'. Similarly, when your friend returns to New York and thinks of what she said to you from Lisbon, she can no longer think of the place to which she referred by her utterance of 'here' in the same way she did when she uttered the sentence 'The weather is lovely here'. If she were now to report what she said to you from Lisbon, she couldn't use the demonstrative

4 Evans (1985a: 306).

'here' again. She'd now say, 'I said that the weather was lovely there', and though she'd be referring to the proposition she said in Lisbon, she'd now have to think of the place to which she referred on the two occasions in a way appropriate to her using 'there', a mode of presentation which identifies that place as one she occupied in the past.

- Actually, the relativity feature can be had by the semantic content of an utterance involving almost any kind of singular term. In a Manhattan restaurant, I direct my companion's attention to a woman at a nearby table and say, 'That's Monica Lewinsky'. Here it's reasonable to suppose that the semantic content of my utterance requires me to think of the person to whom I referred under a current perceptual mode of presentation. Now suppose that hours later I think to myself, 'That was Monica Lewinsky'. As Gareth Evans has forcefully argued,[5] the belief expressed in this utterance isn't one I acquired after acquiring the belief I acquired in the restaurant. Rather, it's the same belief, *a belief I acquired in the restaurant and still retain*. It's therefore reasonable further to assume that the proposition I say in uttering 'That was Monica Lewinsky' is the same as the one I said in my earlier utterance of 'That's Monica Lewinsky'. Yet the proposition expressed in my utterance of 'That was Monica Lewinsky' can't require me to think of Monica Lewinsky under a current perceptual mode of presentation, since the mode of presentation deployed in that utterance is memory-based. The conclusion we are led to is that entertaining the single proposition expressed in both utterances can require a person to think of the referent of the two tokens of 'that' in one way at one time and in another way at another time.

The foregoing gloss of the relativity feature was intended to get the basic idea across, but it needs tightening.[6] We can see the need for this in the following way. I introduced the relativity feature in terms of the example in which I utter 'I'm old enough to vote', and I said that in order for me to believe what I said—viz. the proposition that I'm old enough—I must think of myself in the self-conscious, egocentric way of thinking of oneself associated with the pronoun 'I'. But that needs qualification. For suppose that at some time in the future I hear a recording of my utterance and believe what was said in it but don't recognize that I'm the speaker.

5 Evans (1985a) and (1982).

6 I'm indebted to Sarah Moss for forcing me to face up to the need for tightening.

The belief I form thereby is one that I would report by saying 'I believe that he, the speaker, is old enough to vote'. Here I understand the utterance and believe what's said in it, but I don't, and am not required to, think of myself in the first-person way (the same goes, of course, for any way of entertaining the proposition other than believing it). This demands a more precise statement of the relativity feature. My initial statement of the relativity feature was that

> A proposition has the relativity feature provided it's an *x*-dependent proposition the entertainment of which requires different people, or the same person at different times or places, to think of *x* in different ways.

That isn't wrong, as far as it goes, but we can tighten it this way:

> A proposition *p* has the relativity feature iff *p* is an *x*-dependent proposition such that there are properties Φ and Ψ and ways of thinking *w* and *w*′ of *p* such that (i) $w \neq w'$, (ii) a person who has Φ can entertain *p* only if she thinks of *x* in way *w*, and (iii) a person who has Ψ can entertain *p* only if he thinks of *x* in way *w*′.[7]

When I utter 'I'm old enough to vote', I have the property of thinking of myself as having produced that very utterance. By virtue of my having that property at the time of my utterance, I can at that time entertain the proposition I said in producing that utterance—viz. the proposition that I'm old enough to vote—only if I think of myself in the first-person way. But later when I think of that utterance as not having been produced by me, I don't have that property and it's not a condition of my then entertaining the asserted proposition that I think of myself in the first-person way. At the later time, when I fail to recognize myself as the speaker, the requirement for my entertaining the asserted proposition is the same as the one anyone else must meet: I merely need to think of me in a way that demonstratively identifies me as the speaker. It follows, then, that the proposition that I'm old enough to vote has the relativity feature, and we see how when entertaining that proposition *qua* utterer of 'I'm old enough to vote', I must think of myself in the egocentric way, but not when entertaining the proposition as something said by a person whom I fail to recognize as myself. But how is it so much as possible for a proposition to have such a relativity feature, or, indeed, any relativity feature? That is indeed the crucial question, but one whose answer can't be broached until later in this paper.

7 I'll leave it implicit that Φ and Ψ are properties people might actually instantiate.

So, one thing reference appears to tell us about meaning is that the propositions that are the semantic contents of utterances involving singular terms often have the relativity feature: they are *x*-dependent propositions that require different people, or the same person at different times or places, to think of *x* under different modes of presentations.

III. A Puzzle and Some Possible Resolutions

We have a puzzle. On the one hand, it's plausible that some propositions have the relativity feature. On the other hand, none of the currently dominant theories of propositions can accommodate that feature: if any of those theories is correct, no proposition can have the relativity feature. It's reasonable to suppose there are two possibilities: either one of those theories is correct and it can *explain away* the appearance that propositions have the relativity feature, or else some other account of propositions which can accommodate the relativity feature is correct. My interest in this section is in the possibility of explaining away the relativity feature. I won't need to consider every theory of propositions, since the crucial general point will emerge from a discussion of two of them and a certain other alternative.

Direct-reference semantics and the relativity feature

Direct-reference semantics is a leading theory of linguistic and mental content. According to this theory:

> (1) The propositional contents of our thoughts and speech acts are *Russellian propositions*: structured propositions whose basic components are the objects and properties our thoughts and speech acts are about.
>
> (2) Many singular terms—e.g. pronouns, simple demonstratives, and names—typically function as *directly-referential singular terms*, where a token of a singular term is directly referential provided its only contribution to the proposition expressed is its referent. It's customary for direct-reference theorists to represent the content of a token of ⌜*t* is *F*⌝ as the Russellian singular proposition <*x*, *F*ness>, where *x* is the referent of the token of the

singular term *t* contained in the utterance.[8] Necessarily, <*x*, *F*ness> is true iff *x* instantiates *F*ness, false otherwise.

Thus, according to direct-reference semantics, the semantic content of any literal utterance by me of 'I'm old enough to vote' is (simplifying a little) the singular proposition <SS, the property of being old enough to vote>.

It's obvious that direct-reference semantics can't accommodate the relativity feature. The singular proposition <SS, the property of being old enough to vote >, for example, doesn't require me to think of myself in any particular way when I entertain it. If the direct-reference theory is correct, it's the same proposition as the one expressed by 'Stephen Schiffer is old enough to vote', and, since I might not believe that I'm Stephen Schiffer, I can believe what that sentence says without thinking of myself in the self-conscious way.

None of this is news to the direct-reference theorist. She will deny that the proposition that I'm old enough to vote—or, for that matter, any other proposition—has the relativity feature, and she'll claim that she can account for the data that might seem to suggest otherwise.

She'll begin by claiming that we need to distinguish two questions about my utterance of 'I'm old enough to vote':

(a) How must I, or anyone else, think of me in order to understand my *utterance*?

(b) How must I, or anyone else, think of me in order to understand the *proposition* expressed by my utterance?

It's true, the direct-reference theorist is apt to claim, that the answer to a) entails that in order for me, but not you, fully to understand my utterance, I must think of myself in the self-conscious way. But what explains this, she will claim, isn't that the answer to b) entails that the proposition that I'm old enough to vote has the relativity feature which requires me to think of myself in the self-conscious way when entertaining it. Rather, what explains the answer to a) pertains to the meaning of the pronoun 'I': the meaning of 'I' requires one uttering the pronoun to intend to refer to *himself*, and *that* is why when I utter 'I'm old enough to vote' I must think of myself in the self-conscious way. In order for anyone to understand my utterance *qua* utterance of the

8 I'll be sloppy throughout this paper about use-mention where predicates are concerned, since my focus will be on singular terms. If I weren't being so sloppy, rather than saying that the content of an utterance of ⌜t is F⌝ is <x, Fness>, where x is the semantic referent of the token of t, I'd say it was <x, Φ>, where x is the semantic referent of the token of t and Φ is the property expressed by the token of F.

English sentence 'I'm old enough to vote', she must know that the meaning of 'I' requires one uttering it to think of herself in the self-conscious way.

The direct-reference theorist will acknowledge that the foregoing point about utterance understanding doesn't explain all that needs to be explained in order for her to explain away the appearance that the proposition that I'm old enough to vote has the relativity feature. If the direct-reference theorist is to explain away the relativity feature, then she must, without any appeal to the relativity feature, account for the very strong intuition that it's possible for me to suffer from a form of amnesia which results in its being true that

(*) I believe that I'm old enough to vote, but I don't believe that Stephen Schiffer is old enough to vote.

There is a special onus on the direct-reference theorist to account for the apparent acceptability of (*), since (i) the possibility that (*) is true is entailed by the claim that the proposition that I'm old enough to vote has the relativity feature it seems to have together with certain uncontestable claims, yet (ii) direct-reference semantics together with the fact that I'm Stephen Schiffer entails that the proposition that I'm old enough to vote is identical to the proposition that Stephen Schiffer is old enough to vote, and this makes it look as though the direct-reference theorist is committed to holding that (*) expresses the same proposition as the contradiction

(#) I believe that I'm old enough to vote, but I don't believe that I'm old enough to vote.

Of course, this is a very old conundrum for direct-reference semantics, dating, as it does, from the publication in 1892 of Frege's "On Sense and Reference." For example, common sense doesn't hesitate to suppose that someone who realizes that Wynona Ryder is Wynona Ryder might not realize that Wynona Ryder is Wynona Horowitz, but for the direct reference theorist the proposition that Wynona Ryder is Wynona Ryder is the same proposition at the proposition that Wynona Ryder is Wynona Horowitz. Thus, direct-reference semantics can explain away the relativity feature only if it can give a plausible resolution of its Frege problems.

There are two ways a direct-reference theorist might attempt to diffuse the intuitions—such as the intuition that (*) is true—which gives credibility to the claim that some propositions have the relativity feature. One way is by the theorist's

embracing what I've elsewhere called the hidden-indexical theory.[9] This is a way of being a direct-reference theorist while allowing utterances like (*) to be true. It holds that in a belief report the subject isn't merely being ascribed belief in a Russellian proposition, but belief in such a proposition under a mode of presentation of a contextually determined type. Thus, an utterance 'Lois believes that Superman flies' might be true by virtue of Lois's believing <Superman, being a thing that flies> under a mode of presentation that entails thinking of Superman as a superhero, while an utterance of 'Lois doesn't believe that Clark Kent flies' may also be true by virtue of Lois's not believing <Clark Kent (i.e. Superman), being a thing that flies> under a mode of presentation that entails thinking of Superman/Clark Kent as a nerdy reporter.

The other way a direct-reference theorist might attempt to diffuse the Fregean intuitions is the way of David Braun, Nathan Salmon, Scott Soames, and others.[10] This way holds that belief reports merely ascribe belief in a Russellian proposition, and that, therefore, the semantic content of (*) is the same as that of the explicit contradiction (#), but then, this way continues, the intuition that (*) is true can be explained away in one way or another, perhaps in a way that relies heavily on the mechanisms of Gricean implicature.

I don't think either approach can work, and have argued that elsewhere.[11] But I can't take the space to argue it here, so will assume, as another of this paper's working hypotheses, that since direct-reference semantics is false, it can't be part of any correct way of explaining away the relativity feature.

Fregean semantics and the relativity feature

Frege's theory of propositional content was designed to avoid the problems he thought refuted the direct-reference theory. According to Frege, the propositions we say and believe are structured propositions whose basic components aren't the objects and properties our beliefs and statements are about, but rather *modes of presentation*, or *ways of thinking*, of those objects and properties. Frege's theory can allow that Lois Lane believes that Superman flies but doesn't believe that Clark Kent flies, because it holds that the foregoing two that-clauses refer to distinct propositions owing to the fact that the two names in them introduce distinct modes of presentation

9 Schiffer (1977), (1992), and (2003: 39-46).

10 Braun (1998) and (forthcoming); Salmon (1986), (1989), and (forthcoming); Soames (2002).

of Superman/Kent into those propositions. Frege didn't say what modes of presentation were, but regardless of what he might have had in mind, nothing prevents a Fregean from holding that in the case of paradigmatically referential utterances, the expressed mode of presentation of the referent *x* is an *x*-dependent mode of presentation, a mode of presentation of *x* that wouldn't exist if *x* didn't exist and is perforce a mode of presentation of *x* in every possible world in which it's a mode of presentation of anything. This is how modes of presentation are construed on the neo-Fregean theories of Gareth Evans and John McDowell.[12] Construed that way, the Fregean proposition <*m*, *m*′>, where *m* is an *x*-dependent mode of presentation of *x* and *m*′ is an *F*ness-dependent mode of presentation of *F*ness, would be an *x*-dependent proposition that was truth-conditionally equivalent to, but distinct from, the direct-reference theorist's Russellian singular proposition <*x*, *F*ness>. As noted, the Fregean theory wouldn't incur the direct-reference theory's problems.

It's clear the Fregean theory can't accommodate the relativity feature.[13] If an *x*-dependent proposition has the relativity feature, then there is no one mode of presentation of *x* under which everyone who entertains the proposition must think of *x*, whenever or wherever the entertaining takes place. Fregean propositions, however, do require each *x*-dependent proposition to contain a mode of presentation of *x* under which anyone, at any time or place, who entertains the proposition must think of *x*. But might the Fregean theory accommodate the data that suggest the relativity feature, and in that way explain it away?

Frege felt the relativity feature's pull and famously addressed the problem it created for him as regards two of the kinds of singular terms in play, adverbial demonstratives like 'now', 'today', 'yesterday', 'here', and 'there', and the personal pronoun 'I'. As regards the adverbial demonstratives in question, he wrote:

> If someone wants to say the same today as he expressed yesterday using the word 'today', he must replace this word with 'yesterday'. Although the thought is the same its verbal expression must be different so that the sense, which would otherwise be affected by the differing times of utterance, is readjusted. The case is the same with words like 'here' and 'there'.[14]

11 Schiffer (1987), (1992), (2003), and (forthcoming)

12 Evans (1982), (1985a), and McDowell (1998a).

13 Not to mention it's other problems: see Schiffer (2003: chapter 1).

14 Frege (1967: 24).

Suppose that on day *d* you say 'Today is fine', and that on the day after *d* you say 'Yesterday was fine'. Then, according to Frege, you assert the same proposition in both utterances, and therefore your utterance of 'today' on *d* and your utterance of 'yesterday' on the day after *d* express the same mode of presentation, or way of thinking, of *d*. But what might this single, twice-expressed mode of presentation be? The question demands an answer, because it would seem that when on *d* you say 'Today is fine', you're thinking of *d* in the directly egocentric way "this day" (so to say), but that when on the day after *d* you say 'Yesterday was fine', you're thinking of *d* in the indirectly egocentric way "the day before this day" (so to say). Evidently, Frege meant to deny this, but how?

He didn't say, but Gareth Evans attempted to answer for him:

> Frege's idea is that being in the same epistemic state may require different things of us at different times; the changing circumstances force us to change in order to keep hold of a constant reference and a constant thought—we must run to keep still. From this point of view, the acceptance on d_2 of 'Yesterday was fine', given an acceptance on d_1 of 'Today is fine', can manifest the *persistence* of a belief….
>
> Frege [wrote]: 'The case is the same with "here" and "there".' Indeed it is; our ability to think of a place as 'here' is dependent upon our general ability to keep track of places as we move about …, so … there could not be thoughts interpretable as 'It's ψ here', if they were not entertained by a subject who had the propensity to entertain, as he moves about, thoughts expressible in the words 'It's ψ there'.
>
> These examples suggest that we have to regard the static notion of 'having hold of an object at *t*' as essentially an abstraction from the dynamic notion of 'keeping track of an object from *t* to *t′*'. And the grasp, at *t*, of a thought of the kind suggested by the passage from Frege, a *dynamic* Fregean thought, requires a subject to possess at *t* an ability to keep track of a particular object over time….
>
> Consequently, the *way of thinking of an object* to which the general Fregean conception of sense directs us is, in the case of a dynamic Fregean thought, a *way of keeping track of an object*. This permits us to say after all that a subject on d_2 is thinking of d_1 *in the same way* as on d_1, *despite lower level*

differences [my emphasis], because the thought-episodes on the two days both depend upon the same exercise of a capacity to keep track of a time.[15]

Evans's defense of Frege isn't plausible. Evans is capitalizing on an equivocation in his talk of two thoughts' requiring thinking of a thing in *the same way*. Suppose I look at a ball and think "That's a tennis ball," and you look at a qualitatively identical but different ball and think "That's a tennis ball." Strictly speaking, the way in which I'm thinking of the ball to which I'm referring—the mode of presentation under which I'm thinking about it—isn't the same as the way in which you're thinking of the ball to which you're referring, since my way of thinking is a *my*-tennis-ball-dependent mode of presentation, while yours is a *your*-tennis-ball-dependent mode of presentation. In a more relaxed sense, however, the two ways of thinking of the different tennis balls are the same: they're both qualitatively identical visual modes of presentation. (Similarly, if I think "I'm a paragon" and you think "I'm a paragon," then in the strict sense we're not thinking the same thing, since my thought is true iff *I*'m a paragon, while yours is true iff *you*'re a paragon. But there's another sense in which we are both thinking the same thing—viz. each of us is thinking that he's a paragon.) The two senses are in fact alluded to in the above quotation from Evans, which is why I emphasized '*despite lower level differences*', which I can make sense of only if it alludes to the fact that in the strict sense of 'same way of thinking', the ways aren't the same, whether or not there's a sense in which they're both manifestations of an ability to keep track of the day in question. Here's an analogy. As I move around a physical object, I'm both thinking of the object under distinct perceptual modes of presentation of it *and* keeping track of it. The tracking ability isn't some third mode of presentation; it's no mode of presentation at all but merely an ability that is manifested in the way the distinct perceptual modes of presentation are being deployed. Note that I'm not denying that your utterances of 'Today is fine' and 'Yesterday was fine' have the same semantic content; I agree that they have the same semantic content. What I doubt is whether there is a single mode of presentation that is a constituent of that semantic content.

As regards 'I', Frege took a different tack:

> Now everyone is presented to himself in a particular and primitive way, in which he is presented to no-one else. So, when Dr. Lauben thinks that he has been wounded, he will probably take as a basis this primitive way in which he

15 Evans (1985a: 308-11).

> is presented to himself. And only Dr. Lauben himself can grasp thoughts determined in this way. But now he may want to communicate with others. He cannot communicate a thought which he alone can grasp. Therefore, if he now says 'I have been wounded', he must use the 'I' in a sense which can be grasped by others, perhaps in the sense of 'he who is speaking to you at this moment', by doing which he makes the associated conditions of his utterance serve for the expression of his thought.[16]

Presumably Frege would say something similar about 'you'.

I don't know of any philosopher who has tried to defend Frege's claim about 'I'. Even the most sympathetic of Fregeans distance themselves on this point. Here, for example, is John McDowell:

> [Frege's proposal] is quite unsatisfactory, as becomes clear if we try to construct a parallel account of the role of 'I'-thoughts in receiving communication as opposed to issuing it. Suppose someone says to me, 'You have mud on your face'. If I am to understand him, I must entertain an 'I'-thought, thinking something to this effect: 'I have mud on my face: that is what he is saying'. Frege's strategy for keeping the special and primitive way in which I am presented to myself out of communication suggests nothing better than the following: the 'I'-sense involved here is the sense of 'he who is being addressed'. But this would not do. I can entertain the thought that he who is being addressed has mud on his face, as what is being said, and not understand the remark; I may not know that *I* am he who is being addressed.[17]

I concur.

The no-semantic-content theory

I take it to be obvious that in uttering 'I'm old enough to vote' I said that I'm old enough to vote and that that is known both by me and my hearer. I also take it to be obvious that understanding my utterance requires me, but not my hearer, to think of me in the self-conscious way. One possible explanation of these two obvious truths is that the proposition that I'm old enough to vote both has the relativity feature and is the semantic content of my utterance. One way of attempting to avoid that straightforward explanation is to deny that any proposition is the semantic content of

16 Frege (1967: 25-6).

17 McDowell (1998a: 222).

my utterance and to hold instead that understanding the utterance requires me and my hearer to entertain different but related propositions. That is to say, as regards my utterance of 'I'm *F*', the no-semantic-content theory holds, first, that my literal and serious utterance of 'I'm *F*' has no semantic content—that is, that in uttering the sentence I'm not saying *any* proposition which conforms to the meaning of that sentence—but, second, my utterance (i) has a truth-value; (ii) is fully understood by me and my audience; (iii) is such that that understanding requires me and my audience to think of me in different ways (I, but not you, must think of me under the self-conscious mode of presentation); and (iv) does conform to the meaning of the sentence type 'I'm *F*'(it's just that that meaning can't be a propositional form which requires the literal speaker who utters the sentence to mean a proposition of that form).

Such a position is implicit in the positive proposal John McDowell adds to his objection, quoted just above, to Frege on 'I':

> Frege's troubles about 'I' cannot be blamed simply on the idea of special and primitive senses; they result, rather, from the assumption—which is what denied the special and primitive senses any role in communication—that communication must involve a sharing of thoughts between communicator and audience. That assumption is quite natural, and Frege seems to take it for granted. But there is no obvious reason why he could not have held, instead, that in linguistic interchange of the appropriate kind, mutual understanding—which is what successful communication achieves—requires not shared thoughts but different thoughts that, however, stand and are mutually known to stand in a suitable relation of correspondence.[18]

More recently, the proposal has been explicitly advanced by Richard Heck.[19] Heck writes:

> [A]lthough understanding an utterance of a demonstrative or indexical expression does not require one to think of its referent in the same way the speaker does, there are nonetheless substantial restrictions on how one can think of it and still understand. These restrictions are, of course, determined by context, but it remains the case that, for any utterance of a referring expression, there will be a restricted cluster of ways one may think of the

18 Op. cit.
19 Heck (2002).

> referent and still understand that utterance. What should we make of this fact?[20]

What he makes of it is that for the utterances in question there are no propositions that are the propositions that are said, or expressed, in those utterances:

> If one really wants to find something to call *the* meaning [= the semantic content], then perhaps what is common to the cognitive values the utterance has for different speakers is as good a choice as any. But why do we want to find something to call the meaning? What we (relatively) uncontroversially have are speakers who associate Thoughts [i.e. specific thought contents] with utterances and restrictions upon how the different Thoughts they associate with a given utterance must be related if they are to communicate successfully: to put it differently, we have the fact that utterances have cognitive value for speakers, and we have communicative norms determining how the cognitive values a given utterance has for different speakers must be related if they are to understand one another. Those, it seems to me, are the facts as we find them.[21]

With Heck, I accept the fact of which he implies something should be made, but I don't accept what he makes of it. Will Heck deny that in uttering 'I'm old enough to vote' I said that I'm old enough to vote? If so, then I take that in itself to be a decisive objection to his position: however we might want to analyze the fact that I said that I'm old enough to vote, we may surely take it to be a datum that in uttering 'I'm old enough to vote' I said that I'm old enough to vote. So let's assume that Heck doesn't deny the datum. Given that, he would have to hold either that (a) the that-clause in my utterance of 'I said that I'm old enough to vote' doesn't refer to the proposition that I'm old enough to vote, or that (b) it does refer to that proposition, but that proposition doesn't conform to the meaning of the sentence type 'I'm old enough to vote', and thus can't be the semantic content of my utterance. I take it that (b) is a non-starter: if the proposition that I'm old enough to vote is what I said in uttering 'I'm old enough to vote', then that proposition is *obviously* consonant with the sentence's meaning. But how is Heck to maintain (a)? Of course, (a) would be maintained by a theorist who thinks that propositional attitudes and propositional speech acts are relations not to propositions but to sentential entities of some stripe or

20 Ibid., p. 27.
21 Ibid., p. 31.

other. But such a theorist can't be our present concern. What we need to ask is whether a theorist who accepts that, say, *believing* is a relation between believers and the propositions they believe can reasonably deny that the that-clause in my utterance of 'I said that I'm old enough to vote' refers to the proposition that I'm old enough to vote. I don't think so; at least not without providing reasonable answers to questions such as these:

- Does the that-clause in 'I believe that I'm old enough to vote' refer to the proposition that I'm old enough to vote? It would seem that it must, given that believing is a relation to propositions. For if believing is a relation to propositions, then there must surely be some canonical way of specifying what proposition a person believes, and it's precisely that-clauses which provide our canonical way of specifying what believers believe.
- If the that-clause in 'I believe that I'm old enough to vote' does refer to the proposition that I'm old enough to vote, then how could parity-of-reasoning considerations fail to require saying the same about the that-clause in 'I said that I'm old enough to vote'? And how could the valid inference
    - Fiona said that I'm deranged.
    - I believe what Fiona said.
    - So, I believe that I'm deranged.

  be valid if the that-clause in its conclusion, but not the one in its first premise, refers to the proposition that I'm deranged?

Earlier I said that one thing reference seems to tell us about meaning is that the semantic contents of utterances involving singular terms often have the relativity feature: they are x-dependent propositions the entertainment of which requires different people, or the same person at different times or places, to think of x under different modes of presentations. The upshot of the most recent discussion is a second thing reference seems to tell us about meaning: the propositions that have the relativity feature must be not only fine-grained and object-dependent, but also unstructured. The reason this seems to be the upshot is this. Suppose there were an x-dependent proposition that had the relativity feature and was structured. Then that proposition would have some component that requires different people, or the same person at different times or places, to think of x in different ways, under different modes of presentation. But what could such a component be? It can't be a mode of

presentation or even a mode of presentation kind. Nor can it be something like a rule that lays out the ways in which a person must think of x relative to such-and-such constraints on the person's relation to x, since that way of individuating propositions would have Frege problems like those that plague direct-reference semantics (it would evidently assign the same semantic content to two perceptual-demonstrative utterances of, say, 'That woman is a physicist' when the same woman was referred to in both utterances, even though one could intuitively believe what the one utterance said without believing what the other said). There is no extant theory of structured propositions which allows for such a propositional component, and I have no idea what such a component would be like. Of course, that is no proof that there can't be such a component, and that is why I say merely that the foregoing discussion, if correct as far as it goes, seems to tell us that the propositions which have the relativity feature are unstructured, as well as object-dependent and fine-grained. What kind of propositions can fill that bill?

IV. Pleonastic Propositions and the Relativity Feature

I allude to *the theory of pleonastic propositions*, which I advanced in my recent book *The Things We Mean*. Pleonastic propositions are but one kind of pleonastic entity. Properties, events, states, and numbers are some other kinds. Pleonastic entities are entities whose existence is entailed by what I call *something-from-nothing transformations*. These are conceptually valid inferences that take one from a statement in which no reference is made to a thing of a certain kind to a statement in which there is a reference to a thing of that kind. For example, both the property of being a dog and the proposition that Lassie is a dog are pleonastic entities. From the statement

Lassie is a dog,

whose only singular term is 'Lassie', we can validly infer two of its pleonastic equivalents,

Lassie has the property of being a dog,

which contains the new singular term 'the property of being a dog', whose referent is the property of being a dog, and

That Lassie is a dog is true,

or, more colloquially,

It's true that Lassie is a dog,

which contains the new singular term 'that Lassie is a dog', whose referent is the proposition that Lassie is a dog. It's because something-from-nothing transformations often take one from a statement to a pleonastic equivalent of that statement that I call the entities these transformations introduce *pleonastic* entities.[22]

There is more to the nature of pleonastic propositions than can be discerned merely from the something-from-nothing transformations into which they enter. The full story is too long to tell in this paper, but you can find it—anyway, a fuller story—in my book. Here I'll try to say just enough of what I need to say to address the over-arching issue about the relativity feature.

Pleonastic propositions—the propositions to which that-clauses refer—are individuated by two things: their truth conditions and how one must think about the things involved in those truth conditions in order to believe or otherwise entertain those propositions, where how one must think about any one of those things may depend on one's relation to that thing. It's this last rider which accommodates the relativity feature.

What needs to be explained is how pleonastic propositions can be individuated in a way that accommodates the relativity feature. Since the dominant theories of propositions seem unable to accommodate the feature, one might reasonably wonder whether there is some principled reason why those propositions seem to be individuated in ways that preclude their accommodating the relativity feature. There is such a principled reason, but I think the theory of pleonastic propositions is right to reject it. The principled reason I have in mind pertains to a presupposition theorists of structured propositions—Russellians and Fregeans, for all that presently matters—make about what determines the reference of a that-clause token. It's the application to that-clauses of the principle of compositionality of extension—the principle, to a first approximation, that the extension of a semantically complex expression is determined by its structure and the extensions of its component expressions. Applied to that-clauses, this holds—again to a first approximation—that the reference of a that-clause token is determined by its structure and the references its component expression tokens have in the that-clause token. Given this presupposition, it's hard

22 Something-from-nothing transformations provide the essential characterization of the notion of a pleonastic entity. It's being a conceptual truth that if Fido is a dog, then Fido has the property of being a dog is what, in the first instance, makes the property a pleonastic entity, and thus explains, among other things, how we can have a priori knowledge of that truth. But in order to explain how such conditionals can be conceptual truths the notion of a pleonastic entity is also explained, refinements aside, in terms of the way in which adding to any theory an explicit commitment to pleonastic entities results in a new theory which conservatively extends the theory to which the commitment was added (see Schiffer 2003: §2.2). The explanation in terms of conservative extension is considerably more tentative than the one in terms of conceptual entailment.

to see how an expression in a that-clause—e.g. the second occurrence of 'I' in 'I believe that I'm deranged'—can refer to something that mandates different ways different people, or the same person at different times or places, are to think about the thing to which the expression refers.

I deny the compositionality presupposition. I don't deny that ordinarily the referent of a semantically complex singular term token is determined, roughly speaking, by the singular term's structure and the extensions of its component expressions. But one of the things that make pleonastic propositions special is a certain crucial respect in which the relation between that-clauses and the propositions to which they refer is importantly different from the usual relation between singular terms and their referents, and this difference enables us to explain how the compositionality of reference which usually obtains doesn't obtain in the case of that-clauses.

Let me begin with two claims about that-clause reference which help set the stage for the asymmetry to which I just alluded:

- The referent of a that-clause is always contextually determined, in that no that-clause *type* has a context-independent reference. A corollary of this is that two utterances of any sentence of the form '*A* believes that *S*' may have different truth-values. For example, two utterances of 'Ralph believes that George Eliot was a man' may have different truth-values owing to the fact that in one conversational context but not the other the truth of the utterance requires Ralph to think of George Eliot as a famous author. The same applies to two utterances of 'Ralph believes that she wrote *Ivanhoe*' when the occurrences of 'she' in both utterances refer to George Eliot.
- Belief reports of the form '*A* believes that *S*' are of the form '*a R b*', and every utterance of this form is true just in case the referent of '*a*' bears *R* to the referent of '*b*'. Likewise for saying reports of the form '*A* said that *S*', but for simplicity of exposition I'll run the discussion mostly on belief reports.

Then the asymmetry to which I alluded is as follows.

When '*b*' in an instance of '*a R b*' is *not* a that-clause, we must first identify the referent of '*b*' in order to determine the criteria for truth-evaluating the utterance. For these cases we have the left-to-right direction of

fixing the referent of '*b*' → determining the criteria of evaluation.

For example, in order to truth-evaluate an utterance of 'She kissed him', we must first determine the referents of 'she' and 'him'. It's preposterous to suppose that we fix the referent of, say, 'him' by first determining what must be the case in order for the utterance to be true. The absurdity of that idea is why it's absurd to suppose that one might discover that Picasso ≠ Braque by *first* determining that these two statements may differ in truth-value:

(1a) Henri admires Picasso.

(1b) Henri admires Braque.

On the contrary, we know that (1a) and (1b) may differ in truth-value precisely *because we know that Picasso ≠ Braque. But just the opposite obtains when 'b' in 'a R b' is a that-clause.* For these cases we have the right-to-left direction of

fixing the referent of '*b*' ← determining the criteria of evaluation.

In a propositional-attitude or speech act report of the form '*a R b*' in which '*b*' is a that-clause, we *first* have contextually-determined criteria of evaluation, and *then* those criteria determine the proposition to which the that-clause refers. This is brought home by contrasting the pair (1a,b) with the following two pairs of utterances:

(2a) Nobody doubts that whoever believes that all ophthalmologists are ophthalmologists believes that all ophthalmologists are ophthalmologists.

(2b) Nobody doubts that whoever believes that all ophthalmologists are ophthalmologists believes that all ophthalmologists are eye doctors.

(3a) I believe that I'm deranged.

(3b) I believe that Stephen Schiffer is deranged.

We know that the two members of all three pairs—(1a,b), (2a,b), and (3a,b)—may differ in truth-value, but there's a very important difference between (1a,b), on the one hand, and, on the other hand, (2a,b) and (3a,b). As already noted, it's absurd to suppose we know that Picasso ≠ Braque because we know that (1a) and (1b) may differ in truth-value; rather, we know that (1a) and (1b) may differ in truth-value *because* we know that Picasso ≠ Braque. But just the opposite obtains as regards the

other two pairs. We don't know that (2a) and (2b) may differ in truth-value because we first know that the proposition that whoever believes that all ophthalmologists are ophthalmologists believes that all ophthalmologists are ophthalmologists ≠ the proposition that whoever believes that all ophthalmologists are ophthalmologists believes that all ophthalmologists are eye doctors; rather, we know that the proposition that whoever believes that all ophthalmologists are ophthalmologists believes that all ophthalmologists are ophthalmologists ≠ the proposition that whoever believes that all ophthalmologists are ophthalmologists believes that all ophthalmologists are eye doctors *because we know (2a) and (2b) may differ in truth-value*. Likewise for (3a) and (3b): we first know that they may differ in truth-value, and on this basis we know that the proposition that I'm deranged ≠ the proposition that Stephen Schiffer is deranged. (The contextually-determined criteria for truth-evaluating a belief report pertain in part to the communicative interests of the speaker and her audience, and they are statable without reference to the proposition expressed by the belief report. These criteria determine the truth conditions for the entire belief report, and therewith the proposition to which the belief report's that-clause refers. For example, the operative communicative interests of an utterance of 'Ralph believes that George Eliot was a man' may be such that the utterance won't count as true unless Ralph thinks of Eliot as a famous nineteenth century novelist, and this will determine the referent of the utterance's that-clause to be a proposition the entertainment of which requires thinking of Eliot as a famous nineteenth century novelist.)

If, as I suggested, the referent of a that-clause is directly fixed by the contextually-determined criteria for truth-evaluating the belief report in which the that-clause is contained, it doesn't *follow* that the referent of the that-clause isn't also a function of the references of its parts. For it may be that the contextually-determined criteria for truth- evaluating the report also fix referents for the that-clause's component expressions in a way that permits one to see the referent of the that-clause as a function of the referents its component expressions have in the that-clause.[23] What is important for our purposes, however, is that the suggested asymmetry hypothesis is compatible with the non-compositional determination of

23 The doctrine of "syntactic priority" which Crispin Wright (1983) attributes to Frege has the references of numerals fixed by conceptually prior criteria of truth-evaluation for arithmetical statements, but in this case compositionality of extension is preserved. For affinities between the doctrine of syntactic priority and the asymmetry I'm heralding, see Schiffer (2003: 77-9).

that-clause reference, and therefore presents no barrier to the proposal that that-clauses may refer to propositions that have the relativity feature. Since the reference of a that-clause is determined by contextually-determined criteria for truth-evaluating the entire statement made in the utterance containing the that-clause, there's no *need* for the reference of a that-clause to be determined by its structure and the references of its component expressions.

I think the reference of a that-clause is never determined by its structure and the references of its component expressions. But I also think that semantic properties of the component expressions play two important roles in determining a belief report's criteria of evaluation, and therewith the proposition to which the report's that-clause refers.

The first role is that the structure and semantic values of its component expressions do determine the *truth conditions* of the proposition to which the that-clause refers. Consider an utterance of

(a) Ralph believes that she wrote it.

In order to fix the criteria for truth-evaluating this utterance, we don't need to fix the referent of its that-clause, but we do need to fix the referents of 'she' and 'it' as they occur in the utterance. Suppose the token of 'she' refers to George Eliot and the token of 'it' refers to the novel *Ivanhoe*.[24] Then that would suffice to determine that the that-clause refers to an Eliot- and *Ivanhoe*-dependent proposition that is true in an arbitrary possible world just in case George Eliot wrote *Ivanhoe* in that world. But while those references determine the truth conditions of the proposition to which the that-clause refers, they don't determine the proposition itself, for they don't determine how Ralph, or anyone else, must think of Eliot or *Ivanhoe*. For consider two utterances of (a) in both of which 'she' refers to George Eliot and 'it' refers to *Ivanhoe*. In that event, both that-clause tokens will refer to an Eliot- and *Ivanhoe*-dependent proposition which is true in an arbitrary possible world just in case Eliot wrote *Ivanhoe* in that world. Still, the two utterances may differ in truth-value because in order for one but not the other of them to be true, Ralph must think of Eliot as, say, the author of *Middlemarch*. In this case, the two occurrences of 'that she wrote *Ivanhoe*' refer to distinct but truth-conditionally equivalent Eliot- and *Ivanhoe*-dependent propositions. They are distinct because only one of the propositions is

24 Care is needed here. While we can say that the tokens of 'she' and 'it' refer to Eliot and Ivanhoe, respectively, we can't say that the utterance is true just in case <George Eliot, Ivanhoe> satisfies 'Ralph believes that x wrote y'. See Schiffer (2003: 84-5).

such that entertaining it requires thinking of Eliot as the author of *Middlemarch*. That the two that-clause tokens refer to truth-conditionally equivalent propositions is determined by their common structure and the common semantic values their component expressions have in the two tokens of the same that-clause. That the one that-clause token refers to a proposition that requires Ralph, or anyone else who entertains it, to think of Eliot as the author of *Middlemarch* isn't determined by the semantic value of any expression in the that-clause or by any of the belief reporter's referential intentions; rather, it's determined by the mutual communicative interests of the speaker and her audience (or at least by what the speaker takes those interests to be). In the context of one but not the other utterance of (a), those interests are such that the utterance won't count as true unless Ralph thinks of Eliot as the author of *Middlemarch*.

The second role a semantic property of an expression in a that-clause may play is that the meanings of certain singular terms constrain, and in some cases determine, how *the belief reporter and her audience* must think of the referent of the token of the singular term in a belief report's that-clause. When the that-clause in an utterance of ⌜Ralph believes that … *t* …⌝ refers to an *x*-dependent proposition and *x* is the referent of the token of *t* in the that-clause, then virtually no semantic feature of *t* mandates any particular way in which *Ralph* must think of *x* in order for the semantic content of the report to be true.[25] For example, an utterance of (a) in which 'she' refers to George Eliot may be true even though Ralph believes that the author of *Ivanhoe* is male. The contextually-determined criteria of truth-evaluating the belief report may merely determine a proposition which requires Ralph to think of Eliot as the author of *Middlemarch*. At the same time, however, the occurrence of 'she' does mandate a proposition which requires the *belief reporter and her audience* to think of Eliot as female. Similarly, an utterance of 'Ralph thinks that it'll rain today' doesn't require Ralph to think of the day to which the occurrence of 'today' refers in the egocentric way associated with 'today'. The report might be based on Ralph's having said 'It'll rain on October 21st'. But the semantic content of the belief report does require the belief reporter and her audience to think of the day referred to in the way associated with 'today'. In this way we can see how object-dependent propositions are

25 The 'virtually' qualification pertains to a belief report such as 'Ralph believes that he himself is F', which does seem to require that Ralph think of himself in the self-conscious way associated with 'I' in order for the report to be literally true.

contextually individuated in ways that determine those propositions to have the relativity feature.

In closing this preliminary discussion of the relativity feature (I recognize that a lot more needs to be said to make the notion fully perspicuous), I want to consider what is liable to strike you as a devastating objection to the idea that any proposition has the relativity feature. The problem is most clearly brought out by switching from belief reports to saying reports, and what follows is one way of articulating it.

Suppose Jane utters 'Stephen Schiffer is deranged' and that the semantic content of her utterance—the proposition that Stephen Schiffer is deranged—is such that I can entertain that proposition without thinking of myself in the self-conscious way (I might suffer amnesia and insist 'I'm not Stephen Schiffer'). What am I then to make of the following argument, presented as an objection to my claim that some propositions have the relativity feature?

(1) I can correctly report what Jane said in her utterance by uttering 'Jane said that I'm deranged'.

(2) If (1), then the referent of the that-clause in my utterance is the semantic content of Jane's utterance, the proposition that Stephen Schiffer is deranged.

(3) Trivially, the referent of the that-clause in my utterance is the proposition that I'm deranged.

(4) ∴ The proposition that I'm deranged = the proposition that Stephen Schiffer is deranged.

(5) ∴ Since I can entertain the proposition that Stephen Schiffer is deranged without thinking of myself in the self-conscious way, the same is true of the proposition that I'm deranged.

(6) ∴ Not only is it false that the occurrence of 'I' in a saying report's that-clause mandates that its referent think of himself in the self-conscious way when entertaining the proposition to which the that-clause refers, but, by an obvious extrapolation and generalization, it's false that any proposition has the relativity feature.

What I make of this argument is that it isn't sound: premise (2) is false. By stipulation, *p* is the semantic content of *S*'s utterance of σ iff *S* meant *p* in uttering σ and *p* conforms to the meaning of σ. When *p* is the semantic content of *S*'s utterance of σ, then *S* said *p* in uttering σ; but, we've already noticed (on p.000), a saying report

may be correct even though it doesn't give the semantic content of the utterance of which it's a true report. For example, Al eats a meal prepared by Betty's father, whom he's never met, and says to the guests at Betty's dinner party, 'Betty's father is a great cook'. The semantic content of Al's utterance is the proposition that Betty's father is a great cook. The next day Betty says to her father, 'My friend Al said that you're a great cook' and makes what in the context counts as a true statement. But the proposition to which the that-clause in Betty's saying report refers isn't the semantic content of Al's utterance. There are other kinds of examples. Scott Soames,[26] who also emphasizes that the proposition to which the that-clause in a true saying report refers needn't be the semantic content of the subject of the report's utterance, gives examples involving metaphorical speech, such as Al's saying 'Betty's apartment is a pigsty' and Carla's reporting 'Al said that Betty's apartment was filthy'.

What these examples illustrate is that even though *p* is the semantic content of a speaker's utterance, the speaker may be correctly reported as having said another proposition *q*, if *q* is appropriately related to *p*, where there is more than one way of being "appropriately related." I submit that one way of being appropriately related is illustrated in the example in which Jane uttered 'Stephen Schiffer is deranged', and I correctly report 'Jane said that I'm deranged'. Entertaining the semantic content of Jane's utterance doesn't require me to think of myself in the self-conscious way, but entertaining the proposition to which my that-clause refers—the proposition that I'm deranged—does require me to think of myself in the self-conscious way.

V. Conclusion

What reference has to tell us about meaning, I submit, is that it's plausible that the propositions we believe and mean are pleonastic propositions. It tells us that because:

- In the first instance it tells us that it's plausible that the semantic contents of utterances containing singular terms often have the relativity feature.
- But then it's plausible that propositions can have the relativity feature only if the references of the that-clauses which refer to them aren't compositionally determined.

26 Soames (2002: chapter 6).

- Pleonastic propositions, which are motivated independently of the relativity feature, satisfy the non-compositionality constraint, and are otherwise able to accommodate the relativity feature.
- It's questionable, I submit with only the sketch of an argument, that any non-pleonastic propositions—at least any with which we're currently familiar—can accommodate the relativity feature.

Works Cited

Braun, D. (1998). "Understanding Belief Reports." *The Philosophical Review* 107: 555-95.

_____ (forthcoming). "Illogical, But Rational," *Noûs*.

Evans, G. (1982). *The Varieties of Reference* (Oxford University Press).

_____ (1985a). "Understanding Demonstratives," in Evans (1985b).

_____. (1985b). *Collected Papers* (Oxford University Press).

Frege, G. (1952). "On Sense and Reference," in Geach and Black (1952).

_____ (1967). "The Thought: A Logical Inquiry," in Strawson (1967).

Geach, P. and Black, M., eds. (1952). *Translations from the Philosophical Writings of Gottlob Frege* (Blackwell).

Heck, R. (2002). "Do Demonstratives Have Senses?" *Philosopher's Imprint* 2.

McDowell, J. (1998a). "*De Re* Senses," in McDowell (1998b).

_____. (1998b). *Meaning, Knowledge, and Reality* (Harvard University Press).

Salmon, N. (1989). "Illogical Belief." *Philosophical Perspectives* 3: 243-85.

_____ (forthcoming). "The Resilience of Illogical Belief," *Noûs*.

Schiffer, S. (1977). "Naming and Knowing," *Midwest Studies in Philosophy, Vol. II: Studies in the Philosophy of Language*: 28-41.

_____ (1987). "The 'Fido'-Fido Theory of Belief," *Philosophical* Perspectives 1: 455-480.

_____ (1992). "Belief Ascription," *The Journal of Philosophy* 89: 499-521.

_____ (2003). *The Things We Mean* (Oxford University Press).

_____ (forthcoming). "A Problem for a Direct-Reference Theory of Belief Reports," *Noûs*.

Soames, S. (2002). *Beyond Rigidity: The Unfinished Agenda of Naming and Necessity* (Oxford University Press).

Strawson, P., ed. (1967). *Philosophical Logic* (Oxford University Press).

Wright, C. (1983). *Frege's Conception of Numbers As Objects* (St. Andrews University Press/Humanities Press).