What's the 'Puzzle about Belief'? Revisiting Kripke's Challenge through a Fregean Lens

Ethan Reiter, *University of Chicago*

In "A Puzzle About Belief" (1979), Saul Kripke posits a thought experiment involving a bilingual named Pierre, who comes to what seem like contradictory beliefs about the city of London. On a popular reading, the upshot of how Kripke sets up Pierre's problem is that the challenges of ascribing propositional attitudes such as belief to another person are more complex than suggested in previous philosophical analyses, particularly Gottlob Frege's "Sense and Reference" (1948). In particular, Kripke aims to show that a genuine paradox about belief ascription can arise without the need to presuppose that a thinker with different beliefs about the same object relates to that object through different cognitive perspectives (i.e., what Frege called modes of presentation). In this essay, I take issue with how Kripke draws his conclusion that Pierre's problem must go beyond Frege's notions of sense and reference. I argue that application of Fregean notions of sense indeed 'solves' Kripke's puzzle about Pierre and that Kripke's pre-emptive reply to such an objection does not suffice, casting doubt only on how the Pierre paradox gets off the ground. In sum, I argue that a charitable understanding of Frege's ideas and a closer interrogation of Kripke's set-up reveals that Pierre's 'puzzle' is not necessarily a puzzle.

As the trajectory of analytic philosophy makes clear, describing the thought and belief of others with language is no trivial feat. Often, situations in which Leibniz' law regarding identity[1] apparently fails to hold arise. In "Sense and Reference" (1948), Gottlob Frege highlights these logical problems encountered in situations of thought attribution. He argues that the reason Leibniz' law ostensibly breaks down is that when attributing thoughts to thinkers, we must also attribute their occupying a particular cognitive stance on the objects of their thought. Frege thus introduces the sense-reference distinction to show how intersubstitution of coreferential simple proper names in identity statements and ascriptions of apparently contradictory beliefs to an agent need not jeopardize Leibniz' law. Meanwhile, in "A Puzzle About Belief" (1979), Saul Kripke argues that invoking cognitive perspectives in belief is beside the point. He attempts to show that the same intersubstitution failures can be generated with the disquotation and translation principles alone,[2] not requiring any sort of belief perspective. Kripke puts forward the Pierre puzzle, one version of which supposes that even when a subject adopts the same cognitive perspective on two different but coreferential signs, contradictory beliefs can still arise. This paper reconstructs and juxtaposes Frege and Kripke's puzzles to demonstrate the import of Kripke's critique. However, the paper concludes that Kripke's puzzle does not necessarily amount to a puzzle, given that the Fregean line has its own resources to make sense of Pierre's situation through reiterating the extent to which cognitive perspectives shape thought.

[1] I refer here to Leibniz' statement that "those things are the same of which one can be substituted for another without loss of truth," a principle which Frege also committed himself to.

[2] While the debate over these principles is interesting, it is not the topic of this paper. Thus, this paper will limit itself to critique of Kripke's set-up of his puzzle rather than the principles it presupposes.

1 Preliminary Remarks about Frege and Kripke's Puzzles

For Frege, contexts in which Leibniz's law of identity produces unacceptable results can be resolved through considering thinkers' cognitive stances on their objects of thought. Frege's first puzzle contrasts equality statements like 'a = a' and 'a = b' (provided that 'a' and 'b' designate the same object), stating that while the truth of 'a = a' is trivial and "*a priori,*" "statements of the form *a = b* often contain very valuable extensions of our knowledge and cannot always be established *a priori*" (Frege 1948, 209).[3] That is, in the latter construction, one simply cannot intersubstitute 'a' and 'b' without altering its "cognitive value," despite their coreference (209). Frege's second puzzle, regarding ascription of propositional attitudes like belief, is even more illustrative: statements can differ not only in their informativeness, but also in their *truth value*, even when all which distinguishes them is the substitution of coreferential names. For example, if Venus is in orbit, one might say upon waking up, "I believe I will see the morning star right now" *and* "I do not believe I will see the evening star right now": insofar as one (naturally) resists ascribing irrationality to such an agent for having contradictory beliefs, those coreferential designators of Venus must convey different meanings, such that sense is one thing, and reference another. For Frege, the 'morning star' and 'evening star' represent the different—-and limited—-cognitive perspectives thinkers can take on objects, as seen easily in such cases of discovery, as well as confusion.[4] These distinct "mode[s] of presentation" (210), as Frege calls them, differ in their cognitive value[5]—-"the thought changes," therefore, when intersubstituting distinct senses (215). Consequently, Frege argues that both his first puzzle on identity statements and his second puzzle on propositional attitude ascription involve a thinker taking two irreducibly different cognitive stances on (i.e., senses of) the object of their thought.

Here, it is worth bringing Kripke's views on such cognitive perspectives into focus: he argues that considering sense distinctions is not necessary to create the sort of intersubstitution problems which troubled Frege. Instead, Kripke only presumes the disquotation principle and the translation principle.[6] He poses a problem, by appeal to the imaginary example of Pierre, of a bilingual speaker who holds what seem to be contradictory beliefs about the same proper name, solely by virtue of holding those beliefs separately in his two languages. Pierre grows up a standard French speaker, believing that London is pretty. Later in life, he becomes a standard English speaker by direct method[7] after unknowingly ending up in London (finding, then, that it is *not* pretty). He thus holds the following two beliefs, according to the disquotation principle:

(a) in French: '*Londres est jolie*'

(b) in English: 'London is not pretty'

[3] Frege provides the example of the referent Venus, which is designated by both the names 'the morning star' and 'the evening star.' While 'the morning star is the morning star' strikes one as empty-headed, Frege writes that 'the morning star is the evening star' "was of very great consequence to astronomy. Even today the identification… is not always a matter of course" (209).

[4] A classical example is how readers understand the tragedy of Oedipus Rex: as thinkers like Jerry Fodor have noted, one can attribute to Oedipus the belief that he wants to marry Jocosta but *not* the belief that he wants to marry his own mother, and it is by virtue of this discrepancy that the text is ultimately understood as a tragedy.

[5] Put another way, the names being intersubstituted do not have the same psychological value to their thinker: they are distinct perspectives on, different senses of, the object of thought (i.e., Venus). Thus, the sense-reference distinction is what empowers 'a = b' to be informative, whereas 'a = a' is not, in that it conveys *new* knowledge that two senses are indeed of the same referent.

[6] For the sake of my argument that Kripke's puzzle can be resolved through a Fregean lens, these two principles (in their various weak and strong forms) can be taken for granted and set aside. To briefly summarize them, following Kripke: the disquotation principle holds that when a speaker sincerely assents to a proposition, they believe it (439); the translation principle holds that the truth-value of a sentence in one language is preserved upon translation to another (440).

[7] That is, Pierre learns English entirely without the help of his native language, French. Rather than learning English names through translation from their French equivalents, for example, Pierre learns English as though it is his first language, connecting English expressions directly to his audiovisual experiences of the world. Consequently, he fails to realize '*Londres*' is 'London.'

These beliefs form an outright contradiction, at least if the translation principle is applied to (a) to result in 'London is pretty' and Kripke's ascription of both beliefs to Pierre is accepted.[8]

If both Pierre's French and English beliefs about London are to be upheld, though, then how can Pierre avoid contradictory beliefs? Indeed, Kripke clarifies that "as long as he is unaware that the cities he calls 'London' and '*Londres*' are one and the same, [he] is in no position to see, by logic alone, that at least one of his beliefs must be false" (444). Even if Pierre is a logician par excellence, he would be unable to leverage *modus tollens* to correct his 'contradictory' beliefs until he realizes his beliefs about 'London' and '*Londres*' actually regard the same city. This gives rise to Kripke's thesis: "that the [Pierre] puzzle *is* a puzzle" (433). It is difficult—-paradoxical even, Kripke says—-to state Pierre's true beliefs about the city of London.

Could this Pierre puzzle resemble Frege's second puzzle about belief ascription, inviting clarification that 'London' and '*Londres*' are distinct cognitive perspectives on (i.e., senses of) the city of London? At first glance, Kripke admits, "[o]ne aspect of the presentation may misleadingly suggest the applicability of Frege-Russellian ideas that each speaker associates his own description or properties to each name" (445). However, Kripke swiftly dismisses that '*Londres*' and 'London' could be considered distinct modes of presentation of London in that it is not required to assume that the two translations of the city of London's name satisfy distinct sets of properties. Indeed, Kripke argues that "the puzzle can still arise even if Pierre associates to *'Londres'* and to 'London' *exactly* the same *uniquely identifying* properties," such as in the case that he identifies both as having the very same landmarks, monarchs, etc., insofar as he "regard[s] *both* properties as uniquely identifying" (446). Since Pierre acquires English through the direct method, Kripke maintains that nothing compels Pierre to equate terms like '*Angleterre'* with 'England'; Kripke avoids presupposing "an 'ultimate' level [of language]… where the defining properties are 'pure' properties not involving proper names" (447). Kripke thus concludes that even a shared set of uniquely identifying properties for 'London' and '*Londres*' could not alone compel Pierre to infer any contradiction from his separately held beliefs. Given such perils, Kripke concludes that Pierre's paradox goes deeper than Fregean notions of sense.

2 The Potential for a Fregean Reinterpretation of the Pierre Puzzle

I will argue that Kripke's swift dismissal of Fregean readings does not withstand scrutiny. Specifically, Kripke's rejection that "Pierre believes that *the city* satisfying *one* set of properties is pretty, while he believes that *the city* satisfying *another* set of properties is not pretty" (445) ought to be reconsidered. Kripke was too hasty to assume that Frege's invocation of cognitive perspectives has no place within the puzzle he sets up. Indeed, Pierre's contradictory beliefs, particularly about the prettiness of London, can hardly be accounted for in the first place if Pierre is said to truly occupy the same cognitive perspective in both his English and French belief sets.

Ultimately, the identifying properties of '*Londres*' associated with Pierre's French belief that the city is pretty are fundamentally different from the properties of 'London' associated with his English belief that the city is not pretty. Recall how Kripke sets up Pierre's puzzle. The set of properties informing Pierre's beliefs about '*Londres*' are provided "[o]n the basis of what he has heard of London" in French—-those positive descriptions which make him "inclined to think that it is pretty" (442). Pierre's belief of '*Londres*' thus corresponds to the compliments that he has heard about the city—-they uniquely identify the city's prettiness for him. Yet there are no such compliments which correspond to 'London' for Pierre. There is likewise no such corresponding

[8] I will follow Kripke in ascribing both beliefs to Pierre. If one tries to undermine Pierre's old belief in French that London is pretty, this entails the absurd conclusion that other monolingual French speakers must lack belief about the prettiness of '*Londres*.' But if one tries to undermine Pierre's new belief in English that London is *not* pretty, they mistakenly think "Pierre's French past [can] nullify such a judgment" (444). Kripke suggests considering "an electric shock [which] wiped out all his memories of the French language, what he learned in France, and his French past": but this, too, creates a *reductio* insofar as Pierre cannot *gain* a new belief from destruction of his memory, and Pierre should hold the same belief as his new English countrymen (444). Combining the two denials of belief "in his bilingual stage" would only compounds these difficulties (444).

property, 'this is reputed by my countrymen as pretty,' for 'London'; in fact, Kripke asks readers to imagine the opposite, that Pierre's English neighbors "rarely venture outside their own ugly section" (445). Given such differences, Pierre occupies a specific cognitive stance on '*Londres.*'

Meanwhile, the set of properties informing Pierre's beliefs about 'London' are provided when Pierre "is unimpressed with most of the rest of what he happens to see" while in England (443). Indeed, when Pierre thinks of 'London,' he likely recalls the filth of the part of the city wherein he now lives. There is no corresponding uniquely identifying property, 'part of this city where I lived is filthy,' which Pierre could recall when thinking of '*Londres*' insofar as (like Kripke points out) he does not even realize he is living in '*Londres*.' This explains how Pierre occupies a particular cognitive stance on 'London,' distinct from the stance which he occupies on '*Londres*.' Especially given that Pierre learns English through the direct method, the set of experiences which Pierre uniquely associates with English expressions like 'London' date much later than the set of experiences which Pierre uniquely associates with French expressions like '*Londres.*' Pierre just cannot be said to have become acquainted with the two names for this city in the same way: for all he knows, he does not even live in '*Londres*.' The upshot is that the set of properties on which Pierre bases his beliefs of 'London' simply are *not* shared for '*Londres*.' Insofar as Pierre never realizes that 'London' and '*Londres*' are the same, Kripke must concede that Pierre is in no position to realize that the reputation of '*Londres*' as pretty uniquely identifies the *same* city as does Pierre's uniquely identifying memory of living in an ugly part of 'London.' The set of properties defining 'London' for Pierre, then, are not those which define '*Londres*.'

Consequently, Kripke cannot be correct when he writes that "the puzzle can still arise even if Pierre associates to *'Londres'* and to 'London' *exactly* the same *uniquely identifying* properties" (446): if this was true, which uniquely identifying property shared between both '*Londres*' and 'London' could give rise to contradictory beliefs about whether the city is pretty? As suggested, this property could neither be the city's reputation nor Pierre's living experiences, given that both correspond to Pierre's *distinctive* cognitive stances on 'London' and '*Londres*.'

There is thus simply no potential for these two subsets of uniquely identifying properties (i.e., those which designate '*Londres*' pretty and those which designate 'London' not pretty) to be isomorphic given their irreconcilable implications for the prettiness of the city of London. At some point, something has to give: some uniquely identifying properties of '*Londres*' make it pretty and *other* uniquely identifying properties of 'London' make it *not* pretty if Pierre truly reaches opposing conclusions about the two names in his beliefs.[9] This Fregean reading makes the straightforward concession that Pierre believes that '*Londres*' is pretty while he believes that 'London' is not, grounding the inconsistency in the names' distinctive modes of presentation. On such a view, Pierre's puzzle is no less tractable than Frege's second puzzle concerning failures intersubstituting coreferring names in belief contexts: Pierre's *ostensibly* conflicting beliefs about London's beauty merely reflect the two irreducibly distinct cognitive stances he occupies of it.[10]

Indeed, consider when Kripke compares Pierre's situation to one who comes to different beliefs about the baldness of 'Plato' (English) and '*Platon*' (French), figuring they are different people (446). Kripke introduces this second scenario with the express purpose of pre-empting the Fregean reading this paper defends, illustrating instead how "[t]he puzzle can arise even if Pierre associates exactly the same identifying properties with both names" (446). However, the Fregean counterpoint still remains: what *shared* properties could be the basis for the baldness attributed to

[9] It may be objected at this stage that even if Pierre's belief sets in English and French are the same, he could nevertheless draw different conclusions about the city of London through the explosion principle of classical logic. For the present purposes, this reply can be set aside. It is not a straightforward matter to generalize from classical logic to the psychology of belief, especially in regard to the explosion principle: what it would mean to truly believe both p and $\sim p$—-and in what sense would that entail a believer to believe whatever else they liked? Moreover, this paper follows Kripke in presuming Pierre is a skilled logician; Pierre cannot be admitted to (knowingly) believe in two inconsistent sets of premises, so he could not make use of explosion anyway.

[10] As this paper will explain, the Fregean can analogize Pierre's beliefs that '*Londres*' is pretty and that 'London' is not to the example provided earlier, of one's belief that they will see the 'morning star' but not the 'evening star' at the beginning of the day. In both cases, there is no true contradiction, insofar as the object of one's belief—-and thus the belief—-changes when the sense changes.

'*Platon,*' contra the hirsuteness attributed to 'Plato'? The property on the basis of which 'Plato' has hair is necessarily different from the one on which '*Platon*' is bald. Just as in the Pierre case, it is incoherent to imagine the two names for the Greek philosopher having "exactly the same identifying properties" (446) while at once differing in the very property that makes them different when they are taken as objects of belief—-the crux of the apparent contradiction. If Kripke means to argue that there is truly no psychological difference between the English and French equivalents in the Plato case (just as in the Pierre case), it is unclear how exactly the contradictory beliefs could surface to begin with. Is that not evidence enough of a difference?

Kripke makes the incongruity between Pierre's French and English belief sets seem to stem from a whim—-as though some distinction between just the *valence* of French versus English descriptions of London could have led Pierre to arrive at a different conclusion about the city's prettiness from shared uniquely identifying properties. This is a critical oversight: how can one say Pierre's starting assumptions for arriving at his beliefs about the prettiness of 'London' and '*Londres*' stem from the exact same sets of information? They obviously do not. At some point in cataloging Pierre's French and English belief sets, there *must* be discord (particularly in those beliefs bearing on judgments about the city's prettiness)—-or else, what could possibly account for Pierre reaching different conclusions about its prettiness in the first place? Unless this divergence is recognized for what it is, Kripke would need to make an entirely different argument to contextualize how Pierre arrives at contradictory beliefs. If 'London' and '*Londres*' truly do share all the same uniquely identifying properties, Kripke must pass the explanatory buck for Pierre's contradictory beliefs about the two to intrinsic differences between French and English. After all, only those could plausibly dispose Pierre to judge one thing about the former yet the opposite about the latter if he truly assumes the same cognitive stance toward them both.

In this way, the discrepancy between Pierre's French and English beliefs about London's prettiness is akin to Frege's puzzle, wherein one believes they are seeing the 'morning star,' but not the 'evening star,' upon waking up. Just as how in the latter, the difference between '*a*' and '*b*' is a uniquely identifying property which distinguishes the two modes of presentation (e.g., based on the time of day when Venus is in view or the stage of its orbit), the difference between 'London' and '*Londres*' is those uniquely identifying properties which make the former ugly (e.g., the hideosity of Pierre's neighborhood) but the latter pretty (e.g., the city's reputation of beauty among French speakers). Appreciating this difference further allows for recognizing that Kripke's ultimate question—-"Does Pierre or does he not, believe that London (not the city satisfying such-and-such description, but *London*) is pretty"—-is falsely posed (446). Since Pierre occupies two distinctive cognitive stances on "*London*" in his beliefs, Pierre's beliefs about the city cannot but be understood in Fregean terms of modes of presentation. In the final analysis, Pierre's beliefs of 'London' are irreducibly different from his beliefs of '*Londres*,' as with the 'morning star' and 'evening star.' If Kripke argues that 'London' and '*Londres*' are not two different senses of London to Pierre, he need not only refute the above consideration, but also meet the explanatory burden of *how* Pierre could ever come to contradictory beliefs about the city's prettiness if, in both his languages, the set of uniquely identifying principles are isomorphic. Ultimately, a Fregean can reintroduce cognitive perspectives into Pierre's puzzle, casting doubt on its paradoxicality through concluding that Pierre's beliefs concern different modes of presentation of the city of London after all. Indeed, they stem from different thoughts.

3 Conclusion

This paper suggests a Fregean resolution to the ostensible paradox about Pierre which Saul Kripke puts forward in "A Puzzle about Belief" (1979). I argue that for Pierre, 'London' and '*Londres*' are two distinct modes of presentations, two distinct senses, of the city of London. When making sense of what Pierre believes about the English capital, 'London' and '*Londres*' need not be taken to connect to the same thought: I argue that for Pierre, the names *do* correspond to cities with two *necessarily* distinct sets of uniquely identifying properties. Indeed, the necessary distinction is that

on the basis of which Pierre draws opposite conclusions about whether London is pretty—-the same distinction that engenders his key 'contradiction' in belief.

In Kripke's set-up of the puzzle, Pierre comes to contradictory beliefs about London's prettiness based on an incongruity between what he hears about '*Londres*' from his countrymen while living in France and what he sees on the ground while living in 'London.' I argue that Kripke is wrong to think that both opposing aspects of this inconsistency can be accommodated within the same cognitive stance on the part of Pierre if Pierre never realizes that 'London' and '*Londres*' are coreferring names. Instead, Frege's sense-reference distinction remains crucial: Pierre's puzzle is not paradoxical because he occupies two distinct cognitive stances on the city of London, such that his incongruous beliefs about 'London' and '*Londres*' do not contradict each other—-they stem from altogether different thoughts. Thus, Kripke's 'puzzle' about belief does not amount to a puzzle, so long as Pierre's cognitive perspectives in his beliefs about 'London' and '*Londres*' are distinguished along the line Frege endorses with his second puzzle.

Bibliography

Frege, Gottlob. "Sense and Reference." *The Philosophical Review* 57, no. 3 (May 1948): 209–30. https://doi.org/10.2307/2181485.

Kripke, Saul A. "A Puzzle about Belief." *Meaning and Use*, 1979, 239–88. https://doi.org/10.1007/978-1-4020-4104-4_13.