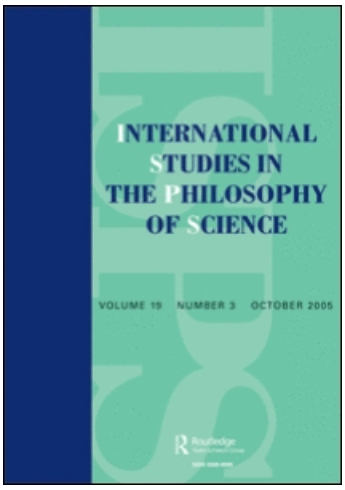

International Studies in the Philosophy of Science
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713427740

Kitcher on reference

Stathis Psillos [a]
[a] Department of Philosophy, London School of Economics, London, UK

Online Publication Date: 01 October 1997

To cite this Article Psillos, Stathis(1997)'Kitcher on reference',International Studies in the Philosophy of Science,11:3,259 — 272
To link to this Article: DOI: 10.1080/02698599708573570
URL: http://dx.doi.org/10.1080/02698599708573570

ARTICLE

Kitcher on reference

STATHIS PSILLOS
Department of Philosophy, London School of Economics, UK

1. Introduction

In his (1978) and parts of (1993), Philip Kitcher advances a new context-sensitive theory of reference which he applies to abandoned theoretical expression-types, such as Joseph Priestley's "dephlogisticated air", in order to show that, although *qua* types they fail to refer uniformly, they nonetheless have referential *tokens*. This piece offers a critical examination of Kitcher's theory. After a general investigation into the overall adequacy of Kitcher's theory as a general account of reference, I focus on the case of abandoned theoretical terms. Kitcher's theory is meant to be able to evaluate and solve disputes about referential continuity and progress in scientific theory-change. To this end, Kitcher employs the principle of humanity and a notion of the "correct historical explanation" of the production of each expression-token. I argue that the application of the principle of humanity does not offer a principled way to show that the historical actors were involved in different modes of reference when they produced different tokens of an expression-type. I also suggest that the principle of humanity, coupled with Kitcher's view that tokens of expression-types may systematically refer to different things, make conceptual progress too easy and thus uninteresting.

2. Context-sensitivity and reference potential

Leaving indexicals aside, the two standard theories of reference—descriptive and causal—treat linguistic reference (or denotation) as a two-place relation between a word and an object. The word's reference constitutes a semantic property of the type in virtue of which the type can then be used in composite expressions to specify their truth-conditions. The prime difference between the two standard theories is that according to descriptive theories, the relation between a word and its referent is mediated by the sense of the word, whereas according to the causal theories à la Kripke (1980) and Putnam (1975b, 1975c) this relation is direct (causal-historical), unmediated by a concept. While on the descriptive theories, an expression acquires its reference (if any) via its sense, the causal theories dispense with senses as reference-fixing devices and suggest that the reference of a word is whatever entity "grounded" the word in a certain

0269-8595/97/030259-14

"dubbing ceremony" or "baptism" in which the word was introduced. The idea of a dubbing ceremony, or of an explicit act of baptism, is clearly an idealisation for words other than proper names. But, according to the causal theory, proper names provide a model for fixing the reference of any other word: what makes it the case that a word refers to a certain entity is the existence of a causal act (a naming ceremony) during which the word was attached to a substance, or a kind, or a physical magnitude where samples of this substance, instances of this kind and paradigmatic effects of this magnitude were present and grounded the word. This model suggests that reference-fixing can be, at least *prima facie*, dissociated from devices such as definite descriptions, analytic definitions, property clusters, necessary and sufficient conditions for kind-membership, etc. Such devices are paradigmatic of the ways reference is fixed according to the description theories of reference. The thrust of these theories is that reference should be understood as best fit: the referent of a word is the entity, if any, that satisfies a set of descriptions associated with this word.

The pros and cons of the two standard theories have been heatedly debated over the last two decades.[1] It is not my intention essentially to contribute to this debate, apart from some general comments and suggestions I shall make in Section 5. What I want to focus on is the fact that on both standard theories, truth and reference are semantic properties of expression-types. The tokens of expressions-types do not have, as it were, autonomous semantic properties, but acquire theirs via their types.

Isn't it, however, a well-known fact that language-users may use tokens of certain expression-types with an intention to refer to things other than that denoted by the type? Or even, that they use a token of an expression-type that does not denote anything to refer to something? This aspect of language-use has been taken to be in the province of pragmatics and not of semantics. In pragmatics—as opposed to semantics—the operative notion is *speaker reference*: the use of an expression to refer one's audience to certain individuals in line with one's relevant intentions. Speaker reference is context-sensitive and belongs to pragmatics, whereas linguistic reference has been seen as context-insensitive and belongs to semantics.[2]

If Kitcher's aim was just to stress that speakers *occasionally* use an expression-type indexically in order to direct their audience to a particular present object (or kind of object, an instance of which is present), there would be no cause of dispute. In such cases—akin to Donellan's (1966) referential uses of definite descriptions—a term-token does refer to the object "at hand" because it's being used to refer to it. What appears to be a semantic property of the token, is just a pragmatic property of the act of production (cf. Bach 1987, p. 86). This sound observation about the pragmatics of language-use would hardly require the advancement of a whole theory of reference such that tokens of the same type may *systematically* refer to different things, depending on the context and the intentions of the utterer. Nor would it require the development of a theory such that referential semantics is applied to tokens rather than to types.

I think Kitcher's aim is more ambitious: it is to show that a context-sensitive theory offers an adequate account of linguistic reference, one yielding the right results when applied to some paradigmatic cases as well as grounding conceptual progress in scientific theory-change. On Kitcher's account, expression-types are no longer associated with single (putative) referents. Instead, each expression-type is endowed with a *reference potential*: a potential such that its tokens may refer to more than one (putative) entity, depending on the event that has initiated the production of each particular token (cf. 1978, p. 145). Such an initiating event is, typically, "either an event in which the referent of the token is causally involved, or an event which involves the singling out, by

description, of the referent of the token" (op.cit., p. 143). In his book (1993, p. 76), Kitcher ties the reference potential of a type to the Fregean notion of "modes of reference", where a mode of reference is "what makes it the case that the token refers to that object". In effect, the potential of an expression-type to refer is a function of the two canonical ways in which a term can refer to anything at all. Since the reference of a word can be specified either by means of a description or in terms of the entity that "grounds" the production of this word, Kitcher proposes two basic modes of reference: a "descriptive mode" and a "baptismal mode".[3] Without over-simplifying the situation, I think that in the final analysis Kitcher associates with each expression-type a two-member reference potential {**d**,**c**} such that the first member **d** picks out its referent by means of a description (or, sets of descriptions), while the second member **c** picks out the causal agent in the presence of which the term was introduced.[4] It may turn out that not only there exists an entity that satisfies **d**, but also that it is identical with the entity dubbed in **c**. In such a case, the type **t** refers uniformly. Suppose, however, that things turn out differently: there is nothing that fits **d** while there exists a causal agent (individual, stuff, substance, magnitude, etc.) dubbed in **c**. Then, instead of having the expression-type **t** uniformly non-referring to anything or uniformly referring to the entity dubbed in **c**, we can have some tokens of **t** refer to something while others do not.

3. Semantic ambiguity

Kitcher motivates his proposal telling the story of the millionairess Eustacia Evergreen who decides to withdraw from public life and leave in her place a "double" to take up her duties. It seems reasonable, he suggests, that the reference of the expression-type "Eustacia Evergreen" must be dealt with at the token-level: some tokens of this type refer to the real millionairess, while some others to the impostor. For instance, when a friend of Eustacia's explains to a bunch of neighbours how immensely wealthy she is, "Eustacia Evergreen" refers to the real millionairess—who owns the wealth. But when she tells them that Eustacia, the woman next door, has invited her for a drink, "Eustacia Evergreen" refers to the impostor—who issued the invitation (cf. 1978, pp. 134–135). Each token of the type "Eustacia Evergreen" refers to one member of the set {millionairess, impostor} each of which features appropriately in the explanation of the production of tokens (cf. op.cit., p. 135).

It is certainly true that some expressions are semantically ambiguous. In point of fact, most proper names have more than one bearer, and therefore same name-tokens may refer to different individuals (e.g. as Devitt has pointed out, name tokens of "Liebnecht" may refer to either the father-Liebnecht or the son-Liebnecht). So, there is the problem of specifying the semantic-type to which a certain name-token belongs. This problem can been dealt with adequately by some token-level causal theories [e.g. Devitt's (1981)]. The thrust of the solution is that classification of tokens to semantic types is a function of what grounds the speaker's ability to produce the token (cf. Devitt & Sterelny, 1987, p. 59). Generally, should there be more than one bearers of a certain name, one should exploit the context in which a certain token is used in order to remove the ambiguity. All this is as it should be. However, Kitcher's example seems to point to the opposite direction. For all we know, there is an apparently uniformly referring name: "Eustacia Evergreen" is the name of the millionairess.[5] When faced with the existence of an impostor, the problem arises how we can classify the tokens of "Eustacia Evergreen" into those that refer to **a** (the real thing) and those that refer to **b** (the impostor). It seems as though context is invoked to determine the semantic properties of tokens of an apparently uniformly referring expression.

I am not sure whether Kitcher wants us to think that some expressions are semantically ambiguous or that all are. The former thesis is certainly sensible (and true), while the latter is much more contentious (and probably false). My guess is the following: Kitcher suggests that we may treat all expressions as semantically ambiguous, until proven otherwise. The evidence for this guess is that on Kitcher's proposal, expression-types are *typically* treated as being endowed with a reference potential. If so, they become semantically ambiguous: tokens of one and the same (apparently uniformly referring) type may systematically have different semantic properties such that grasping and/or determining the semantic properties of the type is never enough to determine those of the tokens. Some extra information is *always* required, viz. which element of the reference potential is actualised on each occasion. On this proposal, when we hear or read an expression-token that contains, say, "Eustacia Evergreen", in order to grasp its sense we must be able to find out which element of the reference potential is presently actualised. Similarly, truth-conditions are no longer associated with expression-types. In order to be able to judge whether an expression-token, e.g. of "Eustacia Evergreen owns a cat", is true or false, we need to be able to identify the truth-conditions of the *particular* token, that being a function of the element of the reference potential being activated. Any single judgement about every single expression-token can go all four ways (true about the millionairess, false about the millionairess, true about the impostor, false about the impostor) depending on what element of the reference potential is being actualised.[6]

If there were factors that could settle the issue of what semantic properties each expression-token has, then the semantic ambiguity of the associated type might not be a problem, after all. The required factors should be such that they pick the mode of reference which is being employed on each occasion of the production of a token, or equivalently, such that they specify what event initiated the production of each token. What we should be after is ways to make the right identifications. Without them, there is no point in saying that we have specified the semantic values of each expression-token of a type.

Kitcher suggests that we look for the event or mode of reference that features appropriately in the explanation of the production of each token. But, notice the following problems. First, it is perfectly possible that we misidentify the initiating event/mode of reference that purports to explain the production of a certain token. Although we might think that the initiating event that explains the production of a token of "Eustacia Evergreen" is linked with the impostor, it might be linked with the millionaires. For instance, when one talks about Eustacia, the woman next door, having invited some close friends to dinner, it might appear that the initiating event picks out the impostor, but it might well be the case that the real millionaires re-appeared for a night in order to have the pleasure to treat some special friends to dinner. Second, it is perfectly possible that there is no way to determine the event(s) which features appropriately in the explanation of the production of each token. For instance, when one talks about Eustacia being a very considerate person, it is likely that it is indeterminate whether the initiating event picks out the millionairess or the impostor, or both. If their reference is settled at the token-level, all standing statements and generalisations are more likely to be indeterminate. In virtue of what does the token of "Eustacia Evergreen" in "Eustacia Evergreen is ugly" or in "All of Eustacia's friends are red-haired" refer either to the millionaires or to the impostor? More generally, it is perfectly common that speakers produce certain expression-tokens which are not initiated by any event in particular, but have their roots in many disparate ones. No factor can

differentially link us with either the millionairess or the impostor when tokens of "Eustacia Evergreen" are employed in expression-tokens such as "Eustacia's cat is so fat", or "Sometimes, Eustacia is a pain in the neck", or "Eustacia is not herself lately", simply because there is no way to associate the occurrence of "Eustacia Evergreen" in such expressions with a particular initiating event/mode of reference. Third, suppose that one uses two different tokens of the same type in the same breath, and says: "Eustacia is a millionairess and Eustacia is not a millionairess". If we treat reference at the token-level, this is no longer an outright contradiction. Depending on the context, it might even be a true statement.

Well, one might say, if "Eustacia Evergreen" is an ambiguous name, then it is to be expected that apparently contradictory statements as the one above are not contradictory. What's the big deal? The big deal is that there is a litmus test for semantic ambiguity to which my last point leads. Whether or not a name is used ambiguously by a community of speakers, or by an individual, shows up in the inferences that they are willing to make. If they endorse "A is an F" and "A is a G", then if they are also willing to infer "A is F and G", then "A" is an unambiguous name for them. The reason is simple: if it wasn't, they would recognise that "A is F and G" might well be false, although all the premises (i.e. "A is F" and "A is G") are true. This point has been recently stressed by Fodor in a different context (1994, pp. 68–70). For our purposes, the thrust is that semantic ambiguity, or its lack, shows up in one's inferential practices. So, if one uses the name "Eustacia Evergreen" unambiguously, one would not be willing to infer "Eustacia is a millionairess and Eustacia is not a millionairess" from "Eustacia is a millionairess" and "Eustacia is not a millionairess". Conversely, if one was willing to perform the later inference, one would recognise that the name "Eustacia Evergreen" is ambiguous. More generally, if one was willing to infer "Eustacia is F and G" from "Eustacia is F" and "Eustacia is G", then "Eustacia Evergreen" would refer uniformly and would not have a reference potential. This test will prove crucial when we discuss Priestley's case in the next section. The only point I want to stress here is that, after all, we seem to have a way to test whether a name is used in an ambiguous way or not, which makes an empirical issue to decide whether it is so used or not.

To sum up the argument so far, if each word is associated with a reference potential á la Kitcher, it seems there is no way to systematically determine the event or mode of reference that features appropriately in the explanation of the production of each token of the word. In other words, there is no way to make sure that the explanation of the production of each token is such that each token uniquely designates one rather than the other individual. I think Kitcher is aware of this problem. For at least he admits that sorting out the reference of each token of an expression-type would require an omniscient observer, who "would be able to spell out the details of the explanation" (of the production of each token) (1978, p. 135). But he does insist that even when we are not as lucky as his envisaged omniscient observer, we might still be able to specify a set of entities such that each token of an expression-type refers to a member of this set, regardless of whether we are able to decide, on every and each occasion, which member is the referent (cf. 1978, p. 135).

At this juncture, it might seem pertinent to connect Kitcher's proposal with Hartry Field's (1973) idea that reference might well be indeterminate. According to this proposal, there is no fact of the matter as to what a certain term denotes (Field discusses the term "mass" in great detail). There are two or more entities (magnitudes in the case of "mass") that the term might be reasonably taken to refer to and there is no way to assert which of them is the denotation. So, one might think that Kitcher's reference

potential is a way to improve on Field's suggestion. Each term is associated with a reference potential, and although we may not be able to decide which member of the potential is actualised on each occasion where a token of the term is used, we can still assert that the term is not denotationless: its tokens do denote some or other member of the reference potential. In fact, Kitcher could rely more on Field's proposal and argue that although a term endowed with a reference potential does not fully denote any of the members of the reference potential, it *partially denotes* more than one thing (cf. Field, 1973, pp. 474–475). Referential ambiguity is not the same as referential indeterminacy,[7] but recognising that words might be indeterminate rather than ambiguous seems, *prima facie*, to be progress. For now we will not have to decide what each token refers to, while we will acknowledge that the type partly denotes more than one entity.

A full discussion of Field's proposal and its connection with Kitcher's would require a separate paper. But the following is worth stressing. We cannot just relax with the thought that reference might be indeterminate. Kitcher rightly wants to advance a theory of reference which can be employed to evaluate and solve disputes about referential continuity and progress in scientific theory-change. If we cannot decide whether past expression-tokens denoted this or that entity, how can we even discuss whether there is referential continuity in theory-change? To press another point made by Field, an adequate theory of reference must be able to show how "denotational refinement" is possible, i.e. how the terms used in a successor scientific theory might referentially overlap with terms of the superseded theory. It transpires that in order to decide referential continuity and refinement we need to establish what past terms referred to and whether what they referred to can be plausibly seen as the referent of terms that feature in successor theories. The decision-problem is here to stay, and in order to address it and solve it Kitcher relies on the heavy machinery of the principle of humanity. To this issue we shall now turn our full attention.

4. Tokens of theoretical terms

A central element of Kitcher's proposal is that "the reference of a token of an expression is the entity which figures in the appropriate way in the correct historical explanation of the production of that token" (1978, p. 133). Kitcher's theory is meant to offer an objective account: the referent of each token is specified by the "correct historical explanation" of its production. The idea is that insofar as we can specify the "correct historical explanation" of the production of each token, we can determine its semantic values, and hence we can decide whether its utterer said something true or false.

To find the "correct historical explanation" of the production of a certain token of a theoretical term, and given that there are no "omniscient observers", Kitcher appeals to the "principle of humanity" according to which

> we attribute beliefs by supposing that our interlocutors have cognitive equipment that is similar to our own, and using what we know about the experiences they have had (1993, p. 101).

Let's see how this works in practice. Kitcher applies his proposal to Priestley's "dephlogisticated air". The correct historical explanation of the production of the token should be such that it "enable(s) us to trace familiar connections among Priestley's beliefs and between his beliefs and entities in the world, ascribing intentions that we would expect someone in his situation to have" (ibid.). The principle of humanity is central because, using what *we* know about the situation Priestley was in, we can reconstruct the situation and attribute to Priestley different *dominant intentions* on

different occasions of his uttering tokens of "dephlogisticated air". More specifically, the principle of humanity gives us the opportunity to single out some occasions on which our subject's behaviour can be best explained, from our own vantage point, by attributing to him dominant intentions to refer to entities we now posit. Take, for instance, Priestley on the occasions in which he employed the phlogiston theory in order to characterise dephlogisticated air. For Kitcher, the correct historical explanation of Priestley's using "dephlogisticated air" on *those* occasions should attribute to him dominant intentions to refer to whatever is left over when phlogiston gets absorbed by the calx of mercury. The mode of reference in those cases, Kitcher says, is fixed by the description "the substance obtained when the substance emitted in combustion is removed from the air" (1993, p. 102). Since there is no such substance, such tokens of "dephlogisticated air" are denotationless. But now, let us take Priestley on some other occasions where, immersed in his laboratory experiments, he isolated a gas which he characterised as dephlogisticated air, and of which he said that it supports combustion better than common air. Then, the correct historical explanation of his productions of tokens of "dephlogisticated air" should be that his dominant intention was to refer to whatever it is that supports combustion, i.e. oxygen. The mode of reference in those cases is fixed by Priestley's dominant intention "to refer to the kind of stuff that was isolated in the experiments (...)" (ibid.) "In both cases", Kitcher said, "our ascriptions of reference are guided by the principle of humanity" (1978, p. 142).

There is something intuitively appealing in this story. For it is certainly true that some tokens of "dephlogisticated air" were uttered in the presence of (and, arguably, because of) oxygen. Still, saying that Priestley was involved in different modes of reference on different occasions of him using the expression "dephlogisticated air" is, if anything, an idealisation. Barring extreme cases, one's intentions to refer are so interwoven that they cannot be naturally broken up into two components, in particular into intentions to refer to a certain object—no matter what this object turns out to be—and intentions to refer to whatever satisfies a certain (possibly theoretical) description. Under normal circumstances, one's mode of reference is a function of both one's intentions to refer to an object (or, to a kind of object, an instance of which is present) *and* one's intentions to refer to *this* object as exemplifying one or more descriptions. Should we rely on the Principle of Humanity—the driving force behind the suggested idealisation—to separate those intentions? In what follows, I shall use Priestley's case to show that the principle of humanity does not offer a principled way to establish systematic correlations between Priestley's use of some tokens of "dephlogisticated air" and his intentions to refer to the stuff he had isolated which are not at the same time intentions to refer to this stuff as phlogiston-free air. The root of the problem is this. The principle of humanity establishes an incoherence between the subject's beliefs and intentions (that is, an incoherence in the subject's *own* perception of the situation he was in) in order to maximise coherence in our judgements of what our subject was doing in light of our knowledge of the situation he was in.

Back in the early 1770s, Priestley did isolate a certain "air", of which he asserted that it was "six times as good as common air", that it supported combustion better than ordinary air, etc. From the perspective of presently accepted theories, there is no doubt that Priestley managed to isolate pure oxygen. In fact, he mentioned to Lavoisier over dinner in Paris in 1774 that he had got an air from *Mercurius Calcinatus* which was "six times as good as common air". Initially, Priestley wasn't able to distinguish the gas that was produced when calces were heated (oxygen) from the gas that was produced when nitrous was heated (nitrogen dioxide), and he thought they were the same "air". He

called them both "dephlogisticated air", for, having accepted the phlogiston theory, he believed that the new "air" was the "air" obtained when phlogiston ("the substance emitted in combustion") was fully removed from common (atmospheric) air.[8] But even when he distinguished nitrous air from dephlogisticated air, he still held to the view that the latter is phlogiston-free ordinary air. Priestley thought that dephlogisticated air and phlogiston were, along with the inflammable air, the only basic substances. He asserted, for instance, that "All the kinds of air with which we're acquainted, except the D^d and inflammable, are all composed of d^d and p^n" (1966, p. 274). Although the gas he called "dephlogisticated air" was identified as oxygen across the channel and in Scotland, Priestley never endorsed the new oxygen-based chemical nomenclature, even long after it became broadly accepted. Nor did he consider that his dephlogisticated air might be the same stuff as oxygen. The issue was far from terminological. Priestley vehemently opposed the idea that water was a compound made up of dephlogisticated air and inflammable air, whereas the *oxygenistes* suggested that water consists of oxygen and hydrogen. "But then I ask", he said once, "where is the oxygen which, according to them [the Antiphlogistonians] constitutes the far greater part of the water? I cannot find it anywhere" (1966, pp. 293–294). Till the end, he insisted that "the doctrine of phlogiston stands firm" (1966, p. 255). Without further going into detail here, there should be no doubt that Priestley thought that the stuff he first isolated was the stuff that is produced when phlogiston is totally removed from ordinary air. Dephlogisticated air, together with phlogiston, formed the basis of all of his analysis of the different kinds of air.[9]

In order to systematically map some tokens of "dephlogisticated air" to oxygen and others to the empty set, the principle of humanity must tamper with this story considerably. It's only natural to think that Priestley's intentions were dominated by his belief that the air-obtained-when-the-substance-emitted-in-combustion-is-removed-from-the-air is identical to the air-that-supports-combustion-better-than-common-air. Although Priestley certainly intended to speak of a certain stuff present in his laboratory, he equally intended to refer to it as the stuff that satisfied a certain theoretical description, i.e. as the stuff that possessed certain theoretical properties. No doubt, some of his beliefs were caused by the presence of oxygen. But this is not sufficient for saying that the intended referent of Priestley's use of "dephlogisticated air" was oxygen. Here, Evans' (1973, p. 207) example might be helpful: my belief that my colleague's partner has nice legs might have been caused by looking at someone else's legs, yet my belief is about my colleague's partner's legs. So, someone's (e.g. Priestley's) beliefs might be dominantly of an item (e.g. phlogiston-free air), even though some of these beliefs have been (partly) caused by another item (e.g. oxygen). Priestley wanted his hearer/reader to understand that he had released a new air which had certain detectable (observable) qualities which made it markedly different from common air. He was so proud of his *discovery* that he went on telling people of the remarkable qualities of the new air. However, believing in the phlogiston theory—which predicted the existence of such an air—he also intended to make known, roughly right from the start, that the *dephlogisticated air* predicted by the phlogiston theory is nothing but the new air he had discovered. In other words, he wanted his audience to think of the new air not just qualitatively, but as an essential element of the phlogiston theory. That's why he chose the expression "dephlogisticated air". And that's why he was unable to make any contact with the *oxygenistes* across the channel.

At any rate, as we saw in the last section, there is a handy test for semantic ambiguity: the conjunction test. Priestley was clearly ready to assert both that

"Dephlogisticated air eases respiration and combustion" and that "Dephlogisticated air is produced when phlogiston is totally removed from ordinary air". As far as the historical evidence goes, he was also ready to infer from these two premisses that "Dephlogisticated air eases respiration and combustion *and* is produced when phlogiston is totally removed from ordinary air". Hence, "dephlogisticated air" was not an ambiguous expression for him: both tokens in the premisses of the above argument refer to the same thing. To consolidate the point, let's take another example. One can easily imagine Priestley—being a committed *contre-oxygeniste*—moving from the premisses "Dephlogisticated air exists" and "Oxygen does not exist" to the conclusion "Dephlogisticated air exists but oxygen does not". But if the token of "dephlogisticated air" in the first premise denoted oxygen, then Priestley would be seen as being willing to endorse a contradiction. If we were to follow the principle of humanity, we would render some tokens referential, but at the price of making Priestley being inconsistent.

I think the last point suggests that we shouldn't rely on the principle of humanity in order to better understand "the judgements and inferences" that Priestley made. This claim to better understanding has been crucial to Kitcher's motivation. He says:

> (O)ur construction aims to make comprehensible the judgements and inference of our subject. To say that an entity "figures in the appropriate way" in the explanation of the production of a token is, I suggest, to claim that the hypothesis that the token was initiated by an event in which the entity was causally involved or singled out by description best explains why our subject makes the assertions and arguments he does (1978, p. 143).

More generally, one should note that attributing different dominant intentions to refer on different occasions makes no difference to the explanation of Priestley's judgements, arguments and assertions. Priestley would make (and in fact did make) the same judgements and assertions about the stuff he isolated regardless of whether he characterised it via theoretical descriptions or by its detectable qualities. What matters is that a premise constant in Priestley's arguments and assertions was that the stuff he isolated and was endowed with certain detectable qualities also answered to a certain theoretical description. In different contexts, he might as well have stressed different qualities of the stuff. But there is no reason to think that there were contexts in which he suspended the foregoing premise. From our own vantage point, all we need in order to understand why Priestley said what he did—as well as, and in order to state about what things he was right and wrong—is the *principle of minimising the attribution of inexplicable error*, as Gareth Evans once put it (1973, p. 196). We minimise inexplicable error by acknowledging that the causal origin of some of Priestley's beliefs about dephlogisticated air was oxygen, while accepting that he was fully immersed in the phlogiston theory. We can still understand and explain Priestley's assertions that dephlogisticated air supports combustion better than ordinary air and that dephlogisticated air is phlogiston-free air, even if we admit that both of them are false (as, I think, we should). For Priestley, the former is simply entailed by the latter: the removal of phlogiston from air leaves the air with greater capacity for absorbing phlogiston, resulting in a better combustion.

One might suggest the following thought-experiment in defence of the principle of humanity. While Priestley released the new "air" and said "Dephlogisticated air supports combustion better than common air", someone went to his lab and asked him: "Joseph, my old mate, suppose it were conclusively demonstrated that there really is no such thing as phlogiston, and thus that there is really no stuff as dephlogisticated air. If

that were to happen, would you be happy to paraphrase what you just said as: 'this stuff supports combustion better than common air'?". Since it seems natural to assume that Priestley would accept the paraphrase, one might think that the application of the principle of humanity is thereby justified because it does, after all, help us specify correctly Priestley's dominant mode of reference.

But such defence of the principle of humanity is inadequate. There is no need to ask counter-factual questions in order to see that on certain occasions Priestley used expressions that did refer to oxygen. Such occasions do not involve him using the term "dephlogisticated air", but rather him saying the likes of "*This* air (or this stuff) makes me feel so light". On these occasions, it is the "this air" or "this stuff" which refer to oxygen and not any token of "dephlogisticated air". The shift from "this air" or "this stuff" to "dephlogisticated air" is crucial. The fact that the former may refer does not entail that the latter refers too. At any rate, even if one conceded that occasional tokens of "dephlogisticated air" did refer to oxygen, it would be plausible to say that they did so accidentally: there was no pattern in Priestley's use of "dephlogisticated air" which would make some tokens refer systematically to oxygen.

Another possible objection here might be that I have assumed an overly strong connection between a speaker's beliefs and a speaker's intentions. Suppose, one might argue, that Priestley believed that the stuff he had isolated satisfies two descriptions, say D and D′ and that he also believed that the stuff-that-supports-combustion-better-than-air (D) is identical to the stuff-obtained-when-the-substance-emitted-in-combustion-is-removed-from-the-air (D′). Isn't it perfectly possible that he might have used "dephlogisticated air" with the dominant intention to refer to whatever satisfies D and not with the dominant intention to refer to whatever also satisfies D′?

Certainly, an expression-type may refer even though certain descriptions associated with it are false. This point is sound and has been brought home by the causal theories of reference. In particular, it is consistent with the view (akin to it Hardin's and Rosenberg's (1982) causal-role theory of reference of theoretical terms) that the *type* "dephlogisticated air" refers to oxygen because the latter grounded the term in a certain dubbing ceremony, and/or because oxygen has some of the sensible qualities attributed to dephlogisticated air. I take it, however, that the main thrust of the objection at hand is different. It is that a speaker may use an expression with the dominant intention to refer to whatever satisfies D and not also to whatever satisfied D′, *in spite of* his belief that the object that satisfies D is identical with the object that satisfies D′ (I abbreviate this by stating D = D′). It is precisely this move that has been made available by the principle of humanity.

Referring expressions are, normally, embedded in a network of descriptions. Speakers may occasionally single out one description, say D, as predominant and use it to identify (or mark) an object. But if a speaker believes that D = D′, then no matter which description the speaker uses, he intends to convey information about one and the same object; an object which, the speaker believes, can be identified in different ways. To see this, imagine that Priestley said to his audience he released a stuff that supported combustion better than common air and that then, he was asked whether this stuff was the same as the stuff that Lavoisier advertised as oxygen. Priestley's expected answer here would be: "Of course not. The stuff I released is obtained when phlogiston is removed from the air". How can we describe Priestley's dominant intention and make sense of it without his beliefs? I think the best way to put the point is that, *because of* his beliefs, he intended to convey information about an object which he wanted his audience to be able to identify in two different ways.

I might sum up the argument so far by saying that the principle of humanity is too strong: it attributes differential dominant intentions where there really are none to be found. But there is also a sense in which it is too weak. In conjunction with Kitcher's claim that each term is endowed with a reference potential, the principle of humanity makes referential continuity too easily available: arguably, all past abandoned expression-types end up having referential tokens. Hence, conceptual progress becomes trivial: no abandoned concept has failed to characterise some natural kind we now posit. For no matter how radical conceptual changes occur, that is no matter whether a concept and its associated descriptions radically change, all abandoned scientific concepts may end up being referential. The problem is this. It seems that, given the flexibility of the principle of humanity, there can be found explanations of why past scientists said the things they did such that *some* tokens of the expressions-types they used refer to entities we now posit. Part, at least, of the rationale for the judgement that Priestley referred to oxygen when he uttered some tokens of "dephlogisticated air" was that on occasions he intended to refer to the stuff he had isolated, and this stuff was oxygen. But one can, too easily, extend this line of thought to all other abandoned expression-types posited in the history of science. For instance, some tokens of "caloric" may be said to refer to internal energy when they were uttered by some caloricists because the latter intended to refer to whatever causes the rise and fall in temperature, and these are due to changes in the internal energy of a gas. Or, some tokens of Aristotle's "seeking its natural place" may be seen as referring to motion-along-a-geodesic since the principle of humanity would not stop us from arguing that, on some occasions, Aristotle's dominant intention was to refer to whatever causes some bodies to follow a certain trajectory; and similarly for tokens of "witch", "sanguine type", "absolute space", etc.

Kitcher might try to avoid too heavy reliance on the principle of humanity and, instead, argue for a different motivation for his view: expression-types are associated with a reference potential because some tokens get their reference in a theory-driven way, while some others get theirs in an experiment-driven way. So, he could say that theory-driven tokens and experiment-driven ones have different mechanisms of reference, e.g. theory-driven tokens are associated with theoretical descriptions, whereas experiment-driven ones relate to the manipulation of causal agents. This may well be a promising line and needs to be further pursued. It may well hold the key to a good account of reference-fixing. But one should be careful here. Even when the mechanism of reference involves experiments rather than theory, experiment-driven tokens of scientific expressions aren't theory-free. The experiment *may* isolate the right causal agent of the phenomena under investigation. But what this agent is taken to be like is, primarily, a theoretical issue. Experiments may isolate causal agents, but only theory can tell us what these agents are. As Worrall pointed out in his work (1994, p. 341), we shouldn't ignore the fact that we never have knowledge about the world by acquaintance; all judgements, especially judgements concerning what kind of entities we refer to, are always theory-dependent. So, although the possibility of an experiment-driven mechanism of reference is certainly worth pursuing further, we shouldn't forget that this mechanism cannot be theory-independent, and therefore not description-independent either.

5. A diagnosis

Let me try to offer a diagnosis as to why Kitcher's descent to the token-level is *prima facie* appealing and why, after all, some of his central underlying insights might well be

preserved by a good account of reference. Why, when for instance it comes to discussing the reference of "dephlogisticated air", it won't do to stick to one of the standard theories of reference?

As has been observed elsewhere (cf. Enç, 1976; Stich, 1991, p. 238), if one adopts a pure causal theory, then it is difficult to avoid the counter-intuitive conclusion that the expression-type "phlogiston" refers to oxygen, since it was oxygen that was causally involved in the grounding of the phenomena of combustion which led to the introduction of "phlogiston". The generic problem with the causal theories is that insofar as certain phenomena are caused, then it turns out that *any* abandoned term will refer, no matter how mistaken and misguided are the descriptions associated with it, given that some thing or other was present in the grounding of the term. As a result, causal theories have a difficulty in explaining why there is referential *failure*, whereas success comes cheap, insofar as the phenomena that led to the introduction of a new term are caused. If, however, one dismisses the causal theories and adopts a descriptive theory, then—given that the description of phlogiston was something along the lines "substance emitted in combustion"—"phlogiston" doesn't refer to anything, since there is nothing that satisfies the foregoing description. The generic problem with the descriptive theories is that they, generally, associate too rich a description with a term: scarcely any theoretical term will end up referring to anything, since it is hardly ever the case that the descriptions associated with the term are going to be satisfied by anything in the world.[10] Not to mention Kripke's main message that it is not necessary (sometimes not even true) that the associated descriptions be satisfied by the referred-to individual or natural kind. So, descriptive theories allow little space for referential *success*, whereas failure comes about too easily.

Kitcher's important insight is that an adequate theory of reference should try to steer between the two extremes (cf. 1993, p. 102, n. 15). More specifically, he wants to avoid the unhappy conclusion that "phlogiston" refers to oxygen. If one appealed to a description theory in order to argue that "phlogiston" failed to refer, the immediate problem would be that expressions such as "dephlogisticated air" would turn out to be denotationless, too. If dephlogisticated air is taken to be the gas that remains when phlogiston is removed completely from ordinary atmospheric air, then since there is nothing like phlogiston to be removed, nor can there be anything like dephlogisticated air remaining.

What if "dephlogisticated air" is a vacuous expression? Well, isn't that the gas that was produced along with mercury when Priestley heated red calx of mercury—the gas that he and his laboratory mice breathed and made Priestley say that he "felt peculiarly light and easy for some time afterwards"—anything other than oxygen, even though Priestley used the expression "dephlogisticated air" to characterise it? It then seems very tempting to find a way to show that "dephlogisticated air" does denote oxygen. Yet, if, in consistency with the pure causal theories, one considered making the foregoing move, one should also have to bracket most (if not all) of Priestley's theoretical beliefs about the substance he had isolated. Hence, unless one were willing to disregard all these beliefs, one could not possibly endorse the claim that "dephlogisticated air" refers to oxygen.

Kitcher's descent to the token-level offers is *prima facie* appealing because it suggests a way to steer between both standard theories. There is no doubt that Kitcher's fundamental insight, the need for a theory of reference that avoids the shortcomings of both the pure causal and the descriptive theories while utilising resources from both, is sound and needs to be further pursued. In fact, the need for such a hybrid theory—

sometimes called, following David Lewis (1984), "causal descriptivism"—has already been pointed out in standard criticisms of the pure causal theories: at least as regards theoretical terms, the burden of reference should be borne by some part of the term's theoretical environment. Reference-fixing should involve some descriptions. These will typically express beliefs about the kind-constituting properties and the explanatory mechanisms attributed to the referent (cf. Enç, 1976). However, this is not the place to discuss the prospects of causal descriptivism, nor some of the obvious challenges that it faces. All I want to conclude with is that Kitcher's token-level semantics and the heavy reliance on the principle of humanity do not offer a satisfactory alternative that sails between the Scylla of the causal theories and the Charybdis of description theories.[11]

Acknowledgements

Many thanks to Chris Daly, Jonardon Ganeri, Peter Lipton, David Papineau and John Worrall, as well as to three anonymous readers, for making incisive and challenging comments on earlier drafts. I am especially grateful to John for bringing Kitcher's views to my attention and for lots of insightful discussions.

Notes

1. The relevant literature is voluminous. An accessible exposition of the debate can be found in Devitt & Sterelny (1987). A more advanced discussion is given in Bach (1987) and Burge (1992). A classic piece which critically examines Kripke's theory is Evan's (1973). The causal theory is defended in Devitt's (1981) and is heavily criticised in Unger's insightful (1983).
2. These issues are developed in detail by Bach in his splendid book (1987), see especially pp. 4–6 & 85–88.
3. Kitcher has also introduced the "conformist" mode of reference, but, as he acknowledges, this mode can be ultimately traced back to either a descriptive or a baptismal mode (1993, p. 77).
4. Kitcher allows that the reference potential of a certain type may comprise several descriptions that members of a scientific community may employ to fix the reference of a token of the type. But these are "presumed to be equivalent", i.e. they are presumed to pick out the same individual (1993, p. 78). Similarly, he allows that there may be many "baptisms" of the same objects, but that they all share "significant common properties" (cf. ibid.) It is in this sense that, I think, the reference potential can be ultimately reduced to a two-member set of modes of reference.
5. I assume the the name "Eustacia Evergreen" is unique. At any rate, it is not the real name of the impostor.
6. Kitcher's example may be usefully compared with Putnam's famous cats/robots case. Putnam (1975a) envisages a situation in which, on a single unnoticed act, Martians remove a number of cats from the earth and replace them with robots that look and behave like cats. Suppose that, in time, we do discover that we are surrounded by feline robots as well as by cats. What does the word "cat" refer to? Putnam, I think correctly, takes it to be a strong advantage of the causal theory that it yields (a) that "cat" still refers to cats, since it was cats—and not feline robots—which grounded the term "cat" and (b) that, after the Martian swindle, we were simply misusing the term "cat" when we applied it to feline robots. Kitcher's account would yield a different result. Some tokens of "cat" refer to cats, while others refer to the feline robots, depending on the context. But it would thereby miss some important true generalisations about cats. "Cats are animals but not robots"—a statement which, at least intuitively speaking, is true—would end up being false.
7. For the relevant differences, one may look at Field's article (1973), especially footnote 12.
8. So he once said (describing an experiment in 1779): "All the p^n [phlogiston] of the charcoal must be exhausted in nitrous before any D^d [dephlogisticated air] can be formed" (1966, p. 171). He later on called this gas "nitrous air" and suggested that it "is hardly to be distinguished from d^d air" (1966, p. 175).
9. This is not meant to be a detailed historical study of Priestley's endeavours. Priestley's full endorsement of the phlogiston theory has been documented in some detail by Kitcher himself (cf. 1978; 1993) and by historians of science (cf. Hankins 1985, p. 106).
10. David Papineau (1996) has recently made an important attempt to bypass this problem based on Ramsey's ideas.

11. Although I am not going to argue for this now, I think that a form of "casual descriptivism" should be the correct account of reference. For relevant arguments, cf. Enç (1976), Berk (1979), Unger (1983) and Kroon (1985; 1987). One can also see my work (1994, chapter 5). Some relevant but still undeveloped arguments in connection with referential-stability in scientific theory-change are offered in the closing paragraphs of my article (1996).

References

BACH, K. (1987) *Thought and Reference* (Oxford, Clarendon Press).

BERK, E. (1979) Reference of theoretical terms, *Southwest Journal of Philosophy*, 10, pp. 139–146.

BURGE, T. (1992) Philosophy of language and mind: 1950–1990, *Philosophical Review*, 101, pp. 3–51.

DEVITT, M. (1981) *Designation* (New York, Columbia University Press).

DEVITT, M. & STERELNY, K. (1987) *Language and Reality* (Oxford, Basil Blackwell).

DONNELLAN, K. (1966) Reference and definite descriptions, *Philosophical Review*, 75, pp. 281–304—reprinted in SCHWARTZ, S. (Ed.) (1977), pp. 42–65.

ENÇ, B. (1976) Reference of theoretical terms, *Noûs*, 10, pp. 261–282.

EVANS, G. (1973) The causal theory of names, *Proceedings of the Aristotelian Society*, 47, pp. 187–208—reprinted in SCHWARTZ, S. (Ed.) (1977), pp. 192–215.

FIELD, H. (1973) Theory-change and the indeterminacy of reference, *Journal of Philosophy*, 70, pp. 462–481.

FODOR, J. (1994) *The Elm and the Expert* (Cambridge, MA, MIT Press).

HANKINS, T.L. (1985) *Science and the Enlightenment* (Cambridge, Cambridge University Press).

HARDIN, C. & ROSENBERG, A. (1982) In defence of convergent realism, *Philosophy of Science*, 49, pp. 604–615.

KITCHER, P. (1978) Theories, theorists and theoretical change, *Philosophical Review*, 87, pp. 519–547—reprinted in D. L. BOYER, P. GRIM & J. T. SANDERS (Eds) *The Philosophers Annual, Volume 2* (1979) [Oxford, Basil Blackwell (all page references are from the latter edition)].

KITCHER, P. (1993) *The Advancement of Science* (Oxford, Oxford University Press).

KRIPKE, S. (1980) *Naming and Necessity* (Oxford, Blackwell).

KROON, F. (1985) Theoretical terms and the causal view of reference, *Australasian Journal of Philosophy*, 63, pp. 142–166.

KROON, F. (1987) Causal descriptivism, *Australasian Journal of Philosophy*, 65, pp. 1–17.

LEWIS, D. (1984) Putnam's paradox, *Australasian Journal of Philosophy*, 62, pp. 221–236.

PAPINEAU, D. (1996) Theory dependent terms, *Philosophy of Science*, 63, pp. 1–20.

PRIESTLEY, J. (1966) *A Scientific Autobiography of Joseph Priestley; Selected Scientific Correspondence.* Edited with Commentary by Robert E. Schofield (Cambridge, MA, MIT Press).

PSILLOS, S. (1994) *Science and Realism: A Naturalistic Investigation Into Scientific Enquiry*, Unpublished Doctoral Thesis, University of London.

PSILLOS, S. (1996) Scientific realism and the "pessimistic induction", *Philosophy of Science*, 63 (Proceedings of PSA 1996), pp. S306–S314.

PUTNAM, H. (1975a) It ain't necessarily so, in *Mathematics, Matter and Method*, Philosophical Papers, Vol. 2, (Cambridge, MA, Cambridge University Press).

PUTNAM, H. (1975b) Explanation and reference, in *Mind, Language and Reality*, Philosophical Papers, Vol. 2 (Cambridge, MA, Cambridge University Press).

PUTNAM, H. (1975c) The meaning of meaning, in *Mind, Language and Reality*, Philosophical Papers, Vol. 2 (Cambridge, MA, Cambridge University Press).

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

STICH, S. (1991) Do true believers exist?, *Proceedings of the Aristotelian Society*, Suppl. Vol. 1991, pp. 229–244.

UNGER, P. (1983) The causal theory of reference, *Philosophical Studies*, 43, pp. 1–45.

WORRALL, J. (1994) How to remain (reasonably) optimistic: scientific realism and the luminiferous ether, in: D. HULL *et al.* (Eds.) *PSA 1994*, Vol.1. East Lansing, Philosophy of Science Association, pp. 334–342.

Note on Contributor

Stathis Psillos was awarded his Ph.D in the Philosophy of Science from King's College, University of London. He is presently a British Academy Postdoctoral Fellow at the London School of Economics. He has published a number of papers in defence of scientific realism. His current research interests include the role of abductive reasoning in philosophy and in Artificial Intelligence, and the philosophy of Rudolf Carnap. *Correspondence:* Department of Philosophay, London School of Economics, Houghton Street, London WC2A 2AE, UK.