Volume 7, Number 2
2013

ISSN 1807-9792

abstracta

Linguagem, Mente & Ação

http://abstracta.oa.hhu.de

d|u|p

abstracta

Linguagem, Mente & Ação

ISSN 1807-9792

Volume 7, Number 2
2013

abstracta
2004 – 2013

Contents

abstracta Linguagem, Mente & Ação ISSN 1807-9792 Volume 7, Number 2, pp. 3–17, 2013

Truths are Valuable, Truth isn't

Alexander auf der Straße

Department of Philosophy, Heinrich-Heine-Universität Düsseldorf, Germany
aufderstrasse@hhu.de

Abstract
This paper deals with the relationship that, according to some, holds between true beliefs and success. It argues for truth-theoretic minimalism. In particular, minimalism will be defended against a particular objection against deflationism raised by Michael Lynch. The paper denies that truth has any non-instrumental value in the sense that truth is pursued for its own sake. Moreover, the instrumental value of true beliefs will be explained in terms of psychological regularities of agents' 'correct' beliefs about the world, rather than in terms of truth as such. The argument concludes with the result that – in the strict sense – truth is valueless because truth is no genuine property. However, the value of individual true beliefs is acknowledged, insofar as they foster one's behavioural success.

1 Introduction

At least since Dummett's seminal paper, 'Truth', dating back to 1959, people have repeatedly remarked that deflationary theories of truth miss the *point* of truth – that truth is an aim. The question whether modern variants of deflationism lack the resources to account for this is still controversial. Among the more recent critics of deflationism, one of the most prominent ones is Michael Lynch, who took up Dummett's original line of argument and developed it further in several respects (Lynch 2004, 2009). He defends a position according to which truth plays a significant role in our everyday life. Truth, from his point of view, has value. We pursue truth for its own sake, he says. This paper aims at undermining this claim by arguing exemplarily against Lynch's particular position. I examine Lynch's argument as a typical example of an argument for the desirability of truth. I shall argue that truth-theoretic minimalism is compatible with the legitimate claims about the value of truth. In contrast to Lynch's view, then, a position will be defended according to which one may acknowledge the value of *particular* truths but can still deny that truth *as such* has any value. In a nutshell, this is because truth is no genuine property.

The following very rough overview will serve as a backdrop against which the relevant theories of truth can be positioned. Briefly speaking, the world of truth theories divides into deflationary views and inflationary ones. Deflationism about truth is a family of varied theories, all of which have in common that some form or other of Tarski's convention T

(T) "p" is true iff p[1]

plays a prominent role. "P" is a variable ranging over declarative sentences of a specified natural language (English, say) and "'p'" (that is, "p" in quotation marks) is a meta-level name for that very sentence. (For sake of simplicity, English serves double purpose here both as meta-

[1] This way of stating the schema ignores some technical niceties, such as the problem that it actually requires quasi-quotation marks. (For "p" here is only a variable.) Still, I adopt this notation since it is one of the most common ones, and the notational technicalities do not affect the present argument.

language and object language.) Depending on what one takes primary truth bearers to be, the schema will be relativised accordingly either to propositions, sentence types, beliefs or utterances. The exact status of the schema varies from theory to theory. Importantly, in their 'metaphysical' part deflationary theories deny that truth has an underlying nature that can be revealed or anyway determined.

Inflationism about truth, on the other hand, is often characterised as the denial of deflationism, i. e. in inflationary theories, convention T plays no significant explanatory role (which, of course, is not to say that it plays no role whatsoever). Moreover, truth is analysable on these accounts: an inflationary theory would state the basic properties that constitute truth. Lynch's theory of truth is of this sort.[2] One of the properties he ascribes to truth is that it is good. It is this claim that I will be focussing on in what follows. In particular, I will defend truth-theoretic minimalism, which is one of Lynch's specific targets (cf. Lynch 2004: ch. 7, therein esp. pp. 107–116, 2004a).

Minimalism[3] says (i) that truth has no underlying nature, (ii) that the truth predicate denotes only a logical property in virtue of being a predicate (but that it denotes no 'real' property), (iii) that all its uses derive – in a yet-to-be-specified way – from (our underived acceptance of) instances of the T schema, and (iv) that everyone who understands how the schema works knows the meaning of "true" entirely. The reason that there is a truth predicate at all in natural languages lies in its usefulness, the minimalist claims. It allows formulating otherwise ungraspable generalisations like "Theory T is true" (read: every single sentence of which theory T consists is true) and so-called blind ascriptions like "What Johnson thought this morning is true", where truth is ascribed to an entity whose identity is potentially unknown to the ascriber (hence the name). Roughly, this is all there is to know about truth, according to minimalism.

Obviously, if minimalism about truth holds, truth is valueless. According to minimalism, the only reason to 'keep' the word "true" in English is because otherwise certain things would either become inexpressible – like "Everything the pope says is true" – or too uneconomical to express – like "The Gravity Theory is true".[4] The truth predicate 'behaves', linguistically speaking, like any other old predicate. In this sense, it denotes a property – truth – because *every* predicate, in a minimal sense, denotes a property qua being a predicate. But since "true" is *not analysable* analogously to, say, "red" or "water", we may say that it denotes only a logical property. To be sure, this claim needs independent justification, which, however, can't be given here.[5] If we only have "true" at the semantic level but no corresponding entity denoted in the

[2] Note that this applies to Lynch's earlier view as exemplified by his 2004 book as well as his more recent 'alethic functionalism' (Lynch 2009, 2012), according to which truth is a functionalistic property, multiply realisable by different properties in different discourse domains. Irrespective of the additional functionalistic analysis Lynch proposes, his latter theory qualifies to be handled on a par with other, more paradigmatic inflationisms (coherentism, correspondence, etc.), since, on his view, truth is similarly analysable in terms of a constituting property. The claim that x is true iff 'x has a property that plays the truth-role' (Lynch 2009: 72) is theoretically on a par with such claims as that x is true iff x is part of a perfectly coherent belief system; that x is true iff x corresponds to reality, and so on. Thanks to an anonymous referee for this journal for pointing out to me that this needed clarification.

[3] By 'minimalism' I mean the deflationary theory of truth defended most prominently by Paul Horwich (e. g., Horwich 2005, 2010). This is not to be confused with Crispin Wright's completely different truth theory of the same name. For the latter, see, e. g., Wright (1994), and the elaborations in Wright (2003).

[4] Examples of the first kind would, as a matter of fact, become inexpressible because their "true"-rectified counterpart involves an infinite conjunction (If he says that God exists, then God exists; if he says that Europe lies south of Canada, then Europe … and so on), but infinite conjunctions can't be expressed by us finite beings. In the latter case, "true" might *in principle* be eliminated, but only at a very high cost (in terms of time and cognitive effort): every time one justifies a certain claim by referring to the truth of the Gravity Theory one would have to recite each of its axioms.

[5] See Horwich (1998: 141–144).

world, then a fortiori this 'thing' – truth – can't bear any properties. In particular, it can't have value.

Now, this flies in the face of common sense. 'Of course', you might think, 'truth has value'. We all pursue the truth. We, especially philosophers and scientists, value truth for its own sake, no matter for what purposes it may turn out to be useful in the future. All this is basically right, and, at least at first sight, it may be doubted whether minimalism is able to explain these alleged 'facts'. But: as always, it highly depends on the details whether the objection actually succeeds. In the following paragraphs, I shall (i) demonstrate that truth does not possess all the properties you might previously have thought it does and (ii) show that all properties that may be reasonably ascribed to truth can me accounted for within a minimalist framework.

In detail, I will proceed as follows. In section 1.1, the 'ceteris paribus' condition in Lynch's truth norm will be discussed; followed by a brief review of the scope of the norm and the exact meaning of "good" in section 1.2. 2.1 analyses the main line of argument as a struggle about the correct 'directionality' of explanations (from generalisations to individual rules, and vice versa). Accordingly, in 2.2 the deflationist position on this issue will be defended. Section 2.3 is a brief discussion of which phenomena should be accounted for by truth theories, and which should rather be left to related theories. Eventually, section 2.4 argues that, in sharp contrast to instrumental beliefs, 'non-instrumental' beliefs are probably not desirable. A short conclusion (3) ends the paper.

1.1 Unqualified CP clauses

Lynch (2004) claims that it is good to believe what is true. He is aware of the problem that the truth predicate serves here only as a device of generalisation, as described above. So, he reformulates the same 'norm' without the predicate. In reconstructing (and undermining) Lynch's argument we need both formulations anyway, so I restate them here in their original form:

> (TN) Other things being equal, it is good to believe that p if and only if it is true that p. (Lynch 2004: 108)
>
> (B) Other things being equal, it is good to believe that p if and only if p. (109)

It is pretty obvious that both formulations amount to the same, since their respective right-hand side is derived by substituting "p" with "it is true that p" (and vice versa), which is just an instance of the T schema.[6] Before discussing Lynch's actual argument, let me very briefly point out some problems of the 'norms' as such.

First of all, there is the ceteris-paribus (CP) clause. Clearly, very often it is good to believe true things. But that means in reverse: sometimes it is *not* good to do so.[7] The CP clause tries

[6] More exactly, substituting "p" for "it is true that p" is an instance of the equivalence schema for propositions, i. e. Tarski's Convention T applied to propositions. (There are equivalence schemata for utterances and sentences as well.) Both Horwich and Lynch agree that propositions are the primary bearers of truth, and both work with the equivalence schema that is relativised to propositions. Accordingly, I shall assume that, on either account, TN is derivable from B, and vice versa, by substituting "p" for "it is true that p", and vice versa.

[7] Oftentimes, arguments to the effect that false beliefs might have positive effects refer to folk psychology. Here is a real-life example:

> [... S]elf-deception tends to lead to positive beliefs about oneself, which in turn trigger the subsequent display of the winner effect [i. e., 'an increased ability to win fights and social conflicts following prior victories']. [... P]ositive beliefs about oneself, and the adaptive benefits that go with it, can be reached through self-deception, past wins, justified true belief, or any number of other sources. The fact that self-deception often leads to these types of beliefs shows that it, like the past wins, can offer some adaptive benefits. (Lopez, J.K. & Fuxjager, M.J. (2012), Self-

to capture this. But it is really just a fig leaf: it plays such an important role for the overall argument that it needs to be spelled out in more detail.[8] An example will help to illustrate what is at stake here: on my (the author's) desk, the keys lie to the right of the candle. So – by the T schema – it is true that the keys lie to the right of the candle. Hence – by TN – it is good to believe that the keys lie to the right of the candle (on the author's desk). Or is it? For almost everyone besides a handful of people this information is irrelevant. Considering the limited cognitive capacities of humans, it is in no reasonable sense 'good' for any of the other seven billion people on planet Earth to have this belief (not to mention beliefs about certain molecule constellations in far-away galaxies).

But maybe Lynch had in mind something different. Maybe the rule is intended to mean it is good to believe that the keys lie to the right of the candle rather than believing that they lie to the left of it (given that, in fact, the keys lie to the right of the candle), so that *if* you believe anything at all about the keys on the author's desk, then you better believe that they are wherever they really are. At least this is not what B (or TN, for that matter) says.[9] But even such a restricted reading of the norm is implausible. Of course, it is very often highly useful to believe falsehoods. And Lynch acknowledges this. This is why he embeds the alleged norm in a CP clause. But to spell out what the CP condition exactly consists in means to commit oneself. And this would reveal the deficits of the norm, for there are counterexamples to even any restricted variant of it. When it comes to norms the burden of proof is not on the side of those who deny their validity.

The difficulties surrounding the CP clause are widely acknowledged in the literature. (One of the most illuminating pieces of work in this regard surely is Heal (1987). For more recent criticism of the ceteris-paribus condition, see Coates (2009); cf. also Piller (2009).) The fundamental problem consists of two aspects. Firstly, there are unbelievably many *trivial* truths. On the one hand, the 'trivially' trivial ones include propositions like 'that the keys lie to the right of the candle on the author's desk', or the contents of perceptual states. On the other hand, there are trivial truths build by running logical operations on atomic truths: for example, the truth of 'that today the weather is nice', combined with any mathematical truth. The CP condition is primarily intended to handle these trivial truths.

The other aspect of the problem is that sometimes it is better not to believe the truth, even in cases where the truth in question is substantive or non-trivial. Lynch's own example

deception's adaptive value: Effects of positive thinking and the winner effect, *Consciousness and Cognition* **21**, 322)

[8] David raises a similar worry against Lynch's strategy:

> It is difficult to criticise claims with 'prima facie/other things being equal'-qualifiers: objections tend to receive the response that that's a case where other things aren't equal. (2005: 297)

For general concerns regarding CP qualifications in (scientific) theories, see Earman, Roberts & Smith (2002). CP laws -— TN is such a 'law' in the relevant sense —- seem to be ultimately flawed, for they can't be tested against *any* available evidence:

> Consider the putative law that *CP*, all Fs are Gs. The information that *x* is an F, together with any auxiliary hypotheses you like, fails to entail that *x* is a G, or even to entail that with probability *p*, *x* is a G. For, even given this information, other things could fail to be equal, and we are not even given a way of estimating the probability that they so fail. (Earman et al. 2002: 293)

[9] Note that Lynch says at one point in the discussion of why believing truths is good: 'it is better to believe something when and only when it is true. Or more loosely: *it is better to believe what is true than what is false*' (2004: 13, emphasis added). This *suggests* a reading like: regarding a particular proposition, if someone is about to form a belief about this very proposition, it is generally better when the belief is true than when it is false. But such a reading is incompatible with TN and B, which are generalisations concerning all propositions, including trivial ones.

of how believing falsities may foster one's success is this: a talent-free climber who thinks that a particular summit is reachable is more likely to get further than the one whose relevant beliefs are accurate in this respect. This is a general phenomenon. Many people reach (certain of) their aims better or equally well when believing falsehoods. Accordingly, the CP condition must be conceived of as also excluding these cases.

Generally, the problem is that the 'ceteris paribus' condition is *unspecified*. Considering all trivial and not-worth-the-effort, non-trivial truths, it is arguably the case that: *other things being equal*, it is good not to believe the truth. This is certainly true (or could be true) if we are allowed to leave the CP clause unspecified, for in that case every justified instance of good, true beliefs would qualify as a case where not everything is equal.

Note that Lynch thinks that it is actually just the other way around:

> [...S]ometimes, believing what is true isn't the best thing—some falsehoods might be better to believe in certain circumstances and some trivial or dangerous truths may not be worth pursuing all things considered. But these cases are the exceptions that prove the rule: *other things being equal*, true beliefs are worth pursuing. (2009: 12)

The idea here is that truths that are not worth pursuing are clearly exceptional and can, therefore, be excluded with the CP condition. Note that Lynch does not have a proof for this. And nor have I for the reverse when playing devil's advocate here. The difference, though, is that the burden of proof is on his side, for the burden of proof is arguably *always* on those who posit norms, not on those who deny them or remain neutral.

1.2 Cognitive Goods

The second minor issue with Lynch's norms is that he is rather silent on the semantics of "good". Good for whom, good in which respect? On this we only read:

> In believing, we operate under the norm of truth: other things being equal, it is good to believe a proposition when and only when it is true. [...] I don't mean that it is necessarily morally better. Things can be better or worse, good or bad in different ways. Clear writing is an aesthetic good; tasty food is a culinary good; and believing true propositions, we might say, is a cognitive or intellectual good. (Lynch 2004: 13, emphasis omitted)

What is a cognitive good? As we have seen, it is not something that lets people reach their respective aims. That believing true propositions is a cognitive good might mean that we praise others for holding true beliefs (like when they have written clearly or cooked well). That is implausible, if anything is. Many of us do not (always) care whether others believe correctly. (Just to state the obvious: this might be true even if most of us care most of the time what others believe.) What, then, is a cognitive good? Or, in other words, in which respects is it good to believe true propositions? Unless this is specified in some more detail, Lynch's theory of truth is trivially false, because it is uncontroversial that believing true propositions *tout court* is not good.

In sum, these are the minor objections against Lynch's norm of truth. Firstly, the ceteris-paribus clause is undefined. Secondly, the scope of the norm is unclear. Does is apply to everyone? In which respect is it good for people to conform to the rule? What does "good" mean in this context? These problems undermine the plausibility of the account but do not affect the actual core of the argument against minimalism. It is to this issue that I now turn.

2.1 Directions of Explanation

The minimalist, remember, says that our concept of truth is constituted by our underived disposition to accept instances of the T schema. From this – it is granted – TN, the truth norm, can't be derived. In the above paragraphs we have seen some reservations one may have about the rule. However, let us now suppose that something like the truth norm in fact holds, i. e. that it is actually good for everyone to believe the truth. A first simple distinction suggests itself: the distinction between, in Lynch's terms, 'being instrumentally good' and the rest. Accordingly, Lynch offers a third norm, which he also claims to hold:

> (BI) It is *more than instrumentally* good to believe that p if and only if p. (Lynch 2004: 111, emphasis added)

This helps understanding how Lynch conceives of TN. Applying the instrumental/non-instrumental distinction, TN is to be understood thus: it is instrumentally good to believe the truth (all else equal). The following can be considered common sense. Humans have certain aims they want to reach. They act according to the beliefs they hold. Simply put, having true beliefs increases one's chances to reach one's aims. For example, you believe, say, that swallowing green pills makes you immortal. It is true (let's suppose) that green pills make immortal. You want to be immortal. Therefore, you swallow green pills, which, by assumption, do make you immortal. Now, you have reached your aim because you truly believed that the pills would have the desired effect. The general phenomenon is: having true beliefs helps, on the whole, reaching one's respective aims. This is meant by 'being instrumentally good': true beliefs are good instruments to succeed in life.

The explanation above is fine as far as it goes. But: what *really* explains success in this case – in terms of causal efficacy – is not the *truth* of the proposition that green pills make people immortal. Rather, the explanation should recur to the *fact* (if it is a fact) that green pills make immortal. No truth is *required* to explain instrumental success.[10] Lynch challenges the claim that this direction of explaining things is correct. It is not, he argues, that a single belief that p helps reaching a single aim if and only if p (under relevant circumstances), from which it is then inferred – after some such occurrences – that, generally, beliefs that p help reaching one's aims if and only if p (viz. iff they are true). According to him, the exact opposite holds:

> [W]hy do we accept the infinite list of little belief norms? Answering this question is crucial, because the nonminimalist [i. e. the inflationist] has a ready and obvious answer. The reason we should accept that it is good to believe that snow is white just when snow is white, and good to believe that Socrates was a philosopher just when he was, is that it is good to have true beliefs. What makes it good to believe a proposition is that proposition's *being true*. (Lynch 2004: 110)

Accordingly, he concludes that 'minimalists must either come up with some other explanation, or admit they can't explain every fact about truth' (2004: 110). Here we see that really, so to say, the 'direction' of explanation is at issue. We could reasonably ask back again: why

[10] As I will argue in section 2.3, a theory of truth is the wrong place to look for an answer to the question why some actions are more successful than others. Please note two important things in this context. Firstly, "fact" is usually defined in terms of 'true proposition'. That itself, however, does not undermine our claim – that truth itself is causally inefficacious – for the relevant point is that, in regard to any particular successful action, its success can be explained *without* invoking the notion of truth. Secondly, facts and truth figure prominently in the area of causation (cf. Schaffer 2008: section 1). In particular, facts (true propositions) are promising candidates for being the relata of causation. Yet, this way of talking can plausibly be traced back to the utility of the truth predicate as the only common device of formulating particular generalisations – a feature that is perfectly explicable in deflationary terms.

is it instrumentally good to believe true propositions? We may visualise the alternatives like this:[11,12]

$$TN \succ G(B\varphi \leftrightarrow \varphi) \succ G(Bp_1 \leftrightarrow p_1), G(Bp_2 \leftrightarrow p_2), \ldots, G(Bp_n \leftrightarrow p_n);$$

$$G(Bp_1 \leftrightarrow p_1), G(Bp_2 \leftrightarrow p_2), \ldots, G(Bp_n \leftrightarrow p_n) \succ G(B\varphi \leftrightarrow \varphi) \succ TN,$$

where "TN" is the truth norm, "B" stands for 'belief', "G()" means '... is good', and the respective second step is a schema that abstracts from particular beliefs. The first option is Lynch's. First a cognitive agent believes that the truth norm (TN) holds, from which she then infers single belief norms (e. g., 'It is good to believe that snow is white iff snow is white'). The alternative is to start from the observation that humans initially form *individual beliefs* that concern particular situations; they then *abstract* from these beliefs ("Every substitution instance of 'φ' in 'G(Bφ ↔ φ)' yields a truth")[13]; in a final step, they *express* this by using the truth predicate, i. e. they endorse the truth norm.[14]

The two alternatives represent possible ways of explaining the genesis of beliefs. By the same token, they represent possible ways of explaining why cognitive agents are justified in having these beliefs. As will become clear in the next section, the important bit is that particular beliefs can serve as an explanatory endpoint, whereas general norms can't. The idea is that people can't be described as being justified in having certain beliefs if there is no plausible story about how their beliefs evolved in the first place. I will come back to this further below.

In order to argue for minimalism, I shall now show why the second alternative is the preferred option. The answer is twofold: I should like to start with detailed comments on the 'order of explanation' issue, before I will formulate some general remarks on the scope of a theory of truth.

2.2 From Individual Beliefs to General Norms

Lynch thinks that the truth norm – "Other things being equal, it is good to believe that p if and only if it is true that p" – explains why people tend to form 'little belief norms' that then guide their action.[15] For the moment, we only consider instrumental value. Think of the 'green pill' example again. Lynch would be inclined to say that what *eventually* explains our success in becoming immortal is the fact that TN holds. But that is absurd. The reason people become immortal (if they do) is *because green pills make immortal*. It is because of this causal chain between pills and their corresponding effects that people become immortal if they

[11] These alternatives illustrate two kinds of explanantia of one single explanandum. They are both supposed to explain the behaviour of people ('us', in Lynch's terms), which is – by assumption – such that people act *as if* they endorse the truth norm. So, although both 'directions' culminate in answering different questions – 'Why do people accept infinitely many little belief norms?' vs. 'Why is believing true propositions instrumentally good?' – they have in common that as a whole they both explain the same pattern of behaviour. Thanks to an anonymous referee for this journal for pushing me in this direction.

[12] An alternative way of formalisation would be to restrict the scope of "G" such that it only aligns the goodness of beliefs about φ with φ: $G(B\varphi) \leftrightarrow \varphi$. This will leave unaffected the general line of reasoning that follows below. Both variants are legitimate formalisations that can be read off Lynch's truth norm.

[13] Note that it already becomes obvious at this stage that the need to *formulate generalisations* is one of the main reasons for introducing the truth predicate in the first place (cf. Horwich 1998: 4, footnote 1, 31–33, and 122–125). The reason is that norm B, as opposed to TN, quantifies over sentences (substitutional quantification) rather than objects, which is unusual. Given only object quantification, the truth norm is the only way to generalise from such individual convictions like, say, that it is instrumentally good to believe that snow is white iff snow is white.

[14] The '≻' indicates an asymmetric relation. Here it represents (from left to right) explanatory priority, i. e. it represents which belief state explains the genesis of which further belief state. Note that the schematic '$G(B\varphi \leftrightarrow \varphi)$' corresponds roughly to Lynch's norm B, cited above, which, by applying the equivalence schema for propositions to its right-hand side, may very easily be converted to TN.

[15] Similar to Lynch, I do not mean that subjects literally believe that these norms hold, in the sense that they explicitly subscribe to them. It suffices that they behave in such a way that their actions show certain regular patterns that are as if these subjects would be following the norms in question.

swallow green pills, if anything. Or, to put it differently, the causal chain between pills and immortality – in conjunction with specific beliefs and wishes – explains the *success* of actions.

The corresponding belief norm – "It is instrumentally good to believe that green pills make immortal iff green pills make immortal" – only holds (if it holds) because the belief that green pills make immortal can – together with the wish to become immortal – lead to a particular type of action (e. g., swallowing green pills), which is especially successful in exactly those situations in which swallowing green pills causes immortality. Note also that it is this *particular* causal connection between pills and immortality and the *particular* success of one's belief about pills and immortality that justifies the further belief that the belief that green pills cause immortality is instrumentally valuable iff green pills cause immortality. One's success, in this example, consists in becoming immortal. And if it was not, given the assumptions, the *pills* that caused this success rather than the *truth* of 'that green pills cause immortality', we would be leaving the grounds of rational discussion.

In a nutshell, what explains instrumental success is, eventually, the contents of our beliefs, and not the truth of the propositions or sentence types used to describe them. Or, more precisely, it *is* truth that explains success, but in an absolutely innocent way. The truth of "Snow is white" consists solely in snow's being white. This is what is expressed by the T schema. So, in this way it is the truth of "snow is white" that explains the instrumental value of certain beliefs involving snow and whiteness. But this way of talking about things leaves unaffected our claim that what is causally efficacious when it comes to instrumental success is not truth – neither truth as such nor particular truths – but the contents of our beliefs.[16]

Stated thus, Lynch's argument seems question-begging. Why should his theory be regarded as an explanatory endpoint? Why not explain the validity of TN in terms of its instances in reverse? To be sure, in terms of generality his explanation is better off. One very general assumption accounts for the vast range of 'little belief norms', of which we conceded, if only for the sake of argument, that they are valid. But generality is not everything.

When it comes to norms, Lynch speaks of 'us' accepting certain norms and that 'we' follow certain rules. 'We' is probably 'we, human beings' or 'we, the average human'. That we accept TN – which we take for granted for the moment – must mean that judging from the observable (including verbal) behaviour of humans, one may conclude that humans implicitly accept this rule in that they act accordingly. Descriptively speaking, there is no difference with respect to the observable behaviour of cognitive agents relative to whether they endorse TN itself or its corresponding instances.[17] Hence, *theoretical* considerations decide between these two otherwise equivalent ways of description (armchair reasoning, if you like).

We assume that knowledge of TN is not innate. Now consider the following arguably simplified story of how we come to learn that true beliefs are instrumentally valuable. From a certain age, children start acquiring beliefs about the outer world. From time to time, they decide which action to take according to the beliefs they hold (in order to reach certain of their aims). Now, at some point in their life they start experiencing that not all of their beliefs correspond(ed) to reality (i. e. they were false). Ex post, they begin to realise that their actions (in terms of the desired outcome) have been structurally more successful whenever

[16] Horwich argues convincingly that in the context of instrumental value the sole reason to utilise the truth predicate is its function as a generalisation device (cf. Horwich 2006).

[17] An interesting difference between the individual belief norms and the general norm TN concerns confirmation. The 'validity' of individual norms may confirm TN, but not the other way around, as only instances can confirm laws (e. g., Hajek & Joyce 2008). In line with this, I assume that individual belief norms can *explain*—both in terms of genesis as well as in terms of justification—why TN holds (if it holds), but not the other way around.

they 'correctly' believed that p iff p. After some such experiences, children might begin to think: 'Uh, maybe this is a general phenomenon, and maybe it's *always* (or most of the time, at least) the case that I'm going to be more successful if my beliefs are correct'. This is to say, they acquire the generalised belief that, in general, it is (instrumentally) good to believe that p iff p. Following common deflationary explanation, they express this (if they do) by using the truth predicate, i. e. they endorse TN.[18] This line of explanation is in accordance with the picture according to which we move from single beliefs to a generalised belief about instrumental value of true beliefs:

$$G(Bp_1 \leftrightarrow p_1), G(Bp_2 \leftrightarrow p_2), \ldots, G(Bp_n \leftrightarrow p_n) \succ G(B\varphi \leftrightarrow \varphi) \succ \text{TN}.$$

Why did we sketch this – arguably highly simplified – picture of how TN might evolve? Because this is what required explanation, given that TN in fact holds (which is the assumption we started with). The important thing to note is that the *basis* of this explanation is individual beliefs, individual actions, and the correlations between successful actions and correct (i. e. true) beliefs that cognitive agents experience.

On Lynch's account, TN is the basic explanation with which we begin. But unlike individual beliefs, we can't take TN as a starting point. We assumed that (the belief in) TN is not innate.[19] For beliefs, we have more or less plausible stories of how they evolve.[20] There is no similar story for TN.[21] The alleged 'advantage' of TN is that it may explain why all true beliefs are instrumentally valuable. But if we *start* our explanation by postulating the norm, the supposed advantage seems to be outweighed by its disadvantages. The problem is that, leaving innateness aside, we have no idea of how TN might crop up in humans. However, if, alternatively, we *end* our explanation with TN (roughly along the lines just sketched), most 'advantages' of the rule remain. The story above ends with the belief that TN holds. The alternative does not even get started because it does not have the resources to explain how TN evolved in the first place.

One might object that the developmental alternatives just sketched do not affect the *justification* of the truth norm and the individual beliefs, respectively.[22] For two reasons this objection does not succeed. Firstly, we have assumed that TN holds. This means that people are right in thinking that true beliefs are instrumentally good. As shown above, truth is not required in order to *explain* instrumental success. (It is only required to express corresponding generalisations.) Hence truth is not required to *justify* beliefs about instrumental success. The ultimate justification for individual beliefs about instrumental success is rooted in the fact that people are able to notice correlations between the truth of their action-guiding beliefs and their success. Noticing these correlations is a plausible reason for people to believe certain things about particular instrumental successes. Asking for further justification would be beside the point. However, there is no such corresponding reason for the explanatory endpoint in Lynch's scenario, i. e. for TN. Hence, even if the focus is restricted to justification-related issues, the theory starting with individual beliefs and ending with TN is better off since its endpoint (individual beliefs) is well justified.

[18] This way of putting things is pretty much inspired by Horwich's line of reasoning (1998: 44–46, 2001, 2006, 2010: 57–77). McGrath (2005) also discusses the relationship between TN and B. Similar to the present account, he argues that the instances of norm B are, relative to one's theory of truth, explanatory basic.

[19] I doubt that Lynch would go as far as denying that TN is not innate in order to save his thesis.

[20] See any good textbook on mental representation.

[21] Apart from innateness, the only possible way out is to assume that the belief that TN holds has proven to be evolutionarily adaptive. Even if that was the case, we would still need to account for *that* by recurring to individual norms that, on an evolutionary time-scale, gave rise to this belief in the first place.

[22] This possible response was pointed out to me by an anonymous referee for this journal.

Secondly, even if Lynch's theory is read as dealing only with justification issues rather than with developmental aspects, the presented sketch excels his theory. Here is why. Both directions of explanation are pretty similar in two important respects. The direction starting from individual beliefs *ends with TN*, which is to say that if TN plays any role in justification, this is in line with this approach. For example, it might be that, once people have acquired TN via the developmental process that I have proposed, they then justify their individual beliefs about instrumental goods by appealing to TN. The present theory sketch explains why they are prima facie right in doing so. Moreover, since both directions of explanation are *equal* in that they both posit TN and in that they both are indistinguishable in terms of observable behavioural patterns they imply, something else must decide between the two. One further such 'something else' is that, *taken as theories that are solely concerned with justification*, one is compatible with a plausible developmental story whereas the other one is not.

We have seen that when it comes to instrumental value of true beliefs, the debate actually is about the appropriate order of explanation. Also, we have seen that taking TN as one's basic explanans is rather problematic. There are even more general considerations that undermine Lynch's approach to truth. It is to these that I now turn. Before one starts construing a theory of truth, one should become clear about what the theory is supposed to be about, what it needs to achieve to be 'successful', and, in general, when it can be considered to be adequate.

2.3 The Primacy of The Phenomena

A very general condition on theory building is what I dub the primacy of the phenomena. A theory, supposed to deal with a certain specified subject, needs to account for the relevant phenomena. In the case of truth, this means that we would begin construing our theory by asking: what are the truth-related phenomena? First of all, there is 'truth-talk', i. e. the fact that our everyday discourses involve the truth predicate. This, I take it, is *the* truth-related phenomenon.[23] It is legitimate to require of a theory of truth that it covers only those aspects of truth-talk that otherwise could not be accounted for, because that is the sole reason for having a truth theory in the first place. Almost everything involving the truth predicate may be explained without referring to truth. In particular, the success of true beliefs may be accounted for by psychology.

Leaving the status of generalisations aside, there are basically two options: saying that it is instrumentally valuable to believe that, e. g., pigs are animals iff *pigs are animals*, or to say that it is so valuable iff *it is true that* pigs are animals. And here we clearly see which option is preferable: we may explain the instrumental value of this belief (or any other) in terms of a psychological story of how beliefs concerning pigs guard one's behaviour in regards to pigs and how believing they are animals helps interacting with them if they really *are* animals. An accompanying theory of truth is only needed to explain our common way of formulating certain generalisations such as TN – just as the minimalist supposes.

The alternative – explaining success in terms of the truth of beliefs – is simply superfluous, since there is this equally general explanation that does not involve truth. Only if no alternative was available, we would need to use 'truth' in explaining things. It is basically the same line of argument that motivates refuting the redundancy theory of truth (if only the other way around; i. e. the theory covers *less* than required by the phenomena). If we only had the redundancy

[23] Acknowledging this is clearly not tantamount to saying that truth talk is the only relevant phenomenon. It is just that truth talk is the most *obvious* truth-related phenomenon, and truth talk is the only *universally accepted* such phenomenon. Still, minimalism, or deflationism more generally, is no linguistic theory but a proper theory of truth. In particular, it is a *metaphysical* theory. (Minimalism says that truth is only a logical, non-genuine property.)

theory of truth – i. e. only a theory that implies that every sentence of the form "T('p')" is identical in meaning to "p" – then a whole bunch of sentences would remain unexplained. And for just these kinds of sentences – generalisations and blind ascriptions – one needs to 'extend' one's theory, which is to say that one ends up with some variety or other of modern deflationism.

Unlike the arguments before, this argument does not presuppose that Lynch's account of truth is a false description of what is going on. Rather, the claim is that his approach goes against the principle of primacy of the phenomena. It is a *psychological* phenomenon that true beliefs foster one's success in reaching aims, but it is a *truth-theoretical* phenomenon that sentences involving the truth predicate are not identical in meaning to their "true"-rectified counterparts. And of course we *could* construe a theory of truth that was general enough as to also cover facts of the former kind. But since in this area we already *have* successful theories, it would be wrong to demand of a theory of truth to account for this aspect as well.

Most people probably tend to think that something like TN holds. Assume it does. How can the deflationary view account for this? We have seen that there are theories that explain the relation between instrumental success and true beliefs (or maybe combinations of theories involving theories of representation, action, and causation). We also have a minimalistic theory of truth that explains all truth-related phenomena. This latter theory does not imply TN. But it is *compatible* with the norm, for it explains why TN – although truth is no genuine property and hence is inefficacious – is the most natural way to *express* the relationship between success and true beliefs, albeit that the *eventual explanation* for this relationship comes from somewhere else (probably from cognitive psychology).

2.4 Deep Normative Facts

What about the even stronger norm BI? Remember that this norm says that

> (BI) it is more than instrumentally good to believe that p if and only if p.

Lynch claims that here we may see even clearer that minimalism lacks the resources to explicate the value of the 'cognitive good', true beliefs. I admit it really does. But unlike TN (and B), in the case of BI it is absolutely controversial whether the norm holds at all. We have already seen in the beginning that TN is formulated quite imprecisely. The CP clause is unspecific; the scope of the norm is vague; and we do not know what "good" is supposed to mean exactly, to name but the most obvious difficulties. The purpose of comparing TN to BI was to get a better grip on the meaning of "good". Assuming that the phrase "instrumentally good" is comparatively unproblematic, that was a success. (We largely ignored the 'ceteris paribus' issue because that would have led us too far afield.) With these provisos, TN seems plausible. BI, however, is much too underspecified to be even evaluated.

What does "more than instrumentally good" mean? On the most charitable reading of the argument, it means that people pursue truth for its own sake, i. e. irrespective of the (potential) instrumental value a true belief might have.[24] Lynch cashes out the validity of norms in terms of actual behaviour. That is to say, a given norm is said to hold depending on whether people in fact act in accord with the norm:

> [. . .] (TN) can't be derived from the purely nonnormative (T). Therefore, contra minimalism, that schema cannot fully capture everything *we believe* is true of truth. It can't capture all the *facts about truth*. (Lynch 2004: 109, my emphasis)

[24] This is not to say that this is a sufficiently detailed description of what "more than instrumentally good" means. Even with this bit of rephrasing it is still unclear what the norm is about, and whether people believe it.

BI is a norm concerning beliefs; it says that 'we' pursue truths (true beliefs) for their own sake. But we do not. There are *unbelievably* more propositions about which most of us never form beliefs – and never want to. Do ever people value the true belief that Kim Kardashian is a superstar? If yes, one may very easily think of further, uncontroversial examples, the number of which will trump by far the number of beliefs people actually value.

This concerns trivial truths as well as non-trivial ones. In effect, everything said in section 1.1 on 'ceteris paribus' conditions applies, mutatis mutandis, to BI as it applied to TN.[25] So I will not repeat those objections here. Rather, I shall evaluate two – arguably still tentative – clarifications by Lynch as to what BI amounts to.

The first argument is this:

> One of the lessons that I take from [Harry] Frankfurt's work is that there are at least two ways something can become important or more than instrumentally good for our lives—and therefore worth caring about. [. . . The second one] is this: something can become worth caring about for its own sake because the very act of caring about it for its own sake is good. [. . . A]rt may well have intrinsic worth; but coming to care about art can make art worth caring about. For caring about art [. . .] can all by itself imbue human life with meaning. Art is important to us, as we might put it, partly because its being important to us is important for us. [. . .]
>
> We can show that caring about the truth as such and for its own sake is part of a flourishing life. And that, surely, is enough to make truth worth caring about. (Lynch 2005: 335, emphasis original)

The argument seems to be that believing that truth is (more than instrumentally) good fosters flourishing lives. I shall not argue against this claim. The reason is that Lynch, claims to the contrary notwithstanding, considers the *instrumental* value of beliefs in this passage. Even if what he says in the last two sentences cited would be true (which I have doubts about), this is irrelevant for BI. A flourishing life is an aim that maybe can be reached by pursuing truth for its own sake. But that makes the pursuit of truth for its own sake (only) instrumentally valuable insofar that it helps reaching that aim.

A second argument in favour of non-instrumental value runs as follows:

> [. . . C]onsider Russell's scenario, that without our knowledge, the world began yesterday [. . .]. If we really lived in the Russell world, as I shall call it, almost all my beliefs about the past would be false. [. . . W]hen I now think about the worlds [actual world and Russell world] in so far as they are identical in instrumental value, there is a difference between the two worlds that matters to me. Even when it has no effect on my other preferences, I, and presumably you as well, prefer true beliefs to false ones. (Lynch 2004a: 503, emphasis omitted)

[25] Again, cf. Piller (2009) for some convincing illustrations of situations in which, even if substantive issues are at stake, it might be good (or better) to believe falsehoods than to believe the truth. Note that in addition to 'situational' reasons for why believing falsehoods is sometimes preferable there might even be, if you like, 'structural' reasons. For example, some people think that ex-post rationalisation is an evolutionary advantage. If that is true, then there are *general*, i. e. evolutionary, reasons for favouring falsehoods over truths in certain circumstances. To be sure, evolutionary advantage is something like an 'aim'. But if believing the truth is sometimes evolutionary disadvantageous, then it can't possibly be that believing the truth is more than instrumentally good. (I am assuming that evolutionary disadvantages are bad.)

If this argument is sound, it shows that truth is non-instrumentally good. Moreover, if it is sound, it shows that minimalism about truth is false, for minimalism lacks the resources to explain the non-instrumental value of beliefs.[26] Fortunately, the argument is unsound.

First of all, the alleged fact that Lynch and his readership prefer true beliefs is insufficient reason to assume that true beliefs are more than instrumentally good. But leave that aside. The argument apparently is based on plausibility considerations. Lynch thinks that his readers, like himself, plausibly prefer living in the actual world over living in a Russell world. I doubt that. We live in the actual world, and the actual world, as Lynch assumes, is no Russell world. Even worse, the Russell scenario is so far-fetched that our theory of truth can't possibly be based on intuitions concerning such a world. Lynch's argument, as I see it, is based on the claim that most (all?) people have the intuition that living in a non-Russell world is better than living in a Russell world.

However, our intuitions are shaped by the world we live in. (Les us ignore cultural and social differences for the moment, which would further complicate matters.) In particular, with respect to many of our intuitions we are trained relative to what is possible in our world. Moreover, our intuitions concerning truth are to a large extent shaped by the specific instrumental successes of true beliefs. The Russell world does not show that believing the truth is more than instrumentally good, for, if truth and falsity do not affect instrumental success of beliefs (as, by assumption, they do not), then there is no reason to favour the truth in the first place.

Probably many people would agree with Lynch that living in the actual world is better than living in a Russell world. That is absolutely plausible, given that their intuitions are shaped by experiences in a world where truth and falsity make a difference. In a Russell world, very many of our beliefs are false. Most theories of physics are false in a Russell word: there has never been a Big Bang, for instance. Still, the universe looks exactly as implied by the Big Bang Theory, i. e. the accuracy of the Big Bang theory in regard to the observable phenomena is the same in both worlds. In other words, the relationship between truth and *instrumental* success is completely disentangled in a Russell world. (That is the point of the thought experiment.) Since our intuitions are shaped by the actual world in which truth and falsity make a difference, these intuitions can't be expected to yield any *reliable* results if applied to scenarios in which that connection between truth and instrumental success is lacking.

Note that Lynch's thought experiment has the interesting result that, in a Russell world, the instrumentally successful beliefs are false ones. The successful beliefs are exactly those beliefs that would be true in the actual world (our world). For example, if I want a hard-boiled egg in a Russell world, then the most efficient way to achieve this would be to believe that the egg-cooking theory of the actual world is true.[27] This applies to all beliefs in a Russell world,

[26] This does not undermine the plausibility of minimalism, for Horwich's argument in favour of (something like) BI is unconvincing anyway. He argues for the non-instrumental value of truth by saying (i) that 'there is a widespread sentiment that certain items of knowledge are desirable regardless of any practical use', (ii) that it would otherwise 'be hard to justify our pursuit of truth in fields of inquiry such as ancient history', and (iii) that 'it is surely no less important to pursue truth [. . .] in normative domains' (Horwich 2010: 65, emphasis omitted). Ad (i): that a *conviction* is widespread does not make it right. Ad (ii): ancient history as a field of inquiry *is* (relatively) hard to justify; justifying its pursuit for truth, on the other hand, is not (in terms of instrumental value). Ad (iii): there is no proof that people do pursue (or should pursue) truths in normative domains *independently from any* instrumental value. Interestingly, Horwich (2013) more recently admits that at least *the belief* that truth is non-instrumentally valuable stems from the frequent instrumental success of true beliefs. So the non-instrumental value of normative truths (if there is such a thing) is not completely independent from the instrumental value of truths more generally.

[27] To be sure, true beliefs would *also* lead to successful actions, but in a less efficient way. That is because each true conviction would be a belief similar to the corresponding belief that is true in the actual world *plus* accompanying

which is to say that one is in general instrumentally more successful if one believes falsehoods. (Although these must be particular falsehoods, of course.)

So, BI and minimalism are, as it is, probably incompatible. (Though Lynch has not presented a proof for this.) But this does not undermine minimalism's plausibility, since (i) the phrase "more than instrumentally good" is, unlike its (instrumental) counterpart, particularly underspecified, and (ii) given that unreliable intuitions about far-away possible worlds is the only support in favour of BI, its justification is highly dubious anyway.

3 Conclusion

In the beginning, I promised to do two things. Firstly, to explain how what you – correctly, maybe – believe about why true beliefs are valuable may be accommodated within a minimalistic theory of truth. To this end, I discussed the 'directionality of explanation' in the context of truth norms. I suggested that the advantages of going from individual beliefs to generalisations, rather than the other way around, outweigh the disadvantages (the need to explain general facts in terms of particular facts). I argued that, generally, a theory of truth would be the wrong place to look for an explanation of instrumental success. Secondly, I promised to show that not all truth norms commonly thought to hold are supported by the available evidence. In particular, I argued that although true beliefs might be instrumentally good (a result that is compatible with minimalism), they are probably not good in a non-instrumental sense.

Several minor issues further undermine Lynch's position. My worries concerned mainly the ceteris-paribus condition involved. It is simply false that believing true things is good tout court. But adding the proviso 'all else equal' is renaming the problem, not solving it. Other worries concerned the term "good", which is largely unspecified on Lynch's account, so that both its meaning and scope remain obscure.

Truth is, after all, non-instrumentally valueless. That is because truth is no genuine property. All plausible reasons for thinking that true beliefs are good stem from the instrumental success of true beliefs in guiding one's action. However, these norms (if true) can be accounted for by a minimalistic, Horwich-style theory of truth. In order to explain instrumental value, we do not need to postulate that truth is a property (and hence something to associate value with). Besides instrumental success, there is no value in the area of truth.[28]

References

Coates, A. (2009), Explaining the Value of Truth, *American Philosophical Quarterly* **46**, 105–115.

David, M. (2005), On 'Truth is Good', *Philosophical Books* **46**, 292–301.

Dummett, M. (1959), Truth, *Proceedings of the Aristotelian Society* **59**, 141–62.

Heal, J. (1987), The Disinterested Search for Truth, *Proceedings of the Aristotelian Society* **88**, 97–108.

Earman, J., Roberts, J. and Smith, S. (2002), Ceteris paribus lost, *Erkenntnis* **57**, 281–301.

Hajek, A. & Joyce, J.M (2008), 'Confirmation', *in* S. Psillos & M. Curd, eds., 'The Routledge Companion to Philosophy of Science', Routledge, Abingdon, pp. 115–128.

Horwich, P. (1998), *Meaning*, Clarendon Press, Oxford.

background assumptions about the peculiarities of Russell worlds. Hence, the fastest – instrumentally most successful – route to a hard-boiled egg is via the false assumption that the egg-cooking theory holds.

[28] This text benefited a lot from the comments that I got when giving a talk about this topic at Professor Schurz's research seminar at Düsseldorf University in November 2013. I would like to thank everyone for their questions and remarks, in particular Alexander Christian, Christian Feldbacher, and Matthias Unterhuber.

Horwich, P. (2001), 'Norms of Truth and Meaning', *in* R. Schantz, ed., 'What is Truth?', de Gruyter, Berlin, pp. 133–145.

Horwich, P. (2005), *Truth*, Clarendon Press, Oxford.

Horwich, P. (2006), The Value of Truth, *Nous* **40**, 347–360.

Horwich, P. (2010), *Truth – Meaning – Reality*, Clarendon Press, Oxford.

Horwich, P. (2013), 'Belief-Truth Norms', *in* T. Chan, ed., 'The Aim of Belief', Oxford University Press, Oxford, pp. 17–31.

Lynch, M. (2004), *True to Life. Why Truth Matters*, MIT Press, Cambridge, MA.

Lynch, M. (2004a), Minimalism and the Value of Truth, *Philosophical Quarterly* **54**, 497–517.

Lynch, M. (2005), Replies to Critics, *Philosophical Books* **46**, 331–342.

Lynch, M. (2009), *Truth as One and Many*, Oxford University Press, Oxford.

Lynch, M. (2012), The Many Faces of Truth: A Reply to Some Critics, *International Journal of Philosophical Studies* **20**, 255–269.

McGrath, M. (2005), Lynch on The Value of Truth, *Philosophical Books* **46**, 302–310.

Schaffer, J. (2008), 'The Metaphysics of Causation', *in* Edward N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy'. Link: http://plato.stanford.edu/archives/fall2008/entries/causation-metaphysics/; accessed 30/10/2013.

Wright, C. (1994), *Truth and Objectivity*, Harvard University Press, Cambridge, MA.

Wright, C. (2003), *Saving the Differences*, Harvard University Press, Cambridge, MA.

abstracta Linguagem, Mente & Ação ISSN 1807-9792 Volume 7, Number 2, pp. 19–39, 2013

The Two-Component Theory of Proper Names and Kripke's Puzzle

Jee Loo Liu

Department of Philosophy, California State University, Fullerton
jeelooliu@gmail.com

Abstract

This paper provides a defense of the description theory of proper names by constructing a 'two-component' theory of names. Using Kripke's puzzle about belief as the stepping stone, this paper first points out problems with Kripke's direct reference theory of names. It then presents the two-component theory of names and defends it against Kripke's general criticisms of the description theory. It also compares the two-component theory of names against other leading description theories and shows how the two-component theory provides a better analysis of names. The paper offers a comprehensive summary of the debate between the description theory and the direct reference theory of names. At the end, it shows how the two-component theory of names can deal with Kripke's puzzle and more.

Introduction

Kripke's puzzle is an old and familiar story. It was put forward in Kripke's "A Puzzle about Belief" (1979). But even today it still has such a charm that people are drawn to it time and time again. In this paper I shall use his puzzle as the stepping stone for developing a refined description theory of proper names.

The debate between the direct reference theory and the description theory is first and foremost related to the issue of *reference*. As Searle puts it, "both theories are attempts to answer the question, 'How in the utterance of a name does the speaker succeed in referring to an object?'"(Searle 1983, 234) According to the direct reference theory, names refer directly; that is to say, nothing that mediates between the name and its referent is semantically significant. On the other hand, according to the description theory, names refer in virtue of the descriptions associated with the use of the name. So the main issue being debated on is this: Is the reference of proper names *mediated* by any description? Secondly, the debate can also be construed as a debate concerning the *meaning* of proper names. The direct reference theorists argue that the *semantic value* of a proper name is simply its referent. The description theorists, on the other hand, argue that the semantic value of a proper name is the set of descriptions associated with the name. This characterization of the core issue demonstrates further that the issue of meaning (or the semantic value) of names is closely related to the issue of reference. Thirdly, Searle (1983) presents the debate as a debate between internalists and externalists. He says, "The issue is simply this: Do proper names refer by setting internal conditions of satisfaction … or do proper names refer in virtue of some external causal relations?" (Searle 1983, 233) Description theories emphasize the speaker's intentional content associated with the name, while direct reference theories appeal to the actual causal chain between the name and its usage outside the speaker's mind. Lastly, Kripke seems to think that this is a debate between subjectivism and objectivism when he says, "It is not how the speaker thinks he got the reference, but the actual chain of communication, which is relevant." (Kripke 1972/1980, 93) All these interpretations of the key issue demonstrate the fact that the issue of reference, albeit a pragmatic issue, is not separable from the issue of semantics, the issue of the objective semantic value of proper names. The last two interpretations (Searle's and Kripke's) further point to the fact that the debate ultimately revolves around the issue whether the speaker's psychological states (what she believes or how much she knows of the referent) play any role in determining the referent of the name (in her usage).

Kripke tries to defend his direct reference theory against the charge that it cannot explain the role of proper names in an epistemic context (such as belief, thought, etc.). There are many famous puzzles involving substitution *salva veritate* for different names of the same referent, and the description theory can easily dissolve them by suggesting that different names have different *senses*. These puzzles were considered to be defeating the direct reference theory of proper names. Kripke thus tries to demonstrate a similar puzzle that does not involve different names, and thus does not involve different *senses*. Using his *principle of disquotation* and *principle of translation*,[1] Kripke presents a puzzle which features a Frenchman Pierre, who is attributed the following set of beliefs:

(1) Pierre believes that London is pretty.

(2) Pierre believes that London is not pretty.

According to Kripke, the two belief reports attribute a contradiction to Pierre, even though Pierre himself cannot be interpreted as being inconsistent.[2]

Kripke also discusses another puzzle, which invokes only the principle of disquotation and no translation is involved. This is the example of Peter's two beliefs concerning the politician/musician Paderewski. In this case, we get a similar set of contradictory belief reports:

(3) Peter believes that Paderewski has musical talent.

(4) Peter believes that Paderewski has no musical talent.[3]

Kripke thinks that these puzzles generate the same difficulty for both the direct reference theory and the description theory. The conclusion he draws from these puzzles is that they reveal a general feature of belief contexts that such contexts resist substitution, and the failure of substitution has no bearing on whether one adopts a direct reference theory or a description theory.

There are numerous approaches in dealing with Kripke's puzzle:[4]

1. Stopping the generation of the puzzle: One could reject one or both of Kripke's principle of disquotation and principle of translation, so as to terminate the generation of these puzzling cases.[5]
2. Biting the bullet: One could simply accept the verdict that Pierre and Peter have inconsistent beliefs and argue that *we all do*, thereby showing that the puzzle is *no* puzzle at all.[6]

[1] The two principles can be stated as follows:**[PD]** If a normal English speaker, on reflection, sincerely assents to '*P*', then he believes that *P*.**[PT]** If a sentence of one language expresses a truth in that language, then any translation of it into any other language also expresses a truth (in that language).

[2] Kripke says that we must say that Pierre has contradictory beliefs, that he believes that London is pretty *and* he believes that London is not pretty, even though Pierre himself "cannot be convicted of inconsistency." (Kripke 1979, 122)

[3] For details of these puzzles, see Kripke (1979).

[4] For discussions on these puzzles, see Marcus (1981); Pettit (1984); Kvart (1987); Over (1983); Corlett (1989); Salmon (1986); McMichael (1987); and Loar (1987).

[5] For example, see Marcus (1981).

[6] For example, see Martinich (1997).

3. Dissolving the puzzle: One could give proper names a different analysis so that the puzzle gets *dissolved* under this new analysis.

My approach is of the third kind. Following Marcus and Katz, I argue that Kripke's puzzle applies only to a direct reference theory such as his own.[7] There are, of course, other versions of the direct reference theory that may avoid generating this kind of puzzle. The new direct reference theorists (such as Nathan Salmon, Mark Richard and Gareth Evans) incorporate some elements of the description theory into their direct reference theories. What I am developing in this paper, on the other hand, is a *new* description theory of proper names that incorporates some elements of the direct reference theory into the description theory. I shall also explain why we should have a description theory rather than a direct reference theory, even though the two sides are meeting in the middle ground. Since the decline of the description theory of proper names follows from Kripke's attack, my paper will treat Kripke's numerous criticisms of the description theory as the main challenge for my new description theory.

In what follows I will first briefly explain why Kripke's theory of proper names does not give us the whole story. I will then introduce my theory which I call the *two-component description theory of proper names*. My proposal will be based on the rejection of the commonly assumed sharp separation between semantics and pragmatics. Using some of the familiar cases Kripke sets up against the traditional description theories, I will explain how my theory gives a different story. Finally, I will go back to Kripke's puzzle and show how my theory can avoid attributing a contradictory set of beliefs either to Pierre or to Peter, and thereby dissolve the puzzle that Kripke poses for the description theory.

1 The Insufficiency in Kripke's Theory of Proper Names

For Kripke, proper names are "rigid designators" in the sense that they designate the same individuals across possible worlds. However, what a theory of names should explain, first and foremost, is not how reference gets fixed *across possible worlds*, but how reference gets fixed *in our actual world*. I think Kripke gives us too simplified a story in the latter respect. For example, Kripke talks about 'Nixon' as being fixed in our world. Kripke seems to assume that we all know which Nixon it is, to whom we then assign all the possible situations. However, it is not the case that if one simply mentions the name 'Nixon,' the name itself will do the job of getting the correct person being discussed. Suppose someone names his dog after the former president Nixon.[8] One day the dog owner is on his way home, and his neighbor, who takes interest in politics, informs him: "Nixon is dead." In this case the dog owner would most likely take it to mean that his dog Nixon had died. Some explanation is required in this context to reveal the fact that it was the *former president Nixon*, not *the dog Nixon*, who died that day. The reason why proper names alone are insufficient is that there might be, and in fact usually is, more than one individual who is called by that name. If a proper name *used in our world* can have more than one bearer, *which* is the one we single out in our discourse? Kripke assumes that context will usually do the job of disambiguating names with multiple bearers.[9] However, there are many contexts in which the usage of a name gets different associations from one speaker to another, or there may be multiple causal chains linking the tokening of a name in a single

[7] Marcus argues that the puzzle is "a predicament that is generated by the theory of direct reference of names taken in conjunction with a plausible disquotation principle relating belief to assent." (Marcus 1981, 501) Katz also argues that Kripke's puzzle "isn't a puzzle for description theorists, since they reject Mill's view of proper names." (Katz 1990, 32)

[8] This example is a spin-off from Kripke's own example of 'Aristotle.' See (Kripke 1980, 8).

[9] Kripke says, "In practice it is usual to suppose that what is meant in a particular use of a sentence is understood from the context." (Kripke 1980, 9)

context to various referents. Our job here is to decipher what "context" consists of and how to analyze cases of single names with multiple bearers.

Kripke appeals to a causal chain picture as a support for his direct reference theory. He says, "Someone, let's say, a baby, is born; his parents call him by a certain name. They talk about him to their friends. Other people meet him. Through various sorts of talk the name is spread from link to link as if by a chain." (Kripke 1972/1980, 91) According to such a picture, names are supposed to designate *the one* originally dubbed by that name, and as the name passed on to us through a chain of communication, we *intend to* use the name to refer to the same person referred to by the previous user of the name. Thus, "a speaker who is on the far end of this chain" is able to, simply by virtue of using the name, refer to the individual initially dubbed that name. The problem with such a picture is that there are often *multiple* causal chains from one name leading back to different objects named.[10] In most cases of proper names, there may be parallel causal chains linking different referents to the same name (type). From one name mentioned, it might not be so clear which one of the objects named was at the beginning of this causal chain of communication. In Kripke's causal theory, communication depends on the audience's intention to use the same reference as the speaker does.[11] But such an intention does not guarantee success.

Suppose someone visits a Swiss museum and the tour guide introduced a painting as one done by Giacometti. Since the visitor only knows of one Giacometti, the one who made those slender-shaped figures (i. e. Alberto Giacometti), the visitor thought that this painting was done by *him*. However, the painting was actually done by Gustav Giacometti, the sculptor's father. When the tour guide announced:

(5) This was painted by Giacometti,

her utterance was associated with Gustav Giacometti. On the other hand, if the visitor tries to report this information to a friend by uttering **(5)**, the utterance would be associated with Alberto Giacometti. Communication via causal chain goes astray in this case.

This is a case of homonymous names, which is a common linguistic phenomenon in many languages. In his amended *Preface* to the 1980 edition of *Naming and Necessity*, Kripke added a brief discussion on the problem of homonyms. He suggests[12] that we treat homonymous names as *distinct* names, just as we typically treat homonymous words as distinct words (Kripke 1980, 8). I agree that it is natural to treat homonyms, such as 'bank' (as riverside) and 'bank' (as financial institution), as different words because they have different meanings. However, to argue that homonymous names should be treated as different *names* simply because they "name distinct objects" (or are connected with different causal chains) is to beg the question. What we are trying to settle here is exactly whether the referent (the object) constitutes the *meaning* of the name. This principle of individuation of names, as Kripke himself acknowledges, also "does not agree with the most common usage" (*Ibid.*) In English, as in other languages, people would say "we have *the same name*," when it is the same word used in their names. We could of course employ the type/token distinction, and say that these people have the same type of name but different tokening. Nevertheless, this distinction still reveals the fact that the various tokenings of the name have *something in common*. It would be contrary to our linguistic practices to

[10] Katz (1994, 18) makes the same criticism of Kripke's theory.

[11] Kripke says, "When the name is passed from link to link, the receiver of the name must, I think, intend when he learns it to use it with the same reference as the man from whom he heard it." (Cited in Erwin, Kleiman & Zemach 1976, 52)

[12] To be sure, Kripke actually does not wish to commit to this particular principle of individuation of names, or such a view that seems to assimilate proper names to demonstratives. He said, "I should stress that I am not demanding or even advocating this usage, but mention it as a possibility to which I am sympathetic." (See Kripke 1980, 8-10, fn. 9-12)

say that names are distinct simply because their referents are distinct, or as Kripke claims, "distinctness of the referents will be a *sufficient* condition for distinctness of the names." (*Ibid.* original emphasis)

The basic assumption behind direct reference theorists is that reference is an objective fact of the use of names in our linguistic community. Kripke posits a causal chain between the mention of a name and the object initially baptized (either explicitly or implicitly) with such a name. Keith Donnellan appeals to a historical connection, viewed by " an omniscient observer of history," between the use of a name and the referent of the name.[13] What the above example shows, is that the occurrence of a proper name in daily discourse does not automatically reveal the chain(s) behind the usage of that name. The first speaker, using the name Giacometti, referred to the painter, while the second speaker, even with the intention to follow the first speaker's usage, actually referred to the sculptor because his other background knowledge superseded the other intention. When there is more than one possible referent of the name, change of speaker associations may take place from one speaker to the next. The objective causal/historical chain itself is insufficient to secure successful transmission of reference. In order for the audience to *fix* the right one that bears the name, some other mechanism is required. According to Edward Erwin, Lowell Kleiman & Eddy Zemach, "Donnellan himself suggests that the speaker's intentions are also relevant." However, they continue, if the historical/causal theory of reference is to be supplemented with the speaker's intention, then the theory will be rendered either "untrue" or "unilluminating" because this simply shows that direct reference, whether it is the historical connection or the causal chain, is insufficient in securing reference (Erwin, Kleiman & Zemach 1976, 54).[14] What we need, I argue, is a description of the speaker's mental associations of the name. In other words, we need an *internalist* theory of names. This is where the call for a description theory comes in.

2 The Two-Component Description Theory of Proper Names

What I propose here is a *two-component description theory of proper names* (in short, TCD). I think two questions should be separated: "What does the name mean?" and "How does the name refer?" According to TCD, descriptions are associated with the use of proper names in two ways: One is the description that gives the meaning of the name, such as "an individual called such and such (by a certain linguistic community)." The other is the set of descriptions that the speaker *would use*, if asked, to specify the intended referent. These two descriptions compose the *sense*, or the semantic value, of a proper name. We shall treat the two kinds of descriptions as an ordered pair:

[P] < "an individual called 'F' (by a certain linguistic community)," Φ = a set of descriptions >

The first description determines the *denotation* of the name. The set of descriptions φ, on the other hand, *fixes the reference* of the use of a name by that speaker. I distinguish 'denoting' and

[13] As Donnellan puts it, "It might be that *an omniscient observer of history* would see an individual related to an author of dialogues, that one of the central characters of these dialogues was modeled upon that individual, that these dialogues have been handed down and that the speaker has read translations of them that the speaker's now predicating snub-nosedness of something is explained by his having read those translations. This is the sort of account that I have in mind by a 'historical explanation'." (Donnellan 1977, 230, my emphasis)

[14] Erwin *et al* further argue that adding the speaker's intention is bringing back a descriptive account of names. They continue, "However, the historical theory of reference was developed to replace such a 'descriptivist' account, i. e. one which holds that successful reference requires that the speaker have such a capacity." (Erwin, Kleiman & Zemach 1976, 57)

'referring' in roughly the following way: *denoting* is a semantic relation; it is something that a name does, *referring* is a pragmatic relation; it is something that the speaker does.[15] Even though the distinction made here is not commonly adopted, it has been adopted by others using different terminology. Kripke distinguishes 'semantic reference' and 'speaker reference' in the case of descriptions. According to him, "If the speaker has a designator in his idiolect, certain conventions of his idiolect determine the referent in the idiolect: that I call the *semantic referent* of the designator...The speaker's referent is the thing the speaker referred to by the designator, though it may not be the referent of the designator, in his idiolect." (Kripke 1977, 256-57) Bach also seems to make a similar distinction when he talks about "speaker reference" and "linguistic reference." (Bach 1984, 141) This is the kind of distinction I intend to capture, but I prefer to separate the two terms. Following Donnellan, I shall say that "referring is not the same as denoting" (Donnellan 1966, 236), and I use "denotation" in roughly the same way that Kripke uses "semantic reference." But with regard to "reference," or "speaker's reference" in Kripke's terminology, I differ from Kripke as well as Donnellan in one major respect: they both think that the speaker can refer to something outside the realm of semantic reference, while I take the speaker's reference to be *constrained by* the semantic reference. (More on this later.)

The *denotation* of a proper name is a set, which consists of members whose qualifying property is that they are all called by that name (by a certain community).[16] The denotation sets the *range* of possible referents, and the speaker's associated descriptions get us to the particular referent *within this range*. Now I shall discuss these two components separately.

The First Component

The first component of the descriptive *sense* of a proper name is its *meaning*, which is different from the *sense*. In contrast to Kent Bach's analysis, I argue that the meaning of a proper name is *not* analyzed as a definite singular term "*the one* called such-and-such" but as an indefinite singular term "*an* individual called such-and-such (by a certain community)."[17] For instance, the meaning of the name 'Nixon' is 'an individual called "Nixon",' and the denotation of the name is the set of all members called 'Nixon.' Since a name is generally used to designate one member of the denoted set, not the whole set, we should further distinguish the denotation of a name itself and the denotation of the name *in use* in the following way: the denotation of the name 'F' in the language is a set, call it λ, and λ = {all individuals called 'F'}, while the denotation *in each particular use* of 'F' is one member of λ.[18]

[15] As Strawson says, "'Mentioning' or 'referring,' is not something an expression does; it is something that someone can use an expression to do." (Strawson 1956, 223-224)

[16] The condition in the parentheses is used to rule out any arbitrary stipulation by one individual. Furthermore, the condition brings in the linguistic community in which the calling relation gets established. This is a three-term relation between the name, the object named, and the linguistic community, and is thus different from the two-term relation between a name and its bearer as championed by Katz. For the contrast of the two views, see Katz (1990, 38-9).

[17] This is basically in agreement with Tyler Burge's *predicate* treatment of proper names. Burge argues that even though proper names are usually used in singular and unmodified form, "they play the role of predicates, usually true of numerous objects–on all occurrences." "[Proper names] play instead the roles of a demonstrative and a predicate. Roughly, singular unmodified proper names, functioning as singular terms, have the same semantical structure as the phrase 'that book.' Unlike other predicates, proper names are usually used with the help of speaker-reference and context, to pick out a particular." (Burge 1973, 431-32)

[18] Katz seems to have made a similar distinction between the set and the individual. He calls the former "type reference" and the latter "token reference." Katz says, "Let us call the referent of a word or expression its 'type reference,' and let us call something referred to in the use of a word or expression its 'token reference'." (Katz 1977, 35-36) In Katz (1990) he also states, "The type-reference of a proper noun is the collection of its bearers," and "the criterion for a literal application of a token of the type is that the referent of the token belongs to the type-reference." (Katz 1990, 48)

In the above analysis of the meaning of proper names, "the calling relation" and the parentheses "(by a certain community)" both need some explication. The *calling* relation is not merely established through one's being addressed to in a certain way. "Being called as" is also one's being mentioned as, one's being referred to as, one is being introduced as, one's being spoken of as, etc.[19] Someone's shouting "Hey, you!" at someone else certainly does not establish "Hey, you!" as the name for that other person. Furthermore, such a calling relation is established through a communal act; it is what the linguistic community jointly has done that establishes the use of a proper name. For the purpose of communication, the use of the name cannot be restricted to only one person (even though it could be stipulated by one person initially). For instance, one may be directly addressed to by the title 'Sir' by some people, but one would not normally be called by and mentioned as 'Sir' by all members in the community.[20] This added condition ("by a certain community") rules out such calling relations as a naming relation. There are, however, no strict linguistic rules, other than conventional usage, governing the use of names. 'Seven' could be a name; 'Moon Unit' could also be a name. Even a title could become a name when it is used by a community as a name. 'Jack the Ripper' is an example at hand.

Furthermore, my theory proposes an *indexical* treatment of the usage of proper names; that is, it treats names as *indexed* to a certain linguistic community. Kripke's causal chain picture gives us a model explaining how an individual falls into the denoted set of a proper name, but one amendment needs to be made to that picture: The community involved does not have to be the original community to which the object named belongs. When we are dealing with proper names in English, the English speaking community is the linguistic community relevant to the case. A proper name, take 'Socrates' for example, should be given the following analysis:

[S] 'Socrates' = "an individual called 'Socrates' by *our* linguistic community"

This treatment does not rely on the fact that Socrates was called 'Socrates' in his times since it is pretty clear that he *was not*. 'Socrates' is a name we use for him in English. Names used in English, however, are not necessarily English names. 'Mitterrand' and 'Chirac' are good examples. Proper names have an interesting status in that they are sometimes not translated and are directly used in a different language. But there are no hard and fast rules about translation for proper names. Many names are not translated (such as different family names in Latinate languages) while many are (such as names of places or first names like 'Peter' and 'Pierre,' 'John' and 'Jon,' etc.). Instead of concluding, as some do, that names are thus not part of a language, I argue that names are *indexed* to the language and it is the language users who determine whether those names are to be translated, or to be incorporated into the language as such. The key point of this analysis is the indexical term 'our.' '*Our*' is indexical to the user of the name, not fixed to the present writer and readers of this paper (that is, not fixed to English).[21] Our analysis preserves this 'semi-independent' status of proper names: they can either be translated or directly used in a different language. When they are translated, the translated names become the names for the individuals in the new language. And when they are directly used, those original names are the names for those individuals in the new language as well. Even if the present linguistic community is not the one that initiated the *calling* relation, as long as the

[19] These descriptions are all included in the definition of the word 'call.'

[20] If, however, everyone in the community calls a person 'Sir' and mentions him as 'Sir' to others, then I would think that the person has adopted 'Sir' as his name. In a similar way, nicknames become proper names.

[21] In the present discussion on proper names, we are dealing only with English and the English speaking community. Thus, the parentheses (by our linguistic community) will sometimes be omitted for simplicity.

name in the initial language is preserved in the present language, we can say that the object named is called such-and-such by our present linguistic community.[22]

Kripke assumes that the usage of a name is passed on through a causal chain to refer back to the individual originally dubbed. But it could very well happen that *somewhere* in the causal chain an error occurred, and a historical person (or a remote object) that is called by our name was not originally so called. With a historical figure, for example, it might well have happened that somewhere in the historical chain there is a translation or even a mis-translation of the original name, such that the name which certain later communities (such as ours) come to use is no longer the same name *given* to the individual by the original community.[23] The same could happen to the name-passing chain of any remote object. Searle (1983) gave us an example presented initially by Gareth Evans. Searle writes, "'Madagascar' was originally the name of a part of Africa. Marco Polo, though he presumably satisfied Kripke's condition of intending to use the name with the same reference as 'the man from whom he heard it,' nonetheless referred to an island off the coast of Africa, and this island is now what we mean by 'Madagascar'." (Searle 1983, 237) Furthermore, there are also cases of broken causal chains of names, such that we would have to conjure up a name to accomplish referring. For instance, we call the prehistoric woman whose remains were recently found 'Lucy' even though it is certain that she was not *named* so initially. When anthropologists talk about *Lucy*, they are talking about that woman. With all these cases, the calling relation allows the names to be legitimate names. As long as our linguistic community *intentionally* uses the name to refer to someone or something, the name is established for the object named.

The Second Component

Taken by itself, a name (in each use) only signifies an indeterminate object that is a member of the denoted set. But in our daily discourse, a proper name is always associated with one particular object. I argue that proper names cannot single out particular objects without the speaker's intention, which is specifiable by a set of descriptions. The second component of the *sense* of a proper name is this set of descriptions *in virtue of which* the speaker accomplishes reference to a particular agent. The importance of the speaker's intention in aiding a proper name to fix the reference is especially obvious in cases of first name reference. In cases of people sharing the same first name, the mere mentioning of the name itself is not going to do the trick to secure reference. In a successful communication when one is talking about a particular person with a common first name, it requires the audience's mental work to grasp whom the speaker *has in mind*. In general, communication is a mental game, in which all participants need to abide by the same linguistic rules. In the case of using proper names, in particular, the participants need to possess common background knowledge, shared intuitions, etc. in order to be successful in reference with proper names. There is no guarantee, however, that reference is always successful in the case of proper names.

More should be said about how this set of descriptions aids in securing reference. A proper way to understand how the associated descriptions fix the reference is to take them to express the speaker's beliefs of the object. The speaker may believe them to be true, but some of the beliefs may turn out false. Even if some or all of the beliefs are false of the referent, they are nonetheless beliefs *about* that referent. This is the speaker's intentionality at work. In gen-

[22] How the name gets passed on from other languages to the present one could be explained by homophonic translation, semi-homophonic translation, or replacement of the original name with a new name, etc. It is generally assumed, though not guaranteed, that the names in the present language have legitimate sources in other languages.

[23] For example, Moses might not have been called 'Moses,' but it is the name *we* come to use to refer to the one that did all those things the Bible attributed to *him*. More on Moses later.

eral cases of using names, the speaker has something/someone in mind, and intends to refer to this particular object by the use of the name.[24] The speaker does not need to have direct acquaintance with the object in order to have some beliefs about it. Naturally, these beliefs may be perceptual beliefs, and thus what the speaker would describe is based either on his/her mental images (Cindy - 'the woman with red hair') or on the conditions under which he/she established physical contact with the object (George - 'the man whom I met yesterday'). However, the descriptions could also be based merely on knowledge by description. In those cases the descriptions would describe a piece of information the speaker previously acquired of the object (Plato - 'the one who wrote *The Republic*'); or the circumstances under which the speaker acquired that information (Osama bin Laden - 'the one I read about in the newspaper'). Therefore, speaker-associated descriptions do not merely describe properties the object presumably has; they describe also the conditions under which the speaker came to know about that object. Even if the descriptions of the speaker's beliefs may be false of the object, the descriptions of the *epistemic conditions* would establish some 'parasitic links' between the speaker and the referent.

The direct reference theorists can point out that all the descriptions the speaker has might turn out false of the referent. This would be the same criticism Kripke has against traditional description theories: It may happen that Moses never went into politics; it may happen that Socrates was not snub-nosed; it may happen that Plato was not Aristotle's teacher, it may happen that Gödel did not prove the incompleteness of arithmetic, etc. If reference is accomplished by a definite description, then it can fail in these cases. According to TCD, however, reference is not merely a two-place relation between the (intended) referent and the supposed property of the referent. In other words, it is not in virtue of any property specified as a definite description of the referent (such as "the tallest spy in the world"), that secures the reference (Sam, if Sam is indeed the tallest spy in the world). On TCD, reference is a *three-place* relationship between the speaker, the referent, and the supposed properties of the referent believed by the speaker. The speaker refers to the object via the set of properties she associates with the object, the same properties that she uses to single out the referent from the denoted set. Even if the property, "being the tallest spy in the world," may be false of Sam, the property "being taken to be the tallest spy in the world by the speaker" is nonetheless true of Sam. In some cases, the speaker may have only skimpy information about the referent that she has heard about and she may be unable to provide a uniquely satisfying set of beliefs. Her reference would be parasitic on those others from whom she hears about the referent. She would still succeed in referring as long as others are successful in referring. It may happen, however, that all of the beliefs that *we* as a linguistic community have about the referent are false, and thus even *parasitically* the speaker cannot refer to the right object since no one would satisfy those descriptions. In those cases, we will have *vacuous* reference or reference to a *fictional* object, but not *meaningless* sentences, since the two sets of description still carry the semantic value of the name as well as the sentences in which the name is used.

If the set of descriptions (the φ) includes both descriptions of the speaker's beliefs and descriptions of the *epistemic conditions* of the speaker's coming to have those beliefs, then it is a rather large set. Only in rare cases would the φ be an empty set. Suppose that the speaker picks up the name from a party conversation without knowing anything about the referent, he/she would at least associate the description such as "the person whom they were talking about,"

[24] Wettstein (1988) calls the speaker's "having something in mind" a "cognitive fix" of the speaker, and he argues that such a cognitive fix is not required for a successful referring act. However, by "cognitive fix," Wettstein means that the speaker can *correctly* distinguish the referent from everything else in the universe, while I claim that correctness is not required here.

"the person who did such and such according to them," etc. with the use of the name. As long as the context is *informative*, one can always acquire new information concerning the referent. Even when one forgets the complete context in which one acquires the use of the name and fails to recall any information concerning the referent, one would have a minimal description such as "is a person," "is a city," etc. If the speaker in using a name fails even in providing those minimal descriptions, then the description he/she associates would be <"an individual called 'F'," Φ= 0>. In this case I would say that the speaker is using the proper name *attributively* in Donnellan's terminology.[25]

Another potential challenge for TCD is this: Even though there are many *Nixons* in our linguistic community, there is only one Nixon to whom we refer by that name when we talk about Nixon in the present English. How does this happen? Here I wish to introduce a *pragmatic* notion "the realm of discourse." A realm of discourse is defined as the set of things discussed by (thought by) a group of people who engage in the discourse. If an individual enters a public realm of discourse, then this group of people would *usually* employ the name in the same way. For instance, even if it is quite likely that there were many people called 'Socrates' ('Σωκράτης' in Greek) around the time that our Socrates was living, he was the only one who was significant enough to enter the realm of our public discourse. Thus if you look up the word 'Socrates' in an encyclopedia, the descriptions you get would be of this particular Socrates. This does not mean that these descriptions are *synonymous* with the name 'Socrates'; it only means that the reference of 'Socrates' is *generally fixed* in our discourse. But the fixing is not done by the name itself; it is done by the general intention of the participants of the discourse. The use of a public name is intended to refer to that particular individual, and the φs different people associate with the use of the name will be largely the same. There is no mystery in how different people come to share largely the same φ: we learn about the world through interaction with others. As a result, in cases of public discourse, proper names are used *as if* they were singularly denotative. It is not because these names really *denote* a singular object, but because the participants share the intention to *refer* to the same individual. This is why we can talk about Socrates, about Nixon, or about Moses without any other specification. It is not, however, a linguistic rule that we have to talk about *this* Socrates, *this* Nixon or *this* Moses. It is rather a rule of pragmatics. A. P. Martinich (1997) defines 'pragmatics' in this way: "Pragmatics is the study of how language is used. . . . Pragmatics focuses on the interaction between speakers and hearers. The major idea that guides research in this area is that speaking is intentional behaviour and governed by rules." (Martinich 1997, 12) Rules of pragmatics depend on the context and the intention of speakers and hearers. Under the present theory, it is not a *wrong* use if one uses a certain name, which normally picks up one particular member in the name-set, to pick out another member. It would be a wrong use, on the other hand, if one uses a certain name to pick out someone by a different name.

To recap, according to TCD, the *sense* (or the *semantic value*) of a proper name should include two components, one determines the name's denotation and the other determines the speaker's reference by using the name. I call the first component the *meaning* of the proper name; the second component, the *associated descriptions* of the proper name. Anyone who satisfies the same calling relation (namely, being-called-by-the-same-name) can qualify as a member of the same set. The *denotation* of a proper name is thus a set, which could have multiple, single, or even no members (if the name is an empty name). The *referent* of a proper

[25] According to Donnellan (1966), a description is used *referentially* if the speaker has the object in mind, and it is used *attributively* if the speaker simply uses a description to pick out *whoever* satisfies the description.

name, on the other hand, is the particular member of that set, the member that is being picked out by the intention of the speaker.[26] The speaker's intention is expressible by descriptions of her beliefs, her mental images, her epistemic relationship with the object, etc. The associated descriptions 'fix' the reference for the speaker's utterance, and thus there is no ambiguity of a proper name in an utterance (as long as the speaker knows what she has in mind). The *semantic value* of a proper name in different utterances, on the other hand, would be different from person to person. Semantic value corresponds to cognitive value. This difference in semantic, and thus cognitive, values explains why one would take 'Cicero is Tully' to be a trivial statement, while another would take it to be informative.

3 Reference and Truth

Under TCD, reference is not *direct*; rather, reference is *mediated* through the speaker's associated descriptions. This feature separates TCD from any form of *direct reference* theory. Let us now turn to the issue of *indirect* reference. The ordered pair **[P]** should determine the actual referent, the α, of each utterance. The first component of **[P]** describes any *indeterminate* member of the denoted set λ = {all individuals called 'F'}. The second component of the ordered pair picks out *that* particular member, the α, and α belongs to λ. Thus α should be a member of the denoted set that is being singled out by the speaker's associated descriptions. I contend that reference cannot be successful without using the right name. As Burge remarks, "A proper name occurring in a sentence used by a person at a time designates an object *if and only if* the person refers to that object at that time with that proper name, *and the proper name is true of that object*." (Burge 1973, 435, my italics) This is simply the linguistic rule of name-using and the social habit of following that rule.[27] For instance, if I intend to refer to Plato by using the name 'Aristotle,' even if I associate all the right descriptions (such as *the one who wrote the Republic*, etc.), I *do not* refer to Plato *by that name*. One may argue that just as Donnellan can use descriptions that don't match to "refer to" a particular individual (what he calls *the referential use* of descriptions), we can also use names that are *not true of the object* to refer to a particular individual. What matters, the line of argument goes, is what the speaker *intends to refer to*. Kripke seems to have taken this line of argument. He gives the example in which two people see Smith in the distance and mistake him for Jones. One person asks: "What is Jones doing?" while the second person answers: "(He is) raking the leaves." According to Kripke, even though the name 'Jones' "*never* names Smith," "in some sense, on this occasion, clearly both participants in the dialogue have referred to Smith."[28] (Kripke 1977, 257) My reply is that there is a sense in which we say that the first speaker succeeds in *referring*, but the speaker does not *use the name* to refer in this case. What the speaker does instead, is to refer by means of other contextual expediency (such as pointing, gazing at, etc.). The use of the name 'Jones' plays no significant role in the referring act. In other words, the only means the two speakers actually use to accomplish their referring to Smith is their contextual relationship to Smith. The name 'Jones' used in this context is simply a "misnomer."

Truth values are assigned to the utterance of a sentence, or we can say, to the proposition expressed by an utterance. According to P. F. Strawson (1956), sentences themselves have

[26] Even though pragmatics relies on the interaction between speakers and hearers, it is mainly the speaker's intention that fixes the reference of an utterance. The hearer may very well have a different set of associated descriptions upon hearing the name mentioned. If the hearer's descriptions would pick out the same referent as the speaker-intended referent, then communication is successful. If not, misunderstanding gets generated.

[27] Katz makes a similar observation. He says, "What makes a word the right name for a thing is that the thing fits or conforms to the meaning of the word in the language." (Katz 1990, 47)

[28] Of course this is an example of "speaker's reference."

meaning but no truth value, and yet we can make use of a sentence to "express a true or false proposition." (Strawson 1956, 223) A sentence of the type "Nixon is dead" in itself cannot be assigned a definite truth value, since it means something such as "Someone called 'Nixon' is dead." In our normal usage where we talk about Richard Nixon, the utterance is true after April of 1994. But the utterance "Nixon is dead" by our dog-owner's neighbor would be true only if the intended referent is the former president Nixon, and would be false if the intended referent is the dog Nixon. The need to assign different truth values to different utterances of the same sentence shows that after the ordered-pair of descriptions fixes an α for an utterance, it is this α that we evaluate when we assign a truth value to that utterance. The semantic content of the proper name is incorporated into the semantic content of the sentence in which the name appears, through the identification (*via* descriptions) of the α in question. Take "Socrates is wise" for example. Our analysis of its truth value will be rendered as such: "Socrates is wise" is true iff an individual called 'Socrates' (by our linguistic community) is singled out in virtue of a set of descriptions in the mind of the speaker, and this individual is wise.

It is not the case that all utterances can be successfully assigned a truth value. There are cases when the set of descriptions cannot generate any real α, such as in the case of empty names. If the fictional name 'Worf,' for example, has never been used by any actual person, then λ = {all individuals called 'Worf'} is an empty set. Most cases of fictional names, however, do not belong to this category. Santa Claus does not exist, but there is a town called 'Santa Claus.' There is no Pegasus, but there might be companies named 'Pegasus.' I argue that the α doesn't get fixed in these cases not because λ is empty, but because Φ fails to pick out any member in λ. In cases where there are entities called by that name, but the speaker's descriptions fail to pick out any member of the set, the reference is vacuous and the name is empty. Under TCD, a sentence containing an empty name is not *meaningless*. In Kent Bach's words, "reference failure does not lead to loss of meaning." (Bach 1984, 174) Depending on the scope reading, in some cases such a sentence would be false while in others it would have no truth value.

Previously I have distinguished *meaning* and *sense*. Now based on what I have said about the assignment of truth values, I wish to introduce a third notion: *content*. The *content* of a proper name is the referent *mediated by* the ordered pair (the two components) of descriptions:

[Q] α < "an individual called 'F'," Φ = a set of descriptions >)

where α is the object referred to, 'F' is the name the speaker uses to refer to α, and Φ is the set of descriptions the speaker *would* use to specify α.[29]

Using**[Q]**, we can analyze the above example of "Nixon is dead" as **(6)** and **(6')**:

(6) Nixon (< "an individual called 'Nixon'," Φ = {is a man, is a former U.S. President, has a large nose and sad-looking eyes,. . . } >) is dead.

(6') Nixon (< "an individual called 'Nixon'," Ψ = {is a dog, has a large nose and sad-looking eyes,. . . } >) is dead.

With the Giacometti case mentioned earlier, TCD would fare much better than Kripke's theory. Our analysis would render the tour guide's remark as **(5')**:

(5') This was painted by Giacometti (< "an individuals called 'Giacometti'," Φ = {a Swiss painter working in the late 19th Century, etc.} >).

On the other hand, when the museum visitor also utters **(5)**, his utterance should be analyzed as **(5")**:

[29] Having the object itself in the proposition does not make the theory *directly referential*, since the object is mediated through the descriptions.

(5") This was painted by Giacometti (< "an individuals called 'Giacometti'," Φ= {a Swiss sculptor who made slender-shaped figures, etc.}>).

I think in this case the tour guide and the visitor use the same sentence-type, which can be analyzed as meaning that the painting was done by someone called 'Giacometti.' However, their utterances do not have the same content in that the tour guide refers to Gustav Giacometti and she is right, while the visitor refers to Alberto Giacometti and he is wrong.

This shift of intention also poses a problem for Kripke's theory concerning speech reports, while TCD can handle this sort of problems easily. Suppose the visitor says

(7) The tour guide said this was painted by Giacometti,

he reports a different *content* of her utterance even though he uses the same words she used. Under Kripke's theory, the visitor would be referring to Gustav Giacometti since he did intend to use the name that the tour guide did. Therefore what he utters would be true as well. Consider the fact that the visitor knows nothing about Gustav Giacometti and his associated descriptions actually pick out the-Giacometti-who-made-those-slender-shaped-figures, it does not seem correct to say that he is referring to Gustav Giacometti simply because he has heard the name 'Giacometti' from someone who did refer to Gustav. On the other hand, TCD can allow us to assign different propositional contents as well as different truth-values to the utterances made by different speakers. There are conceivably many other cases where intentional fixing changes from one speaker to the next. Kripke's causal theory of reference fails to explain these cases.

I now want to show how TCD deals with some of the problems Kripke presents as a refutation of the description theory. One of Kripke's attacks focuses on William Kneale's description theory of names. Kneale's theory is that the meaning of a name is simply 'the individual called by that name.' Kneale argues that statement **(8)** is trifling or non-informative:

(8) Socrates *was* called 'Socrates.'

If **(8)** is trifling, then it must be because the name 'Socrates' itself has already included the information given by the predicate. Therefore, Kneale concludes, 'Socrates' *means* "the individual called 'Socrates'."

Kripke rebukes this argument by pointing out that **(8)** "isn't trifling on any view," because it could happen that the Greeks did not call Socrates 'Socrates' (Kripke 1972/1980, 69). I agree with Kripke on this point. But an important feature in **(8)** is the past tense verb ('was') used by Kneale. I think if **(8)** is stated as

(9) Socrates *is* called 'Socrates,'

then that statement is trifling or non-informative. Under **(9)**, Kneale's argument could support the meta-linguistic analysis of the meaning of 'Socrates.'

How do we explain the difference between the triviality of statement **(9)** Socrates *is* called 'Socrates' and the non-triviality of statement **(8)** Socrates *was* called 'Socrates'? Under TCD, the *calling* relation is indexed to the present language used, thus **(8)** as analyzed in the following way is not trifling:

(8') An individual called 'Socrates' by the present English-speaking community was also called 'Socrates' by the ancient Greeks.

And **(9)** is analyzed in this way which clearly shows how it is a non-informative statement:

(9') An individual called 'Socrates' by the present English-speaking community is called 'Socrates' by the present English-speaking community.

This shows that Kneale is partially correct: the meaning of the name 'Socrates' does include a piece of meta-linguistic information.

Kripke's second criticism of Kneale's description theory is that it violates what he calls "the non-circularity condition":

(C) For any successful theory, the account must not be circular. The properties which are used in the vote must not themselves involve the notion of reference in a way that it is ultimately impossible to eliminate (Kripke 1972/1980, 68).

What is barred by (C) is a circular theory of reference, which uses the notion of reference itself in defining the way to fix the reference. Kneale's theory of proper names, at least as Kripke interprets it, uses "the individual called such and such" *both* as the meaning of 'Socrates' *and* as a way of referring to Socrates. Kripke argues: "Obviously if the only descriptive senses of names we can think of are of the form 'the man called such and such,' ... then whatever this relation of *calling* is is really what determines the reference and not any description like 'the man called *Socrates*'." (Kripke 1972/1980, 70) In this criticism, Kripke seems to take the word 'calling' used in this context to be expressing the same notion as 'referring,' and he thus charges Kneale's theory with the violation of the non-circularity condition. However, under TCD, 'being called such and such' only gives us a descriptive property of members of the name-set (the denotation); it does not determine the exact referent. TCD does not violate the non-circularity condition in that as a theory of reference, what determines the reference is the speaker's intention, which is expressed by the associated descriptions. In contrast to Kneale's theory that uses the same description to give the meaning *and* to fix the reference of a proper name, TCD separates the functions of the two components of descriptions. The associated descriptions in the speaker's mind do not *give the meaning* of proper names. What those descriptions do, is to help identify the referent fixed by the speaker's intention. What determines the meaning, on the other hand, is the denotation of the name. We need to have the two kinds of descriptions to complete both denoting and referring. Once the two components are assigned separate roles, there is no circularity in the definition of reference.

Kripke presents another case that is supposed to be a problem for a particular form of description theory—the cluster concept theory of names (Kripke 1972/1980, 58-59; 64-67). In the example (borrowed from Wittgenstein), Kripke discusses the following statement:

(10) Moses does not exist.

Kripke argues that the Biblical descriptions should not be used to fix the reference of the name 'Moses,' because the failure of satisfaction for the Biblical descriptions of Moses does not lead to the negation of Moses' existence. For one thing, it might happen that Moses did exist but he did not do any of the things that the Bible attributed to him; in other words, the Biblical story might have been a complete fabrication about a real person. Kripke says, "[I]n that case maybe no one would have done any of the things that the Bible relates of Moses. That doesn't in itself mean that *in such a possible world* Moses wouldn't have existed." (Kripke 1972/1980, 58, my emphasis) Kripke's point is that since the properties attributed to *our Moses* in the Bible are not "necessary" properties, we could easily imagine *the same Moses* without having done any of the things that the Bible describes. Therefore, even if all these descriptions were not *true of* Moses in some possible world, we cannot conclude that Moses would not exist in that possible world (in Kripke's conception, a possible world is simply a possible scenario, not a separate realm). What Kripke takes for granted here is that model considerations are built on the actual referent in our world. That is to say, we first use either the Biblical descriptions or the causal chain of the name 'Moses' to fix the reference directly on the person Moses, and then we suppose that in a possible world, this person might not have done any of the things attributed to him in the Bible. However, I argue it that if it were true that our Moses had not done any of the things attributed to him, then our Moses as described in the Bible *does not exist*. In other words, I argue that (10) does mean that *the Moses described in the Biblical story* does not exist. The name 'Moses' is a placeholder for all the Biblical descriptions about *someone*, and if none of the descriptions is true of *anyone*, then the one called 'Moses' in the Bible is simply a fictional character. The putative fact that there was someone named Moses in ancient times is irrelevant to our historical interest and Biblical verification. Therefore, I conclude that even if there were a Moses who had never gone into politics or religion, (10) is still true. The name 'Moses' is a *disguised* set of definite descriptions found in the Bible, and what (10) states is a simplified form of (10'):

(10') Moses (< "an individuals called 'Moses' in our usage," Φ= {the leader of the Exodus, the Hebrew baby boy adopted by the Egyptian royal family, the person who received the Ten Commandments from God, etc.} >) does not exist.

A general skepticism I have about Kripke's causal chain theory is that even in cases such as Moses, he assumes that there was some *invisible* causal chain leading from our current usage of the name back to the *actual* referent of the name. Without knowledge of acquaintance, as in most cases, what we have when we use a proper name is simply the *intended* referent. With legendary figures (and possibly some historical names), there is no guarantee that there was ever this causal chain going back to the actual referent. All we have in our current usage is our *projected* properties of these referents. If there were people called 'Moses' so the name does not denote an empty set, but somehow all the descriptions we associate with this name do not pick out anyone in the set, then there is no individual who is picked out by the name 'Moses' *in our usage*. The negative existential statement (10) should be analyzed as the denial of the fact that anyone satisfies the descriptions we associate with the name:

(10") It is not the case that [there is an α such that α belongs to λ = {all individuals called 'Moses'} and $\Phi\alpha$].

In other words, the negative existential should be analyzed, as Russell suggests, as the secondary occurrence of the *disguised* definite descriptions associated with the name 'Moses.'

Another problem that Kripke attributes to the description theory involves Gödel. This is a case where "the person named by that name did not satisfy the descriptions usually associated with it, and someone else did." (Kripke 1972/1980, 254) If *the only* description we have about Gödel is "the man who proved the incompleteness of arithmetic," and it *could* turn out that Gödel didn't really prove it, but someone else called 'Schmidt' did, then the description we give would fix Schmidt for the name 'Gödel.' With TCD, however, such a situation would not occur. The *descriptions* we give to the name 'Gödel' in this case would be an ordered pair:

(G) < "an individual called 'Gödel'"; Φ= {is the man who proved the incompleteness of arithmetic.}>

If Gödel did not actually prove the incompleteness of arithmetic, *no one else* would satisfy this ordered pair of descriptions. However, such an analysis is still insufficient. (G) gives a more limited *sense* to the name 'Gödel' than what the name usually has in our discourse. The Φ that a speaker associates with the name 'Gödel' is generally something like (G'):

(G') < "an individual called 'Gödel'"; Φ= {is the man who proved the incompleteness of arithmetic, is the person whose name I have read in many logic books,...}>

In other words, the Φ the speaker associates with the name is generally not a single description, but a set that includes descriptions of the speaker's epistemic conditions of the name. This point has been made by John Searle in his *Intentionality*: "At the very least, he [the speaker] has 'the man called "Gödel" in my linguistic community or at least by those from whom I got the name'." (Searle 1983, 251) Searle calls the speaker's cluster beliefs her "Intentional content," and argues that the speaker can still use the name to *parasitically* refer to Gödel even if she is misinformed about who actually proved the incompleteness of arithmetic. The way we succeed in referring to Gödel by using the limited information that we have of *him* is based on the fact that we learned about him in this way. It would not be the case that anyone who happens to satisfy the sole description "the man who proved the incompleteness of arithmetic" would thus become the referent of the name 'Gödel.' It would rather be the case that the man Gödel would satisfy the set of descriptions, because part of the descriptions describe *the epistemic conditions* under which we learned about this name. Kripke's mistake in his attack on description theories is that he assumes that such descriptions fix the referent only via an external, objectively ascertained *true-of* relation, while Searle's cluster theory or my TCD treats descriptions to be specifications of what the speaker has in mind; i. e. the speaker's internal psychological states of belief. As explained earlier, such beliefs would include not only the speaker's belief *of* the attribute of this putative referent, but also the speaker's beliefs about the epistemic conditions under which she acquired such a belief. Some of the external causal chains that Kripke champions can enter the speaker's psychological states, and thereby help secure the right referent. But the point is, external causal chains by themselves do not suffice. If *all* beliefs the speaker has concerning the attributes of the putative referent, as well as about how or from whom she acquired the name, turn out to be false, then I must judge that the speaker simply does not know what she is talking about and her use of the name fails to refer.

Finally, I shall address the Modal Argument that Kripke puts forth against the description theory. Kripke thinks that the descriptivist approach gets the counterfactuals wrong, because "although the man (Nixon) might not have been the President, it is not the case that he might not have been Nixon (though he might not have been *called* 'Nixon')." (Kripke 1972/1980, 49) By this argument, both the qualitative description ("was a President of the United States") and the meta-linguistic description ("was called 'Nixon'") fail to 'fix' the reference across possible worlds. No doubt this is a good argument, but I don't think the description theorist necessarily insists on fixing reference *across possible worlds* purely by the descriptions one uses to fix the reference *in our world*. Let us consider these two statements by Kripke:

(11) It is not the case that Nixon might not have been Nixon.

(12) Nixon might not have been called 'Nixon.'

TCD gives them the following analyses:

(11') It is not the case that [there is an α, α belongs to $\lambda = \{$all individuals called 'Nixon'$\}$ & $\Phi\alpha$, and possibly $(\alpha \neq \alpha)$].

(12') There is an α, α belongs to $\lambda = \{$all individuals called 'Nixon'$\}$ & $\Phi\alpha$, and possibly [α is not called 'Nixon'].

(11') and **(12')** seem to preserve the intuitive distinction that Kripke makes with regard to two kinds of possibility. In other words, we fix the referent of 'Nixon' in our world in the same way TCD describes, *and then* we assign possible counterfactuals to this fixed referent. This fixed-in-the-actual-world brings in *indexicality* to the present language and the realm of discourse. A name used in our present discourse will always pick out the same individual across possible worlds, because it is *this* person whose counterfactuals we are considering. Joseph Almog in "Naming Without Necessity" suggests that naming is naming, and necessity is necessity, and that the connection Kripke tries to draw between the two is unfounded. Almog argues that we should distinguish *two stages* in our semantic theory: the *generation* stage and the *evaluation* stage. In the *generation* stage, we generate the propositional constituent (such as an individual person) corresponding to a name. And the question for this stage is whether the name refers to the individual *via* some descriptive content. In the *evaluation* stage, we evaluate the truth of the proposition in a possible world. And the question for this stage is whether the individual (not the name) bears modal attributes. Almog writes, "The two questions are definitely different. One concerns language. The other is metaphysical, having nothing to do with names. The two questions are not only different; they are independent of each other. First, one could hold the semantical view that names refer by means of descriptive concepts, and yet couple this stand with the metaphysical view that objects ... bear modal attributes.... Conversely, ... we could have naming without necessity. One could believe that names do not refer by means of descriptive concepts, and couple this semantic view with a skeptical metaphysical attitude toward modal individualism." (Almog 1986, 229) Almog himself holds the second view. It should be clear from what has been argued in this paper that I hold the first view, which Almog calls "necessity without naming."

In summary, TCD is a more complete theory than Kripke's direct reference theory, and it does not have the same problems that older description theories do. As A. P. Martinich remarks, "Perhaps behind Kripke's puzzle is an even more general misconception about language: the belief that language is self-contained and that purely linguistic knowledge is sufficient for using language." (Martinich 1997, 31) I think what is being left out in this self-contained view on language is the speaker. In contrast to this picture of a two-term relation between language and the world, the *two-component description theory* is based on the picture of a three-term triadic relation amongst language, speaker and the world.[30] In such a picture, we can *speak about* the world both because we, as speakers, intend to refer to things or events in the world, *and* because the language we use *gives descriptions* of the world. The speaker apprehends the *meaning* of terms in a language prior to choosing the terms for the intended reference. On the other hand, the speaker's intention picks out *the one* called such and such, and it is the speaker's intention that determines the referent within the denoted set. Even if denotation can be established as a verifiable objective fact in most cases, reference is a psycholinguistic act accomplished by the linguistic community. In other words, denoting is *objective* in the sense that it is governed by the *fact* that someone was indeed called such-and-such, whereas referring is a subjective, or *intersubjective*, speech act that could sometimes fail to locate the right referent. Kripke's causal chain theory does not give us a sufficient theory about speaker reference; about how a speaker speaks about a particular individual called by that name. Kripke is right in emphasizing the importance of the relation of causal chain, but this relation simply does not give us the whole story.

There are people who take speaker reference to be in the domain of the *pragmatics*, not *semantics*, of names.[31] I argue, however, that semantics cannot be separated from pragmatics, and speaker reference should be considered as part of the semantics of names in the language. Language in itself sometimes gives only a *partial* proposition, and we have to consider the speaker's intention to complete the content of the proposition. At the same time, language sometimes gives us more than one proposition as in the case of proper names with multiple bearers, and we also need to consider the speaker's intention to pin down the particular proposition expressed. Without the speaker aspect, the semantics of language cannot be either complete or accurate.

[30] The philosophers who hold such a two-term relation would be Tyler Burge, Howard Wettstein, etc. The three-term relation, on the other hand, seems to be explicit or implicit in the philosophy of Donald Davidson, Gareth Evans, etc. Another kind of picture neglected here is a two-term relation between the speaker and the world, which seems to be implicit in the different theories of speaker's meaning.

[31] Examples are Katz (1990) and Fodor (1994, 111-112).

The two components cannot be separated if our usage of a name is to bring us to the right individual.[32]

Kripke says that any theory of beliefs and names must deal with the puzzles about Pierre's and Peter's beliefs, so now I will go back to Kripke's puzzle.

4 An Application of TCD: Back to Kripke's Puzzle and Others

Kripke thinks that the case of Pierre's belief resembles the case of Jones' belief about Cicero and Tully. So we shall begin with the latter case. Why is it that Jones may believe that Cicero is bald while denying that Tully is? With the analysis of the *two-component description theory*, we can explain that it is because "a person called 'Cicero'" and "a person called 'Tully'" are different descriptions. Even when the only description Jones associates with both names 'Cicero' and 'Tully' is merely 'is a Roman orator,' we would still have the difference in names in Jones' mind. So we can have:

(13) Jones believes: Cicero (< "an individual called 'Cicero'," Φ = {is a Roman orator}>) is bald, *and*

Jones believes: Cicero (< "an individual called 'Tully'," Φ= {is a Roman orator}>) is not bald.

This kind of analysis does not require that all propositions involving different names express *distinct* beliefs of the subject. If, for instance, Sally knows that Cicero is Tully, then Sally's belief would be expressed as:

(14) Sally believes: Cicero (< "a man called 'Cicero'," Φ= {a Roman orator, is also called 'Tully,'...}>) is bald.

In this way we would not have too many beliefs individuated by the different names the subject chooses to express her belief.

With the case of Paderewski, Peter uses the same name but he associates different descriptions with the name. Our ascription should be like **(15)**:

(15) Peter believes: Paderewski (< "an individual called 'Paderewski'," Φ= {is a musician,...}>) has musical talent,

and

Peter believes: Paderewski (< "an individual called 'Paderewski'," Φ= {is a politician,...}>) has no musical talent.

This gives us no problem since it is reasonable for anyone to think that there are two 'Paderewski' being referred to in the two utterances. Similarly, Peter can assent to a sentence such as "Paderewski is not Paderewski" by taking it to mean "this Paderewski is not that Paderewski" without violating the law of contradiction.

Kripke asks us to decide whether the sentence "Pierre believes that London is pretty" is true or false. But as I argued earlier, sentences themselves do not have truth value. What we

[32] While putting the same emphasis on the meta-linguistic analysis of the meaning of proper names, TCD is distinguished from Bach's NDT and Katz's PMT in exactly the incorporation of pragmatics into the semantics of proper names. Both NDT and PMT are taken to be *merely* a semantic theory; a theory of sense, not of reference. Bach says, "NDT... does not even purport to be a theory of reference. It is nothing more than a modest theory of the modest meaning of names." (Bach 1984, 161) Katz also says that his PMT "is (part of) a theory of sense, not a theory of reference." (Katz 1990, 40) Both Bach and Katz argue that their theories are therefore not responsible for answering Searle's criticism that the meta-linguistic sense of the name is *insufficient* in terms of fixing the reference in contexts. They also both think that their theories are thus immune to Kripke's circularity argument. I have argued, however, that the issue of reference is an essential part of a semantic theory of proper names.

should do in this case is to find the proper proposition expressed by, or the semantic value of, the utterance. With Pierre's belief, my proposed ascription is this:

(16) Pierre believes: London (< "an individual called 'Londres' (by the French-speaking community)," Φ= {is a city in England, is the city of which I have seen a postcard,. . . } >) is pretty,

and

Pierre believes: London (< "an individual called 'London' (by the English-speaking community)," Φ= {is a city in England, is where I reside at the moment,. . . } >) is not pretty.

The name within the quotation marks will not be translated. Thus, even when Pierre associates the same set of descriptions with 'Londres' and 'London,' his beliefs do not express the same propositions. I thus think that Kripke's puzzle *could not* be generated under a properly laid out description theory such as TCD.

Kripke asks: "What is it about sentences containing names that makes them – a substantial class – intrinsically untranslatable, express beliefs that cannot be reported in any other language?" (Kripke 1979, 129) I think the reason is that proper names are really dependent on the communal usage of a linguistic community. Statements such as 'Londres is London,' 'Eiffel Tower is la Tour Eiffel,' 'Köln is Cologne' are by no means trivial. They convey important information in language acquisition. The way a name is given and used is very much dependent on the conventions of a linguistic community and the sub-groups within. By giving a standard translation of names, we are also changing the context and the epistemic condition of the subject.

Finally, I want to explain why I think the theory of proper names should not be any form of direct reference theory. The main difference between the direct reference theory and the description theory lies in the assertion concerning whether reference is *direct* or *mediated*. In this paper I have argued how reference has to be mediated through the two sets of descriptions, and thus the direct reference theory simply takes the wrong approach. The first component of TCD, the meta-linguistic description of the meaning of the name as "an individual called such-and-such (by a certain community)," is necessary in the mediation of reference. That is to say, the reference of a proper name has to be mediated through social, conventional usage of the name. Secondly, when we talk about an object, the object being discussed always comes into our discourse *via* one perspective (*mode of presentation*) or another. The second component of TCD captures how speaker reference is *mediated* through descriptions of the way (the mode) in which the object is presented to the speaker. Kripke's causal chain or Donnellan's historical explanation take the perspectives out of the speaker's mind and put it in the mind of an "*omniscient observer* of history." But our language is used by people *like us* and we are not omniscient. This fact explains why substitution *salva veritate*, which poses no problem for an omniscient observer, always poses a problem in an epistemic context involving ordinary speakers. This also illustrates the deficiency of direct reference theory in general.

Kripke's conclusion concerning the puzzles involved in the epistemic context seems pessimistic. He says, "When we enter into the area exemplified by Jones and Pierre, we enter into an area where our normal practices of interpretation and attribution of belief are subjected to the greatest possible strain, perhaps to the point of breakdown. So is the notion of the *content* of someone's assertion, the *proposition* it expresses." (Kripke 1979, 135) However, I think the problem of substitution *salva veritate* is a serious problem for the direct reference theory. What it pushes for, is not to abandon the hope of finding an acceptable belief ascription, but to always consider the subject's meta-linguistic beliefs as well as her other relevant beliefs about the object. With a description theory properly laid out that captures those other beliefs, there is no puzzle about beliefs.

References

Almog, J. (1986). 'Naming Without Necessity.' *Journal of Philosophy* 87: 210–242.

Bach, K. (1984). *Thought and Reference*. Oxford: Clarendon Press.

Burge, T. (1973). 'Reference and Proper Names.' *Journal of Philosophy* 70: 425–439.

Burgess, J. (1996). 'Marcus, Kripke, and Names.' *Philosophical Studies* 84: 1–47.

Corlett, J. A. (1989). 'Is Kripke's Puzzle Really A Puzzle?' *Theoria* 55: 95–113.

Donnellan, K. (1966). 'Reference and Definite Description.' In A. P. Martinich (Ed.) *The Philosophy of Language*, first edition. 1985. 236–248.

Donnellan, K. (1977). 'Speaking of Nothing.' In S. Schwartz (Ed.) *Naming, Necessity, and Natural Kinds*, Ithaca: Cornell University Press. 3–31.

Erwin, E., Kleiman, L. and Zemach, E. (1976). "The Historical Theory of Reference." *Australasian Journal of Philosophy* 54:1, 50–57.

Fodor, J. (1994). *The Elm and the Expert: Mentalese and Its Semantics*. Cambridge: The MIT Press. Appendix A: Names.

Kaplan, D. (1975). 'Dthat.' in A. P. Martinich (Ed.) *The Philosophy of Language*, first edition. 1985. 315–328.

Katz, J. (1977). 'A Proper Theory of Names.' *Philosophical Studies* 31: 1–80.

Katz, J. (1990). 'Has the Description Theory of Names Been Refuted?' In G. Boolos (Ed.) *Meaning and Method: Essays in Honor of Hilary Putnam*, New York: Cambridge University Press. 31–62.

Katz, J. (1994). 'Names Without Bearers.' *Philosophical Review* 103: 1–39.

Kripke, S. (1972/1980). *Naming and Necessity*. Cambridge: Harvard University Press.

Kripke, S. (1977). 'Speaker Reference and Semantic Reference.' In A. P. Martinich (Ed.) *The Philosophy of Language*, first edition. 1985. 249–268.

Kripke, S. (1979). 'A Puzzle about Belief.' Reprinted in N. Salmon & S. Soames (Eds.) *Propositions and Attitudes*. Oxford: Oxford University Press, 1988. 102–48.

Kvart, I. (1987). 'Kripke's Belief Puzzle.' In P. A. French, T. E. Uehling Jr. and H. K. Wettstein (Eds.) *Midwest Studies in Philosophy* 10 (1): 287–325. Minneapolis: University of Minnesota Press.

Loar, B. (1987). 'Names in Thought.' *Philosophical Studies* 51: 169–185.

Marcus, R. B. (1981). 'A Proposed Solution to a Puzzle about Belief.' in P. A. French, T. E. Uehling Jr. and H. K. Wettstein (Eds.) *Midwest Studies in Philosophy* 6: 501–510. Minneapolis: University of Minnesota Press.

Martinich, A. P. (Ed.) (1996). *The Philosophy of Language*. Oxford: Oxford University Press, third edition.

Martinich, A. P. (1997). 'Philosophy of Language.' In J. Canfield (Ed.) *Routledge History of Philosophy, Volume X: Philosophy of Meaning, Knowledge and Value in the Twentieth Century*. London: Routledge. 1–19.

McMichael, A. (1987). 'Kripke's Puzzle and Belief 'Under' A Name.' *Canadian Journal of Philosophy* 17: 105–126.

Over, D. E. (1983). 'On Kripke's Puzzle.' *Mind* 92: 253–256.

Pettit, P. (1984). 'Dissolving Kripke's Puzzle about Belief.' *Ratio* 26: 181–194.

Salmon, N. (1986). *Frege's Puzzle*. Cambridge: MIT Press. Appendix.

Searle, J. (1958). 'Proper Names.' in A. P. Martinich (Ed.) *The Philosophy of Language*. first edition, 1985. 270–274.

Searle, J. (1983). 'Proper Names and Intentionality.' In J. Searle, *Intentionality*. Cambridge: Cambridge University Press. 231–261.

Schwartz, S. (Ed.) (1977). *Naming, Necessity, and Natural Kinds*. Ithaca: Cornell University Press.

Strawson, P. F. (1956). 'On Referring.' In A. P. Martinich (Ed.) *The Philosophy of Language*. first edition, 1985. 220–235.

Wettstein, H. (1988). 'Cognitive Significance Without Cognitive Content.' In J. Almog, J. Perry and H. Wettstein (Eds.) *Themes from Kaplan*. New York: Oxford University Press, 1989. 421–454.

abstracta Linguagem, Mente & Ação ISSN 1807-9792 Volume 7, Number 2, pp. 41–59, 2013

Action, Deviance, and Guidance

Ezio Di Nucci

Institute of Philosophy, University Duisburg-Essen, Germany
ezio.dinucci@uni-due.de

Abstract
I argue that we should give up the fight to rescue causal theories of action from fundamental challenges such as the problem of deviant causal chains; and that we should rather pursue an account of action based on the basic intuition that control identifies agency. In Section 1 I introduce causalism about action explanation. In Section 2 I present an alternative, Frankfurt's idea of guidance. In Section 3 I argue that the problem of deviant causal chains challenges causalism in two important respects: first, it emphasizes that causalism fails to do justice to our basic intuition that control is necessary for agency. Second, it provides countless counterexamples to causalism, which many recent firemen have failed to extinguish – as I argue in some detail. Finally, in Section 4 I argue, contra Al Mele, that control does not require the attribution of psychological states as causes.

We should give up the fight to rescue causal theories of action (Davidson 1963; Bratman 1984; Mele & Moser 1994 are some influential examples) from fundamental challenges such as the problem of deviant causal chains; we should rather pursue an account of action based on the basic intuition that control identifies agency. To this end, I propose to revive Harry Frankfurt's concept of guidance (1978). In Section 1 I introduce causalism about action explanation. In Section 2 I introduce Frankfurt's rival idea, guidance. In Section 3 I argue that the problem of deviant causal chains challenges causalism in two important respects: firstly, it reminds us that causalism fails to do justice to our basic intuition that control is necessary for agency. Secondly, it provides countless counterexamples to causalism, which many recent firemen have failed to extinguish. Finally, in Section 4 I argue, contra Al Mele (1997), that control does not in turn require causalism because it does not require the attribution of psychological states as causes.

1 Causalism

The classic version of causalism was first introduced by Donald Davidson in *Actions, Reasons, and Causes* (1963), where Davidson defends the thesis that reasons explanation (rationalization) is "a species of causal explanation" (p. 3). On Davidson's account, then, some action A is intentional under a certain description only if that action was caused by a primary reason of the agent comprising of a pro attitude towards actions with a certain property, and a belief that action A, under the description in question, has that property[1]:

> R is a primary reason why an agent performed the action A, under description d, only if R consists of a pro-attitude of the agent towards actions with a certain property, and a belief of the agent that A, under the description d, has that property (1963, p.5).

[1] Davidson only offers necessary conditions. Any attempt at giving sufficient conditions would, by Davidson's own admission (Davidson 1973), run against the problem of deviant causal chains – see section 3. See also footnote 3 for an example of a full-blown necessary and sufficient account of intentional action (Mele & Moser 1994).

Pro attitudes, says Davidson, can be "desires, wantings, urges, promptings, and a great variety of moral views, aesthetic principles, economic prejudices, social conventions, and public and private goals and values" (p. 3). On Davidson's account, my flipping the switch is intentional under the description 'flipping the switch' only if it was caused by a primary reason composed of a pro attitude of mine towards actions with a certain property, say the property of 'illuminating the room'; and a belief that my action, under the description 'flipping the switch', has the relevant property of 'illuminating the room'.

The crucial element of Davidson's view is that the primary reason, composed of a pro attitude plus a belief, is the action's cause. As Davidson himself points out (p. 12), causes must be events, but pro attitudes and beliefs are states, and so they cannot be causes. Davidson therefore proposes the "onslaught" (or *onset*, see Lowe 1999, p. 1) of the relevant mental state as the cause of action. The difference between a mental state and its onset, which is a mental event, is the same as the difference between believing that there is a bottle on my desk (mental state), and forming the belief (noticing, realizing) that there is a bottle on my desk (mental event). Clearly, while both kinds of mental states, pro attitude and belief, are always needed – on Davidson's view – to rationalize an action under some description, only one mental event is necessary to cause the action.

As Stoutland (1985) emphasizes, the mental states required by Davidson's view must have a very specific content:

> The thesis is a very strong one: it is not saying merely that reasons are causes of behaviour but that an item of behaviour performed for a reason is not intentional under a description unless it is caused by just those reasons whose descriptions yield the description under which the behaviour is intentional. This requires that every item of intentional behaviour have just the right cause (1985, p. 46).

So there must be a content relation between the primary reason and the action description in question. Recall Davidson's definition of "primary reason" (Davidson 1963, p. 5): the belief must make explicit reference to the action description which it rationalizes.

According to Davidson, for example, the following primary reason would not do: a pro attitude towards 'illuminating the room', and a belief that my action, under description 'turning on the light', has the property of 'illuminating the room'. This primary reason makes no mention of the description 'flipping the switch', and therefore it cannot rationalize my action under the description 'flipping the switch'; even though it will rationalize my action under the description 'turning on the light'.

One note of clarification: the content constraint emphasized by Stoutland is on the belief rather than on the pro attitude. That is to say that, as long as the belief has the 'right' content, the pro attitude can have any content. For example, my action of flipping the switch can be rationalized under the description 'flipping the switch' by a very wide selection of pro attitudes – 'turning on the light', 'illuminating the room', 'wasting energy', 'finding some comfort', 'stretching my arm', etc. – as long as the agent believes that her action, under the description in question – 'flipping the switch' – has the relevant property towards which the agent has a pro attitude: 'turning on the light', say.

It must be emphasised that causalism does not depend upon endorsing Davidson's Humean reductionism about motivation: many theorists have proposed versions of causalism that appeal, rather, to a single state of *intention* or *plan*.[2] On these versions of causalism, views will have the following general form: *S A-ed intentionally only if S intended to A*.[3]

In the next section I present an alternative to causal theories of action: Harry Frankfurt's concept of guidance.

2 Guidance

In *The Problem of Action* (1978), Frankfurt puts forward an alternative view according to which what distinguish actions from mere bodily movements are not the movements' causes, but whether or not the agent is in control of her movements. Frankfurt calls the relevant sort of control *guidance*: "… consider whether or not the movements as they occur are *under the person's guidance*. It is this that determines whether he is performing an action" (1978, p. 45).

Frankfurt's proposal does not depend on psychological states as the causes of action, as causal theories do. It focuses, rather, on the relationship between an agent and her action at the time of acting: "What is not merely pertinent but decisive, indeed, is to consider whether or not the movements as they occur are under the person's guidance. It is this that determines whether he is performing an action. Moreover, the question of whether or not movements occur under a person's guidance is not a matter of their antecedents" (1978, p. 45). Frankfurt initially distinguishes between two kinds of purposive movements (p. 46): purposive movements which are guided by the agent, and purposive movements which are guided by some mechanism that cannot be identified with the agent. Through the idea of purposive movement, Frankfurt gives us an insight into what the agent's guidance is:

> Behaviour is purposive when its course is subject to adjustments which compensate for the effects of forces which would otherwise interfere with the course of the behaviour, and when the occurrence of these adjustments is not explainable by what explains the state of affairs that elicits them. The behaviour is in that case under the guidance of an independent causal mechanism, whose readiness to bring about compensatory adjustments tends to ensure that the behaviour is accomplished. The activity of such a mechanism is normally not, of course, guided by us. Rather it is, when we are performing an action, our guidance of our behaviour (1978, pp. 47–48).

For some movement to be under the agent's guidance, then, the adjustments and compensatory interventions don't need to be actualized; it is just a question of the agent being able to make those adjustments and interventions: "whose readiness to bring about compensatory adjustments tends to ensure that the behaviour is accomplished" (1978: 48). This latter point finds confirmation in Frankfurt's famous car scenario, where he stresses that guidance does not

[2] See, amongst others, Searle 1983, Brand 1984, Bratman 1984 & 1987, Thalberg 1984, Adams and Mele 1989, Bishop 1989, Mele 1992, Mele and Moser 1994.

[3] This is actually a statement of the so-called *Simple View*, which not many people endorse (exceptions are, for example, Adams 1986 and McCann 1991). Other views, such as Bratman's (1987, pp. 119–123) or Mele & Moser's (1994, p. 253)) are more complicated. Here is for example the full analysis of intentional action offered by Mele & Moser, which, as I show in Section 3, is also subject to deviant counterexamples: "Necessarily, an agent, S, intentionally performs an action, A, at a time, t, if and only if: (i) at t, S A-s and her A-ing is an action; (ii) at t, S suitably follows-hence, is suitably guided by-an intention-embedded plan, P, of hers in A-ing; (iii) (a) at the time of S's actual involvement in A-ing at t, the process indicated with significantly preponderant probability by S's on bal-ance evidence at t as being at least partly constitutive of her A-ing at t does not diverge significantly from the process that is in fact constitutive of her A-ing at t; or (b) S's A-ing at t manifests a suitably reliable skill of S's in A-ing in the way S A-s at t; and (iv) the route to A-ing that S follows in executing her action plan, P, at t is, under S's current circumstances, a suitably predictively reliable means of S's A-ing at t, and the predictive reliability of that means depends appropriately on S's having suitably reliable control over whether, given that she acts with A-ing as a goal, she succeeds in A-ing at t" (1994: 253).

require those adjustments and interventions to take place; it only requires that the agent be able to make those:

> A driver whose automobile is coasting downhill in virtue of gravitational forces alone might be satisfied with its speed and direction, and so he might never intervene to adjust its movement in any way. This would not show that the movement of the automobile did not occur under his guidance. What counts is that he was prepared to intervene if necessary, and that he was in a position to do so more or less effectively. Similarly, the causal mechanisms which stand ready to affect the course of a bodily movement may never have occasion to do so; for no negative feedback of the sort that would trigger their compensatory activity might occur. The behaviour is purposive not because it results from causes of a certain kind, but because it would be affected by certain causes if the accomplishment of its course were to be jeopardized (Frankfurt 1978, p. 48).

So some movement is under the agent's guidance when the agent "was prepared to intervene if necessary, and that he was in a position to do so more or less effectively" (ibid.); and in such cases the movement in question counts as an action. Guidance captures the idea that one can be in control of x without having to be actively controlling it. Guidance is a passive form of control, as shown by Frankfurt. If we understood control in terms of something we do, and we understood action in terms of control, then we would get a circular picture of agency. That's why we want to be able to describe a form of control that does not depend on the activity of controlling: and that's why we talk, specifically, of guidance.

I'd like to emphasize that the claim is weaker than it might appear at first: I am not suggesting that guidance isn't itself constituted by causal mechanisms; nor am I suggesting that actions do not have causes.[4] My criticism is much more specific than that: by identifying actions' causes with content-specific psychological states causalism runs into difficulties. Also, I am not denying agents' mental phenomenology of intentions, desires, beliefs, etc. Not only do I accept that agents do indeed have intentions, desires, and beliefs; but I also accept that intentions, desires, and beliefs play an important role within agency. Here I am not even disputing that intentions, desires, and beliefs may play some causal role within agency. All I am criticizing is the identification between the relevant psychological states and the action's causes; and the idea that the relevant content-specific psychological states as causes are both necessary and sufficient for intentional action.

This is not the place to develop a full-blown alternative to causalism based on guidance. I just want to touch upon two important points (more on this at the end of Section 4): firstly, if a concept of guidance should be part of an alternative account of agency; and if this alternative view is to be fully naturalistic, then the concept of guidance must not be understood in libertarian terms. Here there are two promising alternatives: one possibility is to develop such a view by going in the direction of Fischer and Ravizza's (1998, p. 31) *guidance control*. Alternatively, the capacity for intervention, correction, and inhibition that characterizes guidance could be accounted for in terms of what has been recently called (by Clarke 2009) *New Dispositionalism*: in brief, the idea (put forward in different versions by Smith 2003, Vihvelin 2004, and Fara 2008) is that having a certain ability to act consists of or depends on having certain dispositions (depending on which of the above versions one takes). Unmanifested dispositions (finkish or

[4] For an idea of the kind of psychological mechanisms that could be appealed to in order to implement guidance, see psychological models of dual control (Norman and Shallice 1986; Perner 2003).

masked dispositions) are compatible with determinism; therefore unexercised abilities are also similarly compatible.[5]

Secondly, if guidance is to be developed into a full account of agency, it must be argued that guidance can be sufficient for agency, and not just necessary. If, then, guidance is to be a sufficient condition for agency, and guidance is to be independent from rationalizing mental states, then we would be offering an account of agency that does not directly appeal to the agent's motivation. Two things here: first, this conclusion might be too quick in overlooking externalism. Explaining agency without appealing to rationalizing mental states does not mean, according to externalists, explaining agency without appealing to reasons or motivation because, crudely put, reasons are facts rather than psychological states (see Stout 1996, Collins 1997, Dancy 2000, Alvarez 2010).

Second, this conclusion would similarly overlook what used to be called the *Logical Connection Argument* (Anscombe 1957, Hampshire 1959, Melden 1961, von Wright 1971) against which Davidson's (1963) original statement of the causal view was addressed. If the relation between an action and the reason why that action is performed is rational, then it cannot be causal – that was the thrust of the old argument. Therefore denying that rationalizing mental states as causes are necessary for agency does not amount to denying the role of motivation simply because the motivational aspect does not entail the causal aspect; just as, in my previous point, the motivational aspect does not entail the psychological aspect.

With these brief remarks about guidance I hope to have shown in which alternative direction I think it would be fruitful to look for an account of agency, given the shortcomings of causalism; but here is not the place to develop such an alternative account in full. Rather, for the rest of the paper I shall motivate the thought that we should look elsewhere by discussing the fundamental weaknesses of causalism. In the next section I argue that deviant causal chains still provide plenty of counterexamples to causalism, despite many attempts at sorting out the problem.

3 Deviant causal chains

Deviant causal chains have long ago been recognised as a problem for causal theories of action.[6] Most attempted solutions assume that we must find a way to reconcile deviance within a causal framework. I argue, rather, that deviant causal chains are symptomatic of a fundamental problem with causalism; and that we should give up the fight to accommodate deviant cases and focus, rather, on developing an alternative to causal views of action which recognises that there can be no action without control, and that control cannot be fully accounted for solely in terms of the content of those motivational states which causalists take to cause action.

The first point to emphasize is that, whether or not one thinks that the problem of deviant causal chains can be solved from within causalism, the strength of deviant counterexamples depends on the absence of control. It is because the climber loses grip on the rope that it would be implausible to insist that she lets go of the rope *intentionally* (Davidson 1973). And it is because a herd of wild pigs can hardly be controlled that it would be implausible to say that I shot dead my enemy intentionally even though my shot only killed her by awakening a herd of wild pigs which trampled her to death (Davidson 1973, Bishop 1989). These two are

[5] On these points, see also Di Nucci 2011b.

[6] Deviant causal chains are, since Davidson, the classic challenge to the *sufficiency* of causalism. There are many challenges to its *necessity* that I don´t have room to discuss here: Dreyfus's skilled activity (1984, 1988, 2005); arational actions (Hursthouse 1991), emotional behaviour (Goldie 2000), passive actions (Zhu 2004), habitual actions (Pollard 2003 & 2006), omissions (Sartorio 2005 & 2009; more on this in section 4), and automatic actions (Di Nucci 2008).

paradigmatic cases, respectively, of *basic* deviance – the climber's – and *consequential* deviance – the wild pigs': both, importantly, are built around lack of control.[7]

What this suggests is that, whether or not we can meet the challenge posed by deviant causal chains, the very fact that we intuitively find these cases challenging tells us that within our intuitions about intentional action and agency more in general there is embedded some kind of *control* condition: such that if a case does not meet this control condition, we won't find it at all plausible that the case can constitute an intentional action. This control condition would, then, appear to be a *necessary* one.

This would have potentially devastating consequences for causal theories of action. If we accept that a control condition is necessary in our account of action, then we cannot also accept the central thesis of causalism according to which whether something is an action depends solely on its causal history. The relevant content-specific psychological states as causes might be necessary for intentional action; but they could not be necessary and *sufficient* if the control condition is also necessary. But then an action – and also, crucially, the difference between an action and any other event – cannot be defined only in terms of its causal history. This would mean, in short, that the causal theory of action – understood as above – is false.

Importantly, we would have just shown that the causal theory of action is false without having to rely on the ultimate success of deviant causal chains as counterexamples; all that is needed is that deviant cases are found to be intuitively challenging – and if the philosophy of action literature of the last 40 years shows anything, it certainly shows that deviant causal chains have some degree of intuitive plausibility.

Not so quick: that some control condition is embedded in our intuitions about intentional action might suggest, but it does not imply, that a control condition also ought to be present in our *philosophical account* of intentional action. Still, it is important to remark that if the former did imply the latter then we would have already shown, in the few paragraphs above, that the causal theory of action is false: not because it is falsified by deviant causal chains, but simply because it does not include a control condition – as emphasized by deviant scenarios. But that a control condition is embedded in our intuitions does not imply that it should also feature in our philosophical account of intentional action because there might be other ways to account philosophically for our intuitions about control: that is what most attempts at 'solving' deviant causal chains have tried to do: articulate a causal theory which at the same time does not renounce its central claim that causal history alone can individuate actions and also accommodates our intuitions about deviant causal chains.

Here I cannot evaluate every attempt at solving the problem of deviant causal chains[8] : but I will analyse some representative proposals, showing that they are ultimately unsuccessful. A standard causalist proposal, as a solution to the problem of deviant causal chains, is the idea that psychological states 'guide' and 'sustain' action (see, for example, Brand 1984 or Thalberg 1984). The already introduced account of intentional action by Mele & Moser (1994) is a good representative of this tradition. Their second necessary condition for intentional action goes as follows: "(ii) at t, S suitably follows – hence, is suitably guided by – an intention-embedded plan, P, of hers in A-ing" (1994: 253).

This is supposed to rule out cases, such as deviant causal chains, in which a 'freak' event interposes itself between intention and action (basic deviance) or between action and intended

[7] Mele & Moser (1994, pp. 47–48) mention these two cases as 'exemplary', referring to basic deviance as 'primary' deviance and to consequential deviance as 'secondary'. Both scenarios are explained in detail within this section for those who are less familiar with them.

[8] For a recent anthology article on deviant causal chains see Stout (2010).

result (consequential deviance), so as to make it implausible that the agent acted intentionally. The 'freak' event, this proposal goes, breaks the guiding and sustaining relationship of the intention with the action or result; so that the action has not, in deviant cases, been guided and sustained by the relevant intention or primary reason; even though the relevant intention or primary reason still causes and rationally explains the movement in question.

Here I argue that emphasizing the guiding and sustaining role of intentions fails to accommodate deviant cases. I will start from cases of consequential deviance because they help illustrate my argument, and then show that my argument applies just as well to cases of basic deviance.

Take the standard scenario of consequential deviance: I shoot to kill you, but you only die because my wayward shot awakens a herd of wild pigs, which trample you to death. I intended for my shot to kill you, and my shot did kill you, so that my intention is satisfied; and my intention did cause its satisfaction. Still, this does not appear to be an example of intentional action; indeed, it isn't even clear that the statement 'I killed you', let alone the statement 'I shot you dead', are true: it is rather the pigs who killed you. But even though my intention is satisfied and it has caused its satisfaction, things did not go according to plan: I meant for the bullet to hit the victim in the chest, killing her. The idea is that the content of my intention has not successfully guided and sustained my movements; otherwise the bullet would have hit the victim in the chest, killing her. So even though my intention has been satisfied, what Mele & Moser (1994) call my 'action-plan' – to hit the victim in the chest, killing her – has not been satisfied; and that's why this is not a case of intentional action.

But the problem with this reply is that we can compare it to one where we would be changing the scenario so that I no longer intend to shoot you dead. If I did not intend to shoot you dead, then this scenario would not be a counterexample to the sufficiency of reasons (or intentions) as causes for intentional action, because it wouldn't be a scenario in which a reason or intention causes its satisfaction but the agent still hasn't acted intentionally. We wouldn't accept a reply to the deviant counterexamples that changed the agent's intentions or reasons; therefore we shouldn't accept this kind of proposal either: because it changes the agent's intentions or reasons.

Superficially, it looks as though the agent's intention hasn't changed, because the agent is still described as having acted with the intention to shoot her victim dead. But by stipulating that the intention contains an action-plan to act in a certain way, the agent's intention *has* actually been changed: the agent no longer simply intends to 'shoot her victim dead'; she now intends to 'shoot her victim dead by hitting her in the chest'. What's the difference between 'shoot her victim dead' and 'shoot her victim dead by hitting her in the chest'? The difference is quite simply that there are other ways to shoot someone dead other than hitting them in the chest. And the deviant counterexample works exactly on the intention to 'shoot her victim dead' being realizable in multiple ways. If we change the content of the intention by narrowing down its conditions of satisfaction, just like if we change the intention altogether, then obviously the deviant scenario no longer shows that the sufficiency claim is false. But given that, in both cases, we have changed the scenario instead of arguing against its supposed implications, then that's no surprise.[9]

[9] I accept that there may be other strategies here which cannot be compared to changing the agent's intentions; but my point is only directed against Mele & Moser's attempt to deal with deviance by further specifying the intention's content through their 'action-plans'; and that particular strategy has the problem I just emphasized. I thank an anonymous referee from pressing me on this point.

And changing the scenario won't do also because the deviance can be changed accordingly, so that a new deviant case can be built around the intention's new, more specific, content of 'shoot her victim dead by hitting her in the chest'. Suppose that, after the events, the shooter is interviewed: "Did you mean to kill her by having a herd of wild pigs trample her to death?" "No, I meant for the bullet to hit her directly and kill her". "Didn't you know she was wearing a bullet-proof vest? Suppose that the bullet had hit her on the chest, and that the vest had protected her. Still, it pushed her to the ground, where she fell on a deadly sharp knife, which killed her. Would you have killed her intentionally *then*?" "No, I meant for the bullet itself to kill her, *directly*". "OK, now suppose...". The regress could continue until we reach action-plans too detailed to be plausibly attributed to agents who act intentionally.

Independently of this regress, we should be in general wary of over-intellectualising planning agents by thinking that their intentions are as specific as Mele & Moser's action-plans. Three points here[10] :

(a) there is probably an indefinite number of micro-descriptions of what we do, but to think that agents where representing all of them would make the intellectual life of the planning agent much more complicated than it is or needs to be;

(b) lots of evidence on automaticity and habitual action suggests that we regularly act purposefully and intentionally without consciously or unconsciously representing our goals; (on this and the previous point see my Di Nucci 2008, Di Nucci 2011c and Di Nucci 2013a);

(c) finally, over-intellectualising may also get agents and their priorities wrong; especially when the means are morally neutral, agents are only bothered by ends and not also by means; stipulating that the end is achieved intentionally only where a specific set of means has been fulfilled may just represent agents' reasoning and priorities in planning and acting.[11]

It could be objected that the above strategy, whatever its merits, was at least able to explain (or at least account for) the relevant sequences being deviant. Why was the way in which the intention was satisfied thanks to the pigs' contribution deviant? Because the agent had in mind a way to satisfy the intention which was different from the way in which the intention was satisfied in reality. And this miss-match between mind and reality explains why those cases cannot count as intentional actions. So then the burden would be on critics of this proposal to be able to explain why these cases cannot count as intentional actions without specifying the intention's content as above.

And it is by recognising that what deviant cases expose is, primarily, the absence of control that we can also explain why those are not intentional actions; in the pigs' case, the agent does not intentionally kill her victim because she is not in control of her victim's death, since she cannot control the pigs. Similarly, in the climber's case the agent does not intentionally let go of the rope because she is not in control of the rope when she lets go of it. So control can explain these cases as non-intentional ones.

Here one could object that control is not necessary for intentional action. Take, for example, the case in which the agent did in fact intend to kill by awakening a herd of wild pigs which would then trample the victim.[12] Here, it could be suggested, the agent can be said to have killed intentionally even though she lacked at least some decisive degree of control

[10] Thanks to an anonymous referee for pressing me on the regress.

[11] For more on this point see Di Nucci 2013b, 2013c, 2013d and (forthcoming).

[12] Thanks to an anonymous referee for suggesting this scenario.

over the satisfaction of her plan – namely she could not control the herd of wild pigs. Here intuitions may indeed differ so I will just defer to the standard literature on the topic in the philosophy of action, where the talk is of rational constraints on intention: I intend A only if I believe I will A is Grice's stronger version of the constraints (1971); and I intend A only if I do not believe that I will not A is Bratman's weaker version of the constraints (1984 & 1987). On both versions the idea would be that if an agent believes that she will not achieve her goal (either because achievement is impossible or because it is improbable or because it is, all things considered, unlikely – as in less than 50 % likely), then she does not intend to achieve it and even if she were to achieve it then the achievement would not be intentional – even though her trying would be intentional. Take the case of someone who has never played golf before but manages a hole-in-one on her first ever time: here it seems that these accounts of rational constraints on intention are in line with intuition in saying that the hole-in-one was neither indented nor intentional.[13]

Specifying the intention's content, on the other hand, does not guarantee control – that's the point of the regress of deviant cases. The only way of doing so is stipulating control within the agent's motivation; so that agents don't simply intend to kill or drink water; agents intend to kill and for the killing to be under their control; and they intend to drink and for the drinking to be under their control.

Indeed, were we to define agency in terms of control instead of in terms of motivation (as causalists traditionally do following Davidson's (1971, 1973, 1978) lead), it would be implied in the content of the intention to 'drink' or 'kill' that the performance must be under the agent's control. If action requires control, then 'kill' can only refer to a true action if it implies control. So that if I intend to perform some action, then I must intend for the performance to be under my control – otherwise I wouldn't intend to perform an *action*.

But, again, specifying the intention's content does not guarantee control. Ian might have an intention to shoot Jen dead by putting a bullet through Jen's forehead, whereby the bullet cuts through her brains destroying systems that are essential for Jen's basic survival – so that it directly causes her death. And suppose that Ian does shoot, and that the bullet does exactly what Ian meant for the bullet to do, and that Jen dies as a direct result of the bullet's trajectory – which was exactly as Ian had planned it. But, unbeknownst to Ian, the bullet only managed to hit the target thanks to an invisible superhero's crucial intervention: it was the invisible superhero that, when the bullet was halfway to its target, took control of it and guided it so precisely where Ian meant it. Ian did everything as planned, but it was only through the superhero's timely intervention that Ian's shot was so precise.

Even though Ian's intention and action-plan were satisfied to the last centimetre, still it looks as though Ian did not intentionally kill Jen – indeed, Ian didn't even kill Jen: the invisible superhero who intervened at a crucial time did. Ian might have fired the shot with the relevant intention and action-plan, but since he did not control his shot, it wasn't *he* who killed Jen. Again, the missing link turns out to be control. Without control there is no action. So Ian killed Jen, and killed her intentionally, only if he controlled the events that proximally caused her death, including the bullet.

What about the case where the invisible superhero does not need to intervene because Ian's shot is precise enough? It may be suggested that guidance has the unwelcome consequence that this case would not count as Ian's intentionally killing because the control is with the

[13] On these issues see also a recent exchange between Di Nucci and McCann in Analysis (Di Nucci 2009 & 2010b, McCann 2010 & 2011).

superhero, but that on Mele & Moser's account the case would count as intentional killing because things went as Ian's action-plan set them up. Two points here: firstly, whether Mele & Moser could claim that this is an intentional case is not obvious, as the superhero's presence and potential intervention was not part of the action-plan. Secondly, I am not sure that one could not claim that this was Ian's intentional killing on a guidance account: after all, that the superhero has guidance does not rule out that Ian may also have guidance; indeed, this may be a case where both have guidance, so that both intentionally kill. And given that the superhero could have easily saved Jen it does not sound implausible to attribute her killing also but not only to the superhero (on these kinds of scenarios, see Di Nucci 2010a, Di Nucci 2011a and Di Nucci 2011b).[14]

What if, the causalist might propose, we build control within Ian's intention and action-plan so that Ian had specified, in formulating his plan, that he meant for no outside intervener to interfere with his murder? Then causalists would be conceding that agents take control to be necessary for intentional action. And also that indeed control is necessary for intentional action – because some movement would then qualify as an intentional action only if it meets some control condition – in this case one stipulated by agents themselves.

But if control is necessary for intentional action, then causalists are wrong. Because then reasons as causes are not sufficient: namely, a movement being caused by a psychological state which rationalizes it isn't sufficient for that movement qualifying as an intentional action – that movement must also be under the agent's control.

These arguments also apply to cases of so-called basic deviance such as Davidson's original climber's scenario (1973):

> A climber might want to rid himself of the weight and danger of holding another man on a rope, and he might know that by loosening his hold on the rope he could rid himself of the weight and danger. This belief and want might so unnerve him as to cause him to loosen his hold, and yet it might be the case that he never chose to loosen his hold, nor did he do it intentionally (Davidson 1973: 79).

A climber formulates the intention to let go of the rope to which her fellow climber is attached so as to kill her fellow climber. Her murderous intention so unnerves the climber that she loses her grip on the rope, thereby letting go of it. The relevant intention caused the movement, but still the movement was no intentional action of the climber: it was an accident. Again, it is lack of control that makes it implausible to argue that the climber let go of the rope intentionally. And, again, we could specify the climber's action-plan so as to rule out the possibility that the intention is satisfied by the climber losing her grip. But at this stage the case becomes equivalent to the one I analysed in this section, so that the previous arguments apply.[15]

3.1 Deviance and Intentional Content

The standard causalist strategy of embedding the guiding and sustaining role in the intention's content fails for both cases of basic deviance and cases of consequential deviance. I will now discuss a more recent proposed solution to the problem of causal deviance, showing that this one comes up short too. It has been recently argued that, assuming "the intentional contents of

[14] I thank an anonymous referee for suggesting this scenario.

[15] It has been argued that the action-plan strategy cannot be applied to cases of so-called basic deviance because some of these cases are too *basic* to think that agents might have planned how to go about them (Bishop 1989, 132–34; see also Schlosser 2007). I am sympathetic with this point; but anyway my arguments in this section show that even if Bishop is wrong and there is a way to apply the action-plan reply to basic deviance, still the reply would be unsuccessful.

reason states are causally relevant and causally explanatory" (Schlosser 2007: 191) of action, then cases of so-called 'basic' deviance can be accommodated within a causal view.

Markus Schlosser's proposed solution to Davidson's climber scenario goes as follows: the climber's intention to rid himself of his fellow climber by loosening his hold on the rope causes him to loosen his hold; but it does not do so in virtue of its content, because part of the causal chain is the climber's nervousness, which is caused by the climber's intention, and which in turn causes the loosening of his hold. But, Schlosser says, "the event of nervousness, trivially, does not cause the movement in virtue of content" (2007: 192). And so the intention could not have caused the movement in virtue of its content, given that it only caused the movement through the state of nervousness. And that is why the movement is not an intentional action even though it is caused and rationalized by the agent's intention to loosen his hold.

Schlosser says that "the reason-states do not explain the occurrence of the particular movement in virtue of their contents – why that particular type of movement occurred, rather than another, cannot be explained by reference to the contents of the reason-states" (2007: 192). That is, according to Schlosser, because the reason-states only cause the movement through a state of nervousness which is, "trivially", a state which lacks intentional content. And it therefore couldn't cause anything in virtue of its intentional content.

Schlosser concludes that "Being caused and causally explained in virtue of content, an action is not merely a response to a cause, but it is a response to a reason-state qua reason-state; it is a response to the content of the mental state in the light of which its performance appears as intelligible" (2007: 192).

It has been recently pointed out (Tannsjo 2009) that even if the interposed state of nervousness is, as Schlosser argues, content-less, still that cannot be enough to account for why those cases do not constitute intentional action. Tannsjo argues that it is often the case, when we act intentionally, that our behaviour is constituted by non-intentional and content-less components:

The problem is that there are some cases where even folk psychology allows for such nonintentional parts of an action. A simple example is when I kick a ball. There are many movements of my legs that are not made in response to the content of my wish to kick the ball; they just happen, and their happening is caused by my desire to kick the ball (2009: 470).

The problem for Schlosser's proposal would then be that many of our movements aren't caused by our reasons or intentions in virtue of their intentional content, simply because it would be both implausible and unnecessary to require that all that we do intentionally is represented within the intentional content of our reasons or intentions. When we kick a ball, we normally do so both successfully and intentionally even though many of the minute performances and movements involved are not represented within the intentional content of our reasons or intentions, and they are therefore not caused by our reasons or intentions in virtue of their content.

The general problem is that we cannot plausibly require that every component of our agency be represented within the psychological states that are supposed to have caused our intentional action. Agents aren't gods; and not only gods act intentionally. Mostly, agents act intentionally even though they could not possibly be aware of every facet of their movements, so that those couldn't be represented within the agent's motivational states.

The problem with Tannsjo's objection is that causalists might very well be happy to concede that these movements, which couldn't be plausibly represented within the agent's reasons or intentions, are not intentional movements. That I intentionally kick a ball does not mean that every aspect, component, or element of my ball-kicking is something that I did intentionally. Kicking a ball might then turn out to be intentional under descriptions such as 'kicking a ball', 'playing football', and 'showing my son how it's done', without thereby having to be intentional under descriptions such as 'moving my foot forward', 'lifting my leg by 12 centimetres', and 'shortening the life-expectancy of the grass'.

And if we accept that the former set of action descriptions can be intentional without the latter set also having to be intentional, then causalists might be happy to concede that the latter set of action descriptions are not intentional; and then they could say that, indeed, these action descriptions are not intentional because they have not been caused by the agent's relevant intention in virtue of its content – since the intention's content makes no mention of them.

The issue, here, becomes foundational: it is argued, on the one hand, that agents do not have to think, occurently, dispositionally, or unconsciously, about every detail, element, and consequence of their actions: that those elements, details, and consequences are intentional even though they were not represented in the content of the agent's reasons or intentions. To demand so much of agents would be absurd. On the other hand, it is argued that no such absurdity is involved, since those details, elements, and consequences are not intentional actions.

But aren't these components still necessary to the performance? And wouldn't agents own up to them if you asked them? "Did you mean to move your foot forward?"; "Did you mean to lift your leg by 12 centimetres?" On the one hand, no agent could have possibly known the exact height at which to lift her leg. But, on the other hand, no agent would deny that they had somewhat meant to do that, since it was required in order to kick the ball – and they definitely meant to kick the ball.

So they hadn't thought about it but, with hindsight, they must have meant to do it if it was part of kicking a ball. What started as a problem for the sufficiency of causal views of action is now starting to look like a problem for the necessary conditions of causal views: are the relevant psychological states really necessary, since agents appear to have meant to do even things that they hadn't thought about, either occurently, dispostionally, or unconsciously? If even those things turn out to have been performed intentionally by agents, then it looks as though the causal view's necessary conditions for intentional action are being challenged – since there is no trace of those performances in the agent's reasons or intentions.[16]

Here the discussion soon becomes fairly technical and complicated if the causal view has to appeal to such things as non-propositional content and sub-personal states in order to show that these performances can indeed be traced back to the agent (see, for example, Bermudez 1995). But here we don't need to take on this major task, because we don't need to accept, as Tannsjo does, Schlosser's assumption that the movements of the climber aren't caused in virtue of the intentional content of the climber's psychological states.[17]

We can accept that the climber's intention to loosen his hold causes the climber to loosen his hold only through a state of nervousness. What we don't need to accept is the bit that Schlosser does not argue for but rather stipulates as 'trivial' (2007: 192): that since the loosening of his hold is caused by a state of nervousness, and since states of nervousness are, by definition, devoid of intentional content, then the loosening of his hold could not have been caused in virtue of content – and therefore it cannot be an intentional action.

Schlosser says that "the reason-states do not explain the occurrence of the particular movement in virtue of their contents – why that particular type of movement occurred, rather than

[16] For an in-depth discussion of these issues, see Di Nucci 2008.

[17] Schlosser's own reply to Tannsjo (2010) is therefore not relevant to my argument here.

another, cannot be explained by reference to the contents of the reason-states" (2007: 192). And also that: "Being caused and causally explained in virtue of content, an action is not merely a response to a cause, but it is a response to a reason-state qua reason-state; it is a response to the content of the mental state in the light of which its performance appears as intelligible" (2007: 192).

Neither of these points is so obvious as not to require argument. Isn't it reasonable that the agent, being nervous, loosened his hold on the rope? Isn't that the sort of thing that would happen to a nervous climber, loosening his hold? Think of the sweat; think of how difficult it would be to maintain the required level of concentration. Furthermore, isn't it reasonable that a person with a conscience would grow nervous at the thought of sacrificing his fellow climber? Wouldn't that be likely to happen to any half-decent person?

Schlosser says that "the state of nervousness… renders it a coincidence that the reason states cause and rationalize the bodily movement" (2007: 191). But it is no coincidence that the climber loosens his hold. And it is no coincidence that the climber becomes nervous. It is in virtue of his intention to loosen his hold that the climber becomes nervous. Another intention, such as, say, the intention to 'have a drink once the climb is over', could hardly have been expected to result in nervousness – it would have been likely to have had a calming influence if anything.

And it is in virtue of his nervousness that the climber loosens his hold. It is because he is nervous that he loosens his hold. Another emotion, such as, say, a sudden rush of affection towards his partner back home, could hardly have been expected to result in the loosening of his hold – if anything, the climber would have tightened his grip on the rope.

The point is that malicious intentions such as the intention to kill a fellow climber are precisely the kind of mental states that normally cause nervousness. And that emotional states of mind such as nervousness are precisely the kind of states of mind that cause loss of control, mistakes, accidents; such as, in these circumstances, the loosening of the climber's hold.

So it is, after all, in virtue of the climber's intention to 'loosen his hold' being an intention to 'loosen his hold' – and not an intention to have a pint later that evening – that the climber grows nervous: it is in virtue of the intention's particular content, 'loosening his hold', that the state of nervousness arises – had the content of the intention been different, it is reasonable to suppose that the agent would not have grown nervous. And it is in virtue of the climber's state of nervousness being that particular state of mind – as opposed to a sudden rush of affection or love – that the climber loosens his hold: had the climber been in a different emotional state of mind, it is reasonable to suppose that he would not have loosened his hold.

Schlosser's solution depends on the idea that, on top of a causal relation, there is also a rational relation between 'normal' pairs of reason (or intention) + action. And that in deviant cases this breaks down: there is no rational relation between the climber's intention to loosen his hold and his loosening his hold, because there is no rational relation between the climber's nervousness and his loosening his hold. So even though the causal relation still holds, the rational relation is interrupted by the state of nervousness.

But I have just shown that there are rational relations both between the climber's intention and his nervousness, and between his nervousness and his loosening his hold. Each pair of events is neither randomly nor coincidentally connected: we would reasonably expect them to be connected in just the way in which they are connected.

Naturally, this is not the same kind of rational relation: because the agent does not loosen his hold in light of his state of nervousness; but isn't his state of nervousness the reason why he loosens his hold? To say this is to misinterpret what 'reasons' are, a causalist ought to reply.

And that's true. The climber does not grow nervous *in order to* satisfy his intention; nor does he let go of the rope *in order to* satisfy his nervousness either. But that's just to re-state the agreed upon data: deviant cases are different from *normal* cases. The point is that the difference is not where Schlosser places it: namely in the idea that the presence of the intermediary contentless state brakes down the normative relation between the intention to let go and letting go; because not only does a normative relation between intention and action still stand, but it also runs, importantly, through the very intermediary state of nervousness.

Let's be perfectly clear here: I am not arguing that the relation between the agent's intention to let go and the state of nervousness, and the relation between the state of nervousness and letting go, are the same kind of rational relations as, say, the relation between my desire for a cup of tea and my boiling the kettle. Even though these are both kinds of explanatory relations, they are different kinds of explanatory relations. So my argument does not amount to equating them; I am only denying that the difference between these two kinds of explanatory relations is that only the latter is a rational/normative relation in which two events are causally connected in virtue of content. This much – which is the distinctive feature upon which Schlosser rests his argument – both kinds of relations have in common.

We can now see the same argument from a different point of view. Schlosser claims that "why that particular type of movement occurred, rather than another, cannot be explained by reference to the contents of the reason states" (2007: 192). We can now see that this is not true. It is exactly the fact that the content of the climber's intention is 'loosening his hold' that explains why the climber grows nervous. Indeed, we couldn't reasonably have expected the climber to suddenly feel *gratitude* towards his fellow climber as the result of his intention to 'loosen his hold'. Similarly, it is exactly the fact that the climber grows nervous that explains his loosening his hold. Loss of control is often explained by nervousness – and it is reasonable that a nervous person would lose control. As a result of nervousness, for example, we would not have expected the climber to, say, take a novel out of his rucksack.

This alternative fails too, then. In this section I have argued that the problem of deviant causal chains cannot be accommodated by the causal theory of action. I now turn to the relationship between causalism, psychological states, and control.

4 Control, causalism, and psychological states

It could be thought that guidance isn't really an alternative to causal theories of action because guidance itself depends on the attribution of psychological states as causes. Al Mele (1997) has gone in this direction. In this section I challenge his arguments, arguing that guidance, as opposed to causal theories, does not require psychological states. Before analysing Mele's argument, I should emphasize the generality of my discussion in this section: in arguing against the need to necessarily attribute psychological states as causes, I also provide an another independent general reason against causalism as contrasted to guidance, namely that it needs the attribution of psychological states as causes. And while deviant causal chains are a challenge to the sufficiency of the causal view, arguing that psychological states as causes are not necessary to account for control is a challenge to the necessity of the causal view (see footnote 6 for literature that challenges the necessity of the causal view).

Mele applies his argument directly to Frankfurt's coasting scenario:

> In the absence of a desire or intention regarding 'the movement of the automobile', there would be no basis for the driver's being 'satisfied' with the speed and direction of his car. So we might safely attribute a pertinent desire or intention to the driver, whom I shall call Al. What stands in the way of our

> holding that Al's acquiring a desire or intention to coast down hill is a cause of his action of coasting, and that some such cause is required for the purposiveness of the 'coasting'? ... his allowing this [the 'coasting'] to continue to happen, owing to his satisfaction with the car's speed and direction, depends (conceptually) on his having some relevant desire or intention regarding the car's motion (1997, p. 9).

Mele thinks, then, that we can "safely attribute" the relevant psychological states, and that nothing stands in the way of thinking that those psychological states are causing the agent's behaviour. "Then it is natural to say that Al is coasting in his car because he wants to, or intends to, or has decided to – for an identifiable reason. And the 'because' here is naturally given a causal interpretation. In a normal case, if Al had not desired, or intended, or decided to coast, he would not have coasted; and it is no accident that, desiring, or intending, or deciding to coast, he coasts" (1997, p. 9).

My argument against Mele in this section will develop in two directions: first, I will argue that the issue is not the possibility of the attribution of the relevant psychological states, but rather its necessity. Secondly I will argue, following Carolina Sartorio (2005 & 2009), that these cases cannot be explained by appeal to 'reasons as causes'.

It has already been noticed (Zhu 2004, p. 304) that arguing for the attribution of the relevant intention is not enough for the causalist. What the causalist needs is to argue for the attribution of the relevant intention as a cause. But one might think that the relevant intention is necessary without thinking that the relevant intention is necessarily causal: "the explanation that Al allows the car to continue to course because 'he wants to, or intends to, or has decided to' for certain reasons, does not imply that it must exclusively be a causal explanation. Some philosophers contend that reasons explanations of action can be non-causal explanations as well" (p. 304).

Also, it is not enough for Mele to show that it is possible to attribute the relevant intention to the agent – namely, the agent's intention to coast. What Mele needs to show is that the attribution of the intention to coast is necessary in order for the agent to coast intentionally. If Mele doesn't show that, then he leaves room for an alternative account, one on which there is no intention to coast. It might be, for example, that all the agent intends to do is get home: and that, because coasting doesn't undermine the satisfaction of that intention, the agent doesn't intervene. The agent's intention to get home doesn't imply the agent's intention to coast: it might be that the agent's intention to get home leaves room for the agent's intention to coast, given that coasting is, admittedly, one of many ways in which the agent can satisfy her intention to get home.

But, again, that is not enough: what Mele needs is to show that the intention to coast is necessary. That, namely, the agent could not have coasted without an intention to coast; rather than just that the agent could have been coasting as the result of an intention to coast. Mele has only shown the latter, but not the former, and that is why Frankfurt's account stands.[18]

Mele's point might show that the agent doesn't intend not to coast – because if she had intended not to coast, presumably, since her behaviour was under her guidance, she would not

[18] Obviously a general intention or plan to 'get home' is not enough for a causalist. Let us explain that in Davidson's terms: if we analyse the general plan to 'get home' in terms of a desire to 'get home' and a belief that 'driving will get us home', for example, that belief-desire pair does not rationalize 'coasting' because there is no mention of 'coasting' in either the content of the desire or the content of the belief. And that is the same reason why a general intention to 'get home' which makes no mention of 'coasting' in its content will not do. That is why, if the intentional action in question is 'coasting', Mele needs to argue that an intention to coast (or a desire to coast, or a similarly suitable belief) is necessary.

have coasted. But showing that the agent doesn't intend not to coast falls short of attributing any intention to the agent: it doesn't show that the agent intends to coast. So that isn't enough either. And it is important to emphasise that it is not open to a causalist to argue that 'intending to φ' and 'not intending not to φ' are equivalent, since only the former points to an *actual* psychological state: and the causalist needs actual psychological states because she needs causes.

Mele is looking for a reason not to attribute psychological states to the agent; and a reason not to take them to cause the agent's movements. But what Mele needs, in order to refute Frankfurt, is to show that there cannot be guidance without those psychological states causing movement. Frankfurt's challenge is exactly that guidance doesn't depend on causal antecedents. Because all that Mele shows is that it is possible to attribute those psychological states, Mele does not show that guidance isn't possible without those psychological states. In order to show the latter, Mele should have argued that the attribution of those psychological states is necessary, and not merely possible.

So far I have been granting to Mele the possibility of attributing the relevant psychological states, arguing that to reduce guidance to causalism is not enough that it is possible to attribute these psychological states; the relevant psychological states need to be necessary, but they are not, and therefore Mele's case fails. But recent work on omissions suggests that the attribution might be problematic, so that the argument against Mele would be even stronger.: not just that Mele fails to show that the attribution is necessary. More importantly, the attribution would not be warranted.

If we take Frankfurt's scenario to be a case of omission (omission to actively drive; omission to intervene; omission to grab the wheel), then it is not clear that the psychological states required by Mele's argument can explain the driver's behaviour. Sartorio has recently argued (2005 & 2009 – see also Clarke 2010) that causal theories of action cannot accommodate omissions because omissions cannot be explained in terms of 'psychological states as causes'. Focusing on an example involving a drowning child and a passive by-stander, she argues that the failure of the by-stander to intervene to save the child – which constitutes an omission – isn't causally explained by the by-stander's psychological states (Sartorio focuses specifically on the state of intention). She claims that the following causal explanation is false: (A1) 'My forming the intention not to jump in' causes (O2) 'my failure to jump in'. Granting the possibility that omissions can belong to causal chains, Sartorio claims that this causal explanation – which exemplifies the kind of causal explanations provided by causalism – fails; and that therefore, generalizing, causalism fails with regards to omissions.

According to Sartorio the truth of 'My forming the intention not to jump in causes my failure to jump in' is challenged by the following being true: (O1) 'My omitting to form the intention to jump in' causes (O2) 'My omitting to jump in'. Sartorio's claim is a conditional: If 'O1 causes O2', then it is false that 'my forming the intention not to jump in causes my failure to jump in'. Sartorio argues for the antecedent by arguing that (O1) is a better causal explanation of (O2) than (A1), my forming the intention not to jump in. Indeed, the claim is even stronger: I omitted to jump in because of O1 and not because of A1: "I failed to jump in because of what I omitted to intend to do, not because of what I intended to do" (2009: 519), where what I omitted to intend to do refers to (O1) and what I intended to do refers to (A1). So I failed to jump in not because of my intention not to jump in. Therefore my intention not to jump in does not explain my omitting to jump in. It follows that the claim that 'my forming the intention not to jump in causes my failure to jump in' – which is an example of causalist explanation – is false.

Therefore, following Sartorio's argument, it would not just be, as I have argued above, that the psychological explanation as causal explanation in Frankfurt's scenario is merely possible but not necessary; also, the psychological explanation as causal explanation fails because the explanatory work is done by what I don't intend and not by what I do intend.

Here it might be insisted, on behalf of Mele, that at least the actual intervention, if not the coasting, isn't possible without the agent being in some mental state; and that if the agent is not able to intervene, then she hasn't got guidance over her actions. So guidance does depend on the agent being in some psychological state. But, again, all that is needed, if anything, for the agent's intervention is some intention to get home. If something happens or is about to happen that might undermine the satisfaction of such an intention, then the agent might intervene. But her intervention doesn't require an intention to coast, nor does her intervention show that the agent had an intention to coast.

Mele might have been hinting, rather, at the intention to coast being necessary in order to *explain why* the agent is coasting. Two points here: firstly, the difficulties faced by causalism that I emphasized in this paper are a direct result of the ambition to offer, all-in-one, a definition of intentional action together with a reasons explanation: that's at the root of the problem of deviant causal chains. Secondly and more importantly, as I pointed out in my discussion of Mele, we can actually make rational sense of the agent's coasting without attributing an intention to coast. If, for example, all the agent intended was to go home; and the agent did not intend not to coast, then his coasting makes perfect rational sense; and we have then explained why he is coasting. And we have done so, contra Mele, without attributing an intention to coast.

We have here rebutted Mele's attempts to reduce a form of control such as guidance back to the causalist model of psychological states as causes. In conclusion, let me just summarize what this article has achieved: I have argued that we should abandon the long struggle to patch up causalism, and that we can make sense of control independently of causalist commitments.

References

Adams, F. (1986), 'Intention and Intentional Action: The Simple View', *Mind & Language* 1: 281–301.

Adams, F. and Mele, A. (1989), 'The Role of Intention in Intentional Action', *Canadian Journal of Philosophy* 19: 511–31.

Alvarez, M. (2010), *Kinds of Reasons*. Oxford UP.

Anscombe, G.E.M. (1957), *Intention*. Basil Blackwell.

Bermudez, J. (1995), 'Nonconceptual Content: From Perceptual Experience to Subpersonal Computational States', *Mind and Language* 10: 333–369.

Bishop, J. (1989), *Natural Agency. An Essay on The Causal Theory of Action*. Cambridge University Press.

Brand, M. (1984), *Intending and Acting*. MIT Press.

Bratman, M. (1984), 'Two Faces of Intention', *Philosophical Review* 93: 375–405.

Bratman, M. (1987), *Intention, Plans, and Practical Reason*. Cambridge, Mass.: Cambridge University Press.

Clarke, R. (2009), 'Dispositions, Abilities to Act, and Free Will: The New Dispositionalism', *Mind* 118: 323–351.

Clarke, R. (2010), 'Intentional Omissions', *Nous* 44 (1): 158-177.

Collins, A. W. (1997), 'The psychological reality of reasons', *Ratio*, X: 108–123.

Dancy, J. (2000), *Practical Reality*. Oxford UP.

Davidson, D. (1963), 'Actions, Reasons, and Causes', *Journal of Philosophy* 60: 685–700.

Davidson, D. (1971), 'Agency', in Binkley, R., Bronaugh, R., and Marras, A. (eds.), *Agent, Action, and Reason*. University of Toronto Press.

Davidson, D. (1973), 'Freedom to Act', in Honderich, T. (ed.), *Essays on Freedom and Action*. Routledge and Kegan Paul, 137–56.

Davidson, D. (1978), 'Intending', in Yovel, Y. (ed.), *Philosophy of History and Action*. The Magnes Press, The Hebrew University.

Di Nucci, E. (2008), *Mind Out of Action*. VDM Verlag.

Di Nucci, E. (2009), 'Simply, false', *Analysis* 69/1: 69–78.

Di Nucci, E. (2010a), Refuting a Frankfurtian Objection to Frankfurt-Type Counterexamples. *Ethical Theory and Moral Practice* 13 (2): 207–213.

Di Nucci, E. (2010b), Rational constraints and the Simple View. *Analysis* 70 (3): 481–486.

Di Nucci, E. (2011a), Frankfurt counterexample defended. *Analysis* 71 (1): 102–104.

Di Nucci, E. (2011b), 'Frankfurt versus Frankfurt: a new anti-causalist dawn', *Philosophical Explorations* 14 (1): 1–14.

Di Nucci, E. (2011c), Automatic Actions: Challenging Causalism. *Rationality Markets and Morals* 2 (1): 179–200.

Di Nucci, E. (2013a), *Mindlessness*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Di Nucci, E. (2013b), 'Embryo Loss and Double Effect', *Journal of Medical Ethics 39* (8): 537–540.

Di Nucci, E. (2013c), 'Double Effect and Terror Bombing', in Hoeltje M. Spitzley T. & Spohn W. (eds.), *Was dürfen wir glauben? Was sollen wir tun? Sektionsbeiträge des achten internationalen Kongresses der Gesellschaft für Analytische Philosophie e. V.* (DuEPublico ISBN 978-3-00-042332-1).

Di Nucci, E. (2013d), 'Self-Sacrifice and the Trolley Problem', *Philosophical Psychology* 26 (5): 662–672.

Di Nucci, E. (forthcoming). *Ethics without Intention.* London: Bloomsbury.

Dreyfus, H. & Dreyfus, S. (1984), 'Skilled Behavior: The Limits of Intentional Analysis', in Lester, E. (ed.), *Phenomenological Essays in Memory of Aron Gurwitsch*. The University Press of America.

Dreyfus, H. (1988), 'The Socratic and Platonic Bases of Cognitivism', *AI & Society* 2: 99–112.

Dreyfus, H. (2005), 'Overcoming the Myth of the Mental: How Philosophers Can Profit from the Phenomenology of Everyday Expertise', *APA Pacific Division Presidential Address.*

Fara, M. (2008), 'Masked Abilities and Compatibilism'. *Mind*, 117, pp. 843–65.

Fischer, J.M. & Ravizza, S.J. (1998), *Responsibility and Control.* Cambridge UP.

Frankfurt, H. (1978), 'The Problem of Action', *American Philosophical Quarterly* 15: 157–162.

Goldie, P. (2000), 'Explaining expressions of emotions', *Mind* 109: 25–38.

Grice, H. P. (1971), 'Intention and Uncertainty', *Proceedings of the British Academy* 57, 263–79.

Hampshire, S. (1959), *Thought and Action*. Chatto and Windus.

Hursthouse, R. (1991), 'Arational Actions', *Journal of Philosophy* 88 (2): 57–68.

Lowe, J. (1999), 'Self, Agency, and Mental Causation', *Journal of Consciousness Studies* 6: 225–239.

McCann, H. (1991) 'Settled Objectives and Rational Constraints', *American Philosophical Quarterly* 28: 25–36.

McCann, H. (2010), Di Nucci on the Simple View. *Analysis* 70: 53–59.

McCann, H. (2011), The Simple View again: a brief rejoinder. *Analysis* 71(2): 293–95.

Melden, A. I. (1961), *Free Action*. Routledge & Kegan Paul.

Mele, A. (1992), *Springs of Action*. Oxford UP.

Mele, A. (1997), *Philosophy of Action*. Oxford UP.

Mele, A. and Moser, P. K. (1994), 'Intentional Action', *Nous* 28: 39–68.

Norman, D.A. & Shallice, T. (1986), 'Attention to Action: willed and automatic control of behaviour', in Davidson, R.J., Schwartz, G.E. & Shapiro, D. (eds.), *Consciousness and Self-Regulation*, iv. New York: Plenum, 1–18.

Perner, J. (2003), 'Dual control and the causal theory of action', in Roessler, J. & Eilan, N. (eds.), *Agency and Self-Awareness.* Oxford: Clarendon Press.

Pollard, B. (2003), 'Can Virtuous Actions Be Both Habitual and Rational?', *Ethical Theory and Moral Practice* 6: 411–425.

Pollard, B. (2006), 'Explaining Actions with Habits', *American Philosophical Quarterly* 43: 57–68.

Sartorio, C. (2005), 'A new asymmetry between actions and omissions', *Nous* 39: 460–482.

Sartorio, C. (2009), 'Omissions and Causalism', *Nous* 43: 513–530.

Schlosser, M.E. (2007), 'Basic deviance reconsidered', *Analysis* 67 (3): 186–194.

Schlosser, M.E. (2010), 'Bending it like Beckham: movement, control, and deviant causal chains', *Analysis*: 70 (2): 299–303.

Searle, J. (1983), *Intentionality*. Cambridge UP.

Smith, M. (2003), 'Rational Capacities, or: How to Distinguish Recklessness, Weakness, and Compulsion', in Stroud, S. & Tappolet, C. (eds.) *Weakness of Will and Practical Irrationality*. Oxford UP.

Stout, R. (1996), *Things that happen because they should*. Oxford UP.

Stout, R. (2010), 'Deviant Causal Chains', in O'Connor, T. & Sandis, C. (eds.), *A Companion to the Philosophy of Action.* Blackwell: 159–165.

Stoutland, F. (1985), 'Davidson on Intentional Behaviour', in LePore, E. And McLaughlin, B.P. (eds.), *Actions and Events*. Basil Blackwell.

Tannsjo, T. (2009), 'On deviant causal chains – no need for a general criterion', *Analysis* 69: 469–473.

Thalberg, I. (1984), 'Do our intentions cause our intentional actions?', *American Philosophical Quarterly* 21: 249–260.

Vihvelin, K. (2004), 'Free Will Demystified: A Dispositional Account'. *Philosophical Topics*, 32, pp. 427–50.

von Wright, G.H. (1971), *Explanation and Understanding*. Cornell UP.

Zhu, J. (2004), 'Passive Action and Causalism', *Philosophical Studies* 119: 295–314.

abstracta *Linguagem, Mente & Ação* ISSN 1807-9792 Volume 7, Number 2, pp. 61–68, 2013

Silogísticas Keynesianas: As Inferências Imediatas[1]

Frank Thomas Sautter[1] and Isac Fantinel Ferreira[2]

[1]Universidade Federal de Santa Maria (UFSM) / Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)
ftsautter@ufsm.br

[2]Universidade Federal de Santa Maria (UFSM)
isac_fferreira@hotmail.com

Palavras-chave: Silogística. Termos negativos. J. N. Keynes. Método diagramático. Extensão *versus* Intensão.

Keywords: Syllogistics. Negative terms. J. N. Keynes. Diagrammatic method. Extension *versus* Intension.

Resumo

John Neville Keynes utiliza um método diagramático, adaptado do método diagramático de Euler, no qual o conteúdo semântico de um juízo categórico é associado a um subconjunto próprio de um conjunto de diagramas básicos. Diferentes silogísticas caracterizam-se por distintos conjuntos de diagramas básicos. Comparamos, mediante o método diagramático keynesiano, três silogísticas (todas elas com pressuposição existencial e com pressuposição "não universal" dos termos) quanto à validade de inferências imediatas: a silogística sem termos negativos, a silogística com termos negativos em que um termo e seu correspondente termo negativo se complementam em relação ao universo do discurso, e a silogística com termos negativos em que um termo e seu correspondente termo negativo não necessariamente se complementam em relação ao universo do discurso.

Abstract

John Neville Keynes uses a diagrammatic method, adapted of Euler's diagrammatic method, in which the semantic content of a categorical judgment is associated to a proper subset of a set of basic diagrams. Different syllogistics are characterized by different sets of basic diagrams. We compare, by Keynesian diagrammatic method, three syllogistics (all of them with existential presupposition and with "non universal" presupposition of terms) as to validity of immediate inferences: the syllogistic without negative terms, the syllogistic with negative terms in which a term and its corresponding negative term complement each other in relation to the universe of discourse, and the syllogistic with negative terms in which a term and its corresponding negative term does not necessarily complement each other in relation to the universe of discourse.

[1] Este trabalho é parcialmente baseado na dissertação de mestrado "John Neville Keynes e a silogística com termos negativos" de autoria do segundo autor sob orientação do primeiro autor, defendida no Programa de Pós-Graduação em Filosofia da Universidade Federal de Santa Maria (UFSM).

Introdução

John Neville Keynes, pai do renomado economista[2], é autor de um manual de lógica popular no final do século XIX e início do século XX: *Studies and Exercises in Formal Logic, including a Generalization of Logical Processes in their Application to Complex Inferences*[3]. A primeira edição desse manual, de 1884, não utiliza termos negativos; esses estão presentes na quarta edição de 1906.

Ambas edições apresentam uma versão do método diagramático de Leonhard Euler[4] que, aplicado às inferências imediatas, consiste nas seguintes etapas: primeiro, a cada tipo de juízo categórico é associada uma coleção de diagramas básicos representando as combinações das extensões dos termos do juízo admitidas pela verdade do juízo. Essa coleção deve ser entendida como uma disjunção: ou a verdade do juízo é devida à situação representada por um diagrama básico de sua coleção, ou por outro diagrama básico de sua coleção, e assim por diante, até a consideração de todos os elementos da coleção associada ao juízo. A totalidade dos diagramas básicos de todos os quatro tipos de juízo categórico constitui o universo dos diagramas básicos. A coleção de diagramas básicos associada ao juízo é a informação semântica veiculada por ele.

Segundo, uma inferência imediata é válida se, e somente se, cada diagrama básico associado à premissa também é um diagrama básico associado à conclusão. No lugar da preservação de verdade, a validade é entendida, pelo método diagramático de Keynes, como não criatividade da informação semântica veiculada pelos juízos, ou seja, a informação semântica veiculada pela conclusão – o contido – também é veiculada pelas premissas – o continente.

Keynes concebe o termo negativo, do ponto de vista extensional, como sendo o complemento absoluto do seu correspondente termo positivo relativamente ao universo do discurso, assim, a negação de termos, ou negação terminística, obedece, para Keynes, ao Princípio do Terceiro Excluído. Quando consideramos um termo positivo e o seu correspondente negativo o universo do discurso fica dividido em, apenas, duas partições: a partição que contém os objetos denotados pelo termo positivo, e a partição que contém os objetos denotados pelo termo negativo. Em contrapartida, sob um ponto de vista intensional, Keynes entende que o termo negativo e seu correspondente positivo envolvem apenas um conceito, ou seja, envolvem a consideração de apenas uma conotação, a saber, a conotação constituída pelas notas características do conceito do termo positivo. Os conceitos sob os quais caem os objetos denotados pelo termo negativo são marcados pela ausência de uma ou mais notas que compõem a conotação do termo positivo correspondente. Neste sentido, a concepção de Keynes acerca dos termos negativos apresenta, por um lado, uma "diversidade" extensional: o termo negativo e seu correspondente positivo dividem os objetos do universo do discurso em dois conjuntos mutuamente excludentes entre si; e por outro, uma "unidade" intensional: o termo negativo e o positivo envolvem a consideração de um único conceito. (Keynes, 1906: 57–65)

[2] John Maynard Keynes, filho de John Neville Keynes, é considerado um dos maiores economistas do século XX e fundador da macroeconomia moderna.

[3] Em 1884, Keynes publicou a primeira edição desta sua obra que teve quatro edições: a segunda em 1887, a terceira em 1894 e a última em 1906. Para a última edição desta sua obra Keynes contou, inclusive, com a ajuda de seu filho John Maynard, evidentemente ainda jovem na época: "Em 1906, Maynard ajuda o pai a atualizar seu livro *Studies and Exercises in Formal Logic* (de 1884), quando se prepara sua quarta reedição." (Gazier, 2011: 39).

[4] O método diagramático criado por Euler tornou-se bastante conhecido na Lógica, contudo não é incomum encontrarmos nos manuais de Lógica posteriores a Euler uma exposição do seu método na qual os diagramas não correspondem às quatro figuras originais de Euler para a silogística. As quatro figuras originais de Euler encontram-se em quatro cartas escritas por ele à princesa alemã de Anhalt-Dessau, são elas as cartas cento e dois, cento e três, cento e quatro e cento e cinco (*Letter CII, Letter CIII, Letter CIV* e *Letter CV*), segundo a organização realizada por David Brewster, em uma edição inglesa, na obra *Letters of Euler: On Different Subjects In Natural Philosophy Addressed to a German Princess*, de 1833. A leitura daquilo que ficou conhecido como "Diagramas de Euler" se caracteriza pela apresentação extensional das proposições representando cada termo através de um círculo (figura fechada).

Uma das principais motivações para a introdução de termos negativos ao sistema lógico é a maior simetria entre as coleções de diagramas básicos associados aos juízos categóricos. Contudo, isso é uma motivação muito menos importante do que o possível ganho em poder expressivo e em poder inferencial. O argumento é simples: o aumento no número de diagramas básicos permite, em geral, uma descrição mais fina da realidade.

Nesse trabalho serão comparados três sistemas de silogismos assertóricos categóricos com pressuposição existencial dos termos envolvidos: um sistema sem termos negativos (Seção 1); um sistema com termos negativos complementares, ou seja, tal que um termo e seu correspondente termo negativo se complementam em relação ao universo do discurso (Seção 2); e um sistema com termos negativos não complementares, ou seja, tais que, em geral, um termo e seu correspondente termo negativo não se complementam em relação ao universo do discurso (Seção 3). Esse último sistema não foi explorado por Keynes, pois, como dissemos, para ele o termo negativo corresponde ao complemento extensional do termo positivo. A motivação para a investigação de sistemas com termos negativos não complementares é similar à motivação para a passagem de sistemas sem termos negativos para sistemas com termos negativos complementares: o aumento do número de diagramas básicos pode, em princípio, alterar a quantidade de inferências válidas. Keynes (1906: 59–61) discute brevemente os termos negativos não complementares. Ele esclarece que no discurso cotidiano empregamos termos negativos não complementares e, inclusive, dispomos de meios linguísticos para distinguí-los de termos negativos complementares; por exemplo, Keynes (1906: 61) sugere que expressar que uma mesa é não-moral é aceitável, mas é inaceitável expressar que uma mesa é imoral[5].

Antes de prosseguir, três observações são necessárias.

A determinação do conjunto de diagramas básicos de uma silogística é bastante simples, porque a negação judicativa utilizada respeita o Princípio do Terceiro Excluído. É suficiente determinar os diagramas básicos associados a um juízo e os diagramas básicos associados ao juízo oposto contraditório desse juízo[6]. Por comodidade, nas três silogísticas determina-se o conjunto dos diagramas básicos a partir do subconjunto próprio de diagramas básicos associados ao juízo universal afirmativo e do subconjunto próprio de diagramas básicos associados ao juízo particular negativo.

Simboliza-se “$\underline{X}$” ao termo negativo associado ao termo positivo simbolizado por “X”[7]. Para simplificar a discussão, “X” será referido como o par literal de “$\underline{X}$”, e esse será referido como o par literal daquele. Além disso, “U” é o universo do discurso.

A pressuposição existencial de um termo corresponde à pressuposição de não universalidade de seu par literal, ou seja, a denotação de seu par literal não é idêntica ao universo do discurso. Portanto, para fins de comparação de silogísticas com pressuposição existencial – uma delas sem termos negativos e duas delas com termos negativos – considera-se a pressuposição existencial e a pressuposição de não universalidade de cada termo.

[5] Neste exemplo, “não-moral” é um termo negativo complementar, enquanto que “imoral” é um termo negativo não-complementar.

[6] Uma verdade lógica é associada ao conjunto de todos os diagramas básicos, enquanto que uma falsidade lógica é associada ao conjunto vazio de diagramas básicos.

[7] A notação mais frequente para expressar o termo negativo associado a um termo positivo, empregada nas figuras constantes neste trabalho, consiste em acrescentar um traço *acima* do termo positivo. Por comodidade tipográfica, expressaremos o termo negativo, no corpo do trabalho, mediante o acréscimo de um traço *abaixo* do correspondente termo positivo.

1 Silogística sem termos negativos

A silogística sem termos negativos é caracterizada pelo conjunto dos cinco diagramas básicos dados nas duas figuras abaixo.

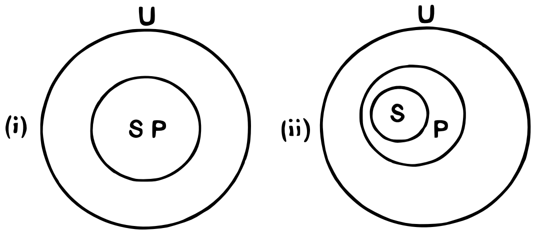


Figura 1: Diagramas básicos associados ao juízo universal afirmativo na silogística sem termos negativos (Keynes, 1906: 158)

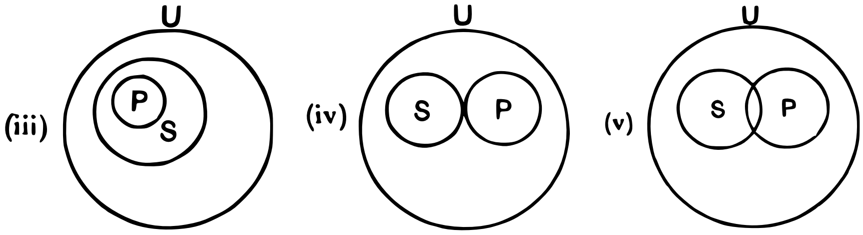


Figura 2: Diagramas básicos associados ao juízo particular negativo na silogística sem termos negativos (Keynes, 1906: 158)

O diagrama básico (iv) é o único associado ao juízo universal negativo e, portanto, os diagramas básicos (i), (ii), (iii), e (v) são associados ao juízo particular afirmativo (Keynes, 1906: 158).

A silogística sem termos negativos admite seis inferências imediatas não triviais válidas, ou seja, inferências imediatas válidas nas quais a premissa é distinta da conclusão. Estas inferências imediatas válidas para a silogística sem termos negativos estão sintetizadas no Quadro 1 abaixo:

Denominação da inferência imediata	Especificação da inferência imediata
Subalternação[8]	De "Todo S é P" infere-se "Algum S é P"
Subalternação	De "Nenhum S é P" infere-se "Algum S não é P"
Conversão[9]*per accidens*[10]	De "Todo S é P" infere-se "Algum P é S"
Conversão *simpliciter*[11]	De "Nenhum S é P" infere-se "Nenhum P é S"
Conversão *per accidens*	De "Nenhum S é P" infere-se "Algum P não é S"
Conversão *simpliciter*	De "Algum S é P" infere-se "Algum P é S"

Quadro 1: Inferências Imediatas válidas para a silogística sem termos negativos

[8] A subalternação consiste na manutenção dos termos sujeito e predicado em suas posições originais.

[9] A conversão consiste na permuta de posição do termo sujeito com o termo predicado. Portanto, na consideração dos diagramas básicos associados à conclusão, os termos sujeito e predicado devem ser permutados, ou seja, o diagrama básico (ii) é o diagrama básico (iii) do juízo converso, e vice-versa.

[10] "*per accidens*" indica mudança na quantidade do juízo.

[11] "*simpliciter*" indica a permanência da quantidade do juízo.

A subalternação do juízo universal afirmativo e a conversão *per accidens* do juízo universal afirmativo não são independentes dos demais, porque dispomos da conversão *simpliciter* do juízo particular afirmativo: a subalternação do juízo universal afirmativo resulta da aplicação da conversão *per accidens* do juízo universal afirmativo, seguida da aplicação da conversão *simpliciter* do juízo particular afirmativo; a conversão *per accidens* do juízo universal afirmativo resulta da aplicação da subalternação do juízo universal afirmativo, seguida da conversão *simpliciter* do juízo particular afirmativo.

2 Silogística com termos negativos complementares

A silogística com termos negativos complementares é caracterizada pelo conjunto dos sete diagramas básicos dados nas duas próximas figuras, onde o universo do discurso corresponde ao interior do círculo[12].

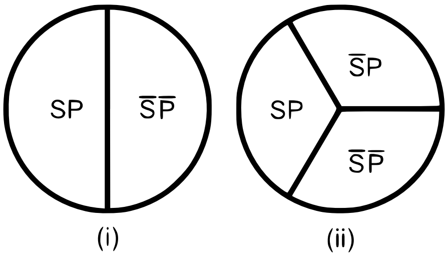


Figura 3: Diagramas básicos associados ao juízo universal afirmativo na silogística com termos negativos complementares

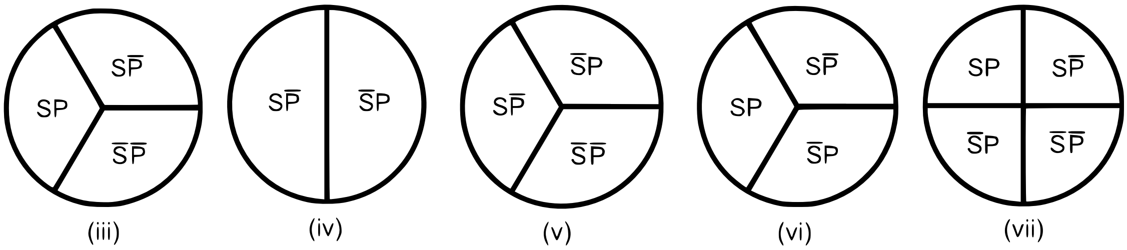


Figura 4: Diagramas básicos associados ao juízo particular negativo na silogística com termos negativos complementares

O diagrama (iv) da silogística sem termos negativos desmembra-se nos diagramas (iv) e (v) da silogística com termos negativos complementares, e o diagrama (v) da silogística sem termos negativos desmembra-se nos diagramas (vi) e (vii) da silogística com termos negativos complementares. Portanto, os diagramas básicos (iv) e (v) são associados ao juízo universal negativo e os diagramas básicos (i), (ii), (iii), (vi), e (vii) são associados ao juízo particular afirmativo[13]. (Keynes, 1906: 173–174)

A introdução de termos negativos multiplica as possibilidades de inferências imediatas não triviais válidas. Oito tipos de inferência imediata são admissíveis. Se "S" é o termo sujeito

[12] A representação keynesiana por círculos concêntricos em diversos diagramas básicos é enganosa (Keynes, 1906: 171–172), porque sugere uma relação de inclusão entre extensões. Utilizou-se, em seu lugar, uma representação por gráfico de pizza.

[13] Keynes destaca o aumento de simetria ocasionado pelo acréscimo de termos negativos. Sem termos negativos, os juízos têm as seguintes quantidades de diagramas básicos associados a eles: universal afirmativo são dois, universal negativo é um, particular afirmativo são quatro, e particular negativo são três. Com termos negativos, os juízos passam a ter as seguintes quantidades de diagramas básicos associados a eles: universais são dois, e particulares são cinco.

da premissa e “P” seu termo predicado, esses tipos são caracterizados do seguinte modo[14]: na subalternação (etiqueta “I_1”) o termo sujeito da conclusão é “S” e o seu termo predicado é “P”[15]; na obversão (“I_2”) o termo sujeito da conclusão é “S” e o seu termo predicado é “$\underline{P}$”; na inversão parcial[16] (“I_3”) o termo sujeito da conclusão é “$\underline{S}$” e o seu termo predicado é “$\underline{P}$”; na inversão total (“I_4”) o termo sujeito da conclusão é “$\underline{S}$” e o seu termo predicado é “$\underline{P}$”; na conversão (*simpliciter* ou *per accidens*) (“I_5”) o termo sujeito da conclusão é “P” e o seu termo predicado é “S”; na conversão obvertida (“I_6”) o termo sujeito da conclusão é “P” e o seu termo predicado é “$\underline{S}$”; na contraposição parcial (“I_7”) o termo sujeito da conclusão é “$\underline{P}$” e o seu termo predicado é “S”; e na contraposição total (“I_8”) o termo sujeito da conclusão é “$\underline{P}$” e o seu termo predicado é “$\underline{S}$”.

As inferências imediatas não triviais válidas na silogística com termos negativos complementares pode ser resumida no Quadro 2 abaixo, no qual: “A” indica um juízo universal afirmativo, “E” indica um juízo universal negativo, “I” indica um juízo particular afirmativo, “O” indica um juízo particular negativo, “I_n” são as etiquetas dos distintos tipos de inferências imediatas, e “Ø” indica a inexistência de inferência imediata do tipo indicado:

Premissa	A	E	I	O
I_1	I	O	Ø	Ø
I_2	E	A	O	I
I_3	O	I	Ø	Ø
I_4	I	O	Ø	Ø
I_5	I	E	I	Ø
I_6	O	A	O	Ø
I_7	E	I	Ø	I
I_8	A	O	Ø	A

Quadro 2: Inferências imediatas não triviais válidas na silogística com termos negativos complementados

3 Silogística com termos negativos não complementares

A silogística com termos negativos não complementares é caracterizada por um conjunto de quatorze diagramas básicos, desde que cada diagrama básico da silogística com termos negativos complementares é duplicado, originando dois diagramas básicos da silogística com termos negativos não complementares.

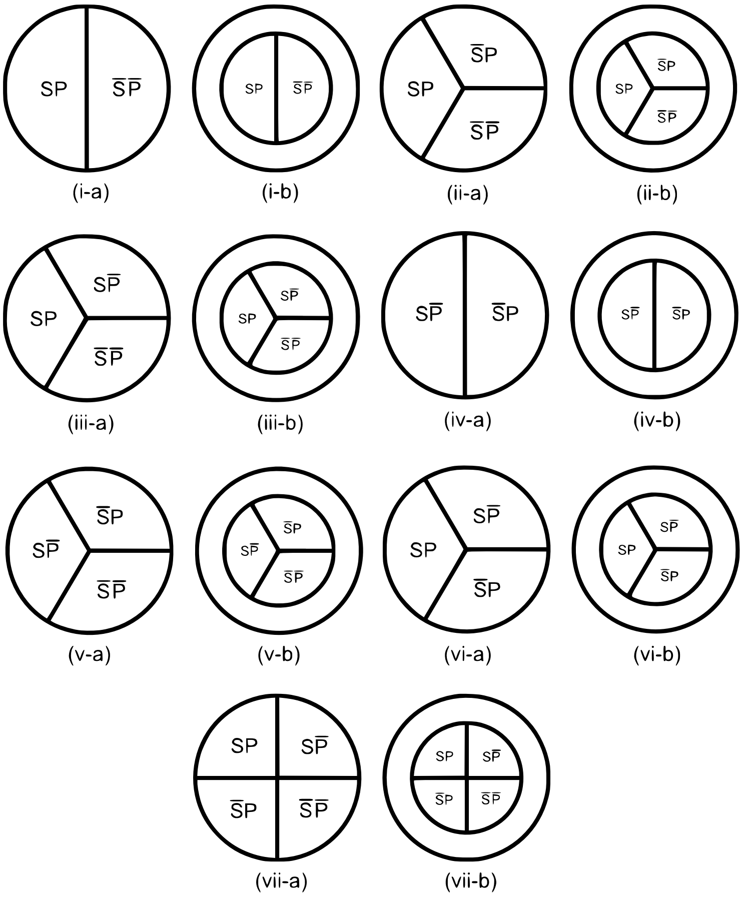
Cada diagrama da silogística com termos negativos complementares corresponde, nesta silogística com termos negativos não complementares exposta pela Figura 5 logo abaixo (onde o universo do discurso é representado pelo círculo mais externo), a dois diagramas. Um é idêntico a aquele, quer dizer, os diagramas assinalados pela letra “a”, ou seja, (i-a), (ii-a), (iii-a), (iv-a), (v-a), (vi-a) e (vii-a) são iguais a (i), (ii), (iii), (iv), (v), (vi) e (vii), respectivamente. E outro é

[14] Para facilitar a referenciação posterior, será fornecida uma etiqueta para cada tipo de inferência imediata.

[15] Keynes não a reconhece como uma inferência imediata, talvez porque o único papel que ela joga na determinação da validade dos modos seja na obtenção de modos mais fracos, por exemplo, de “BARBARA” se obtém “BARBARI”.

[16] Keynes (1906: 134–140) distingue dois tipos de inversão, a inversão parcial e a inversão total; do mesmo modo como distingue dois tipos de contraposição, a contraposição parcial e a total. Tanto na inversão quanto na contraposição o que determina que estas inferências imediatas sejam “totais” (inversão total e contraposição total) é a obtenção de um termo negativo na posição de predicado da proposição inferida.

semelhante a aquele mas possui uma área a mais, que correspondem aos diagramas marcados pela letra "b": (i-b), (ii-b), (iii-b), (iv-b), (v-b), (vi-b) e (vii-b).



(i-a) (i-b) (ii-a) (ii-b)

(iii-a) (iii-b) (iv-a) (iv-b)

(v-a) (v-b) (vi-a) (vi-b)

(vii-a) (vii-b)

Figura 5: Diagramas básicos da silogística com termos negativos não complementares

Essa duplicação implica a manutenção das mesmas inferências imediatas não triviais válidas obtidas na silogística com termos negativos complementares, nem mais, nem menos. A prova é simples: seja uma inferência imediata não trivial válida na silogística com termos negativos complementares. Se um "novo" diagrama básico estiver associado à premissa, o correspondente "antigo" diagrama básico[17] também estará associado à premissa. Por ser uma inferência válida, o diagrama básico "antigo" estará associado à conclusão. Mas, nesse caso, o "novo" diagrama básico também estará associado à conclusão.

A prova no caso de uma inferência imediata não trivial inválida com termos negativos não complementares é ainda mais simples, pois o mesmo diagrama básico "antigo" que serve de contraexemplo na silogística com termos negativos complementares, também serve de contraexemplo na silogística com termos negativos não complementares.

A diferença, se houver uma, entre a silogística com termos negativos complementares e a silogística com termos negativos não complementares reside nas inferências mediatas, nos modos válidos.

Considerações finais

A introdução de termos negativos produziu um ganho no poder expressivo e nas inferências imediatas válidas. Contudo, restrito à mesma linguagem, não houve acréscimo de novas infer-

[17] Por exemplo, (i-a) é um diagrama básico "antigo" e (i-b) é o seu correspondente diagrama básico "novo".

ências imediatas válidas. A mesma questão se coloca a respeito das inferências mediatas. Que há um acréscimo de inferência mediatas válidas não há dúvidas; a questão realmente interessante é saber se, restrito à mesma linguagem, há acréscimo de novos modos válidos? A silogística com termos negativos é uma extensão conservativa da silogística sem termos negativos? Curiosamente, no seu *Symbolic Logic* (Carroll, 1986: 246–247), Lewis Carroll demonstra a validade de uma inferência mediata na qual as premissas são expressas em uma linguagem sem termos negativos e a conclusão é expressa, necessariamente, numa linguagem com termos negativos[18].

O ganho ou perda de modos válidos também se coloca a respeito da passagem de uma silogística com termos negativos complementares para uma silogística com termos negativos não complementares. Um resultado de Luiz Carlos Pereira *et al.* (Pereira, 2008: 105–111) pode ajudar a decidir parte dessas questões. Segundo eles, os modos válidos da silogística aristotélica são intuicionisticamente válidos[19]. Provavelmente as duas silogísticas com termos negativos sejam extensões conservativas da silogística sem termos negativos, e os termos negativos operam, nesse caso, apenas como elementos ideais.

Finalmente, o tratamento heterogêneo aqui dispensado para a negação – a negação judicativa é clássica, a negação terminística é intuicionista, ao menos em uma silogística examinada – sugere a seguinte questão: o mesmo tratamento poderia ser produzido no interior da lógica contemporânea? Por exemplo, poder-se-ia investigar o comportamento de uma lógica quantificacional em que as negações aplicadas a fórmulas abertas não respeitam, em geral, o Princípio do Terceiro Excluído, mas as negações aplicadas a fórmulas fechadas o respeitam.

A silogística aristotélica, como se vê, está muito distante de ser um terreno estéril!

Referências

Carroll, L. (1986) *Symbolic Logic*, New York: Clarkson N. Potter.

Euler, L. IN Brewster, D. (ed.) (1833) *Letters of Euler: On Different Subjects in Natural Philosophy Addressed to a German Princess*, vol. 1, New York: J. & J. Harper.

Gazier, B. (2011) *John Maynard Keynes*, Tradução de Paulo Neves, Porto Alegre: L&PM.

Keynes, J. N. (1884) *Sudies and Exercises in Formal Logic, including a Generalization of Logical Processe in their Application to Complex Inferences*, London: Macmillan.

Keynes, J. N. (1906) *Sudies and Exercises in Formal Logic, including a Generalization of Logical Processe in their Application to Complex Inferences*, Fourth edition re-written and enlarged, London: Macmillan.

Pereira, L. C. et al. (2008) 'Alguns Resultados sobre Fragmentos com Negação da Lógica Clássica', *O que nos faz pensar*, 23, 105–111.

[18] As alegações e exemplos de Lewis Carroll devem ser admitidas *cum granu salis*, porque ele tem uma concepção heterodoxa acerca da pressuposição existencial dos juízos: no juízo universal negativo os termos não tem pressuposição existencial, nos demais têm.

[19] Esse trabalho prova que diversos fragmentos da lógica quantificacional clássica são intuicionistas, inclusive um fragmento capaz de acomodar a silogística aristotélica.