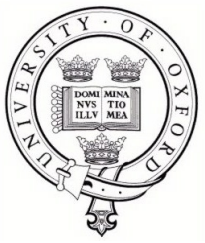
Mental Files

Thea Goodsell

A thesis submitted for the degree of

Doctor of Philosophy



Magdalen College

University of Oxford

Hilary Term 2013

Abstract

Mental Files

Thea Goodsell

Thesis submitted for the degree of Doctor of Philosophy,

University of Oxford, Hilary Term 2013

It is often supposed that we can make progress understanding singular thought about objects by claiming that thinkers use *mental files*. However, the proposal is rarely subject to sustained critical evaluation. This thesis aims to clarify and critique the claim that thinkers use mental files.

In my introductory first chapter, I motivate my subsequent discussion by introducing the claim that thinkers deploy modes of presentation in their thought about objects, and lay out some of my assumptions and terminology. In the second chapter, I introduce mental files, responding to the somewhat fragmented files literature by setting out a core account of files, and outlining different ways of implementing the claim that thinkers use mental files. I highlight pressing questions about the synchronic and diachronic individuation conditions for files.

In chapters three and four, I explore whether *de jure coreference* can be used to give synchronic individuation conditions on mental files. I explore existing characterisations of de jure coreference before presenting my own, but conclude that de jure coreference does not give a useful account of the synchronic individuation conditions on files. In chapter five, I consider the proposal that thinkers must sometimes *trade on the coreference* of their mental representations, and argue that we can give synchronic individuation conditions on files in terms of trading on coreference. In chapter six, I bring together the account of files developed so far, compare it to the most developed theory of mental files published to date, and defend my account from the objection that it is circular. In chapter seven, I explore routes for giving diachronic individuation conditions on mental files. In my concluding chapter, I distinguish the core account of files from the idea that the file metaphor should be taken seriously. I suggest that my investigation of the consequences of the core account has shown that the file metaphor is unhelpful, and I outline reasons to exercise caution when using 'files' terminology.

Approximate number of words: 74,556

Acknowledgements

I am pleased to acknowledge that this research was made possible by an AHRC doctoral studentship. And I am grateful to the fellows and staff of Magdalen College, for providing me with my Oxford home for so many years, and for financial support over the course of my studies. Particular thanks are due to Lizzie Fricker, my teacher, mentor and friend throughout my time at Magdalen.

I would like to take this opportunity to acknowledge the influence of four other inspiring teachers: Chris Brooke, Emma Brown, Russell Dudley-Smith and Jenny Locke. I would also like to thank the administrative staff at Oxford University's Philosophy Faculty for their help over the years. And I must thank my team of proof-readers, spread over three continents: Simona Aimar, Courtney Cox, Jessie Greengrass, Oliver Lyttelton, Karin Maraney, Amy Monroe, James Studd and Bobby Talalay.

A very considerable debt of thanks is due to my supervisor, John Hawthorne. His patience, encouragement and good humour were invaluable aids as I wrote this thesis, to say nothing of his philosophical insight and advice. Rather than try to quantify my appreciation and gratitude, I'll simply say: thank you.

I am grateful to Simona Aimar, Cian Dorr, Michael Murez, François Recanati, Emanuel Viebahn and Tim Williamson for helpful conversations or comments on my work. And I am grateful to my examiners, Bill Brewer and Anandi Hattiangadi for the interesting discussion in my viva.

Special thanks go to Daniel Morgan for so many enjoyable conversations about modes of presentation. They also go to James Studd, for advice and philosophical discussion instrumental in moving this thesis from 'almost a first draft' to 'completed'. And special thanks go to Robert Watt, whose encouragement and friendship were invaluable, most notably during this thesis's extended teething-period.

Finally, it would be disingenuous to limit these acknowledgements to the academic sphere. For their unwavering and good humoured support, I must thank Courtney Cox, Jessie Greengrass, Alexandra Jenkins, Benjamin Kent, Sarah Lane Smith, Karin Maraney, Amy Monroe, Christopher Namih, Rupert Paines, Bobby Talalay, Alexandra Scott and Lizzie Wells, all friends in the very strongest sense of the word. And I am happy to take this opportunity to thank my sister Hester and my parents, Helen and Robert Goodsell, for, among other things, their material and non-material support throughout my studies, their forbearance in the months before I submitted, and for their unfailing faith in me.

Costaging and Cofiling Theses

Costaging theses

CoStaging-1 r_t and s_t are costaged iff r_t and s_t deploy the same MOP

CoStaging-2 r_t and s_t are costaged iff there is some mode of presentation that both r_t and s_t deploy

CoStaging-3 r_t and s_t are costaged iff r_t and s_t contain the same node

CoStaging-4 r_t and s_t are costaged iff the thinker who holds r_t and s_t cannot sensibly doubt that r_t and s_t share referential-value

CoStaging-5 r_t and s_t are costaged iff the thinker who holds r_t and s_t takes r_t and s_t to be about the same thing

CoStaging-6 r_t and s_t are costaged iff the thinker who holds r_t and s_t treats r_t and s_t as about the same thing

CoStaging-7 r_t and s_t are costaged iff r_t and s_t assumedly-corefer

CoStaging-8 r_t and s_t are costaged iff r_t and s_t are de jure coreferential

CoStaging-9 r_t and s_t are costaged iff r_t and s_t are RR-coreferential, or r_t and s_t are putatively RR-coreferential

CoStaging-9* r_t and s_t are costaged iff c_{r_t} is part of r_t and c_{s_t} is part of s_t, and either c_{r_t} RR-corefers with c_{s_t} or c_{r_t} putatively RR-corefers with c_{s_t}

CoStaging-10* r_t and s_t are costaged iff c_{r_t} is part of r_t and c_{s_t} is part of s_t, and either c_{r_t} AP-corefers with c_{s_t} or c_{r_t} AP-pseudo-corefers with c_{s_t}

CoStaging-11* r_t and s_t are costaged iff c_{r_t} is part of r_t and c_{s_t} is part of s_t, and c_{r_t} is p-linked with c_{s_t}

CoStaging-12* r_t and s_t are costaged iff c_{r_t} is part of r_t and c_{s_t} is part of s_t, and either c_{r_t} IK-corefers with c_{s_t} or c_{r_t} IK-pseudo-corefers with c_{s_t}

CoStaging-13* T's record-stages r_t and s_t are costaged iff r_t and s_t contain component-stages c_{r_t} and c_{s_t} respectively, and T can trade on the coreference of c_{r_t} and c_{s_t}

CoStaging-14* r_t and s_t are costaged iff r_t and s_t contain c_{r_t} and c_{s_t} respectively, and the thinker can trade on the coreference of c_{r_t} and c_{s_t}, and r_t and s_t are belief-records or records of belief-like information

CoStaging-15 r_t and s_t are costaged iff r_t and s_t causally originate in the same object, and enter through the same acquaintance relation

CoStaging-16 Record-stages r_t and s_t are costaged iff r_t and s_t are not insulated from one another

CoStaging-17 i and j are costaged at t iff at t T takes i and j to be about the same thing

CoStaging-18 i and j are costaged at t iff there is a presumption of identity between i and j at t

CoStaging-19 i and j are costaged at t iff there is a information-gathering presumption of identity between i and j at t

CoStaging-20 i and j are costaged at t iff there is a current-reasoning presumption of identity between i and j at t

Cofiling theses

CoFiling-1 File-stages F_{t_1} and G_{t_2} are cofiled iff there is a record r such that F_{t_1} contains r and G_{t_2} contains r

CoFiling-2 File-stages F_{t_1} and G_{t_2} are cofiled iff F_{t_1} S* G_{t_2}
x S* y iff there is a chain of file-stages F_1, $F_2 \ldots F_k$ (k >0) such that x shares a record with F_1, F_1 shares a record with $F_2 \ldots F_k$ shares a record with y

CoFiling-3 File-stages F_{t_1} and G_{t_2} are cofiled iff F_{t_1} and G_{t_2} are suitably causally connected

CoFiling-4 File-stages F_{t_1} and G_{t_2} are cofiled iff F_{t_1} and G_{t_2} are causally connected in the way appropriate for cofiled file-stages

CoFiling-5 File-stages F_{t_1} and G_{t_2} are cofiled iff F_{t_1} contains a record-stage r_{t_1}, and G_{t_2} contains a record-stage s_{t_2}, and r_{t_1} and s_{t_2} [AP-/IK-corefer] or [AP-/IK-pseudo-corefer]

CoFiling-6 File-stages F_{t_1} and G_{t_2} are cofiled iff F_{t_1} contains a record-stage r_{t_1}, and G_{t_2} contains a record-stage s_{t_2}, and r_{t_1} and s_{t_2} either RR-corefer or putatively RR-corefer

CoFiling-7 T's file-stages F_{t_1} and G_{t_2} are cofiled iff F_{t_1} contains a record-stage r_{t_1}, and G_{t_2} contains a record-stage s_{t_2}, and T can trade on the coreference of r_{t_1} and s_{t_2}

CoFiling-7* T's file-stages F_{t_1} and G_{t_2} are cofiled iff a record r_{t_1} is a member of F_{t_1} in virtue of containing $c_{r_{t1}}$ and a record s_{t_2} is a member of G_{t_2} in virtue of containing $c_{s_{t2}}$, and T can trade on the coreference of $c_{r_{t1}}$ and $c_{s_{t2}}$

Contents

1 Introduction: Modes of Presentation 1
- 1.1 Modes of presentation . . . 2
  - 1.1.1 A solution for a problem about belief . . . 2
  - 1.1.2 Singular thought . . . 5
  - 1.1.3 MOPs . . . 8
  - 1.1.4 A further use for MOPs . . . 14
- 1.2 Preliminaries . . . 15
  - 1.2.1 Outline . . . 15
  - 1.2.2 Assumptions and terminology . . . 16

2 Mental Files 21
- 2.1 Mental files . . . 22
  - 2.1.1 The metaphor . . . 22
  - 2.1.2 Mental files . . . 23
  - 2.1.3 The core account of files . . . 25
- 2.2 Further details . . . 35
  - 2.2.1 Picturing files . . . 35
  - 2.2.2 Uses of files . . . 46
- 2.3 Costaging and cofiling . . . 59
  - 2.3.1 Two questions about mental files . . . 60
  - 2.3.2 Preliminary attempts to answer the costaging question . . . 63
- 2.4 Final remarks . . . 68

3 Assumed-Coreference and De Jure Coreference 69
- 3.1 Introduction . . . 70
- 3.2 External-coreference . . . 72
- 3.3 Assumed-coreference . . . 73
  - 3.3.1 Introducing assumed-coreference . . . 73
  - 3.3.2 Assumed-coreference and costaging . . . 76
  - 3.3.3 Problems and limitations . . . 77
- 3.4 De jure coreference . . . 79
- 3.5 RAS-coreference . . . 81
  - 3.5.1 Introducing RAS-coreference . . . 81

3.5.2 The test for RAS-coreference . . . 82
3.5.3 RAS-coreference in thought . . . 84
3.6 SR-coreference . . . 85
3.6.1 The relational framework . . . 86
3.6.2 Types of semantically required coreference . . . 87
3.6.3 Putative SR-coreference . . . 91
3.7 De jure coreference and thought . . . 94
3.7.1 RR-coreference . . . 94
3.7.2 Putative RR-coreference . . . 95
3.7.3 An answer to the costaging question? . . . 96
3.8 Accessibility, explicitness and costaging . . . 99
3.9 Manifest consequence and multiple takes . . . 104
3.10 Final remarks . . . 109

4 Explicitly Guaranteed Coreference 111
4.1 Introduction . . . 112
4.2 AP-coreference . . . 113
4.2.1 Introducing AP-Coreference . . . 113
4.2.2 The non-transitivity of AP-coreference . . . 117
4.2.3 Further evidence for the non-transitivity of AP-coreference? . . . 123
4.3 Concerns with AP-coreference . . . 129
4.4 An alternative account: IK-coreference . . . 132
4.4.1 Is de jure coreference *explicitly* guaranteed coreference? . . . 132
4.4.2 IK-coreference . . . 134
4.4.3 Is IK-coreference non-transitive? . . . 136
4.5 De jure coreference and costaging . . . 137
4.5.1 AP-coreference and costaging . . . 137
4.5.2 IK-coreference and costaging . . . 139
4.6 Final remarks . . . 142

5 Trading On Coreference 144
5.1 Introduction . . . 145
5.2 The argument for trading on coreference . . . 146
5.2.1 Trading on coreference . . . 146
5.2.2 The argument for trading on coreference . . . 148
5.2.3 Reasoning as if components are coreferential is ubiquitous . . . 150
5.2.4 A metarepresentational alternative . . . 152
5.2.5 When do thinkers trade? . . . 154
5.3 Trading on Coreference and Transparency . . . 155
5.3.1 The problem . . . 155
5.3.2 Possible responses . . . 160
5.3.3 Breaking the stalemate . . . 168
5.3.4 Referential-value in the puzzle-cases . . . 171
5.4 Trading on coreference and costaging . . . 174

5.5 Final remarks . . . 177

6 Costaging 179
6.1 Introduction . . . 180
6.2 A partial theory of files . . . 180
6.2.1 The theory . . . 180
6.2.2 Remaining issues . . . 185
6.3 Recanati's answer to the costaging question . . . 187
6.4 Circularity objections . . . 196
6.4.1 First objection . . . 196
6.4.2 Second objection . . . 198
6.4.3 The status of CoStaging-13* . . . 199
6.5 Final remarks . . . 207

7 Cofiling 208
7.1 Introduction . . . 209
7.2 Cofiling considerations . . . 210
7.2.1 Records, referential-components and files . . . 210
7.2.2 Are there cross-temporal examples of Frege-cases? . . . 210
7.2.3 Why allow for persisting files? . . . 213
7.2.4 Shifts in referential-value . . . 216
7.3 Cofiling options . . . 218
7.3.1 Shared records . . . 218
7.3.2 Causal connectedness . . . 219
7.3.3 De jure coreference . . . 219
7.3.4 Trading on coreference . . . 223
7.3.5 Identifying characteristics . . . 229
7.4 Final remarks . . . 232

8 A Cautionary Note 235
8.1 Introduction . . . 236
8.2 The thick and thin accounts . . . 236
8.2.1 Beyond material objects . . . 236
8.2.2 Distinguishing the thick and thin accounts . . . 240
8.2.3 The value of the file account . . . 241
8.3 The dangers of 'files'-terminology . . . 242
8.3.1 Status and the dangers of 'files'-terminology . . . 243
8.3.2 Files in psychology and linguistics . . . 244
8.4 Final remarks . . . 250

Bibliography 252

Chapter 1

Introduction: Modes of Presentation

1.1 Modes of presentation

1.1.1 A solution for a problem about belief

Here's a familiar line of thought: as a first pass, believing (or having any other propositional attitude) is a binary relation between a thinker and a proposition. However, assuming that propositions are coarse-grained, this first pass account of belief leads to problems. To resolve these problems, we posit modes of presentation.

But simply positing modes of presentation doesn't tell us much about what modes of presentation are. Explaining what modes of presentation are is a principal motivation for claiming that thinkers use *mental files*. This thesis aims to clarify and critique the claim that thinkers use mental files. So a good starting point is the idea of a mode of presentation.

Propositions are coarse-grained iff utterances U_1 and U_2 express the same proposition, when U_1 and U_2 are utterances in the same context of atomic sentences differing only in cointensional expressions. In contrast, propositions are fine-grained iff utterances U_1 and U_2 express different propositions. So utterances of (1) and (2), made in the same context, express the same coarse-grained proposition.

(1) Hesperus is bright.

(2) Phosphorus is bright.

By this test, Russellian and Stalnakerian propositions are coarse-grained. Russellian-propositions are structured entities built out of the referents of expressions or thoughts. Utterances of both (1) and (2) express a proposition constructed in the same way from the planet Venus and the property of being bright.[1] An alternative account of coarse-grained propositions is given by Stalnaker (e.g. 1976; 1984), who argues that the proposition expressed by an utterance is the set of possible worlds

[1] See Russell (2010 [1903]).

at which the utterance is true. Utterances of (1) and (2) are true at all the same possible worlds, so express the same proposition.[2]

The problem with treating propositional attitudes as a relation between a thinker and a coarse-grained proposition is that there are data demonstrating a conflict between this understanding of propositional attitudes and our ordinary practice of ascribing attitudes and using attitudes to explain behaviour. For example, our usual practice is to use a disquotational principle in ascribing beliefs. A disquotational principle DP can be borrowed from Kripke (1979). For an English sentence 'S', containing no indexical, pronominal or ambiguous terms:

DP If a normal English speaker, on reflection, sincerely assents to 'S', then she believes that S.[3]

Suppose Thales sincerely assents to (1), and also sincerely assents to (2′).

(2′) It is not the case that Phosphorus is bright.

According to DP, we can ascribe Thales the beliefs expressed by (1) and (2′). But utterances of (1) and (2′) express contradictory coarse-grained propositions, that *Venus is bright* and *it is not the case that Venus is bright*. So assuming that beliefs are a binary relation between a thinker and a coarse-grained proposition, DP leads us to ascribe Thales contradictory beliefs. Nonetheless, we have no reason to suppose that Thales is anything other than rational.

A second kind of datum relates to practical reasoning. For example, suppose that Thales assents to (3) and (4).

(3) I want to see Hesperus.

(4) If I go into the garden, I will see Phosphorus.

[2]Because of how they treat necessary truths, Stalnakerian propositions are coarser than Russellian. However, my interests in this thesis license my drawing the coarse/fine-grained distinction where I do.

[3]Adapted from Kripke (1979, pp248-249).

Using ordinary attitude-ascription practices,[4] we ascribe to Thales a desire that *he sees Venus* and a belief that *if he goes into the garden he will see Venus* — nonetheless, everything being equal we don't expect Thales to be immediately motivated to go into the garden, even though we suppose he is rational.

A third kind of datum relates to theoretical reasoning. For example, suppose Thales assents to (1) and (5).

(5) Phosphorus is large.

Using DP and the first-pass account of propositional attitudes, we ascribe Thales a belief that *Venus is bright*, and that *Venus is large*. However, all things being equal, we would not expect him to be able immediately to infer that *something is both bright and large*.

These data conflict with our ordinary way of thinking about propositional attitudes. We expect rational and reflective thinkers not to believe both that P and $\neg P$. We expect that if someone desires that P and believes that Φing leads to P, then all things being equal she will immediately be motivated to Φ. And we expect that if someone believes that P, and believes that Q, she will be able to immediately infer what is is obviously logically entailed by P & Q.

A popular response to these data is to abandon the first-pass account of propositional attitudes, and to replace it with an account of propositional attitudes supplemented with a theory of *modes of presentation*. There are different ways of incorporating modes of presentation into the theory. One route is to give a fine-grained account of propositions featuring modes of presentation — for example, claiming that propositions are not built out of referents, but rather out of the modes of presentation of those referents.[5] A propositional attitude is still a binary relation, but this time between a thinker and a fine-grained proposition. Another alternative is to

[4] Though DP cannot be used here, because (3) and (4) contain indexical expressions.

[5] E.g. Frege (1948 [1892]).

continue to give a coarse-grained account of propositions, and to treat propositional attitudes as a ternary relation between a thinker, a coarse-grained proposition, and a mode of presentation for that coarse-grained proposition.[6]

Whichever approach we take, the thought is that introducing modes of presentation can be used to remove the impression of a puzzle from the data discussed above. Take the first kind of case as an example: if we suppose attitudes are relations between thinkers and fine-grained propositions, then we don't ascribe Thales beliefs in the inconsistent propositions that *Venus is bright* and *it is not the case that Venus is bright*. Rather, we ascribe the thinker beliefs in propositions we can express as *Hesperus is bright* and *it is not the case that Phosphorus is bright*. It may be that there is no world at which both these propositions are true, but nonetheless the propositions are logically consistent with one another.

Alternatively, suppose that attitudes are relations between thinkers, coarse-grained propositions and modes of presentation. Then we only claim that it is irrational to believe a proposition and its negation of propositions if those beliefs deploy the same mode of presentation for that proposition. So there is nothing inconsistent in believing that *Venus is bright* and *it is not the case that Venus is bright*, so long as beliefs use different modes of presentation for the proposition *Venus is bright*.

1.1.2 Singular thought

The puzzles I have outlined for the first-pass account of propositional attitudes all involve *singular thought*. 'Singular term', 'singular thought' and 'singular mode of presentation' are philosophers' terms of art which are hard to adequately define without begging any questions — particularly because some suggest that what makes a thought or term *singular* is its association with a special kind of mode of presentation

[6]E.g. Salmon (1986, pp111-113); Braun (1998, p565).

(e.g. Jeshion 2010). Nonetheless, the idea of singular thought will occur repeatedly through this thesis, so it is useful to indicate the distinction between singular and non-singular thought about objects.

There tends to be agreement at least that the central cases of singular thought are attitudes that would be expressed using sentences containing referring *proper names*, and attitudes that would be expressed using sentences containing referring *demonstrative terms* (such as 'that').[7]

The disagreement arises over what characterises these central cases as *singular*, and therefore over which further cases also count as singular thought. Hawthorne and Manley (2012, pp16-19) provide a useful summary of the principal means of distinguishing singular from non-singular thought. These include:

i *Reference determination*: Non-singular thoughts have their reference fixed *satisfactionally*, i.e. the thought is about whatever fulfils some description. Whereas singular thoughts have their reference fixed relationally, i.e. their reference is fixed by a relation between the thinker and an object in the world. For example, I can think that James is wearing a striped jumper in at least two ways. One of these I would express (6), the other (7):

(6) The tallest logician in the room is wearing a striped jumper.

(7) He is wearing a striped jumper.

In the (non-singular) belief I express (6), the referent is determined by James satisfying the description 'the tallest logician in the room'. In the counterfactual situation where some taller logician is, unknown to me, hiding in the room but not wearing a striped jumper, my belief would turn out to be false. In the (singular) belief I express (7), the referent is determined in virtue of

[7]Note that the puzzles for the first-pass account of belief can be replicated using attitude-reports containing demonstratives rather than proper names. See p13 for the 'Illustrious' example illustrating this point.

some relation I stand in to James — for example, currently visually attending him. We can suppose that even if I am mistaken in my belief that he is the tallest logician in the room, the belief I express (7) is about James.

ii *Content*: Singular thoughts are relations to singular propositions. Singular propositions are the kind of proposition that would be expressed by utterances containing referring noun phrases. So (assuming a non-referential semantics for definite descriptions), the type of content expressed by utterances of (6) is non-singular, and the type of content expressed by utterances of (7) is singular. I have a singular thought when I stand in an attitude-relation to the second type of proposition.

iii *Type of mental representation*: Singular thoughts deploy a special kind of mental representation. Any thought deploying that kind of mental representation counts as singular. So my belief expressed (7) uses this special kind of mental representation, but my belief expressed (6) does not.

There is room for combining these kinds of distinction. For example, one might suppose the special singular mental representations have their reference fixed relationally rather than satisfactionally, and that beliefs using these mental representations are relations to singular propositions.[8]

Depending on which of these distinctions is taken as fundamental, and which incidental, we get different accounts of what can count as singular thought. For example, we might wonder whether the belief Charlie (sincerely) expresses with (8) is a singular thought.

(8) Father Christmas likes sherry.

If the crucial distinction between singular and non-singular thought is in the kind of proposition expressed, assuming for the sake of argument a Russellian account

[8] At least where those beliefs successfully refer.

of propositions, one might well think that the belief expressed (8) cannot be a genuine relation to a proposition because there is no such thing as Father Christmas. Therefore, it isn't a relation to a singular proposition, and so isn't a singular thought. But if the crucial distinction is in the kind of mental representation, then one might think that Charlie can have the relevant kind of mental representation whether or not (8) expresses a proposition, and so the belief Charlie expresses (8) is a singular thought just if he has that kind of mental representation.

A further recurring question about singular thought is: what are the conditions on having a singular thought? Some suppose that singular thought requires some kind of acquaintance with the object thought about (e.g. Evans 1982; Bach 1994 [1984]). Others suppose that singular thought comes cheap, and can be achieved (for example) by using a 'dthat' operator in thought, converting a description to a demonstrative (Kaplan 1996 [1975]). And others argue that singular thought neither requires acquaintance, nor comes as cheap as simply using a dthat operator (e.g. Jeshion 2010).

This is not the place to launch a full investigation of these questions about singular thought. They are mentioned here because they shape the landscape within which this thesis is situated.

1.1.3 MOPs

In 1.1.1, I suggested that *modes of presentation* do apparently valuable work in disparate theories of belief. But this does not much help us understand the thing doing that work. Instead, there are many different proposals about what modes of presentation are, and what their characteristics are. They are said to be descriptions (e.g. Searle 1958), terms in the language of thought (e.g. Fodor 1992), objects in a "third realm" (e.g. Frege 1956 [1918]), and mental files (e.g. Recanati 2012). And some claim that modes of presentation are public, i.e. sharable between individuals

(e.g. Frege 1993), others that they are private (e.g. Crimmins and Perry 1989).

I am interested in just one of these proposals: that modes of presentation are mental files. Mental files are most commonly introduced as solution to puzzles about singular thought about objects, and as such my interest in modes of presentation will generally be limited to singular modes of presentation for objects.[9] For brevity, I will call a singular mode of presentation for an object a 'MOP'.

MOPs are variously discussed in terms of 'senses', 'guises', 'ways of believing', 'modes of acquaintance', and 'modes of presentation'. Some of these terms are sometimes thought to carry more theoretical baggage than others. For example, if one uses the term 'sense', one might be thought to be committed to the idea that the thing playing the MOP role is a public rather than private object. However, I suggest that (with or without this baggage) if it is playing the MOP role, a 'sense' counts as a 'MOP'.[10] When discussing others' work on singular modes of presentation for objects, I will frequently substitute 'MOP' for 'sense', 'guise' and so on. It is convenient to use only one terminology, and generally there are few *relevant* theoretical commitments built into the use of one term or the other. Where those commitments are relevant, I avoid substituting my terminology of MOPs for the terminology used in the original work.

I don't presume that there is anything special about MOPs compared to other modes of presentation. A MOP is simply a mode of presentation for an object deployed in singular thoughts about that object. It is a further question whether MOPs differ from other modes of presentation, except in being deployed in singular

[9] I use 'object' in a narrower sense than some. In the logical sense of 'object', an object is simply anything within the range of first-order quantifiers. Here I use 'object' to discuss the subjects of singular thought, such as material objects, and perhaps abstracta such as events, times and numbers. I will follow the mental files literature in focussing on singular thought about material objects.

[10] It is helpful for me to talk both of 'MOPs' and the 'MOP role'. Compare this way of talking with talk about *water*. Sometimes, one wishes to simply talk about water. Other times, one wants to discuss the 'water role', for example when investigating whether the water role on Twin Earth is played by H_2O or XYZ (see Putnam 1975). I will say more about the MOP role below.

thought about objects. I also don't assume that the problems that arise for singular thought about objects do not arise for thought about properties or relations. In 8.2.1, I will discuss extending the mental files account of singular modes of presentation for objects to accommodate other kinds of thought. However, in order to engage fully with existing work on mental files, my primary focus is on singular modes of presentation for objects (i.e. MOPs).

In 1.1.1, I discussed modes of presentation for propositions, as well as things propositions are about (for example) objects, properties and relations. My interest in this thesis is on mental files — which is a theory of MOPs rather than modes of presentation for propositions. Nonetheless, mental files should still be of interest to those who analyse belief as a ternary relation between a thinker, a coarse-grained proposition, and a mode of presentation for that proposition. This is because (assuming a coarse-grained account of propositions) we need an account of modes of presentation for what propositions are about as well as for propositions. Except in recherché cases, one has attitudes towards propositions in virtue of being able to think about the things those propositions concern.[11] And the reason Thales has two modes of presentation for the proposition *Venus is bright* is because he has two MOPs for *Venus*.[12]

We also need an account of modes of presentation for what propositions are about in order for modes of presentation to play their role explaining practical and theoretical reasoning.[13] Thales assents to (1) and (5), and so believes the propositions *Venus is bright* and *Venus is large*. But we cannot explain his failure to infer that *something is both bright and large* just in virtue of his having different

[11] (See Salmon 1986, p108). A thinker may have an attitude towards a proposition without being able to think about the things that proposition concerns by having beliefs they would express with a phrase like "Mary's favourite proposition".

[12] Though Thales might have had two modes of presentation for the proposition in virtue of having two modes of presentation for the property *brightness*.

[13] This point may have been neglected because of the typical focus just on cases of belief in contradictions.

modes of presentation for those propositions — the proposition that *Venus is bright* is different from the proposition that *Venus is large* , and so it is in itself uninteresting that Thales has different modes of presentation for these propositions. He would have different modes of presentation for these propositions even in the counterfactual situation where he assents to (9).

(9) Hesperus is large.

To explain Thales' reasoning, we cannot think just in terms of modes of presentation for propositions. Instead, we need to turn to his MOPs. And an account of MOPs will inform answers to a further question: what is the relationship between modes of presentation for propositions, and modes of presentation for the things propositions are about? I suggest that there is therefore value in investigating MOPs even if we take belief to be a three-place relation between thinkers, coarse-grained propositions, and modes of presentation for coarse-grained propositions.

The constraint on a theory of MOPs is their role in explaining the kinds of puzzling examples given above, i.e. in explaining *Frege-cases*.

Frege-cases A Frege-case is a case in which:

i. Describing propositional attitudes as a binary relation between a thinker and a coarse-grained proposition would lead to ascribing a rational thinker *prima facie* irrational contradictory beliefs, or *prima facie* irrationality in her theoretical and practical reasoning.

ii. A natural way of explaining the thinker's mistake is to make a claim of the following sort: "she does not realise that a=b".

Clause ii limits Frege-cases to where the focus should be on thoughts about *objects*. It means, for example, that we don't count as a Frege-case an example where we explain

the thinker's mistake by saying (for example) "she doesn't realise that couches are sofas" (see Burge 2007 [1979]). Nor do we count cases explained by saying "that's a very long piece of reasoning that the thinker hasn't had time to complete yet", nor cases we explain by saying "the thinker has views about logic which make her think it is acceptable to assert simultaneously that P and $\neg P$".

As I define them, Frege-cases may arise from a puzzling *presence* of beliefs — for example, beliefs in the propositions P and $\neg P$. They also arise from a puzzling *absence* of motivation or ability to form inferences. Suppose a thinker is disposed to assent to both (1) and (5). We suppose there is a Frege-case when the thinker *fails* to be able immediately to infer that *something is bright and large*.

Crucially, a theory of MOPs is constrained by the need for MOPs to explain Frege-cases. In particular, there is the *Frege-constraint* on a theory of MOPs.

Frege-constraint Frege-cases are explained by claiming that a thinker has attitudes deploying more than one MOP for a single object.

Meeting the Frege-constraint is central to the role of MOPs, so if we cannot use what we posit as MOPs in this kind of explanation, then that theory of MOPs is inadequate.

Given the diverse claims made about what MOPs are, we need some way of identifying what MOPs are. We can use Frege-cases as our primary means of identifying MOPs, by asking what it is about the victim of a Frege-case or her attitudes that results in her *prima facie* irrational attitudes and reasoning. And we can use Frege-cases to test hypotheses about what distinguishes one MOP from another. Being able to use Frege-cases in this way is convenient, because there is disagreement about what role MOPs play in other phenomena (for example, whether they have a place in a correct account of the semantics of natural language). Moreover, MOPs were introduced in response to Frege-cases, making Frege-cases the canonical means of investigating MOPs.

I have already discussed one kind of Frege-case — in 1.1.1, I discussed a Frege-case involving attitudes a thinker would express using distinct proper names. Two further Frege-cases will be referred to frequently.

> **Paderewski**:[14] Peter is a rational, reflective, competent English speaker. He has heard of a politician called 'Paderewski', and a musician called 'Paderewski', but does not realise that the politician and musician are the same person. In contexts where he takes 'the musician' to be under discussion, he sincerely and reflectively assents to (10).
>
> (10) Paderewski had musical talent.
>
> However, Peter is prejudiced against politicians, and in contexts where he takes 'the politician' to be under discussion, he sincerely and reflectively assents to (10′), and denies (10).
>
> (10′) It is not the case that Paderewski had musical talent.

The Paderewski case is interesting because it involves beliefs that Peter expresses using what is apparently a single name, referring to a single individual.[15] It therefore disrupts theories that attempt to explain Frege-cases by emphasising that the problematic attitudes are expressed using different names.

A second important Frege-case uses demonstratives rather than proper names.

> **Illustrious**:[16] Standing on shore, Tony sees the aircraft carrier HMS Illustrious, partly obscured by several buildings. He makes out the ensign on the stern, and pointing at the stern sincerely utters (11).

[14] Case adapted from Kripke (1979).

[15] There are many sensitive issues around what counts as a single name or a single word — for example, whether 'Littleton' and 'Lyttelton' count as different names or orthographic variations on a single name (see e.g. Kaplan 1990, 2011; Hawthorne and Lepore 2011). In this thesis, I will largely put such issues aside. However, I will assume that 'Anne Hathaway' is in at least some important sense the same name when it is used to name Shakespeare's wife and an actress who has played Catwoman. It is this sense that I will be using when I talk about words or names being 'the same'.

[16] Case adapted from Perry (1977, p483).

(11) That ship is British.

Not realising how big aircraft carriers are, Tony takes the bow and stern he sees to be the bow and stern of different ships. Knowing there are few British aircraft carriers, he infers the belief he sincerely expresses with (11′), whilst pointing at the bow of the ship.

(11′) It is not the case that that ship is British.

There are ways of responding to Frege-cases without positing MOPs.[17] But as my objective is to explore the idea of mental files rather than evaluate the more general claim that there are MOPs, I will assume that there are MOPs.

1.1.4 A further use for MOPs

The central focus of this thesis will be on the mental representations involved in propositional attitudes, rather than in the semantics of propositional attitudes ascriptions (or semantics more generally). However, one might hope that an account of MOPs may be useful in resolving problems in semantics similar to the problems about belief discussed in 1.1.1.

Suppose the meaning of a proper name is simply the object named — so the meaning of 'Gottlob Frege' is simply Gottlob Frege. Call this the 'Millian' theory of names.[18] We see an example of a problem with the Millian theory of names if we remember that Thales sincerely assents to (1) but also sincerely assents to (1′). On the one hand, it appears that utterances of (12) are true and utterances of (13) are false, so there does not appear to be substitution of proper names *salva veritate* in belief contexts.

(12) Thales believes that Hesperus bright.

[17] See e.g. Fine (2007).
[18] Mill (1904 [1843], chII).

(13) Thales believes that Phosphorus is bright.

But on the other hand, according to the Millian theory names, we cannot explain this failure of substitution by appeal to differences in meaning between the proper names.

Those who suppose MOPs are an indispensable part of our account of propositional attitudes generally suppose that MOPs are equally useful in resolving puzzles for Millianism. One route is to abandon Millianism, and to suppose that the meaning of a name isn't simply the object named, but rather is (at least) a MOP for the object named.[19] Another is to suppose that the meaning of a name is *generally* the object named, but in some contexts, for example in that-clauses of attitude statements, the MOP is semantically expressed.[20] A further route is to suppose that MOPs never enter the semantics of utterances containing names, but we can explain away the puzzles by supposing that MOPs non-semantically influence our response to sentences containing proper names, for example because utterances containing names often pragmatically convey information about a thinker's MOPs.[21]

In this thesis, I will largely put aside these semantic questions in favour of investigating MOPs themselves, rather than their use in semantic theory.

1.2 Preliminaries

1.2.1 Outline

The aim of this thesis is to clarify and critique an increasingly popular suggestion — that thinkers use *mental files*. In chapter two, I introduce the idea of mental files, and highlight pressing questions about synchronic and diachronic individuation condi-

[19] E.g. Frege 1948 [1892]; Forbes 1990.
[20] E.g. Richard (1990).
[21] E.g. Salmon (1986).

tions on files. In chapters three and four,[22] I explore whether *de jure coreference* can be used to give synchronic individuation conditions on mental files. I explore existing characterisations of de jure coreference before presenting my own, but conclude that de jure coreference does not give a useful account of the synchronic individuation conditions on files. In chapter five, I consider the proposal that thinkers must sometimes *trade on the coreference* of their mental representations, and argue that we can use trading on coreference to give synchronic individuation conditions on files. In chapter six, I bring together the account of files developed so far, compare it to the most developed theory of mental files published to date, and defend my account from the objection that it is circular. In chapter seven, I explore routes for giving diachronic individuation conditions on mental files. In my concluding chapter, I distinguish the core account of files from the idea that the file metaphor should be taken seriously. I suggest that investigating the consequences of the core account shows that the file metaphor is unhelpful, and I outline reasons to exercise caution when using 'files' terminology.

Before embarking on my discussion of mental files, it is helpful to clarify some of the assumptions and terminology I will be working with.

1.2.2 Assumptions and terminology

Mental representations I assume there is a value in talking of thinkers having *mental representations*. Mental representations are mental particulars representing things (such as objects or properties or relations), or representing the world as being in a certain way.[23] Having an explicit attitude towards P is, in part, a matter of having a mental representation that P. One might imagine having a belief that P is

[22] Parts of these chapters form Goodsell (forthcoming).

[23] On some theories of propositions, the belief-representation Charlie would express "Father Christmas likes sherry" fails to express a proposition. Nonetheless, Charlie still represents the world as being in a certain way.

a matter of having a mental representation that P in one's belief box, or associating that representation with some other kind of belief marker. Thinking or reasoning is the result of operations over mental representations, and having an implicit attitude is, in part, a matter of being disposed to form a particular mental representation.

In most of this thesis, I will take for granted not only that there are mental representations, but that we can talk of them persisting through time. In later chapters, I will call into question our understanding of what the criteria are for sameness and difference of mental representation across time — but to facilitate early discussion I will assume an intuitive understanding of what it is to have a persisting mental representation.

Not all mental representations which represent the world as being in a certain way realise propositional attitudes. I assume that thinkers have representations throughout their cognitive system, but that belief-representations are associated with higher levels of thought. So we might suppose that a frog's actions are guided by its representations of its environment, but these representations are belief-like whilst falling short of being belief-representations. And we can suppose that if a thinker is knowingly subject to a Müller-Lyer illusion, she will have a belief representation that she is seeing lines of the same length, but she will also have *another* representation, a visual representation, that the lines are different lengths. I also allow that mental representations need not be conscious. For example, one might think that a blind-sighted patient has no phenomenal consciousness of their visual representations.

I take for granted that mental representation is concatenative, i.e. that a mental representation that Fa is constituted by components representing F and a, and the final mental representation preserves those components (van Gelder 1990, p360).[24] This assumption is implicit or explicit in much that it is written about mental rep-

[24] I use the following labelling convention: Predicates F, $G\ldots$ may be predicated of objects a, $b\ldots$. Unless otherwise stated, F and G may have the same extension, and it may be the case that $a = b$.

resentation, only seriously challenged by connectionist accounts of representation.[25] Within the context of a mental files theory, this assumption is appropriate. If a mental representation of a state of affairs is in a mental file, an important part of that representation's content is determined by its membership of a file. The representation can be thought of has having at least two components, 1) the file, and 2) the rest of the representation. The complete representation is made by combining these elements, and each element is preserved within the final representation.

Information and records Many writing on mental files discuss *information* as well as propositional attitudes. 'Information' is potentially a theoretically loaded term. For example, 'information' is treated by some as factive. However, there is no consistent set of theoretical commitments associated with the term, and generally the commitments are irrelevant to the questions at hand. When considering the work of others, I take 'information' simply as an alternative way of talking about mental representations of states of affairs, unless these theoretical commitments are in play.

I introduce the term 'record', to discuss those mental representations representing that the world is a certain way, associated with some attitude (or some attitude-like sub-attitude state). Introducing a new term allows me to generalise across different accounts of what mental representations are, to abstract away from theoretical commitments associated with the term 'information', and to distinguish representations that the world is a certain way from representations of objects or properties. Where relevant, I can specify different kinds of record associated with attitudes. So if a thinker T has an non-dispositional belief that P, this implies that the thinker has a *belief-record* that P, which in turn implies that T has a record that P.

In everyday speech, we switch between using 'belief' as a term for a kind of men-

[25] And only by those which argue that the whole story of mental representation can be given in connectionist terms. Contrast these with theories arguing for concatenative mental representations realised by a connectionist implementation (see e.g. Fodor and Pylyshyn 1988).

tal representation, and as a term for a relation between a thinker and a proposition. Similarly, I am relatively relaxed in how I use the term 'belief', though I reserve 'records' for mental representations realising attitudes or attitude-like states. Because records are the realisation of attitudes, I talk of 'records' being the product of inferences, just as I talk of 'beliefs' being the product of inferences.

The Simplifying Assumption I often make a provisional *simplifying assumption* that all records are 'single-reference'. Single-reference records contrast with 'multi-reference' records, which (if they have content) have content that can be stated in forms such as aRb, or aRa, or $Fa\&Fb$. In contrast, single-reference records have contents that can only be stated in the form Fa.

The simplifying assumption facilitates exposition of ideas that would otherwise be complicated by multi-reference records.[26] It is therefore useful when testing whether an idea is potentially helpful. However, it is dramatically over-simplifying, forcing the theory to neglect commonplace representations such as *Jack loves Mary* or more problematic ones such as *Hesperus is bigger than Phosphorus* and *Jack thinks himself funny*. It must always be abandoned before a proposal is built into the final theory.

Referring and referential Some tokenings of words refer. These tokenings are *referring*.[27] *Referential* tokenings are used as if they refer, but may not in fact refer.

(14) Andreas is from Denmark.

(15) Thumbelina is from Denmark.

(16) Thumbelina does not exist.

[26] In making the simplifying assumption, I follow the practice, common amongst proponents of mental files, of neglecting multi-reference records. Unlike many, I make this practice explicit.

[27] I assume that some tokenings of words and some components of mental representations refer, and that we cannot give an adequate semantics of language or thought in terms of predication alone (cf. Quine 1960).

Tokenings of the underlined NPs in (14) and (15) and (16) are referential. However, only tokenings of the underlined NPs in (14) are referring.[28]

The same distinction between referential and referring can be drawn in components of mental representations. *Referring* representations successfully refer. *Referential* representations are used as if they refer, but may not successfully refer.

The phrase 'as if they refer' is purposefully imprecise — we have an adequate intuitive understanding of what it is to use a word as if it refers, and with it an adequate understanding of what it is for a component of a mental representation to be deployed 'as if' it refers.

I will sometimes talk of records being *about* an object. Many records are about more than one thing — arguably, the record with content that *Anne, Charlotte and Emily are sisters* is about three different people. But where I am making the simplifying assumption, I can safely talk about 'the object' the record is about.

Referential-value Referential NP-tokenings or referential components of mental representations have a *referential-value*. This referential-value may be a single reference to an object or mereological sum, a plural reference, or simply reference failure. Two referential NP-tokenings or components of mental representations share a referential-value iff (a) they singly refer to the same object (or mereological sum), (b) they both plurally refer to the same plurality of objects, (c) they both fail to refer at all, or (d) they are related in such a way that they satisfy one of (a)-(c) but it is indeterminate which one.

[28] Assuming the utterances of (14) are in contexts where 'Andreas' successfully refers, and assuming that tokenings of 'Thumbelina' never successfully refer.

Chapter 2

Mental Files

2.1 Mental files

2.1.1 The metaphor

Imagine the filing system used in a philosophy department to keep records on its students. The filing system is made up of a series of folders, each labelled with the name of a student and (in ideal circumstances) only containing records about that student. Call the folder together with its label and contents the 'file'. For convenience, we can talk of the file both containing records and also partially consisting of the records it contains. When a new student joins the department, the administrator opens a new file on that student. If new records are collected about a student, they can be added to the relevant file, and if some records become obsolete these can be removed and destroyed. The file persists through these changes in what it contains. And when the administrator wants to find something out about a particular student, she can find the file labelled with that student's name, and access the relevant records.

Several organisations might store records about the same student — for example, there might be files on a student in her college office and her dentist's office, as well as in the philosophy department. Each file would collect records on the same student, but the information recorded would probably be somewhat different. Even imagining the college office and philosophy department have records with exactly the same information about the student, the college office and philosophy department would still have different files about the student.

The administrator aims to keep a single file on each student.[1] However, things might go wrong. A sloppy administrator might create two files on a single student,

[1] At least for the filing system I am imagining. There is nothing in the metaphor itself that is committed to an ideal of a 1:1 mapping from files to individuals. We can imagine a department that keeps a separate file for each student on a course, resulting in several files on a single student if she takes several courses. However, following normal practice, I start by supposing that the ideal is a 1:1 mapping from files to individuals.

and file some records pertaining to that student in one file, and some in the other. Or she might open only one file on two students, and file records pertaining to both students together. Inaccurate records might get added to a file, or records from one student might get added to another student's file. All these mistakes will have repercussions when the administrator wants to find out something about a student — accessing the relevant file may result in the administrator only having use of some of the records collected about the student, or may lead the administrator to use records pertaining to a different student.

2.1.2 Mental files

This is the kind of picture held in mind by those who suppose that we can describe thought about individuals in terms of 'dossiers' or 'mental files'.[2] The increasingly popular idea is that a useful analogy can be drawn between how the mind handles information about the objects it encounters, and the way the philosophy department handles information about its students. These similarities license us to claim that there are *mental files* implementing the MOP role.

There are numerous examples, but these are typical:

> When we receive what we take to be *de re* information which we have interest in retaining, our operating system may create a locus, or dossier, where such information is held; and any further information which we take to be about the same object can be filed along with the information about it we already possess... The role of a name is to identify a file for a particular object — as I shall put it, we use names to "label" dossiers. In sum, then, on coming across a new name, one which is taken to stand for some particular individual, the system creates a dossier labelled with that name and puts those classified conditions into it which are associated with the name.
>
> (Forbes 1990, p538)

Mental files are files on individuals that we may or may not have directly

[2]Early examples include Lockwood (1971), Evans (1985 [1973]), Strawson (1974), Perry (1980). The fullest published account given in terms of 'mental files' to date is Recanati (2012). Other examples will be cited in what follows.

> perceived... A single mental file is a repository of information that an agent takes to be about a particular individual. An agent's set of mental files partly constitutes her perspective on the world insofar as the individual mental files capture the agent's way of individuating and identifying objects, and the objects she has mental files on are the objects that are available for her to think about... [M]ental files are typically labeled with mental names that serve as long-term representations of the individual that the file is about. We think about the individual that the file is about by thinking with the mental name, and we use mental names as our mode of accessing the file contents.
>
> (Jeshion 2009, p393)

> When a person understands a name, they have a dossier that includes everything they know about, or believe to be true about, the referent of the name. When you learn a new name, you create a dossier. Every time you come to believe something new about the bearer of the name, you add this piece of information to the dossier. If you change your mind about something, you delete a piece of information (or misinformation) from the dossier.
>
> (Larson and Segal 1995, pp188-190)

These examples are typical in describing mental files as collecting together records, in accordance with the thinker's understanding of the sameness and difference of the objects those records are about. They are also typical of all but the earliest uses of file-talk, in that although the account of files is schematic, 'files' seem to be more than just a metaphorical way of talking about something else. Rather, the impression is that mental files will be a central component in an account of the mental representation of objects.

The literature which uses the idea of 'mental files' is relatively uncoordinated. Mental files are appealed to in answers to a variety of questions. These different uses of files will be discussed in more detail in what follows, but they include playing a role in: an account of the analysis of identity judgements (e.g. Strawson 1974); the appearance of 'de jure' coreference (e.g. Lawlor 2001); the semantics of proper names (e.g. Forbes 1989, 1990); the conditions on singular thought (e.g. Dickie 2010; Jeshion 2010); the meta-semantics of proper names (e.g. Evans 1985 [1973]; Dickie

2011); and the semantics of proper names (e.g. Forbes 1989, 1990). However, all who use files talk agree or assume that mental files play the MOP role.[3]

Adding to the lack of coordination, those using the terminology of 'mental files' rarely compare their use of the idea to others. They may orient their theory by pointing to the developing habit of appealing to files,[4] but this does not translate into discussion of the different ways mental file accounts may go.[5]

This means that there is no established 'theory of mental files' to evaluate. Rather, there are many rough sketches, overlapping only in places, but agreeing on certain key features of mental files. This introduces a primary task of this thesis — to fill out these sketches and give a theory of mental files. Once we have got a clearer understanding of what it is to claim that thinkers use mental files, it will be possible to assess the value of this claim.

Discussing the idea there are mental files will involve discussing records. If mental files are associated with singular thought, it may be that many of a thinker's records are not in files. However, my interest is in those records which are in files. For convenience, except where I explicitly broaden the scope of my discussion, I will restrict my discussion of records to records which are in mental files. I will take that restriction as read, discussing just 'records' when, strictly speaking, I should discuss 'records in a mental file'.

2.1.3 The core account of files

Before I can start developing a theory of files, I need to identify what it is that the overlapping sketches agree on — this will be the core account of files. I will use this core account as guide when filling in a theory of files.

[3]See 1.1.3. One exception is Grice (1975 [1969]) who discusses files exclusively in relation to descriptive thought.

[4]E.g. Crane (2011, p37).

[5]Again, Recanati (2012) is the exception — he does discuss some alternatives to his way of developing the mental file idea.

Five slogans form the core of the mental file account:

i The MOP role is played by files.

ii If records are in a single mental file, then they are treated as records about about the same object.

iii Files are mental particulars.

iv Files fix the reference of records in the file.

v Two records treated as about the same object are members of the same mental file.

As they stand, these slogans are very imprecise. The first task is to identify what precisely these slogans mean.

MOP role played by files The first slogan is that the MOP role is played by files. Thinkers maintain their files in accordance with their understanding of the identity and difference of objects. So when a thinker finds herself in a Frege-case, she has two mental files, containing records on a single object. Thales (see p2) has opened two mental files on the planet Venus — one he associates with the name 'Hesperus', and one he associates with the name 'Phosphorus'. Or, in the Paderewski example (see p13), Peter opens two files on Paderewski, both of which associates with a single name 'Paderewski' — but he adds records 'about the pianist' to one file, and records 'about the politician' to the other. In each case, the thinker fails to recognise that records in different files are records about a single individual. The thinker believes contradictory coarse-grained propositions about the object, but this is not irrational because the corresponding records deploy different mental files.[6]

The first part of the core account implies that thinkers cannot find themselves in Frege-cases as a result of having just one mental file, but taking themselves as

[6]I discuss what is involved in deploying a file in 2.2.1.

having two. If the MOP role is played by files, the Frege-constraint[7] means that in order for a thinker to be in a Frege-case, she must have two mental files on the object. It cannot be enough that the thinker has just one mental file but supposes herself to have two.

Claiming that the MOP role is played by files does not mean that all files are MOPs. Files may play other roles, as I discuss in 2.2.2. However, any adequate account of files must allow them to play the MOP role.

Records in the same file are treated as records about the same object The second slogan is that holding records in a mental file (on an object)[8] is a sufficient condition on the thinker treating those records as about the same object.[9] As much of this thesis will be spent discussing the various ways one might treat records as being about the same object, for now, I rely on an intuitive understanding of what it is to treat records as about the same object. However, it is worth noting that the sense under investigation is a relatively loose sense of what it is to treat records as about the same object. File theorists allow for mental files 'about' subjects that are known not to exist,[10] so I must be able to treat my records 'about' Thumbelina as about the same object, just as I can treat my records about Gottlob Frege as about the same object, even though I know that my records 'about' Thumbelina are about nothing at all.

However, there are sensitive cross-temporal issues that should be taken into account from the start. We can imagine a thinker having two records r and s in the

[7] See p12.

[8] For now, I take this qualification as read. Unless I state otherwise, assume the files under discussion are files about or purporting to be about objects.

[9] I assume that the object is a single object. Azzouni discusses files on plurals, for example files on the Young Hegelians, rejecting the "metaphysically awful" (2011, p50) suggestion that this is a file on a set or a mereological sum. However, I will continue to talk as if files are always about single objects. I have fewer qualms than Azzouni about discussing the Young Hegelians as a mereological sum, and from an expositional perspective it is simpler to talk of files on single objects. If necessary, it should be straightforward to extend my account to files on plurals.

[10] E.g. (Segal 2001, p553), see also 2.2.2.

same persisting mental file F, but due to the thinker changing her mind or being forgetful, r and s are never in F at the same time. We might wonder to what extent a thinker could treat r and s as about the same object. To resolve this problem, it is helpful to start distinguishing temporal stages of records and mental files. We should say if r_t and s_t are in the same file-stage (are *costaged*), then at time t the thinker treats those record-stages as about the same object.[11]

This statement of the slogan makes the simplifying assumption, that all records are single-reference.[12] Once we allow that records might represent relations between different objects, we can't require that a thinker treat all costaged records as about the same thing. Instead, we should say that if r_t and s_t are in the same file-stage, then r_t contains c_{r_t}, which is a stage of a record-component and s_t contains a stage of another component c_{s_t} , and at t the thinker treats c_{r_t} and c_{s_t} as about the same thing.

Files are mental particulars The third slogan of the core account of mental files is that mental files are concrete mental particulars. Files are not abstract, or shared between individuals. Rather, they belong to particular thinkers. Two thinkers may have files containing records with identical coarse-grained content, but this does not mean that they share a mental file.

Files fix the reference of records in the file The fourth slogan is that mental files fix the reference of records in the files. Files are sometimes introduced explicitly to give an account of reference-fixing (e.g. Evans 1985 [1973]; Dickie 2011). In these discussions, the reference-fixing role of mental files is made particularly apparent. But the idea is evident in the rest of the mental files literature. There is a long

[11] I use the following labelling convention: Thinker T possesses various mental files, F, G, H, etc. Files contain records, r, s, u, etc. Where only one time needs mentioning, I label it t. Otherwise, I use t_1, t_2, etc. t_1 precedes t_2, and t_2 precedes t_3. t may or may not be identical to t'. The t_1 stage of file F is F_{t_1}. The t_1 stage of record r is r_{t_1}.

[12] See p19.

tradition (starting with Frege 1948 [1892]) of treating MOPs as presenting or determining a referent. So files are often assumed to determine the reference of thoughts deploying that file in virtue of being MOPs (e.g. Forbes 1989, p85; Recanati 2012, pVIII). Where the reference-fixing role of files is not made explicit by writers it is implicit in pictures drawn of files that imply that all the records in a file must be about a single thing (e.g. Larson and Segal 1995), even if the records have their causal origin in different objects.

To ease exposition, make the simplifying assumption that all records are single-reference.[13] There are two ways we might think that record r (contained in file F) is 'about' object *o*. One is that r is derived as the result of a causal interaction with an object *o*. The other is that we check whether r represents the world correctly by checking the facts about *o*. We have reason to suppose that these might come apart.

> **Example I:** I see a stranger, and mistake her for Julia, an acquaintance of mine. As a result of seeing the stranger, I form a belief I express with (17).
>
> (17) Julia was drinking a glass of wine.

It is natural to say that the belief expressed (17) should be checked for accuracy against the facts about Julia, rather than about the stranger. And this would mean that the record realising that belief is about Julia in the second sense. Nonetheless, the record causally derives from my interaction with the stranger, and so is 'about' the stranger in at least the first sense. In future, I limit myself to using 'about' in just the second sense — I will discuss 'causal origin' when the first sense is under discussion.

The picture is that records in a mental file get their reference fixed collectively, in virtue of their file membership.[14] At first pass, this means that whatever the

[13] See p19.

[14] I will discuss the options for how reference is fixed in a little more detail in 2.2.2, but as a

causal-origin of records in a file, all the records a single file share referential-value. In other words, they are all either about the same thing, or all fail to be about anything at all.

However, this only does as a first pass account. We may want to allow files, over time, to change referential-value (for example, because the dominant causal source of information in the file changes). So we cannot simply assume that r_{t_1}, which is in F_{t_1} at t_1, has the same referential-value as s_{t_2}, which is in F_{t_2}. We can say that F_{t_1} always delivers a single referential-value to the record-stages within F_{t_1}, but there is no guarantee that F_{t_2} will deliver the same referential-value to the record-stages F_{t_2}, even though F_{t_2} is a later stage of F_{t_1}.

So at a second pass, we might say that the reference-fixing role of files means that whatever the causal origin of record-stages in F_{t_1}, they share referential-value.

Finally, we need to drop the simplifying assumption, and remember that not all records are single-reference. This leaves us with the claim that the reference-fixing role of files means that if record-stages r_t and s_t are members of file-stage F_t, each r_t and s_t contain component-stages c_{r_t} and c_{s_t} respectively, and the referential-value of c_{r_t} and c_{s_t} is determined by r_t and s_t's membership of F_t, and the referential-value of c_{r_t} and c_{s_t} is the same. It is component-stages c_{r_t} and c_{s_t} that the thinker treats as about the same (see part ii of the core account). And we can say that the file is about whatever it is c_{r_t} and c_{s_t} are about.

Two records treated as about the same object are members of the same mental file All file-theorists agree that if records are in a single file, then the thinker treats those records as about the same object. But many writing on mental files assume that if a thinker treats records as about the same object, then those records belong to the same mental file (e.g. Strawson 1974; Dickie 2010; Jeshion

holding theory suppose that the reference of the file is the dominant causal source of belief-records in the file (see Evans 1985 [1973]).

2010). In other words, files are maintained in accordance to an ideal of a 1:1 mapping of files to objects thought about.

However, we should not accept that whenever records are treated as about the same object, they are in the same mental file.

There are several problems with this idea. One is that a thinker may treat records as about the same thing in virtue of an idle supposition. For example, Mary might speculate that *John Cleese is Lord Lucan*, and in doing so treat her records about Cleese as if they are about Lucan. But we would not expect her to merge her files on Cleese and Lucan.

Another relates to what happens when thinkers form identity judgements. A thinker has formed an identity judgement when she comes to judge that what she had hitherto taken to be two objects is in fact just one object.[15] Suppose Jack forms the identity judgement that *Mars is Saturn*. If Jack formed such a belief, we would expect him to start treating his records about Mars and Saturn as about the same object. We would see this manifested in Jack forming various inferences (for example, that *Mars has rings*), and questioning beliefs he now supposes to be inconsistent (such as that *Mars is small* and that *Saturn is large*). However, at least when Jack has only recently come to believe that *Mars is Saturn*, we expect Jack to retain the ability to think independently of both Mars and Saturn. But this means Jack must have separate mental files on Mars and Saturn. Files fix the reference of records in the file, so if Jack had immediately merged his mental files, he would not be able to think independently of both Mars and Saturn. Rather, his thoughts 'about' Mars and Saturn would share referential-value. Because Jack is still able to think independently of Mars and Saturn, we know that Jack has not merged his files, and so is treating records in *different* mental files as about the same thing. Hence it is not the case that all records treated as about the same thing are in the

[15] Normally this judgement will be in the form $a = b$, but in a reflective thinker may take a form like *that file is about the same thing as that file*.

same mental file.

This case is evidence that we should supplement our theory of mental files with the idea that files can be *linked*.[16] A link between between files is a representation that the files are about the same thing, so a thinker can treat the linked files (and records in the linked files) as about the same thing in virtue of the link. However, linked files maintain their distinct identity and so may have different referential-values.

Supposing that suppositions of identity and identity judgements result in files being initially linked rather than merged does not preclude the files eventually merging. After all, we might think that after falsely believing that *Mars is Saturn* for long enough, Jack will lose the ability to think independently of Mars and Saturn. We might hope that we will be able to adjudicate under what circumstances this happens once we have given a satisfactory answer to the account of files' synchronic individuation conditions.[17]

Here's another case that conflicts with the claim that two records treated as about the same object are members of the same mental file: Mary is introduced to

[16] *Linking* is proposed by Lawlor (2001), Perry (2001) and Recanati (2012), among others.

[17] See 5.4. The counterpart of file-merging is splitting: if a thinker comes to suppose that she is confused, and has added records from two different objects into a single mental file, then she won't treat those records as about the same thing so must split the the existing mental file into two (though, in practice, it may be hard for her to work out which records belong in which part of the split file). Talk of merging and splitting suggests a picture where there can be fusion and fission of a persisting object. But this leads to just the same puzzles about the transitivity of identity we get for fusion and fission in personal identity cases. Take an example of fission: at t_1, there is a single file-stage F_{t_1}. But between t_1 and t_2 the file is split, resulting in file-stages G_{t_2} and H_{t_2}. If we take the language of fission seriously, we suppose that G_{t_2} and H_{t_2} are both stages of the same file as F_{t_1}. But as identity is transitive, then G_{t_2} and H_{t_2} should be stages of the same file, and it makes no sense to say that F_{t_1} underwent fission.

There are various solutions available. We might take a four-dimensional view of files (see e.g. Sider 1997) and say that F_{t_1} is a temporal part of two distinct files. G_{t_2} is another temporal part of one of these files, and H_{t_2} is a temporal part of the other. Alternatively, one might suppose that files do not persist through fusion and fission, and merging and splitting operations result in the creation of new mental files. We might hope that assessing the relative merits of these proposals will be easier with a clearer understanding of the diachronic indivdiuation conditions on files (see 7.3.4). For now, I will remain neutral as to whether 'merging' results in a merged file which is numerically identical with its antecedents, or whether 'merging' results in a new file containing records which are descendants of records in two defunct files.

Superman in such a way that she knows from first hearing about Clark Kent that he is Superman, but can still make easy sense of the idea that Superman might turn out not to be Clark Kent. It is natural to say that Mary treats her records about Superman as her records about Clark Kent (for example, she might infer that *Superman sometimes works as a journalist*). Nonetheless, it is also natural to say that she has two files on Superman (one associated with 'Superman', one with 'Clark Kent'). Similar cases arise when a thinker entertains the possibility that she might have been treating records as about a single object when they originate in two different objects. We can allow for this by saying that when a thinker has entertained the supposition that a single object is in fact two objects, she might maintain two linked mental files rather than a single file. Where the thinker maintains linked files, she (perhaps cautiously) treats records r and s as if there's a single object they are both about, but r and s are in distinct files.

Revising the slogan to take account of these issues, and to make the slogan compatible with the idea that there are multi-reference records, we say: if records r and s are treated as if there is some object that both r and s are about, then either there's some file that both r and s are members of, or there are two linked files that r and s are members of. There are two scenarios when we expect thinkers to form linked files (though we do not say that these scenarios always result in linked files rather than a single file). These are when (a) the thinker treats r and s as about the same object in virtue of a prior identity judgement (or supposition of identity); (b) the thinker treats r and s as if there's a single object they are both about, but has entertained the idea that that single object is in fact two objects.

The five revised slogans of mental file theory are:

I The MOP role is played by files.

II If r_t and and s_t are costaged, then r_t contains a component-stage c_{r_t} and s_t

contains a component-stage c_{s_t} , and at t the thinker treats c_{r_t} and c_{s_t} as about the same object.

III Files are mental particulars.

IV If r_t and s_t are in F_t, then r_t and s_t contain component-stages c_{r_t} and c_{s_t} respectively, and the referential-value of c_{r_t} is the referential-value of c_{s_t}, and this referential-value is determined by F_t.

V If records r and s are treated as if there is some object that both r and s are about, then either there's some file that both r and s are members of, or there are two linked files that r and s are members of.

I will assume that to count as adequate, a theory of mental files must allow that files and records display these characteristics, and that all theories which count as genuine file theories claim that files and records display these characteristics.

To this, I can add that the final and most obvious commonality across accounts of mental files is the thought that there is some useful analogy to be drawn between non-mental filing systems, such as the one described in 2.1.1, and the mental representation of objects. Any theory will count as a mental file theory if it both subscribes to these five slogans, and also draws an analogy between non-mental filing systems and the mental representation of individuals.

Identifying the core of the mental file account introduces some structure into an otherwise chaotic-seeming mental files literature. However, in light of the previous lack of coordination, it is inevitable that identifying the core account of file theory will result in a few writers being discounted as mental file theorists, even though they use the language of 'files'. For example, identifying file-accounts as subscribing to these five slogans leads to disqualifying Schroeter (e.g. 2007; 2008) from counting as a file-theorist, because even though she uses the language of 'mental files', she

argues that the file does not fix the referential-value of the records in the file.

2.2 Further details

There is much more that must be said in order to develop a complete theory of mental files. Merely claiming that files are mental particulars and that the MOP role is played by them tells us little about what these mental particulars are. We are offered the imagery of files containing records, but this alone does not help matters much. We may have a good grasp of what a paper folder is, and what spatial relations are involved in a paper folder containing paper-records. But we don't have a clear understanding of what the cognitive equivalent of a paper folder is, and we would be naive to expect that anything in the mind could be said to spatially contain attitude-records. More must be said about files and records before we can fully understand the claim that thinkers use mental files in representing objects.

2.2.1 Picturing files

One can picture or describe mental files in different ways. Each corresponds to different ways one might think of a file being deployed in thought, and different ways of understanding the mental representations that files are said to contain.

The first distinction is between *global* and *conception* versions of files. *Conception* versions say that the only records in a file are belief-records[18] and records of belief-like information that falls short of being belief. Other attitudes and thoughts are associated with the file in that they deploy it, but are no part of the file. The core-account of mental files should be adjusted to acknowledge that only *belief*-records treated as about the same thing are added to the same mental file.[19] Conception

[18]And perhaps other attitudes that are belief-like in that they represent the object as being a certain way, e.g. regrets.

[19]There's also a possible view where of all the belief-records a thinker treats as about a single object without any prior identity judgement, only a privileged selection of those belief-records are

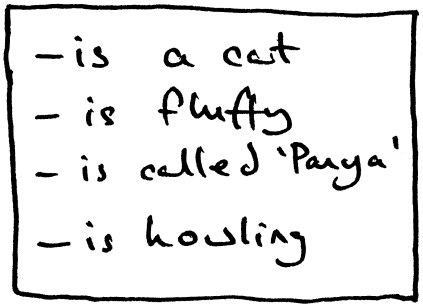
versions of files are entertained by Woodfield (1991) and Fodor (2008).

Global versions draw no distinction between records that are contained in the file and records that deploy the file without being contained in it. Instead, all records associated with the file are contained in the file. This means that files contain desire-records, speculation-records and so on. Global versions are supported by Forbes (1990) and Lawlor (2001).

It is often hard to determine whether a file theorist has a global or a conception picture in mind. Mental files are frequently introduced where the only records under consideration are belief-records or records of belief-like information. Hence, the fact that a writer only mentions belief-records may signal a commitment to a conception picture of files, or simply that they are not considering other attitudes.

Orthogonal to the distinction between conception and global pictures of files, are differences in the descriptions of themselves.

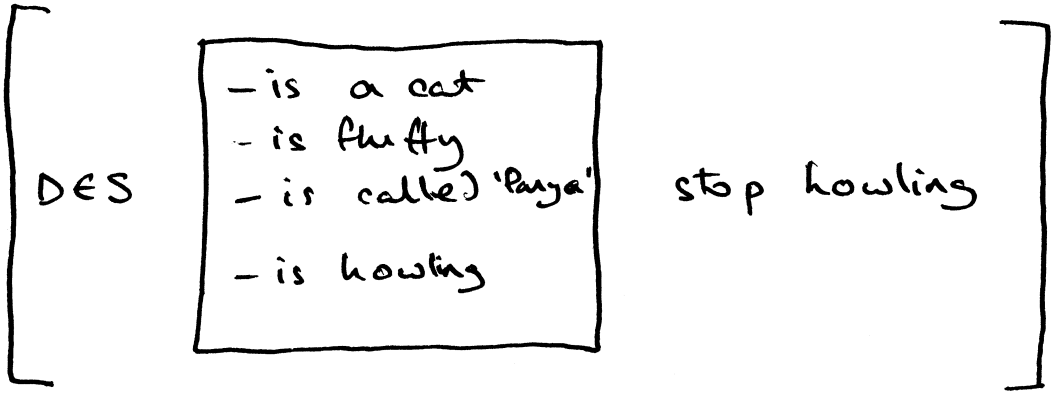
The *simple* picture of files imagines that files are non-overlapping clusters of records about a particular object. The simple picture is that usually suggested by the most thinly sketched descriptions of files (e.g. Evans 1985 [1973]; Strawson 1974). On the simple model, we might picture my file on Panya like this:



On conception versions of the simple picture, deploying a file in a thought means

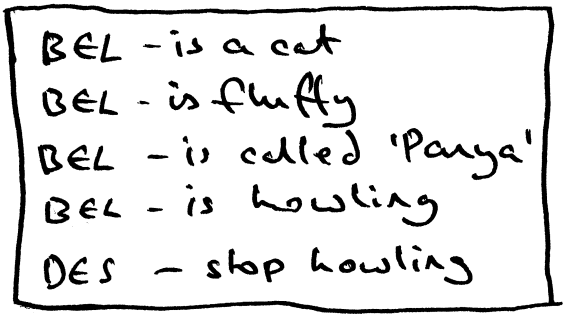
contained in the mental file. This kind of picture is generally ignored by those using files to respond to puzzles about singular thought, and conflicts with the core account of files, even rewritten to limit file-membership to belief-records. Because it conflicts with the core account, I discount this option.

either entertaining a record in a file, or using the whole file as a constituent of the thought-record, so my desire-record that *Panya stop howling* can be represented as:



However, this seems implausible. We don't suppose that all our belief-records on a single individual are deployed every time we think of that individual (Fodor 2008, pp94-98).

On global versions of the simple picture (e.g. Lawlor 2001), deploying a file in a thought is simply a matter of entertaining a record in a file.



However, whilst this is less implausible than the conception version of the simple picture, the simple picture is still untenable. On the simple picture, there is no way of deploying the file in thought other than deploying the whole file or entering a record into the file, so the simple picture cannot handle multi-reference records.

Suppose that T has two relational belief-records, *Jack loves Mary* (record r) and *Mary thinks Jack is a fool* (record s). We can suppose r is in T's *Jack*-file, and s is

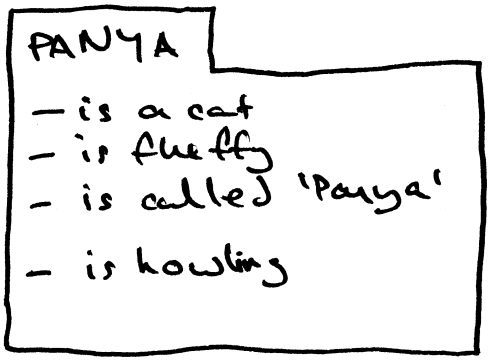
in T's *Mary*-file. The problem is explaining how r refers to Mary as well as to Jack. One option is that r is in the *Mary*-file as well as the *Jack*-file. But then a single record is in two files at once, and we have abandoned the simple picture's claim that files are non-overlapping clusters of records. The other is that r is just in the *Jack*-file, and r contains the *Mary*-file. But equally, we would have to say that s is in the *Mary*-file, and contains the *Jack*-file. But two files cannot both contain one another, so this proposal is incoherent.

An alternative might be to allow that records can contain mental names. So r is in the *Jack*-file, and contains a name for Mary. But the simple picture gives no account of how mental names can connect with files (except by being contained in the file). For a record in the *Jack*-file to contain a mental name referring to Mary, we need some account of how this name is associated with the *Mary*-file. But the simple account doesn't provide this.

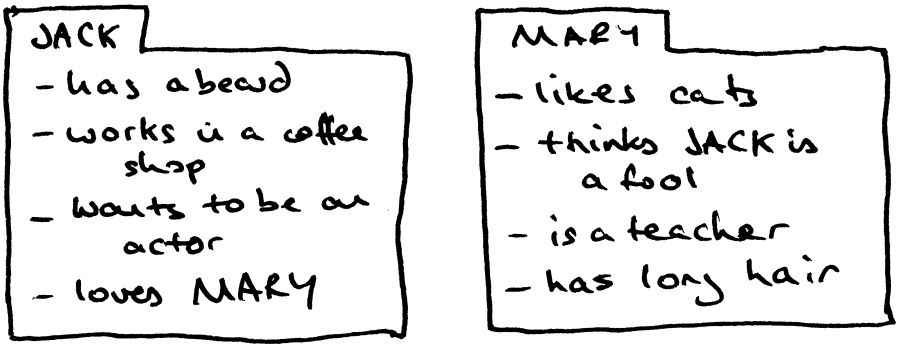
In contrast, the *label* picture of files does give an account of how mental names connect with files. The *label* picture of files is much like the *simple* picture, but is supplemented by the suggestion that the file has a label (e.g. Bach 1984; Forbes 1990; Larson and Segal 1995; Dickie 2010; Jeshion 2010). One role for the label is that it can operate as a mental name or marker for the subject of the file.

On the label view, deploying a file in a thought involves either entertaining a record in a file, or using the file's label in other records. Using the label allows the subject of the file to be thought about, without involving the whole contents of the file. The label automatically acquires its reference from the file labeled, and the thinker will treat records containing labels for the same file as about the same thing, even if those records are not in the same mental file.

Conception versions of the label picture suggest that only belief-records and records of belief-like information will be contained in the file. Records of other attitudes will deploy the file by containing the file's label.



And on both conception and global versions, we can describe relational beliefs by allowing that a file can contain a label for some other file.



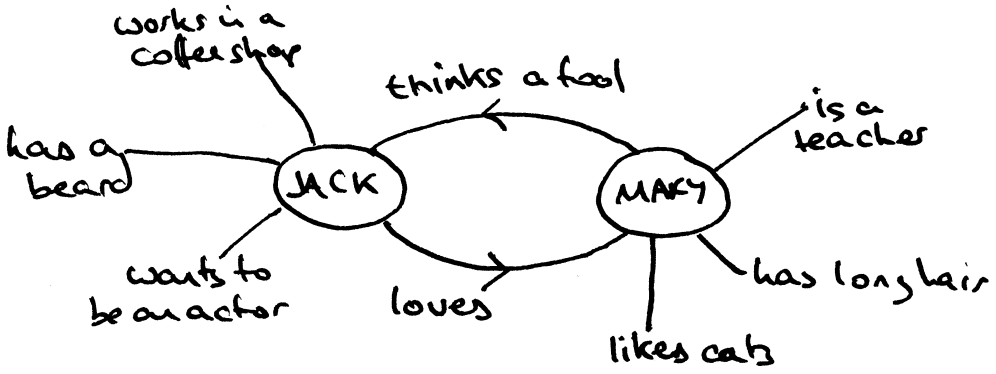
Both the simple and label pictures are *container* pictures of files. Container pictures suggest that files are bundles of records, contained *inside* the file. Translating into less spatial terms, this means that container pictures only allow for records to be in one file at once, no matter how many different objects they are about.[20] In contrast, *non-container* models loosen the analogy between mental files and non-mental files, abandoning the idea that we should picture records *inside* the file. Discussing files using the terminology of 'notions', Crimmins writes:

[20]Unless we allow nesting of files, a single piece of paper can only be inside one file at once. We might want to allow for some nesting of files, for example, I might represent Thales' belief-states by storing models of Thales' files in my file on Thales. However, none of the points I make about the container picture would be undermined by supposing that files are nested, so for ease of exposition I will assume there is no file-nesting.

> Though files have their information literally *inside* them, *containing* this information, it may be misleading to think of notions as *containing* information. Notions are *parts of* beliefs. The file analogy can lead one to get the issue of what contains what backwards.
>
> (Crimmins 1992, p87n)

Non-container models allow that multi-reference records exist in more than one file at once.

There are two principal non-container views. The *node* view treats a mental file as a network of records connected by a node (e.g. Crimmins 1992; Perry 2001).[21] Records are made up of nodes, the predicates wired to nodes, and wires between nodes.[22] Multi-reference records contain several nodes, and so might count as 'contained in' several files at once.



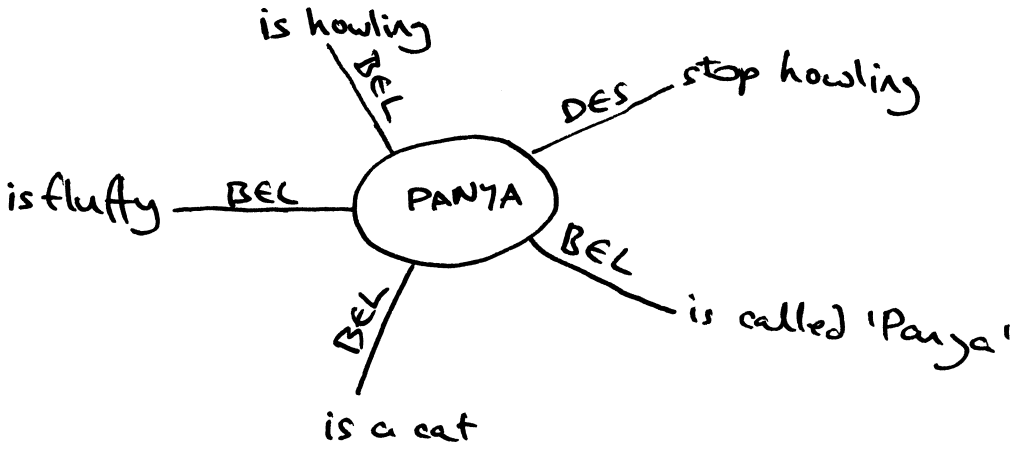
Records representing records with more than two relata are difficult to draw — but can be thought of as wires connecting multiple nodes.

On the node picture, there is just one option for what it is to deploy a file in a thought: the record deploys a file iff it contains the relevant node. One might adopt a conception version of the node picture by specifying that only records of belief and

[21] Strawson's (1974, p54) 'dot model' of identity judgements is much like the node picture. Strawson offers the dot model as an alternative to a file model. The fact that the node picture is now presented as a version of file theory is a sign of how far the file metaphor has developed.

[22] Wires between nodes are not links between nodes. Links are a representation that the nodes are about the same thing, but wires represent some non-identity relation. Murez (MS) argues that file theorists are committed to an asymmetric treatment of identity and other relations.

belief-like information count as contained in the file, but this is a conception view by fiat.[23] There is no difference in implementation between cases where the record is contained in the mental file and cases where it is merely deployed. Alternatively, we might give a global version of the node picture, allowing that any record partially constituted by the node is contained in the file. We can picture the global version something like this:



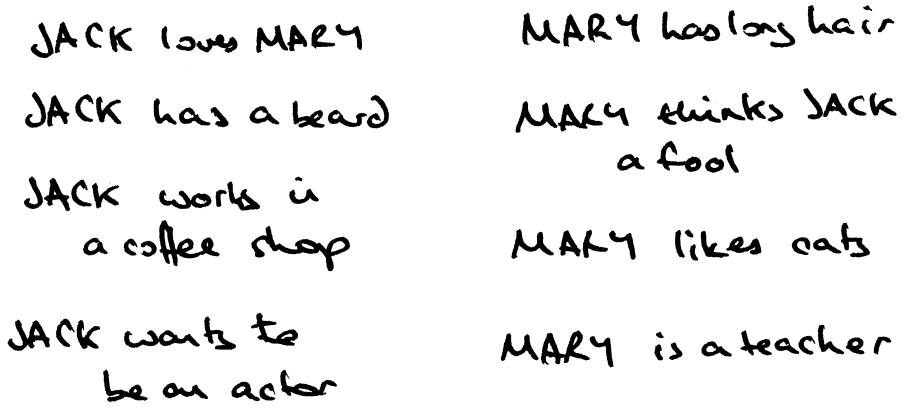
One might be discomforted by treating a file as merely a node in an evolving network of records, worrying that as the records in the file change, we lose track of the identity of the node across time. To circumvent this worry, one might adopt a *token* picture of mental files. To the best of my knowledge, there are no examples of the token picture explicitly described in terms of 'files'. However, it seems no great stretch to the node picture (which is an established version of file theory) to the token picture, suggesting that the token picture is a viable option for mental files theorists.

According to the token picture, records consist of various representational component-stages, each of which can occur only once in a single record (call these 'tokens'). Some tokens are referential. Two records deploy the same mental file, iff

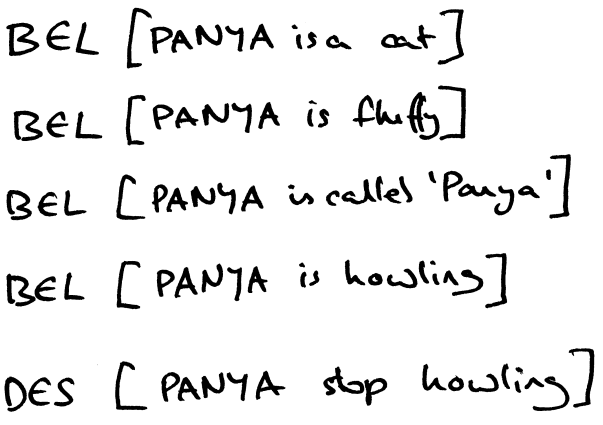
[23] A conception version by fiat may be attractive if we are primarily interested in a causal theory of reference-fixing, and if only belief-records and belief-like information have a reference-fixing role.

they contain referential tokens of the same type.[24]

On this view, there is little difficulty describing mutli-reference records. A record can contain many referential tokens, and so be in as many files as it contains referential tokens of distinct types.



As in the node picture, there is only one way a record can associated with a file — this time, in virtue of that record being partially constituted by a token of the appropriate type. We can generate a conception version of the token picture by fiat if we claim that files contain only belief-records and records of belief-like information. But on a global version, files contain all records containing tokens of the relevant type, so all records containing tokens of the relevant type are in that file. We can picture the global version something like this:



[24] I say more later about what it is for tokens to be of the same type.

These different pictures are not interchangeable. As I have begun illustrating here each picture presents a distinct way of thinking about the mental representations of objects, and as I demonstrate in 7.3.4, the choice of picture influences answers to further questions about mental files. The choice of picture is therefore not irrelevant or incidental to the file theory. Nonetheless, the normal approach is simply to present one of these pictures without defending this choice.

I have already argued that the simple picture is untenable because it cannot handle multi-reference records. And we should also reject the label picture (and any container picture). The core account of files says if r and s are treated as if there's some object that both r and s are about, then there's some file that both r and s are members of (except when r and s are treated as about the same thing because they are in linked files). Consider the cotemporal beliefs T would express (18), (19) and (20).

(18) Mary is a teacher.

(19) Jack is a fool.

(20) Jack loves Mary.

At first pass, T has record-stages r_t, s_t and u_t, recording the beliefs expressed (18), (19) and (20) respectively. T treats r_t as about the same thing as u_t, and treats s_t as about the same thing as u_t, but does not treat r_t and s_t as about the same thing. So there should be a file-stage F_t that r_t and u_t are both in, and a file-stage G_t that s_t and u_t are both in.[25] But because r_t and s_t contain no component-stages that T treats as about the same object, r_t and s_t should be in different files, hence $F_t \neq G_t$. But on container pictures like the label picture, a single record cannot be in more than one mental file. So on container pictures, T must have two record-stages representing the belief expressed (20), a record v_t in the Jack-file (containing

[25]I assume there is no Frege-case, and that T forms these beliefs about Jack and Mary without having to form an identity judgement or deploy more than one mental file for each individual.

a label for the Mary-file), and a record w_t in the Mary-file (containing a label for the Jack-file).

As many multi-reference records are treated as about the same thing when we have no reason to suppose that this is the result of linking, adopting a container picture of mental files leads to a counter-intuitive result: that there is massive repetition of content across mental representations. For example, there must be two records that *Jack loves Mary* , and three records that *Anne, Emily and Charlotte are sisters.*[26]

We might accept this as an interesting consequence of file-theory if we had any reason to suppose that the best way to model thought about individuals is in terms of records stored *inside* files. The container picture is the picture most obviously suggested by the analogy with non-mental filing systems, but this isn't reason enough to accept the counter-intuitive implications of the label picture. Other pictures are available, and we have no reason to suppose mental files have all the features suggested by the analogy with non-mental filing systems. Responding to my earlier statement of this objection, Recanati (2012, p50) is correct to point out that the point of adopting a model is to see where it leads. But that does not mean we should follow the model *wherever* it leads. When following the model requires developing an increasingly implausible theory just to rescue one interpretation of an analogy, we should step back and ask whether there are more plausible alternatives available. And in this case there are. The node and token pictures allow records to exist in more than one mental file, so on these pictures there is no need to posit duplication of information to accommodate relational beliefs.

[26]We might attempt to avoid this conclusion by adapting the label picture to allow records to exist in two files. But (depending on exactly how we implement the idea), the label view becomes functionally tantamount to the node or token view. The key thing distinguishing the label view is that it allows for records to be contained in at most one mental file.

In light of my discussion of pictures of files, there are some general points worth bearing in mind.

First, on conception versions of node and token pictures, a file is a bundle of belief-records and records of belief-like information which all contain the same node (or tokens of a particular type). On global versions of node and token pictures, a file is the bundle of records which all contain some particular node (or tokens of some particular type).

Second, a record is associated with a file in virtue of containing some component, either a node or a token of some particular type. All records associated with that file contain that node, or tokens of that type.

This means that costaged record-stages r_t and s_t will be costaged in virtue of component-stages of r_t and s_t. Suppose a thinker has records as follows: r_t represents aRb and s_t represents bRc, and r_t and s_t are associated with file-stage F_t. r_t and s_t are associated with F_t in virtue of the component-stages representing b, not the component-stages representing a or c. Thinking back to the core account of mental files, if record-stages r_t and s_t are associated with F_t in virtue of containing component-stages c_{r_t} and c_{s_t} respectively, it will be c_{r_t} and c_{s_t} which share referential-value, and c_{r_t} and c_{s_t} which are treated as about the same thing.

Third, records represent (or potentially represent) complete contents. This means that a record is not a predicate like "— *is a cat*" which can be wired to any node or associated with any referential token. Rather, the record is the predicate "— *is a cat*" complete with the node to which it is wired (or together with some referential token). In line with the file metaphor, we talk of files *containing* records, but records are associated with a file by containing a node or some referential token, and are only complete in virtue of partially consisting in that node or token. This means records in files are only complete in virtue of partially consisting in the same thing which also makes them a member of a particular file.

Fourth, whilst non-container pictures of files allow that a multi-reference record may exist in more than one file, they don't allow single-reference records to exist in more than one file. For a record to be contained in some file, it must contain the appropriate node (or a token of the appropriate type). So for a record to be contained in multiple files, it must contain multiple nodes (or tokens of multiple types). But as the content of a record is generated by the components of that record, as soon as the record contains multiple nodes (or tokens of multiple types) then the record will have content that can be represented as aRb, and will no longer be a single-reference record.

Fifth, because records can exist in more than one file, *being costaged* is not an equivalence relation.[27] Suppose T has three record-stages, r_t, s_t and u_t. r_t has the content *Jack loves Mary*, s_t has the content *Mary admires Anne Bronte* and u_t has content *Anne Bronte had two sisters*. Assuming T has just one MOP per object, then we expect r_t and s_t to be costaged in virtue of the component-stage referring to Mary. And we expect s_t and u_t to be costaged in virtue of the component-stages referring to Anne Bronte. But we don't expect r_t and u_t to be costaged at all.

Nonetheless, the picture is that records are costaged in virtue of containing a particular node, or tokens of some particular type. This means that if r_t and s_t are costaged in virtue of component-stages c_{r_t} and c_{s_t} respectively, and s_t and u_t are costaged in virtue of c_{s_t} and c_{u_t} respectively, then r_t and u_t should be costaged in virtue of component-stages c_{r_t} and c_{u_t}.

2.2.2 Uses of files

Files are supposed to be useful in accounting for various features of the mental representation of objects. Setting out the supposed role of files in accounting for these features draws out further details of the file-account.

[27] Record-stages r_t and s_t are costaged iff they are members of the same file-stage.

Confusion Confusion can be thought of as the reverse of a Frege-case. Example I is one example of confusion, and example II gives another typical kind of case.

> **Example II:** [28]Sally shares her home with two mice. Because the mice are similar looking, Sally thinks only one mouse lives in the flat, and calls it 'Reepicheep'. She forms various beliefs 'about Reepicheep', for example the belief she expresses with utterances of (21).
>
> (21) Reepicheep likes apples.

We can describe confusion using mental files. Making the simplifying assumption that all records are single-reference: when a thinker is confused, she adds information originating in more than one object to a single mental file. Files determine the referential-value of records in the file, so the reference of confused beliefs (like that expressed (21)) is determined by the file in which the beliefs are recorded.

Reference-fixing Mental file theories are generally part of a tradition that suggests the reference of a singular thought is determined by a causal relationship between the thinker and the object thought about.[29] Subsequently, files are associated with the idea that singular terms in language have their reference determined by a causal relationship between the file that the term is associated with, and the object the file is about. Usually, the idea is that the causal source of the records in a file (or the causal source of some subset of those records) determines the referential-value of the node (or tokens of the relevant type), and every record in a file has a component with that referential-value.[30]

There are different ways of filling out the details of reference-fixing. For example, one might take a *dominant causal source* view (e.g. Evans 1985 [1973]). On this

[28]Example adapted from Camp (2002).

[29]See Hawthorne and Manley (2012, pp27-35) for criticisms of the position that if T can think singularly about o, then T has mental representations caused by o.

[30]See also 2.1.3.

view, the referential-value is determined by the causal origin of the belief-records and belief-like information-records in the file. Whatever is the dominant causal origin of these records is the reference of the file — i.e. the object referred to by each record in the file in virtue of that record's membership of the file. Files can refer to a single object despite containing confused information, so long as there is a single dominant causal source of the information in the file. If no causal source is dominant, this might lead to the file failing to refer. Alternatively, the lack of dominant causal source may lead the file to refer to a mereological sum of the causal sources, or it being indeterminate what referential-value the file has.

One might supplement a dominant causal source view with further constraints. For example, Dickie (2010) adds a further condition to the dominant causal source view. The file contains a record of what kind of thing the subject of the file is. What kind of thing the subject of the file is supposed to be determines the file's 'template' — the set of possible ways for the file to develop as the thinker continues to gather information from the object. For example, a template of an ordinary physical object will include the possibility that the file will contain a record that the object moved towards the thinker. An object o is the referent of the file iff o is the dominant causal source of information in the file, *and* it is possible for things in the same category as o to fill the file's templates. Suppose o is a dot on a computer screen, but information from o is added to a file with a template appropriate to an ordinary physical object. Then the file won't be about o, because o cannot move towards the thinker (Dickie 2010).

Alternatively, one might abandon the idea of dominant causal source when making claims about the way the causal-origin of a file's records fix the file's referential-value. Recanati (2012) argues that files are 'based on' certain acquaintance relations to objects. A file may be based on a simple acquaintance relation (for example, visually attending an object), or a composite relation (for example, the relation of

visually attending an object and remembering having visually attended it in the past), or the abstract 'encyclopedia' relation, which is the relation a thinker is in to an object whenever she is in some kind of acquaintance relation to an object.[31] Recanati accepts a dominant causal source view for files based on the encyclopedia relation, but suggests that for files based on simple and composite ER-relations, the reference of the file is the unique causal origin of records gathered through the acquaintance-relation the file is based on.[32] This is not simply a proposal about how we cash out what it is for a causal source to be 'dominant'.[33] This is because Recanati (2012) makes the demanding claim that a file F, based on relation R, refers to *o* only if *o* is the unique causal source of information gathered into F via R. If there is no unique source, then the file fails to refer (Recanati 2012, p61).

File theory is associated with causal metasemantic theories. However, there is no commitment to all mental files having their reference determined by the causal source of the information in the file. I focus on files on material objects, but if we allow for files on theoretical objects, it is harder to give a sensible account of causal reference-fixing. Even considering material objects, we should allow that a thinker can have a mental file on an object that has not yet causally impacted on the thinker, as an admiral might have a file on a warship he has ordered to be built, even when he has not seen it or heard testimony about it yet.[34] And we can allow for files whose reference is fixed by some canonical description. For example, I have many records about Jack the Ripper. I treat these records as about a single object, their reference is fixed collectively, and they seem a good candidate for being contained

[31] See also Recanati (1997, pp122-130).

[32] At the heart of this proposal is an account of what it is for a file to be 'based on' some ER-relation or other. Recanati's explanation of this 'basing' relation is brief (2012, pp66-67), and one might well be dissatisfied with it. However, there is no space to explore these issues thoroughly.

[33] Evans (1985 [1973], p16) suggests that dominance may not simply be a matter of volume of information, but instead results from several factors including the thinker's interest in the object.

[34] Example from Hawthorne and Manley (2012, p28). Jeshion (2010) suggests we have files in this kind of case, and it is natural to think that these files refer, though they cannot refer in virtue of the ship *causing* the admiral's records.

in the same mental file. Nonetheless, I suppose that the reference of these records is fixed not by a causal relationship but rather by the canonical description *the actual unique perpetrator of the Whitechapel murders*.

The core account of files says that the MOP role is played by files. If a thinker has to be acquainted with o to have a singular thought about o, then these examples are cases where files are not playing the MOP role. If a thinker need not be acquainted with o to have a singular thought about o, then these cases are further examples of files playing the MOP role. Without a clearer account of the conditions on singular thought, we cannot hope to distinguish the cases definitively.

There is clearly much work to be done. Firstly, in working out the details of these metasemantic proposals,[35] and secondly, in identifying which metasemantic proposals are appropriate for which kinds of thought. However, these issues are extremely complex, and arise across various accounts of MOPs. It is impossible to give an adequate treatment of them within the confines of a thesis on mental files. Rather, in what follows, I assume that for most singular MOPs some kind of causal account of reference fixing is appropriate. I will leave the details of the causal story, and of alternative means of singular reference fixing, as a project for future research. Where the issue is relevant, I adopt the theory that the reference of a file is generally determined by the dominant causal source of information in the file, using this as a place-holder metasemantic theory.

However, even without working out the details, there are a few salient points worth bearing in mind. First, file theories are associated with a causal account of reference-fixing for singular thought, but there is nothing incoherent in the idea that some files have their reference fixed non-causally. Second, we should allow that files might fail to refer. Third, we should allow that files might change their referential-value over time — suppose we adopt a dominant causal source view and say that

[35] We need to answer questions such as: what is it for a causal source to be dominant? And what do we count as the causal source of a record?

there is reference failure when there is no dominant causal source. Imagine a case where a thinker opens a file on a, but then confuses a and b so adds records from b to the a-file. At the beginning of the file's life the dominant causal source of records in the file is a, but over time more records from b are added so first there is no dominant causal source, then the dominant causal source is b. In this case, the file will have three different referential-values — first it will refer to a, then it will have no referent, then it will refer to b.

Empty singular thought Sometimes we say that thinkers have thoughts about an object, without realising that this object does not exist, as when we say that a credulous child has thoughts about Father Christmas. And sometimes we say that thinkers have thoughts about an object they realise does not exist. One example is saying a reader has thoughts about the motivations of Heathcliff. In both these cases, we might say that the thinker experiences those thoughts as indistinguishable from or similar to standard examples of singular thought (for example, thoughts about the motivations of Emily Bronte).

If attitudes are realised by mental representations, then we need an account of the mental representations involved in thought about non-existents. A natural way of accounting for thought about non-existents is to say that thinkers can open mental files 'about' non-existents (e.g. Larson and Segal 1995, p189; Segal 2001, p553).[36] We can describe the thinker's reasoning 'about' the non-existents by saying the thinker has mental representations of these objects (even if she cannot stand in a relation to a coarse-grained proposition about them).[37] And we can explain the similarity with standard cases of singular thought by saying that the same kind of

[36] I introduced the idea that files may lack a referent in my discussion of how files fix reference.

[37] I make no commitments as to whether there is a viable account of propositions that allows for propositions about non-existents. However, if we think there is no such account, then a useful consequence of supposing that there are empty files is that we can describe 'attitudes' about non-existents just in terms of records in mental files, without having to claim such attitudes are a relation between a thinker and a proposition.

mental representation used to represent Emily Bronte is used to represent Father Christmas and Heathcliff — i.e. thinkers can have mental files on Father Christmas and Heathcliff as well as Emily Bronte. These files contain records representing the attributes the thinker believes or supposes the non-existent to have. Nonetheless, the file and its records fail to be about anything.

Constituting singular thought Some go further than claiming that the MOP role is played by files. Rather, they claim that the use of mental files constitutes the singularity of singular thought.[38] The strongest conclusion of this form is drawn by Jeshion, who claims that a thought is singular iff it deploys a mental file (2010, p130). Jeshion adopts a label picture of mental files, and claims that as only file labels act as mental names and demonstratives, only thought deploying file labels is singular.[39]

The disadvantage of Jeshion's (2010) approach to defining singular thought is that it excludes files from being deployed in non-singular thought. This excludes files from playing a role in language-understanding (see below), and means we cannot think of thinkers having files on properties.[40]

Recanati (2010, 2012) takes an alternative position. He suggests that there is a norm linking mental files with singular thought, and that one can have a singular thought-vehicle (i.e. a mental file) whilst not meeting the conditions on thinking a singular thought-content (2010, pp184-185). However, mental files also have non-singular uses, such as their use in language-understanding (2010, pp177-181, 2012, pp171-177). These uses are derived from the core singular use of mental files.

[38]This kind of claim relies on defining singular thought in terms of a particular kind of mental representation, rather than e.g. as a relation to a particular kind of proposition. See 1.1.2.

[39]Jeshion's claims are supplemented by a somewhat confusing discussion of the 'object files' posited in the psychological literature on visual attention. Jeshion suggests that there is an 'ontogenetic' connection between mental and object files, and this is the origin of the singularity of thought involving mental files. Jeshion's remarks are highly speculative, making her position difficulty to evaluate. See also 8.3.2.

[40]Fodor (2008) and Crane (2011) are among those who suggest there are files on properties. See also 8.2.1.

Language understanding The familiarity theory of definiteness proposes that we distinguish definite and indefinite NPs in terms of familiarity: a definite NP refers to an already familiar referent, whereas an indefinite NP introduces an unfamiliar referent.[41] Suppose (22) and (23) are discourse initial utterances.

(22) A mouse kept me awake last night.

(23) The mouse kept me awake last night.

(22) is entirely felicitous. However, for (23) to be felicitous, there would have to be some mouse familiar at the beginning of the discourse, perhaps because it is visible to the speaker and her audience.

Early versions of the familiarity theory of definiteness treats indefinite NPs as referring expressions. This allows for a relatively simple solution to another problem associated with indefinites — how they can act as the antecedents of anaphora, as in (24).

(24) A mouse kept me awake last night because it was scampering about the room.

However, this does not account for all the data. There are cases that raise difficulties for the strategy of treating indefinite NPs as referring expressions.

(25) Every cat caught a mouse and killed it.

In one scope-reading of (25), there is no mouse, familiar or unfamiliar, being referred to by the indefinite NP, so the indefinite NP cannot refer let alone introduce an unfamiliar referent to be the referent of the later anaphor.

Karttunen (1979 [1969]) developed a new way of resolving these problems. Karttunen's strategy was to introduce discourse referents.

[41] See Heim (1988 [1982], p298-299) for more discussion of the history of familiarity theories of definiteness.

> Let us say that the appearance of an indefinite noun phrase establishes a 'discourse referent' just in case it justifies the occurrence of a coreferential pronoun or a definite noun phrase later in the text.
>
> (Karttunen 1979 [1969], p366)

Discourse referents are not genuine referents, but act as place-holders for genuine referents. The indefinite NPs in utterances of (24) and (25) introduce discourse-referents.

Recanati (2005, 2012) suggests that when an NP introduces a discourse-referent, it results in a thinker opening a mental file.[42] That file then stores information about the referent or putative referent, gleaned from the discourse.

Continued Belief Perry (1980) suggests that we use mental files existing across time to explain what it is to have a continued belief.

Perry's idea is that it is not sufficient for continued belief that a thinker merely retain a belief towards a coarse-grained proposition P.

> **Example III:** Tony is looking at a partially occluded aircraft carrier, and takes himself to be looking at the bow and stern of different ships. Colin points towards the bow, and claims "That is the HMS Illustrious", and Tony believes him. A few minutes later, John points towards the stern and claims "That is the HMS Illustrious". Tony supposes that John is better informed than Colin, and believes John, assuming that Colin was wrong.

In III, Tony retained a single attitude (belief) towards a single coarse-grained proposition (*o is the HMS Illustrious*), but nonetheless we would not say that Tony had a

[42] Inspiration for this view comes from Heim's development of the idea of 'discourse referents' as 'files' (1988 [1982]; 1983), and Kamp's development of Karttunen's ideas using other file-like entities in his Discourse Representation Theory (e.g. Kamp 1993, 2001, but for similarities to file-theory see especially Kamp 1990). See also 8.3.2.

continued belief that *o* is the HMS Illustrious. Rather, he 'changed his mind', whilst retaining belief in a single proposition.[43]

Nor is retaining a belief towards a coarse-grained proposition clearly necesary for continued belief.

> **Example IV:** In a train travelling through Europe, Courtney retains the belief she would express (26).
>
> (26) This country is very flat.
>
> But unknown to her, during the course of the train journey, Courtney has passed through several countries.

In IV, we might say that Courtney continues to have the same belief, even though its coarse-grained content changes over the course of the journey.

Perry's suggestion is that to have a continued belief is to retain a record in a mental file. In example III, there is no continued belief because although Tony's records share coarse-grained content, they are members of different files. In example IV, there is continued belief because a record is retained in a single file.

De jure coreference Some records are about the same thing — for example, the belief-records expressed (1) and (2). But some records are about the same thing in some particularly explicit and guaranteed way — call these records *de jure coreferential*.

I will discuss de jure coreference in considerable detail in chapters 3 and 4. For now, just note that some explain de jure coreference in terms of sameness of mental file (e.g. Lawlor 2001; Recanati 2012). The suggestion is that records are de jure coreferential iff they are contained in the same mental file (Lawlor 2001), or iff they are part of the same mental file as a result of being gathered through the same acquaintance relation (Recanati 2012).

[43] See also Evans (1982, pp308-309). To make these points simply, I ignore issues of tense.

Subpersonal thought Some proponents of mental files extend the suggestion that there are files beyond mature human conscious reasoning, and suggest there are files at the level of subpersonal thought (e.g. Recanati 2012).

One reason for supposing there are subpersonal files is that psychologists studying the subpersonal mechanisms of visual attention in adults discuss 'object files' (e.g. Kahneman and Treisman 1984; Kahneman et al. 1992). These 'object files' are frequently mentioned by file-theorists (e.g. Dickie 2010; Jeshion 2010; Recanati 2012) as the subpersonal analogs of the mental files used in personal-level thought, and the shared terminology of 'files' lends itself to supposing that the idea of 'files' is useful throughout the levels of mature human thought.

Nonetheless, it is difficult to fully comprehend the suggestion that there are subpersonal mental files, or that 'object files' are subpersonal analogs of personal-level mental files, without a clearer idea of what mental files are supposed to be. But one might suppose that the core account of files suggests it is on the right track. The core account of files says that records treated as about the same object are added to the same mental file, except in cases where they are added to linked files. And there is evidence of mental representations being treated as about the same object throughout the levels of the cognitive system.[44]

Animals and infants There are also reasons to think that animals and infants have mental files. My remarks will be highly speculative, and aim to be only an indicator of how the argument could go.

First, there are suggestions that infants also deploy object files (e.g. Carey and Xu 2001) adding to the idea that there are files or analogs of files in infant and perhaps even animal thought, as well as in mature human thought.

Second, naturalistic accounts of the mind favour theories that suggest no sharp distinctions between mature human thought and animal or infant thought. Hence,

[44]E.g. early vision motor control (Pylyshyn 2003, p150).

there is a presumption in favour of ascribing animals and infants the same kind of psychological structures as mature humans. This argument is strengthened by seeing that animals and infants have the same types of mental representations that are supposed to result in files in mature humans. For example, one of the paradigm cases of singular thought in mature humans is visually attending an object. Animals and infants also visually attend objects, and subsequently are attributed singular thoughts about them. Furthermore, we suppose there are Frege-cases or cases analogous to Frege-cases, that arise for animals and infants. We can suppose a dog growls at someone it is normally friendly towards when she is wearing an extremely strong new perfume, and we explain the dog's behaviour by saying it fails to recognise someone it normally recognises. And it is just this kind of failure to recognise an object as the same that we see in Frege-cases.

If we account for singular thought and Frege-cases in mature human thought using mental files, then there is a presumption in favour of explaining the same phenomena in animal and infant thought using mental files.

Frege-cases We can allow that a thinker might have more than one record with the same coarse-grained content.[45] So because we think of files as bundles of records, there is no barrier to claiming that thinkers have two files containing records with identical coarse-grained content.[46] This allows us to use mental files in a response to Fine's (2007) objection to using MOPs in resolving Frege-cases.

Fine's argument starts by imagining

> a universe that is symmetric around the thinker's centre of vision. Whatever she sees to her left *is* and *looks* qualitatively identical to something she sees on her right (not that she conceptulizes the two sides as "left"

[45]For example, if a thinker sincerely assents to both (1) and (2) without recognising that *Hesperus is Phosphorus*.

[46]Thinkers can also have two records in the same file with identical coarse-grained content. For example, Thales might have two records with the content *Mary loves Venus* in his *Mary*-file, one he would express "Mary loves Hesperus", and one he would express "Mary loves Phosphorus".

> and "right' since that would introduce an asymmetry).
>
> (Fine 2007, p36)

He then suggests that the thinker sees a person, Bruce 'in double' and takes him to be two people. The thinker starts to have simultaneous thoughts about what each of the supposed two people is like, attributing exactly the same properties to 'each' person (Fine 2007, p71).

Fine points out that if we attribute MOPs, we must claim that the thinker has two MOPs for *o*. On my understanding of MOPs, this seems right. We can imagine a Frege-case in this example — for example, suppose Bruce is wearing pyjamas. The thinker might then infer that there exist at least two things wearing pink pyjamas. This inference would not appear rational were we to describe their beliefs about 'each person' just as a two-place relation between the thinker and a coarse-grained proposition. But Fine then argues:

> there is nothing sensible we can say as to what these modes of presentation might be. There can be no purely descriptive difference between them, since there is no purely descriptive difference in the way that our thinker conceives of the two Bruces; and there is no plausible non-descriptive difference in the two modes of presentation. If, for example, we take the difference to lie in the original sightings of Bruce, we then implausibly relate the content of the thoughts to the sightings and also make it impossible for the thinker to continue to have coordinated thoughts about Bruce once she has lost all memory of the sightings.
>
> (Fine 2007, p71)

However, once we introduce mental files, we have a response to Fine: the thinker has two mental files on Bruce, each containing records with identical coarse-grained content. A thinker can have two records with identical coarse-grained content (for example, Thales' records that *Hesperus is a planet* and *Phosphorus is a planet*). And there is nothing in the file picture that suggests a thinker can't have two bundles of records, each with identical coarse-grained content. In fact, once we allow that a thinker can have two records with identical coarse-grained content, it is natural

to claim that it is at least possible that a thinker have two bundles of records with identical coarse-grained content. Moreover, the Bruce case suggests that the file picture should accommodate distinct files containing records with identical coarse-grained content. This proposal enables us to say what the different MOPs are in the Bruce case — the thinker has two files with identical coarse-grained content, and each one is a MOP for Bruce.

However, one might still wonder whether what has been given so far counts as a 'sensible answer' to the question of what the MOPs are.

2.3 Costaging and cofiling

The core account of mental files does not give an account of the individuation conditions on mental files. But it is particularly important that a theory of mental files gives some account of their individuation conditions. First, we have no grasp on mental files independent of giving a theory of mental files. And for any particular kind of thing, an account of its individuation conditions will be part of any complete theory of that thing. Second, we cannot suppose giving the individuation conditions on mental files is of secondary importance relative to other parts of mental file theory. A key claim made about files is that they play the MOP role. So we explain Frege-cases by making claims about whether or not a thinker's records are members of the same mental file. And this puts the individuation of mental files at the heart of one of the key explanatory roles they are supposed to play.

Mental file theorists hope to learn something about the mental representation of objects by thinking of this mental representation as analogous to more familiar non-mental filing systems. However, describing the mental representations using the same filing terminology used to describe the non-mental system can disguise places where we do not understand the claims made about mental representation as well as we might like. The individuation conditions of mental files are a case in point. In

the case of non-mental filing systems, we have an adequate understanding of what is required for records to count as members of the same file — this is a matter of location in space. But spatial explanations are no help in the mental case. We need some further account of what it is required for mental records to count as members of the same mental file. And just talking of 'bundles' of mental records belonging to the same mental file, as we talk of bundles of non-mental records belonging to the same non-mental file, disguises the fact we have an incomplete understanding of the individuation conditions on mental files.

2.3.1 Two questions about mental files

Files are thought of as persisting through time. As they persist, they undergo various changes — records are added and removed, and the referential-value of the entire file may shift. This suggests that two records might be members of the same mental file whilst never *simultaneously* being members of that file. And whilst collective reference-fixing by files ensures that at a single moment in time all the records in a file share a referential-value, possible changes in referential-value means that records in a single file at different times may have different referential-values.

These considerations suggest that we should break down questions about individuation conditions on mental files into two distinct components — a synchronic and diachronic element.

I think of persisting mental files as made up of a succession of individual stages of mental files, or *file-stages*. Similarly, persisting records are made up of a succession of *record-stages*. I suggest we should investigate the synchronic component by asking the *costaging question*.

> **Costaging question** Under what conditions are record-stages r_t and s_t members of the file-stage?

And we can explore the diachronic component through the *cofiling question*.

> **Cofiling question** Under what conditions are file-stages F_t and $G_{t'}$ stages of the same file?

When discussing individuation conditions, one can look for one-level or two-level individuation conditions. Looking for a one-level individuation condition for costaging, I would ask: *under what conditions is file-stage F_t numerically identical to file-stage G_t?*[47] However, it is hard to see how we could make much progress answering this question. One option might be to claim: F_t and G_t are numerically identical iff any record-stage r_t in F_t is in G_t. However, not only does this tell us very little about what files are, it may in fact be wrong. Imagine Mary's complete set of beliefs about Felix and Oscar would be expressed with utterances of (27), (28) and (29).

(27) Felix and Oscar are twins.

(28) Felix and Oscar are the same height.

(29) Felix and Oscar live together.

Mary should have two file-stages, one on Felix and one on Oscar. But if relational records are shared between file-stages, we should expect that Mary's *Felix*-file-stage and *Oscar*-file-stage to contain exactly the same record-stages. By the proposed one-level individuation condition on costaging, we would ascribe Mary just one file-stage, when we should ascribe her two.

Instead, I ask for a two-level criterion.

> A two level criterion of identity for objects of some kind involves a function to such objects from what may not be objects of that kind.
>
> (Williamson 1990, p146)

In particular, I ask what it is for two record-stages to be members of the same file. It seems likely we can make more progress discussing two-level individuation condi-

[47] See Williamson (1990, ch9).

tions, and are less likely to deliver an uninformatively circular answer (Williamson 1990, p147).

The costaging and cofiling questions are not the only questions we might ask about individuation conditions. For example, we might also be interested in cross-world individuation conditions on mental files. However, this gets us into cross-world identity issues that are beyond the scope of this thesis. Nonetheless, it is worth noting that we frequently accept individuation conditions that say little about cross-world identity. For example, Frege's individuation condition on directions is that lines have the same direction iff they are parallel.[48] This individuation condition tells us little about direction's cross-world individuation conditions, because we have little idea of what it is for lines in different worlds to be parallel with one another.

We might also want to ask about the individuation conditions on records, or on nodes, or on types of token. However, I suggest that these are closely related to the individuation conditions on files. On the files pictures I am considering, files are bundles of records containing the same node, or tokens of the same type. So learning what it is for records to be members of the same file can be expected to help us understand the individuation conditions on nodes and what it is for tokens to be of the same type. And as records are partially constituted by referential-components, i.e. nodes or tokens, learning about the individuation conditions on these referential components will advance our understanding of the individuation conditions on records. I choose files as my starting point simply because this thesis is about the idea that thinkers use mental files.

[48] This too is a two-level individuation condition.

2.3.2 Preliminary attempts to answer the costaging question

I will start by focussing on the costaging question. I will work through some unsuccessful attempts to answer the costaging question, setting the stage for further discussion of the costaging question and strong-coreference phenomena, and drawing out some further details of the mental file account. I will turn to the cofiling question in chapter 7.

Throughout the rest of this section, I use the simplifying assumption that all records are single-reference.

Same MOP One might be tempted to give CoStaging-1 as an answer to the costaging question:

CoStaging-1 r_t and s_t are costaged iff r_t and s_t deploy the same MOP

However, this option doesn't advance our understanding of files or MOPs at all. Part of the core account of mental files is that the MOP role is played by files. So it is trivial that deploying the same MOP indicates that two records are costaged.

Moreover, for CoStaging-1 to provide a complete account of the conditions on costaging, we would need what we don't have: a complete account of the identity and difference of MOPs. The Frege-constraint[49] gives a sufficient condition for a thinker having two different MOPs, but we cannot use it to extract a necessary condition on having different mental files. We should not suppose that a thinker will be in a Frege-case whenever she has more than one file on a single object (after all, the thinker may treat the records in those files as about the same thing in virtue of a link between the files).

The problem is worse when we think that CoStaging-1 limits files to singular thought. Suppose we think there may be files on non-singular thought, then we can

[49] See p12.

at best hope for something like CoStaging-2.

CoStaging-2 r_t and s_t are costaged iff there is some mode of presentation that both r_t and s_t deploy

However, we understand modes of presentation even more poorly than we understand MOPs. The Frege-constraint is a constraint on MOPs rather than modes of presentation more generally. We as yet have no idea what are the conditions on thoughts deploying the same or different modes of presentation.

Same cognitive particular Another option is to think that we can find our answer to the costaging question in 2.2.1. Having records in the same mental file is simply for those records to be partially constituted by the same node, or referential tokens of the same type.

Considering just one example of this view will reveal the problems for all views of this type. Assuming a global node picture of mental files, we might be tempted to propose CoStaging-3.

CoStaging-3 r_t and s_t are costaged iff r_t and s_t contain the same node

There are two problems here. First, we cannot expect to open the head to see records and nodes, and from this learn what it means to claim that two records share a node. The problem of giving an account of the conditions on sameness and difference of file is much the same as giving an account on the conditions of sameness and difference of node. If the node picture is correct, then CoStaging-3 may be trivially true, but it is far from illuminating.

Second, imagine the counterfactual situation in which we open the head and find something resembling the pictures of nodes from 2.2.1. We might be tempted to point to certain intersections of wires and call these the nodes. But this does not mean that these intersections are nodes. Quoting Millikan:

> Not everything that falls out of a representational system is necessarily read or readable even by its primary interpreters... [T]he mere being the same of two thoughts or percepts does not accomplish anything all by itself even when the fact of this sameness is a natural indication of sameness in content, or when this sameness is an implication of the content represented. The fact of sameness must be *read* somehow if it is to represent, rather than just be, a sameness. This sameness must appropriately interact with or *move* the thinking system in some way if it is to represent itself.
>
> (Millikan 2000, p133)

Even assuming we can identify node-like structures, and perhaps even allowing that these structures have activity patterns that reliably covary with the presence or absence of a particular individual, this does not give us evidence that the thinker uses these structures as a representation of a particular individual. Rather, we need some account of what costaging enables the thinker to *do*. It is this that will warrant our claim that some mental structure or other is a node. Merely identifying plausible candidates for being a node is not enough.

Cannot doubt about the same One of the traditional Fregean ways of individuating modes of presentation is considering what a thinker can doubt about the referent of names or sentences (see Schellenberg forthcoming, §3). This suggests CoStaging-4:

CoStaging-4 r_t and s_t are costaged iff the thinker who holds r_t and s_t cannot sensibly doubt that r_t and s_t share referential-value

However, CoStaging-4 is not satisfactory. A non-decisive difficulty is that even if we suppose that the idea of doubting one's record-stages share referential-value is highly idealised, CoStaging-4 seems to presuppose that thinkers with mental files have relatively sophisticated meta-representational abilities. It is difficult to see how we could make sense of a dog or infant having mental files if we were to adopt CoStaging-4.

A second more decisive difficulty is that CoStaging-4 seems to generate the wrong results. Suppose I visually track a gull as it flies across the sky. When I first see it t_1, I form the belief *that is a gull*. At t_2, I form the belief *that is a yearling*. At t_2, I retain the belief I formed at t_1, realised as record-stage r_{t_2}, and I also have the belief formed at t_2, realised as s_{t_2}. Between t_1 and t_2, I do not stop attending the gull. This is a paradigm case of opening a single mental file on an object, and adding information to that file over the course of the object-tracking. Therefore, I take it as a datum that the beliefs formed at t_1 and t_2 enter the same mental file, and r_{t_2} and s_{t_2} are costaged.

But suppose I fantasize that I am frequently tricked, and come to doubt that I managed to keep track of a single bird between t_1 and t_2. Once I doubt that I have successfully tracked the gull, it seems fully sensible for me to ask whether the beliefs realised as r_{t_2} and s_{t_2} are about the same bird — after all, I am hypothesising that they were formed on the basis of encounters with different birds.

I have stated it as a datum that r_{t_2} and s_{t_2} are costaged. By the core account of files, this means they must share referential-value.[50] But there is a clear sense in which I can doubt whether they are about the same thing. I will explore this tension in greater detail in 5.3.4. However, for now, I simply observe that it undermines CoStaging-4 as a potential answer to the costaging question.

Treated as about the same Looking for some account that emphasises what the thinker *does* with records, we might be tempted by CoStaging-5:

CoStaging-5 r_t and s_t are costaged iff the thinker who holds r_t and s_t takes r_t and s_t to be about the same thing

Even considering adult humans, we should be cautious attributing widespread meta-representation of mental representations: few people have meta-representations

[50] See 2.1.3.

about their record-stages. So one must understand what it is to *take* records as about the same thing in fairly idealised terms. Perhaps to *take* records as about the same thing is simply to act as if they are about the same thing. Reflecting the core-account of files,[51] we can restate CoStaging-5 as CoStaging-6:

CoStaging-6 r_t and s_t are costaged iff the thinker who holds r_t and s_t treats r_t and s_t as about the same thing

It is common to find accounts of mental files that imply CoStaging-6. In particular, any account of files which suggests that files are maintained in accordance with a norm of a 1:1 mapping of files to individuals (singularly thought about) suggests CoStaging-6. Accounts of files that suggest a norm of a 1:1 mapping of files to individuals claim that when a thinker comes to accept that two files are about the same thing,[52] she merges those files. And when a thinker comes to believe she has confused two individuals, she attempts to separate out the file on those individuals into two distinct files.[53]

However, as I argued in 2.1.3, we should not accept an account where files are maintained according to an idea of a 1:1 mapping of files to individuals. Rather, we should allow that thinkers sometimes link mental files rather than merging them. So long as thinkers sometimes treat r_t and s_t as about the same in virtue of r_t and s_t being in linked files rather than merged files, CoStaging-6 cannot give necessary and sufficient conditions on costaging.[54]

[51] See 2.1.3.

[52] She may come to accept this in an idealised way, for example by coming to believe that $a = b$ when she has files both a and b. There is no need for the thinker to have beliefs about her files.

[53] This kind of picture is in present in accounts of files as divergent as Strawson (1974) and Dickie (2010).

[54] Though my eventual answer to the costaging question in 5.4 may be thought of as a refinement of CoStaging-6.

2.4 Final remarks

In this chapter, I have described the core account of mental files, which gives merely a rough sketch of what mental files are. I have also begun the project of filling in the sketch to develop a theory of mental files. I have set out various decision points, and where possible I have argued in favour of one or other position. In doing this, I have provided an overview of the fragmented literature on mental files, and countered the tendency of mental files theorists to develop a plausible picture of mental files without considering alternative ways of developing the idea.

I have laid out two important questions, the costaging and cofiling questions. I have addressed only the costaging question, but I have so far been unable to give a satisfactory response to it. The core account of files says that records in a mental file are treated as if they are about a single thing, but as my discussion of CoStaging-6 showed, just treating record-stages as about a single thing is not enough to ensure those record-stages are costaged.

However, this does point to where we should look next for an answer to the costaging question. There has been a recent surge of interest in the different ways in which a language user may treat two linguistic tokens as coreferential. We may well hope for one of the kinds of coreference discussed to shed light on how we answer the costaging and cofiling questions. So it is to recent work on linguistic coreference that I now turn.

Chapter 3

Assumed-Coreference and De Jure Coreference

3.1 Introduction

It is undeniably important for language users to be able to refer to objects. However, for communication to take place, not only must language users be able to refer to objects, they must be able to refer to something previously referred to (i.e., to corefer), and they must be able to corefer in such a way that both the language user and her audience is aware that coreference is taking place. It is this awareness of coreference that enables language to be used in building up a detailed understanding of a topic through testimony alone.

There has recently been a proliferation of philosophical interest in the different ways in which language users can be aware that they are coreferring.[1]

There is as yet no consensus on the characteristics of these different types of coreference, nor on how to test for them, nor on what explains them. This makes an investigation of coreference in language an independently interesting project. Moreover, there is good reason to hope that an exploration of coreference in language will advance the project of this thesis.

My aim is clarify and critique the claim that thinkers use mental files. A significant question remaining is: what are the individuation conditions on mental files? I suggested we should split this question into the costaging and cofiling questions.[2]

In 2.3.2 I explored some approaches one might take to answering the costaging question, i.e. to giving the conditions under which record-stages r_t and s_t are members of the same file-stage. However, I did not find an adequate answer. I argued that it is not sufficient for costaging that a thinker merely treat record-stages as about the same object. However, part of the core role of files is that a thinker treats costaged records as about the same object. So we might just be looking for some particularly strong kind of 'treating as if about the same object' sufficient for

[1] For example, Perry (1988); Fiengo and May (1994, 2006); Taylor (2003); Fine (2007); Pinillos (2011).

[2] See 2.3.1.

costaging.

It is this line of reasoning that suggests that exploring coreference in language will advance our understanding of mental files. In accounts of linguistic coreference, we have a detailed study of the different ways in which language users can use the fact that NP-occurrences are coreferential, including some ways which *guarantee* that the NP-occurrences are about the same thing if they are about anything at all. We might well hope to find some account of coreference that can be used to answer the costaging question.[3]

In this chapter, I start by drawing a neglected distinction between two types of coreference, assumed-coreference and de jure coreference. I argue that of the two, it is de jure coreference that might provide an answer the costaging question. This sets up the task for the rest of this chapter and a large part of the next: critiquing two influential accounts of de jure coreference, and assessing them as potential answers to the costaging question. In 4.4, I use my findings to develop my own account of de jure coreference, and then show how it points towards the answer to the costaging question.

I use the following terminology: *Utterances* are sentences of a language, indexed to a particular context. If utterances are produced, they are indexed to the context at which they are produced, but we can also consider utterances that are not produced by any language user. Utterances are made up of *occurrences* of words and phrases. Occurrences of noun phrases are *NP-occurrences*.

In this chapter and the next I will assume that anaphoric pronouns are primarily referential devices rather than bound variables or impoverished definite descrip-

[3]This hope is reinforced by those who suppose that the same coreference phenomena exist in thought as exist in language (e.g. Fine 2007; Recanati 2012). Additionally, some suppose we can explain files in terms of de jure coreference. For example Lawlor (2001, passim.) suggests that file-stages are constituted by the appearance of de jure coreference, and Dickie (2010, p214) entertains the idea that de jure coreference is explanatorily prior to files.

tions.[4] My objective is to engage with discussions of coreference rather than argue for some particular way of drawing the line between referential and non-referential NP-occurrences, and a full discussion of the semantics of anaphora is beyond the scope of this thesis.

3.2 External-coreference

When enumerating the varieties of coreference, it is convenient to start with the simplest: *external-coreference*[5] (or simply 'coreference'). Where O_t and $O_{t'}$ are NP-occurrences occurring at possibly distinct times t and t' respectively:

O_t *externally-corefers* with $O_{t'}$ iff O_t and $O_{t'}$ refer to the same object

External-coreference is an equivalence relation holding only between referring NP-occurrences. The externally-coreferential NP-occurrences may be produced in very different contexts. There is no requirement that they are produced by the same speaker, in the course of the same discourse, or that the speakers have any knowledge of each other or each other's language.

Later, I will discuss external-coreference in thought. Use the simplifying assumption, that all records are single-reference.[6] Where r_t and $s_{t'}$ are record-stages occurring at possibly distinct times t and t' respectively:

r_t *externally-corefers* with $s_{t'}$ iff r_t and $s_{t'}$ refer to the same object

External-coreference defined as a relation between record-stages is also an equivalence relation. And as with linguistic external-coreference, there is no requirement that the externally-coreferential record-stages are cotemporal, or occur in the mind of the same thinker.

[4] In this, I follow e.g. von Heusinger (2002), rather than e.g. Reinhart (1986).
[5] The term 'external-coreference' is taken from Perry (1988).
[6] See p19.

3.3 Assumed-coreference

3.3.1 Introducing assumed-coreference

There must be more to coreferential communication than just external-coreference. Notoriously, the coreference of a speaker's NP-occurrences is no guarantee that either the speaker or her audience realise that they corefer, and hence no guarantee that the discourse participants can exploit that coreference.

Nonetheless, participants in a discourse do have ways in which they exploit the coreference of NP-occurrences in communication. One of these is *assumed-coreference*.

Assumed-coreference occurs when:

> [T]o understand the internal structure of the discourse, and the emotions to which it gives rise, one must see that the various referring expressions are *supposed* to be about the same thing.
>
> (Perry 1988, p13)[7]

Rather than use the under-explained idea of the 'internal structure of the discourse' in defining assumed-coreference, I suggest we define assumed-coreference in terms of the more familiar idea of the discourse's *common ground* .

The common ground of a discourse is a set of propositions that the participants in the discourse take for granted, or act as if they take for granted.[8] Moreover, each participant in the discourse believes the other participants take these propositions for granted (or will act as if they do). There is no requirement for the propositions in the common ground to be true. The common ground is updated throughout the conversation, in part because utterances made during the course of a conversation contribute to the information in the common ground. Even if the asserted content

[7]Perry (1988) uses the expression 'internal coreference' rather than 'assumed-coreference'. 'Assumed-coreference' better reflects the nature of the phenomenon, and avoids confusion between assumed-coreference and other phenomena identified as 'internal coreference' (e.g. Recanati 2012, ch8).

[8]See Soames (e.g. 1989); Stalnaker (e.g. 2002).

of an utterance isn't added to the common ground, the fact that the utterance was produced will be added. When the common ground is updated with new information, "obvious consequences" (Stalnaker 2002, p708) of that information are also added to the common ground.

Provisionally, assumed-coreference can be defined as follows:[9] Where O_{t_1} and O_{t_2} are NP-occurrences occurring in the same discourse, at times t_1 and t_2 respectively (and t_1 is not later than t_2):

> O_{t_1} and O_{t_2} *assumedly-corefer* iff it is common ground at t_2 that O_{t_1} and O_{t_2} externally-corefer.

Of course, the fact that NP-occurrences corefer cannot be part of the common ground before those NP-occurrences are produced. Instead, when O_{t_2} is produced, the common ground is immediately updated so that it is part of the common ground that O_{t_1} externally-corefers with O_{t_2}.

The idea of assumed-coreference can be clarified using examples. Some NP-occurrences both externally-corefer and assumedly-corefer, as in example V.

Example V:

MARY: [Pointing to the New Bodleian]
That is one of the most beautiful buildings in Oxford.

JACK: [Looking at the New Bodleian]
No, it is hideous.

Some NP-occurrences externally-corefer but do not assumedly-corefer:

[9] I will outline some difficulties with this definition in 3.3.3. However, as I argue in 3.3.2 that assumed-coreference cannot be used to answer the costaging question, I will not attempt to improve on this provisional definition.

Example VI:

MARY: [Pointing to the New Bodleian]
That is one of the most beautiful buildings in Oxford.

JACK: [Thinking Mary is looking at the Sheldonian, and himself pointing to the New Bodleian]
And that is one of the ugliest.

But as Perry (1988, p13) points out, there can also be cases where NP-occurrences assumedly-corefer but do not externally-corefer.

Example VII:

MARY: [Looking at the New Bodleian]
That is one of the most beautiful buildings in Oxford.

JACK: [Thinking Mary is looking at the Sheldonian, and looking at the Sheldonian himself]
Yes it is.

MARY: That's surprising. I didn't think you liked modern architecture.

JACK: It isn't modern at all, it was designed by Wren.

Considering the discourse as a whole, it is clear that the occurrences of the underlined NPs do not externally-corefer. But they do assumedly-corefer: to understand Mary and Jack's responses, we must see that at the time of Jack's first utterance of 'it', it was common ground that the NP-occurrences did externally-corefer.

In VII, the assumedly-coreferring NP-occurrences refer but fail to externally-corefer. In VIII, the assumedly-coreferring occurrences do not externally-corefer because they fail to refer at all.

Example VIII:

[James and Jack are in a house they falsely believe to be haunted by a ghost]

JAMES: I saw her out of the corner of my eye last night.

JACK: And I heard her breathing in the kitchen.

Again, to understand the discourse, one must see that at the time Jack produces his 'her'-occurrence, it is common ground that the 'her'-occurrences externally-corefer, even though the occurrences do not in fact refer at all.

I can now summarise some of the main features of assumed-coreference. The relation holds only between NP-occurrences that are part of the same discourse, but can hold between NP-occurrences uttered by different speakers. The relation can hold between referential NP-occurrences, regardless of whether those NP-occurrences refer or externally-corefer.

3.3.2 Assumed-coreference and costaging

Using the simplifying assumption,[10] we might hope for some answer to the costaging question along the lines of CoStaging-7.

CoStaging-7 r_t and s_t are costaged iff r_t and s_t assumedly-corefer.

There are two difficulties with this proposal. First, assumed-coreference has only been defined as a relation between NP-occurrences. But CoStaging-7 presupposes that assumed-coreference can be defined as a relation between record-stages. However, it is not clear we can define assumed-coreference for this type of relatum, as there is no clear mental correlate of the common ground of a discourse.

Second, if the objective is to answer the costaging question, it is clear that it is not worthwhile trying to work out a definition of assumed-coreference for mental

[10] See p19.

relata. Assumedly-coreferential relata need not share referential-value. We saw this in example VII, in which the assumedly-coreferential NP-occurrences referred to different objects. However, costaged records must share referential-value thanks to the reference-fixing role of mental files.[11] So any relation that does not guarantee shared referential-value cannot be sufficient for costaging.

3.3.3 Problems and limitations

There are difficulties with my definition of assumed-coreference independent of difficulties with CoStaging-7. Principal among these is the fact that, in most conversations, the exact words used to communicate a particular point are not remembered by the participants. But the given definition of assumed-coreference requires that it is common ground that particular NP-occurrences were produced. One option would be to say that there can only be assumed-coreference where propositions about the particular words used in each utterance are part of the common ground, and so assumed-coreference is a relatively rare phenomenon. Alternatively, we might try to build on the proposal that it can be part of the common ground that discourse-participants are talking about the same thing without propositions about the exact NP-occurrences used being part of the common ground. Spelling out the details of this proposal will be complicated by cases in which several different individuals are being talked about in a single conversation.

Even if these issues can be overcome, assumed coreference faces a further problem: it cannot be used to draw a distinction between cases where it seems to be luck that two NP-occurrences externally-corefer, and cases where it seems to be guaranteed that they externally-corefer. Contrast example V with example IX.

[11] See 2.1.3.

Example IX:

MARY: [Looking at the New Bodleian]
That is one of the most beautiful buildings in Oxford.

JACK: [Not identifying what Mary is talking about, but feeling contrary]
Well, if you like it, it must be hideous.

In both V and IX, the underlined occurrences of 'that' and 'it' externally- and assumedly-corefer. However, in example V, the external-coreference results from the good fortune of two independent demonstrations sharing the same referent. But in IX, the occurrences seem guaranteed to externally-corefer (if they refer at all) simply in virtue of the linguistic rules governing anaphora. Assumed-coreference, in focussing on what is common-ground between discourse participants rather than on linguistic rules, cannot shed light on the difference between these examples.

Similarly, we might hope for some explanation of the differences between utterances of (30) and (31).

(30) I became fascinated by Leningrad as a child, so when the Cold War ended I got on the first St Petersburg flight I could afford.

(31) I became fascinated by St Petersburg as a child, so when the Cold War ended, I got on the first St Petersburg flight I could afford.

It seems likely that in most discourses where they would be uttered, utterances of (30) and (31) would contain NP-occurrences that assumedly-corefer. But in utterances of (31) and not (30), the occurrences seem guaranteed to corefer (if they refer at all) in virtue of the linguistic rules governing proper names. Assumed-coreference cannot be used to describe the difference between these cases.

3.4 De jure coreference

This suggests we need to identify a new variety of coreference that can be used in identifying this difference. This is *de jure coreference*.[12] De jure coreference carries with it some kind of guarantee of coreference. Paradigm examples of de jure coreferential terms include anaphoric pronouns and their antecedents, and occurrences in a single use of a name.[13] Because de jure coreference carries with it some kind of guarantee of coreference, it is an excellent candidate for supplying an answer to the costaging question.

Much of the recent work on coreference by philosophers focusses on the definition, characteristics and explanation of de jure coreference. Therefore, I cannot simply propose CoStaging-8 as an answer to the costaging question (making the simplifying assumption that all records are single-reference).[14]

CoStaging-8 r_t and s_t are costaged iff r_t and s_t are de jure coreferential.

Instead, I need to pay careful attention to disagreements about what de jure coreference is. To facilitate this, I will discuss various more precisely defined relations (such as RAS-coreference and AP-coreference). These are all proposals about how we should characterise de jure coreference. Studying these proposals will allow me to clarify what de jure coreference is, and to evaluate different versions of CoStaging-8.

[12] I use the term 'de jure coreference' because this terminology is beginning to gain traction in the literature (e.g. Pinillos 2011; Recanati 2012; Schroeter Forthcoming), and because it is not associated with any prior commitments as to how the relation should be characterised or explained. De jure coreference is variously called 'representing as the same' (Fine 2007), 'coordination' (Fine 2007), 'grammatically determined coreference' (Fiengo and May 2006), and 'explicit coreference' (Taylor 2003).

[13] It seems likely that a precise account of what it is to have a single use of a name will be given in terms of de jure coreference (e.g. Fine 2007, p108). But there is an intuitive sense of what it is to have a single use of a name that can be grasped by thinking that in the Paderweski case (see p13), Peter has *two* uses for the name Paderewski, even though those uses are coreferential. When Peter sincerely asserts "Paderewski is a musician", the occurrence of 'Paderewski' is part of one use of the name. And when Peter sincerely asserts "Paderewski is a politician", the occurrence of 'Paderewski' is part of his other use of that name.

[14] See p19.

The first account of de jure coreference I will consider is given by Fine (2007). Fine's account of de jure coreference is part of an extended development of a *relationist* semantics. A relationist semantics for NP-occurrences grants semantic significance to relations between NP-occurrences, as well as to the NP-occurrences' 'intrinsic' semantic features (such as having a referent, having a sense, being assigned a particular discourse referent or variable, or being of a particular semantic type).[15] One of the supposed advantages of relationist semantics is an explanation of Frege-cases and related puzzles for the semantics of language and mental representation that does not rely on modes of presentation. In 3.8, I will raise considerations that suggest that Fine is not successful in explaining Frege-cases just using semantic relations. Nonetheless, Fine's relationist semantics contains many ideas that should be of interest to anyone interested in MOPs or mental files.

Fine's discussion of de jure coreference is distinctive in that it contains an account of de jure coreference in thought as well as in language. And de jure coreference in both thought and language are species of a wider phenomenon of *coordination*. Fine initially glosses coordination as "the very strongest relation of synonymy or being semantically the same" (Fine 2007, p5).[16] Coordination is the principal semantic relation, obtaining between occurrences of referring expressions uttered by a single speaker (2007, ch2), referring expressions uttered by different speakers (2007, ch4), non-referring referential expressions (2007; 2010), general terms (2007, pp129-131), variables (2003; 2007, ch1), and thoughts (2007, ch3).

Given my interest in answering the costaging question, I will discuss Fine's account of de jure coreference in thought. However, much of what Fine says about de jure coreference in thought presupposes the fuller account Fine gives of de jure coref-

[15]This use of 'intrinsic' stretches the ordinary sense of the term (see also Rattan 2009, p1127). In the context of semantic relationism, 'intrinsic' means 'not related to another NP-occurrence' rather than 'not related to anything else at all'.

[16]It is not clear this gloss is appropriate. On Fine's picture, pronominal anaphora coordinate with their antecedents (2007, pp122-123). But it is far from clear that they are 'synonymous' or 'semantically the same' as their antecedents.

erence between NP-occurrences,[17] and so my discussion starts with Fine's account of de jure coreference in language. I clarify Fine's account of de jure coreference between NP-occurrences, and discuss how this account is extended to thought. I criticise some elements of Fine's account, and show that on Fine's definition of de jure coreference, de jure coreference cannot be used in answering the costaging question. However, this failure reveals an important decision point in how we think about de jure coreference. Whether de jure coreference has any hope of answering the costaging question will depend on which side we come down on.

3.5 RAS-coreference

3.5.1 Introducing RAS-coreference

Fine's preliminary account of de jure coreference between referring NP-occurrences is in terms of referential occurrences *representing as the same* (Fine 2007, p40).[18] Although Fine gives no complete definition of representing as the same, it is a fairly intuitive idea that is widely used by others writing on coreference.[19] However, Fine's introduction to the idea is not without problems, so I start by treating representing as the same as a stand-alone account of de jure coreference, before considering Fine's final account in 3.6.

Fine introduces representing as the same by distinguishing it from *representing as being the same* (Fine 2007, p40). Utterances of (32) and (33) both represent as being the same, but only (32) contains occurrences which represent as the same.

[17] Fine initially presents his account in terms of sentences and expressions rather than utterances and occurrences (2007, ch2), though he is happy switching to talk of occurrences of expressions (2007, ch4). I present Fine's account in terms of utterances and occurrences because nothing substantive is altered by this change, and because it allows for consistency with my discussion of de jure coreference in chapter 4 whilst avoiding difficulties about the individuation of expressions and sentences (see p13n).

[18] Fine only discusses examples where *referring* NP-occurrences represent as the same. However, he says nothing that suggests that non-referring referential NP-occurrences cannot represent as the same. For now, I follow Fine in discussing only referring NP-occurrences.

[19] See, for example, Pinillos (2011), Recanati (2012).

(32) Cicero is Cicero

(33) Cicero is Tully

All and only utterances asserting identities contain NP-occurrences representing as being the same. But only some identity utterances contain NP-occurrences representing as the same. NP-occurrences which represent as the same *RAS-corefer*.

Utterances can contain RAS-coreferring NP-occurrences even if they don't assert an identity, as in (34).

(34) Last time I saw Mary, Jack had just sent Mary an email.

Occurrences of different words can be RAS-coreferential (Fine 2007, p41), as in (35).

(35) Last time I saw Mary, Jack had just sent her an email.

And just using the same word and having the same referent is not enough to ensure that occurrences RAS-corefer (Fine 2007, p41), as when, in a version of the Illustrious case,[20] Tony utters (36):

(36) That isn't that.

[Pointing, first at the bow, then at the stern of the HMS Illustrious]

3.5.2 The test for RAS-coreference

Fine characterises RAS-coreference by offering a test for it:

> But a good test of when an object is represented as the same is in terms of whether one might sensibly raise the question of whether it is the same. An object is represented as the same in a piece of discourse only if no one who understands the discourse can sensibly raise the question of whether it is the same.
>
> (Fine 2007, p40)

However, as it stands, this test is not adequate as a characterisation of RAS-coreference.

[20] See p13.

A first difficulty is that there are different ways of thinking about what can be 'sensibly questioned'. For example, one might think that if it is common ground that a conversation is taking place in London, it might not be sensible to question this. However, it is clear from Fine's examples that this is not the intended sense of 'sensibly question'. Instead, his examples suggest that what matters is what it is *nonsensical* to question, rather than merely *non-sensible*.

A second difficulty is that, as it stands, this test is not passed by the utterances Fine (2007) gives as examples of RAS-coreference. There are two ways of questioning whether, for example, the occurrences of 'Cicero' in an utterance of (32) externally-corefer. On encountering an utterance of (32) a thinker might question whether there is any such person as Cicero.[21] Alternatively, a thinker might understand an utterance of (32), but still worry that she has not understood it, and so think that it is possible that the occurrences of 'Cicero' do not externally-corefer. Presumably these two forms of questioning must be ruled out, so Fine's own examples of RAS-coreference pass the test for RAS-coreference.

A third difficulty is that, as it is given, Fine's test provides only a necessary condition on RAS-coreference, and no indication is given of what extra is needed before we can state necessary and sufficient conditions. If, like Fine, we are treating RAS-coreference as a merely a preliminary account of de jure coreference, this is unproblematic. But if we are treating RAS-coreference as a serious proposal about how to characterise de jure coreference we need necessary and sufficient conditions. However, there are no obvious reasons not to suggest that this test can provide both necessary and sufficient conditions on RAS-coreference.

Taking these considerations together, I propose the following definition of RAS-coreference between referring NP-occurrences: Where O_{t_1} and O_{t_2} are referring NP-

[21] So long as knowing that P is compatible with being able to question that P, it is possible to fully understand an utterance, and to know the referential-value of the NP-occurrences in the utterance, whilst being able to question whether the referring NP-occurrences do in fact refer.

occurrences produced in the course of a single discourse D, and using the appropriate understanding of 'sensibly to question'

> O_{t_1} and O_{t_2} *RAS-corefer* iff it is not possible for someone who understands D sensibly to question that O_{t_1} and O_{t_2} externally-corefer, except by questioning whether O_{t_1} and O_{t_2} refer at all, or by questioning her own understanding of D.

3.5.3 RAS-coreference in thought

Fine does not offer a test or definition of RAS-coreference in thought. He does give an example.

> Suppose, for example, that I continuously observe an object — say a snake. I first see it coiled and later see it uncoil. The various momentary observations that make up the continuous observation then all represent the snake as the same. It is not like seeing a snake on two separate occasions and judging that it is the same.
>
> (Fine 2007, p67)

We might try to define RAS-coreference in thought in the same manner it is defined for referential NP-occurrences: Where r_t and s_t are single-reference record-stages possessed by thinker T

> r_t and s_t *RAS-corefer* iff it is not possible for T sensibly to question that r_t and s_t externally-corefer, except by questioning whether r_t and s_t refer at all, or by questioning her own understanding of her mental representations.

However, if the snake example is to count as an example of RAS-coreference, this definition of RAS-coreference cannot be used.[22] Suppose the thinker in the

[22]The problem is similar to the problem giving CoStaging-4 as an answer to the costaging question (see p65).

snake example has cotemporal record-stages r_t and s_t, that she would express using utterances of (37) and (38) respectively.

(37) That snake was coiled.

(38) That snake is uncoiled.

It seems open to the thinker to doubt that r_t and s_t are about the same snake, just by supposing that the snake was switched mid-way through her observation of it. The thinker does not question her understanding of r_t and s_t. She assumes r_t refers to the snake she saw coiled, and s_2 refers to the snake she saw uncoiled. However, she worries that these are different snakes. So by the proposed test for RAS-coreference in thought, r_t and s_t are not RAS-coreferential.

Fine offers no other lead on characterising RAS-coreference in thought. We might agree with him that there is some intuitive difference between cases like his snake case, and cases where we form a judgement representing something as being the same. But there is little to go on here in establishing an adequate account of RAS-coreference beyond merely asserting this intuitive difference. So I suggest we look beyond Fine's preliminary characterisation of de jure coreference in trying to answer the costaging question.

3.6 SR-coreference

RAS-coreference is just Fine's preliminary characterisation of de jure coreference. Fine fleshes out the idea with a semantic relationist account of de jure coreference, which aims to both characterise and explain de jure coreference in terms of semantic relations. It is this semantic relationist account that should be taken as Fine's considered contribution to the debate, rather than his more widely adopted idea that de jure coreferential NP-occurrences RAS-corefer.

3.6.1 The relational framework

To understand Fine's explanation of de jure coreference, we need to work within his semantic relationism. In other words, we have to accept (a) the fact that two NP-occurrences are de jure coreferential is a semantic feature of the NP-occurrences;[23] and (b) this semantic feature exists in virtue of relations between each occurrence, rather than because of their intrinsic features such as their referential-value or sense (Fine 2007, pp41-42).

Fine claims that NP-occurrences are de jure coreferential when there is a *semantic requirement* that they externally-corefer (2007, p50).

Semantic requirements are a subclass of *semantic facts*. Like all semantic facts, they are semantic as to *status* rather than semantic as to *topic*. Facts which are semantic as to topic concern semantic properties and relations, but are made true partly in virtue of non-semantic properties and relations (for example, the fact that *an utterance of "snow is white" expresses a truth* is semantic as to topic). Facts which are semantic as to status are true only in virtue of semantic considerations, (for example, the fact that *"snow is white" is true iff snow is white*). Semantic requirements, like all semantic facts, are factive. That is, if it is a semantic requirement that P, then P.

All the classical consequences of semantic requirements are semantic facts. But only the *manifest* consequences of semantic requirements are semantic requirements.[24]

A fact is a manifest consequence of another fact or group of facts if it can be deduced by an idealised thinker from these facts, even if the idealised thinker has a different 'take' on an individual whenever it appears in the requirements. Assume,

[23] Rather than being e.g. a syntactic feature (see Fine 2007, pp40-41).

[24] One might well ask which semantic facts count as semantic requirements in the first place (independent of considerations of closure). Fine gives no explicit discussion of this. Given Fine gives priority to what is accessible to the understanding, one might be tempted to claim that known semantic facts are semantic requirements. But I know that it is a semantic fact that *'Hesperus' and 'Phosphorus' corefer*, but it is not a semantic requirement that they corefer. Instead, it is better to suppose that semantic requirements are those semantic facts a language-user is required to know in order to count as a competent language-user, and manifest consequences of those semantic facts.

as Fine does, a coarse-grained account of propositions. On Fine's account of the difference between semantic facts and semantic requirements, if it is a semantic fact that Fa, and a semantic fact that Ga, then it will be a semantic requirement that $Fa\&Ga$, but not a semantic requirement that $\exists x(Fx\&Gx)$. This is because the thinker could not have deduced $\exists x(Fx\&Gx)$ if she had different takes on object a in her knowledge of the semantic facts that Fa and Ga, but she could have deduced the proposition $Fa\&Ga$ whilst having different takes on a. Hence, if it is a semantic fact that *'Cicero' refers to o* , and that *'Tully' refers to o* it is not a semantic requirement that *'Cicero' and 'Tully' corefer* .

The shift to semantic requirements from semantic facts is in line with Fine's conception of semantics as giving priority to what is accessible to the understanding, rather than what is a logical consequence of what is known but nonetheless not accessible to the understanding (Fine 2007, p50).[25]

This allows Fine to claim that referentialism is compatible with the transparency of meaning, that is, with the claim that semantic requirements are accessible to the understanding.[26] So long as we understand the transparency claim in terms of semantic requirements rather than semantic facts, then we can maintain the transparency of meaning without having to suppose that Thales[27] knows that his occurrences of 'Hesperus' and 'Phosphorus' externally-corefer just because he knows what his occurrences of 'Hesperus' and 'Phosphorus' refer to.

3.6.2 Types of semantically required coreference

With this terminology of semantic requirements, I can present Fine's account of de jure coreference.

Fine treats de jure coreference as just one species of coordination. And as he

[25] As I show in 3.9, this means there are serious questions about the idea of manifest consequence.
[26] I discuss this further in 3.8.
[27] See 1.1.1.

supposes that there are different forms of coordination, he describes different forms of de jure coreference, each with different characteristics.

Proper names, uttered by a single speaker Where O_t and $O_{t'}$ are referring occurrences of proper names, produced by a single speaker:

> O_t *SR_n corefers* with $O_{t'}$ iff it is a semantic requirement that O_t and $O_{t'}$ externally-corefer[28]

In utterances of (32), the occurrences of 'Cicero' are semantically required to corefer. In contrast, in utterances of (33), the occurrence of 'Cicero' is semantically required to refer to Cicero, as is the occurrence of 'Tully'. But as semantic requirements are closed only under manifest consequence, it is not a semantic requirement that the occurrences of 'Cicero' and 'Tully' externally-corefer.

Fine claims that SR_n-coreference is an equivalence relation (2007, p55). Because semantic requirements are factive, only externally-coreferential NP-occurrences are SR_n-coreferential.

Derived reference There are cases where one NP-occurrence *derives* its reference from another NP-occurrence, but not vice versa (Fine 2007, p122), for example when an occurrence of an anaphoric pronoun derives its referent from its antecedent, as in utterances of (35). In these cases, there is an asymmetric relation holding between the NP-occurrences. Where O_t and $O_{t'}$ are referring occurrences produced by a single speaker in the course of a single discourse:[29]

> $O_{t'}$ *SR_d-corefers* with O_t iff it is a semantic-requirement on $O_{t'}$ that it externally-corefer with O_t, and it is not a semantic-requirement on O_t that it externally-corefer with $O_{t'}$

[28] See Fine (2007, p51).

[29] $O_{t'}$ may or may not precede O_t. 'Antecedents' may occur after the anaphor, as in "Before he started speaking, I thought Jack was interesting".

An utterance of (34) does not contain SR_d-coreferential NP-occurrences, because the 'Mary'-occurrences are semantically required to corefer with one another (these occurrences do SR_n-corefer). But an utterance of (35) contains an occurrence of 'her' that SR_d-corefers with the occurrence of 'Mary', because there is a semantic-requirement on the 'her'-occurrence that it externally-corefers with the 'Mary'-occurrence, but no semantic requirement on the 'Mary'-occurrence that it externally-corefers with 'her'.

Interpersonal cases Additionally, Fine discusses *inter*personal semantic requirements of coreference, defining a relation of SR_i-coreference. Where O_t and $O_{t'}$ are occurrences of names, uttered by different speakers:

> O_t *SR_i-corefers* with $O_{t'}$ iff it is a semantic requirement that O_t and $O_{t'}$ externally-corefer

SR-coreference Fine's breakdown of different types of de jure coreference may be helpful when thinking about the distinct characteristics of each type of case. However, it is sometimes useful to think in more general terms. To enable this, I define a relation of 'SR-coreference' *simpliciter*: Where O_t and $O_{t'}$ are referring NP-occurrences:

> O_t *SR-corefers* with $O_{t'}$ iff it is a semantic requirement that O_t and $O_{t'}$ externally-corefer

All SR_n-, SR_d- and SR_i-coreferential NP-occurrences are SR-coreferential.

Fine characterises de jure coreference as SR-coreference. This characterisation carries with it a definition of de jure coreference, as well as a semantic relationist explanation of the relation.

Fine's discussion of SR-coreference suggests that the characteristics of the relation depend on the relata. In particular, it is worth noting (given my subsequent

discussion of whether *intra*personal de jure coreference is non-transitive) that Fine uses a variation on the Paderewski case[30] to argue that SR-coreference is a non-transitive relation when holding between occurrences of proper names uttered by different individuals (Fine 2007, pp105-109).

> **Example X:** Sarah is a competent user of the name 'Paderewski', who has never thought that there are 'two Paderewskis'. Peter has never heard of Paderewski until he hears Sarah utter (39), and a short time later utter (40).
>
> (39) Paderewski was a talented musician.
>
> (40) Paderewski was a notable politician.
>
> Peter supposes that Sarah is talking about two different people, forming beliefs about the attributes of 'each Paderewski'. In a subsequent conversation about pianists, Peter produces his own utterance of (39). And in another conversation about politicians, Peter produces his own utterance of (40).

Call the 'Paderewski'-occurrences in Sarah's utterances of (39) and (40) S_1 and S_2 respectively. Call the 'Paderewski'-occurrences in Peter's utterances of (39) and (40) P_1 and P_2 respectively. S_1 SR_n-corefers with S_2, P_1 SR_i-corefers with S_1, and P_2 SR_i-corefers with S_2. This means that S_1 and S_2 SR-corefer, P_1 and S_1 SR-corefer, and P_2 and S_2 SR-corefer. However, P_1 and P_2 do not SR_n-corefer, and we have no reason to think they SR-corefer at all. Hence, this case is evidence for the non-transitivity of coreference in the interpersonal case.[31]

However, there is currently no reason to suppose that SR-coreference is anything other than an equivalence relation in the *intra*personal case. There is an asymmetric

[30] See p13.

[31] Contrast Fine's conclusion with Taylor (2003, pp13-14). Taylor considers a similar case, concluding that as de jure coreference must be transitive, Peter fails to produce occurrences of 'Paderewski' that de jure corefer with Sarah's.

kind of *intra*personal SR-coreference, SR_d-coreference. But suppose that $O_{t'}$ SR_d-corefers with O_t. This is itself an asymmetric relation, but it produces a semantic requirement that O_t and $O_{t'}$ corefer, hence O_t and $O_{t'}$ are SR-coreferential. This subsequent semantic relation is itself symmetric, even though it results from an asymmetric semantic relation.

3.6.3 Putative SR-coreference

I have been assuming that only referring NP-occurrences can be de jure coreferential. But there is are also cases of de jure coreference, or something very like it, between non-referring referential NP-occurrences. Compare (41) and (42).

(41) The best thing about Mary is that when Mary visits she brings presents.

(42) The best thing about Father Christmas is that when Father Christmas visits he brings presents.

Cases like (42) are examples of *de jure pseudo-coreference*. It is a further question whether the similarities between de jure coreference proper and de jure pseudo-coreference are such that de jure pseudo-coreference can be counted as a species of de jure coreference.

Many accounts of de jure coreference give a unified treatment of (41) and (42). If we explain de jure coreference in terms of pre-semantic features like syntax or logical form, then pre-semantically there is no difference between referring and non-referring cases. And if we explain de jure coreference in terms of an object-independent sense or a non-factive semantic relation, then a unified account of (41) and (42) is also possible. However, Fine's account of de jure coreference uses factive semantic requirements of coreference. And so he cannot give a unified account of de jure coreference proper and de jure pseudo-coreference: whilst there may be semantic requirements of coreference between the occurrences of underlined words in an utterance of (41),

there can be no semantic requirements of coreference between the occurrences of underlined words in an utterance of (42), because the occurrences fail to refer.

Fine's discussion of de jure pseudo-coreference is somewhat telegraphic, but he suggests that there is a "backup" semantics of "putative semantic requirements" (Fine 2007, p126). If the putative semantic requirement is met, as in utterances of (41), there is a genuine requirement. If it is not met, as in utterances of (42), the putative requirement provides a "backup" semantic requirement "suitably related" to the original semantic requirement (Fine 2007, p127). Sophisticated speakers can know that they are using terms related by putative semantic requirements, even when they are not in a position to know whether these requirements are met, as when a speaker utters (42) whilst being unsure whether the 'Father Christmas'-occurrences refer (Fine 2007, p127).

Fine (2010, pp498-500) says a little more about de jure pseudo-coreference. He proposes that there is a putative semantic requirement that two occurrences corefer iff "it is taken to be a semantic requirement that they corefer" (2010, p499), i.e. iff they are taken to SR-corefer. And this is where Fine leaves the topic.

However, it is not enough to explain putative semantic requirements just in terms of 'taking' occurrences to SR-corefer. There are various ways one might wish to gloss what it is to 'take' occurrences to SR-corefer. One might 'take' two occurrences to SR-corefer just when one believes they SR-corefer.[32] But this may be an adequate account of (42) when uttered by someone who believes in Father Christmas, it cannot account for the de jure pseudo-coreference in an utterance of (42) by someone who doesn't believe in Father Christmas.

Alternatively, one might 'take' two occurrences to SR-corefer just when one accepts for the purposes of the discourse that they SR-corefer. This may account for

[32]I assume that the attitudes involved in 'taking' something to SR-corefer are highly idealised, so the fact that most people don't have the conceptual resources or introspective inclination to formulate an assertion that P doesn't disbar them from believing, accepting or pretending that P.

de jure pseudo-coreference in utterances of (42), but not in utterances of (43).

(43) Hob and Nob are both fools, because although there is no such thing, Hob thinks Baba Yaga has cursed his mare, and Nob thinks she has blighted his crops.

We cannot describe the de jure pseudo-coreference of the occurrences of the underlined words in utterances of (43) in terms of the speaker accepting that the occurrences SR-corefer (for the purposes of the discourse), because the speaker claims that there is no such thing as Baba Yaga.

So we might suggest that a thinker 'takes' two occurrences to SR-corefer during part of the discourse, and allow that thinkers can switch in and out of accepting something for the purposes of the discourse, just as on Evans' account of fictional discourse, thinkers can switch in and out of a game of make-believe during the course of a single utterance (Evans 1982, ch10). Using asterisks to mark where one is accepting something for the purposes of the discourse, rewrite (43) as (43′):

(43′) Hob and Nob are both fools, because although there is no such thing, *Hob thinks Baba Yaga has cursed his mare, and Nob thinks she has blighted his crops*.

But this proposal still needs refining. It seems open for me to accept what I like for the purposes of part of a discourse. For example, it seems open for me to refuse to accept that occurrences of the underlined NPs in 43 are SR-coreferential. Merely talking in terms of what a thinker accepts makes the de jure pseudo-coreference too much a voluntary matter. Instead, we should say that, for NP-occurrences O_t and $O_{t'}$ which occur in a single discourse D:

> O_t and $O_{t'}$ are *putatively SR-coreferential* iff a speaker must accept for the purposes of at least part of the discourse that O_t and $O_{t'}$ are SR-coreferential in order to count as understanding D

This definition is of little use as a test for putative SR-coreference if we need to know whether occurrences are putatively SR-coreferential before we can decide whether a thinker needs to accept that those occurrences are SR-coreferential in order to count as understanding the discourse. However, Fine doesn't give a test for identifying semantic requirements, presumably because competent language users are supposed to be able to recognise semantic requirements. Similarly, we don't need a test for identifying putative semantic requirements: competent language-users just recognise them, and so will recognise when a speaker must accept occurrences are SR-coreferential in order to count as understanding the discourse.

3.7 De jure coreference and thought

3.7.1 RR-coreference

Fine's account of de jure coreference is distinctive in that he provides a distinct account of de jure coreference in thought as well as language.[33] To facilitate discussion, start by making the simplifying assumption that all records are single-reference.[34]

Fine's suggestion is that just as there are semantic requirements governing the content of language, there are *representational requirements* governing the mind's content, which like semantic requirements are closed under manifest rather than classic consequence (Fine 2007, pp72-73). As such "representational requirements constitute a body of information accessible in principle to the thinker" (Fine 2007, p73). And just as there are semantic requirements that NP-occurrences externally-corefer, there are *representational* requirements that records externally-corefer. Like semantic requirements of coreference, representational requirements are factive: if it is representationally required that P, then P.

[33] Contrast this with Pinillos (2011) who does not discuss thought, or Recanati (2012) who simply assumes that there is a mental equivalent of linguistic de jure coreference.

[34] See p19.

We can characterise de jure coreference between records as RR-coreference:

r and s *RR-corefer* iff r and s are representationally required to corefer.

Repeating an example from 1.1.1, Thales has records r and s he would express with utterances of (1) and (3) respectively.

(1) Hesperus is bright.

(2) Phosphorus is bright.

That r and s externally-corefer is a classical consequence of the fact that r refers to *Venus*, and the fact s refers to *Venus*. However, r and s do not RR-corefer, because representational requirements are closed only under manifest consequence.

Thales has a record-stage u he would express using (3).

(3) I want to see Hesperus.

r and u do RR-corefer, just as occurrences in utterances of (1) and (3) SR-corefer. Considering the snake example discussed in 3.5.3, we can say that the belief-records expressed (37) and (38) also RR-corefer.

Fine claims that Frege-cases arise when a thinker has externally-coreferential records that do not RR-corefer (Fine 2007, p83).

3.7.2 Putative RR-coreference

Fine does not discuss the mental analogue of putative SR-coreference. However, given the close similarities between RR-coreference and SR-coreference, we can reconstruct an account of putative RR-coreference holding between non-externally-coreferential records.

Using the conclusion of my discussion of putative SR-coreference as guide, I suggest the following definition of putative RR-coreference: where r and s are records belonging to thinker T:

> r and s are *putatively RR-coreferential* iff T must accept for the purposes of at least part of her reasoning that r and s are RR-coreferential in order to count as rational

Again, we have to assume a highly idealised understanding of what it is to 'accept for the purposes of reasoning' as thinkers rarely have non-dispositional attitudes about their records. Presumably, this highly idealised acceptance is realised in how the thinker treats records when reasoning with them.

3.7.3 An answer to the costaging question?

With a clearer understanding of de jure coreference, and still using the simplifying assumption, it's possible to consider a revised version of CoStaging-8, CoStaging-9:

CoStaging-9 r_t and s_t are costaged iff r_t and s_t are RR-coreferential, or r_t and s_t are putatively RR-coreferential.

We should abandon the simplifying assumption, and consider relations between stages of record-components. I suppose that just as NP-occurrences (components of utterances) can be SR-coreferential or putatively SR-coreferential, component-stages can be RR-coreferential or putatively RR-coreferential. This gives CoStaging-9*, which is compatible with a single record-stage being about more than one thing.

CoStaging-9* r_t and s_t are costaged iff c_{r_t} is part of r_t and c_{s_t} is part of s_t, and either c_{r_t} RR-corefers with c_{s_t} or c_{r_t} putatively RR-corefers with c_{s_t}.

What kind of thing is the potentially RR-coreferential part of a record depends on what picture of files we are working with.[35] On the node picture, c_{r_t} and c_{s_t} are a shared node. On the token picture, c_{r_t} and c_{s_t} are referential-tokens of the same type.

[35] See 2.2.1.

CoStaging-9* seems to have considerable potential as an answer to the costaging question. Like CoStaging-7, CoStaging-9* is influenced by a study of linguistic coreference. But unlike CoStaging-7, CoStaging-9* gives an answer to the costaging question in mental rather than linguistic terms.

Moreover, my objection to CoStaging-6[36] was that a thinker can treat record-stages as about a single thing even when those record-stages are about different things. In contrast, we know that if record-stages are RR-coreferential, then they are about the same thing. So CoStaging-9* seems compatible with the reference-fixing role of mental files.[37]

And CoStaging-9 seems compatible with the other parts of the core account of files. The core account implies that costaged record-stages are treated as records about the same object. We know that representational requirements are accessible to thinkers, so we currently have no reason to suppose that a thinker would not treat RR-coreferential or putatively RR-coreferential record-stages as about the same object.[38] This allows for files to play the MOP role. We can suppose that if a rational thinker has costaged records r_t and s_t, it is accessible to her that those records are costaged and so she will not end up in a Frege-case in virtue of reasoning as if r_t and s_t are about different things. And if a thinker has records r_t and u_t that are about the same thing but are not costaged, then it is not necessarily accessible to her that r_t and u_t are about the same thing, explaining why she might reason as if r_t and u_t are about different things, and so end up in a Frege-case.

One worry might be that representational requirements of coreference are not equivalence relations. Allowing for multi-reference records, we don't expect costaging to be an equivalence relation, but if r_t and s_t are costaged in virtue of component-stages c_{r_t} and c_{s_t} respectively, and if s_t and u_t are costaged in virtue of component-

[36] See p67.
[37] See 2.1.3.
[38] Though in 3.8, I suggest this first impression is incorrect.

stages c_{s_t} and c_{u_t} respecively, then we expect that r_t and u_t are costaged in virtue of component-stages c_{r_t} and c_{u_t} respectively.[39] So we may be concerned that because SR-coreference is not always transitive, RR-coreference is non-transitive, and so cannot be used in an answer to the costaging question.

However, whilst Fine claims that in the *inter*personal case, SR-coreference is non-transitive,[40] this does not give us reason to think that SR-coreference is non-transitive in the *intra*personal case. And when we are thinking about RR-coreference, clearly our model should be *intra*personal SR-coreference rather than *inter*personal SR-coreference. The interpersonal phenomena that lead to Fine claiming that SR-coreference is non-transitive in the interpersonal case have no correlate in the *intra*personal case.

Interestingly, although Fine presents this consideration as part of an objection to mental files accounts,[41] Fine himself considers something like CoStaging-9*:

> [I]t is hard to know what talk of mental files is meant to convey. Perhaps one thing it may reasonably be taken to convey is that certain items of information are *stored* together in a single "location," while other items of information are not...
>
> [I]in virtue of what will information be stored in the same location or in a different location?...surely the answer to the question is that the location will be the same when the information represents its objects as the same.
>
> (Fine 2007, pp67-68)

Nonetheless, despite CoStaging-9*'s potential as an answer to the costaging question, we should reject CoStaging-9*. There are two reasons for this. First, Fine is committed to a picture where representational requirements may not be explicit to the thinker. As a result, CoStaging-9* is not compatible with the core account of files after all. Second, both SR-coreference and RR-coreference are defined in terms of manifest consequence, and it is possible to raise serious questions about the idea

[39] See 2.2.1.
[40] See 3.6.2.
[41] I discuss his objection in 6.4.2.

of manifest consequence.

3.8 Accessibility, explicitness and costaging

Fine argues that we should accept a properly formulated principle of *transparency*, i.e. we should accept that semantic requirements are accessible to the understanding (Fine 2007, pp61-62). This is why his semantic relationism is stated in terms of *manifest* rather than *classical* consequence.[42]

Explaining what it is for a fact to be accessible to the understanding, Fine writes:

> [A] fact concerning a given language, or portion of language, L is *accessible to the understanding* if any rational and reflective individual who understands L is thereby in a position to know that the fact obtains.
>
> In saying that the cognizer would be in a *position* to know, I mean that he would know as long as nothing short of further empirical knowledge stands in the way of his knowing. He gives the matter some thought, the right questions are put to him, he is not confused, etc.; and, of course, he should be allowed to have whatever concepts are required to reflect on his own use of the language.
>
> (Fine 2007, p60).

However, Fine's conception of accessibility means that T might not be able to report or act on P even though it is accessible to T that P. Fine's own example of this is in a case involving coordinated general terms. Suppose various general terms are introduced by stipulation, and all are given long definitions. Among these new general terms are the terms 'glub' and 'flox'. The definitions of 'glub' and 'flox' are so long that even someone who remembers and is able to use both definitions may not immediately recognise that the terms have the same definition, and may in fact doubt that 'glub' and 'flox' have the same definition. Fine claims that from the fact that it is a semantic requirement that a 'glub' is something which is *F or G or H... or N*, and from the fact that it is a semantic requirement that a 'flox' is something

[42] See 3.6.1 and (Fine 2007, pp48-49).

which is *F or G or H... or N*, then under just manifest consequence it is a semantic requirement that 'glub' and 'flox' are coextensive (Fine 2007, pp130-131).

A natural response to the glub/flox case would be to say: because it is possible for users of the terms 'glub' and 'flox' to doubt that they have the same definition, then it is not accessible to the thinker that 'glub' and 'flox' have the same definition, and hence Fine is mistaken in claiming that it is a semantic requirement that 'glub' and 'flox' are coextensive.

But this misinterprets Fine's understanding of 'accessible'. There is nothing in Fine's understanding of 'accessible' which means P is not accessible to T if it would take T some time to work out that P. This means that there can be semantic requirements in a speaker's language that the speaker cannot immediately access, either to report on or to act upon.[43]

Imagine a simple task — sorting a pile of cards containing sentences like 'a is a chicken' and 'b is a glub' into two piles, one of glubs and one of things that aren't glubs. Suppose the person carrying out the task is someone who understands and can use the terms 'glub' and 'flox'. We might expect her to put cards reading 'o is a flox' into the 'not-glub' pile, especially were she to be carrying out the selection task under time pressure — despite the fact that it is accessible to the thinker that o is a flox iff o is a glub.

We can describe the problem by saying that although the semantic requirement that 'glub' and 'flox' is accessible to the card-sorter, this requirement isn't *explicit* to her. I don't stipulate a sharp cut-off between cases where P is explicit to T, and cases where P is not explicit to T, but rather I suggest using the intuitive distinction between cases where we would say that P is available for T to report on,

[43]I distinguish being able to act on and report semantic requirements, because some semantic requirements may be immediately accessible to act upon even if one cannot report on them. For example, if I say to a child "If you like Mary, give her a biscuit", the child can demonstrate that she accesses the semantic requirement on the occurrence of 'her' that it corefer with 'Mary', by giving Mary a biscuit (if she likes Mary). The child can demonstrate this access to semantic requirements without having the conceptual sophistication or inclination to report on semantic requirements.

use in reasoning, or act upon, without T having to take time to recall or work out that P, and cases where we would say P is only available for T to report on, use in reasoning, or act upon, if T takes time to recall or work out that P.[44] On this account of what it is to be 'explicit', it is not explicit to me that the Prime Minister of Australia is called 'Julia Gillard' when I take a few seconds to recall her name (even though given time I can recall her name without prompting). Though having recalled the Australian Prime Minister's name, her name may be explicit to me for some time afterwards. In the card-sorting case, a language-user would sort a card reading 'o is a flox' into the 'not-glub' pile only if it wasn't explicit to her that 'glub' and 'flox' are coextensive, even if the thinker would be able to work out that 'glub' and 'flox' are coextensive given a little time and the appropriate prompting.

To allow for false beliefs, I use 'explicit' non-factively, so it can be explicit to T that P whilst it is also the case that $\neg P$.

It's possible to develop examples of semantic requirements of coreference that may not be explicit to a competent language user. Suppose we introduce various new words that function semantically like proper names but have their references fixed by long definite descriptions. Among these are two names, 'Euphemia' and 'Oswald', which have their reference fixed by the same definite description 'the unique F, G, H... N'. It will be accessible to someone who understands the stipulations by which they were introduced, that 'Oswald' and 'Euphemia' are semantically required to corefer, and so SR-corefer.[45] Although it is accessible to competent users of 'Oswald' and 'Euphemia' that they SR-corefer, it may well not be explicit to all users of 'Oswald' and 'Euphemia' that they SR-corefer, or that they externally-corefer. In fact, just as thinkers may doubt whether 'glub' and 'flox' have the same definition, users of 'Oswald' and 'Euphemia' may doubt whether 'Euphemia' and 'Oswald' have

[44] We might think that it always takes a little time (recorded in milliseconds rather than seconds) to recall even the most explicit semantic facts.

[45] I assume that the definite description 'the unique F, G, H... N' is satisfied by exactly one object.

the same reference.[46]

It's also possible to develop examples of representational requirements of coreference that are not explicit to the thinker, i.e. that are accessible but are not immediately available for the thinker to report on or use in reasoning. Make the plausible assumption that a thinker who can understand occurrences of 'Euphemia' and 'Oswald' can have thoughts about Euphemia and Oswald, with the reference of those thoughts fixed by the same long descriptions. Those thoughts about Euphemia and Oswald are RR-coreferential, but it may well be the case that it is not explicit to the thinker that those thoughts are externally-coreferential. And so it may be open to the thinker to treat her thoughts about Euphemia and Oswald as thoughts about different things.

And this means that we must reject CoStaging-9*, because CoStaging-9* is after all incompatible with the core account of files.[47] The core account states that costaged records r_t and s_t contain referential component-stages $\mathrm{c}_{\mathrm{r}_t}$ and $\mathrm{c}_{\mathrm{s}_t}$ respectively, such that the thinker treats $\mathrm{c}_{\mathrm{r}_t}$ and $\mathrm{c}_{\mathrm{s}_t}$ as about the same object. Suppose Mary is familiar with the names 'Oswald' and 'Euphemia', and is able to have thoughts about Oswald and Euphemia, though it is not explicit to her that her thoughts about Oswald and Euphemia are about the same object.

Mary hears a trusted informer claim (44) and (45):

(44) Euphemia is tall.

(45) It's not the case that Oswald is tall.

On the basis of these utterances, by t Mary has formed the beliefs she would herself express using utterances of (44) and (45). These beliefs are recorded by record-stages u_t and v_t respectively.

[46] This should raise concerns with the fact that Fine treats SR-coreference as a development of the idea of representing as the same. A necessary condition on two NP-occurrences representing as the same is that anyone who understands the discourse in which they occur cannot sensibly question whether the NP-occurrences corefer.

[47] See 2.1.3.

As it is not explicit to Mary that her thoughts about Oswald and Euphemia are about the same thing, then u_t and v_t will not contain any component-stages that Mary treats as about the same object. So u_t and v_t fail to meet a necessary condition on being costaged. Nonetheless, u_t and v_t contain RR-coreferential component-stages, indicating that CoStaging-9* is unsuccessful as an answer to the costaging question.

This case is also evidence that Fine is not successful in explaining Frege-cases just using semantic relations. Fine's idea is that Frege-cases arise when thinkers have non RR-coreferential thoughts about a single object. However, if we allow that we explain Mary's mistake by claiming "she doesn't realise that Euphemia is Oswald" then we can say that Mary is in a Frege-case, even though the component-stages representing Euphemia and Oswald in u_t and v_t are RR-coreferential.

The general point is clearest when we make the simplifying assumption that all records are single-reference. If it is not explicit to a thinker that record-stages r_t and s_t are about the same object (if about anything at all) then the thinker might treat r_t and s_t as about different objects, and as a result might end up in a Frege-case involving r_t and s_t. If a thinker does not treat r_t and s_t as about the same object, r_t and s_t cannot be costaged. And if a thinker is in a Frege-case involving r_t and s_t, r_t and s_t cannot be costaged.[48] Hence, we cannot expect any answer to the costaging question from a relation between r_t and s_t that is compatible with it not being explicit to the thinker that r_t and s_t are about the same object (if about anything at all).

Classical and manifest consequence (as defined by Fine) both fail to preserve the explicitness associated with costaging, so we cannot expect an answer to the costaging question to be given in terms of a phenomenon closed under classical or manifest consequence.

[48] By the Frege-constraint on MOPs (see 1.1.3), and the claim that the MOP role is played by mental files (see 2.1.3).

3.9 Manifest consequence and multiple takes

A further reason to be dissatisfied with CoStaging-9* is also a reason to be dissatisfied with Fine's explanation of Frege-cases in terms of semantic and representational requirements of coreference. Semantic and representational requirements are defined in terms of manifest consequence. However, there is reason to think that the idea of manifest consequence as Fine defines it is unsatisfactory, and that it is impossible to recover an idea of manifest consequence adequate for use in a semantic account based on semantic requirements.

The motivation for moving from semantic facts to semantic requirements, and with it the motivation for considering manifest rather than classical consequences, was to give priority in semantics to what is accessible to the understanding. However, if this is the motivation for supposing that semantic requirements are closed under manifest consequence, it is difficult to see why Fine characterises manifest consequence as he does. Fine gives the following definition of manifest consequence:[49]

> Say that p' is a *differentiation of* the proposition p if it is the result of replacing distinct occurrences of the same object by distinct objects (this corresponds to the possibility that even though the objects are in fact the same they may not appear to be the same to the cognizer.) The proposition q will then be a manifest consequence of the propositions $p_1, p_2, p_3, \ldots$ if, for any differentiation $p'_1, p'_2, p'_3, \ldots$, there is a differentiation q' of q for which q' is a classical consequence of $p'_1, p'_2, p'_3, \ldots$.
>
> (Fine 2007, p48)

But it is difficult to see why, when providing differentiations of propositions, we only replace *objects*. The point of differentiation is to acknowledge that if we take a coarse-grained view of propositions, an ideal cogniser might have different takes on a single object, and that if this were the case, the ideal cogniser would not be in a position to infer all the classical consequences of the propositions she knows.

[49] Fine's propositions are Russellian — structured entites built from from the referents of expressions.

However, there is no reason to think that the problem for a coarse-grained view of propositions stops with having different takes on different objects.

Once we allow that a thinker might have different takes on the same object, we must allow that the thinker has different takes on certain properties and relations. For example, if we suppose Peter has different takes on Paderewski, we should allow he has different takes on properties such as *being Paderewski's favourite piano*, and relations such as *being preferred by Paderewski to*.

Additionally, on many plausible theories of semantics, it is possible to have multiple takes on things other than objects. For example, we might have multiple takes on the denotation of 'cat' (see e.g. Loar 1985) and 'sofa' (Burge 2007 [1979]). And we might think that the expressions 'twelve inches from' and 'one foot from' are associated with different takes on the same relation. Different takes on properties and relations will prevent an ideal cogniser from inferring all the classical consequences of the propositions she knows, just as having different takes on a single object prevents an ideal cogniser from inferring all the classical consequences of the propositions she knows. So if the point of the move to manifest consequence was to identify what an ideal cogniser could infer even with multiple takes, then we should differentiate not just objects, but anything one might have multiple takes on.[50]

This consideration should force us at least to adjust the definition of manifest consequence, to require differentiation to involve not just replacing objects, but anything that an ideal cogniser might have multiple takes on.

Adjusting the definition in this way leads to different manifest consequences. For example, under Fine's definition of manifest consequence, it is a manifest consequence of $\forall x(Fx \to Gx)$ and Fa that Ga. But suppose a thinker can have multiple takes on F. Then on the adjusted definition of manifest consequence, it is not a

[50]One might worry that 'object' can be taken to mean anything within the range of first-order quantifiers, so this objection only holds if we give 'object' the narrow interpretation I have adopted. But Fine's own examples show he differentiates only individual constants, indicating he uses 'object' in much the same way I do. See p9n.

manifest consequence that *Ga*.

Alternatively, the general terms 'glub' and 'flox' are both introduced using the stipulation that something which is a 'glub'/'flox' is *F or G or H... or N*, and we allow that a thinker might have two takes on *G*, then *contra* Fine, it is not a semantic requirement that 'glub' and 'flox' are co-extensive.[51]

To what extent this adjusted definition of manifest consequence delivers different results from Fine's definition depends on the range of things an ideal cogniser can have different takes on. The greater the number of things an ideal cogniser might have different takes on, the fewer classical consequences turn out to be manifest consequences.

One might think that this is not a significant a problem for Fine's idea of manifest consequence, because we can adjust the definition and still get a usable account of semantic requirements. However, there are considerations that suggest adjusting the definition of manifest consequence disrupts our ability to use semantic requirements in giving a semantics of English.[52]

To give an adequate semantics for English, we need an account of how the 'her'-occurrence in an utterance of (35) comes to refer to Mary.

(35) Last time I saw Mary, Jack had just sent her an email.

Fine's approach is to allow for the *chaining* of semantic requirements.[53] Chaining means that if it is a semantic requirement on the occurrence of 'her' that it derives its referent from the 'Mary'-occurrence, and it is a semantic requirement on the

[51] See 3.8. This kind of consideration disrupts my argument that the Euphemia/Oswald example shows that CoStaging-9* is false. However, my general point remains: manifest consequence does not preserve the explicitness requisite for costaging.

[52] A more general concern, which I don't consider, are the implications of supposing that a thinker might have different takes on relations like *refers to* or *represents*.

[53] I focus on anaphora because Fine's fullest explanation of chaining is in terms of anaphora. Fine also suggests that it is chaining that gets us the compositionality of semantics, for example how the meaning of 'even prime' is derived from the meanings of 'even' and 'prime' (Fine 2007, p126). Compositionality is crucial to language, so if chaining can't be made to work, the problem for semantics is worse than simply lacking an account of how anaphoric pronouns get their reference.

'Mary'-occurrence that it refer to Mary, then it is a semantic requirement on the 'her'-occurrence that it derive its referent from the 'Mary'-occurrence referring to Mary.[54]

However, we might wonder how chaining can work. If we carry out a process of differentiation on all repeated objects to work out manifest consequences, then we should differentiate the repeated object *the 'Mary'-occurrence*. And if we differentiate *the 'Mary'-occurrence*, then it is a classical but not manifest consequence of the requirements *the 'her'-occurrence derives its reference from the 'Mary'-occurrence*, and *the 'Mary'-occurrence refers to Mary* , that *the 'her'-occurrence derives its reference from the 'Mary'-occurrence which refers to Mary*. This reflects the fact that, if an ideal cogniser had one take on the 'Mary'-occurrence for the proposition *the 'her'-occurrence derives its reference from the 'Mary'-occurrence*, and a different take on the 'Mary'-occurrence for the proposition *the 'Mary'-occurrence refers to Mary*, she could not infer that *the 'her'-occurrence derives its reference from the 'Mary'-occurrence which refers to Mary*.

We might think that Fine offers considerations that mean we don't have to differentiate NP-occurrences. Fine claims that:

> [S]yntax is transparent, even if semantics is not; and one's take on the expressions of the language should always be presumed to be the same, even if one's take on their referents is not.
>
> (Fine 2007, p109).

If we read this as discussing 'occurrences' rather than expressions, then Fine is claiming that we cannot have different takes on occurrences. And if this is true, then in the example above we don't have to differentiate the 'Mary'-occurrence, and hence it is a manifest and classical consequence that *'her'-occurrence derives its reference from 'Mary'-occurrence which refers to Mary*.

[54]See Fine (2007, p123). Fine's describes chaining in terms of 'expressions' rather than occurrences, but he must be taken to be talking of occurrences, because there is no requirement on the expression 'her' that it corefer with the expression 'Mary', or 'Sarah', or any other possible antecedent.

The difficulty is that it is possible to have different takes on a single occurrence. There are many different examples. A competent lipreader may be looking at and listening to a speaker. But that lipreader is wary of tricks, and remains agnostic as to whether she is looking at the speaker she is listening to. Hence, when the speaker says 'I am Jack', the lipreader has two takes on the 'I'-occurrence, allowing for the possibility that the occurrences are produced by different speakers. Alternatively, imagine hearing a live radio broadcast from a rally several hundred metres away, and hearing the same speech a second or so later broadcast from loud-speakers at the rally.[55] Again, a competent but cautious language-user might be agnostic about whether she is in fact hearing the same speech twice, and so maintain two different takes on each occurrence. Alternatively, imagine a competent language-user coming across a sign reading 'Market here 21st December, 2012'. Later, she comes across the same sign, but fails to recognise that she is somewhere she has been before, and hence that the sign is the same one she saw earlier. She will have two takes on the utterance, and on the occurrences that make up the utterance.

These examples demonstrate that there is no bar on a competent language-user having two takes on a single occurrence. In each case, we cannot accuse the language-user of misunderstanding, only of caution, or not knowing where she is. And so these examples show that there is not some special feature of occurrences that mean we don't have to differentiate occurrences when working out the manifest consequences of semantic requirements. So it is not a manifest consequence of the semantic requirements governing an utterance of (35) that *the 'her'-occurrence derives its reference from the 'Mary'-occurrence' which refers to Mary*, and so this is not a semantic requirement.

This dilemma is this: if we want to account for how anaphora get their reference

[55]Those who have listened to the radio near the Houses of Parliament will be familiar with hearing the chimes of Big Ben live on the radio before hearing them in person.

in terms of some kind of principle of chaining,[56] we have to weaken the definition of manifest consequence so we don't differentiate every object a thinker might have different takes on. We could instead use some weaker idea, for example that to work out the manifest consequences of a thinker's known propositions, we only differentiate those objects the thinker does in fact have different takes on. But if we switch to this weaker account of manifest consequence, we get inaccurate claims about semantic requirements. I take it as a datum that the occurrences of 'Cicero' and 'Tully' in utterances of (33) don't SR-corefer. But now suppose that (33) is uttered by someone who knows that it is a semantic requirement that 'Cicero' refers to o, and that it is a semantic requirement that 'Tully' refers to o, and has just one take on o. We should accept that this thinker will know that it is a semantic fact that the occurrences of 'Cicero' and 'Tully' corefer, but this should not make it a semantic requirement that they corefer. However, on our weaker account of manifest consequence, because the thinker has just one take on o, it is a semantic requirement that the occurrences corefer.

Resolving these issues risks taking me outside the scope of this thesis. However, we should see that this is a second serious reason to be disatisfied with CoStaging-9* as an answer to the costaging question — CoStaging-9* relies on the idea of representational requirements, and with it the idea of manifest consequence. However, it is far from clear that Fine (2007) has given an adequate account of manifest consequence.

3.10 Final remarks

In this chapter, I distinguished de jure coreference from assumed coreference, a relation with which it might otherwise be confused. I considered an influential account of de jure coreference, and argued that it is not able to provide an answer

[56] Or we want chaining to explain semantic compositionality. See p106n.

to the costaging question. In doing so, I clarified several aspects of Fine's account of de jure coreference, and raised concerns with the idea of manifest consequence, and with Fine's explanation of Frege-cases.

Two points are worth highlighting. First, although Fine attempts to give an account of Frege-cases that avoids appeal to modes of presentation, he treats the idea that a thinker can have multiple 'takes' on a single object as a given. This emphasises the importance of studying what is involved in having multiple takes (i.e. MOPs) on a single object, and hence on investigating the proposal that mental files play the MOP role.

Second, CoStaging-9* fails as an answer to the costaging question because on Fine's account of de jure coreference, two NP-occurrences or records can be de jure coreferential without it being *explicit* to the thinker that the occurrences corefer. This introduces an important decision point in theories of de jure coreference: whether de jure coreference is explicitly guaranteed coreference, or merely guaranteed coreference. If de jure coreference is to have any hope of being used in an answer to the costaging question, then de jure coreference must be explicitly guaranteed coreference. Otherwise, the proposed answer to the costaging question will fail for just the same reason CoStaging-9* failed.

In the next chapter, I will discuss accounts of de jure coreference which treat de jure coreference as explicitly guaranteed coreference.

Chapter 4

Explicitly Guaranteed Coreference

4.1 Introduction

In 3.4, I suggested that we might give CoStaging-8 to answer the costaging question: under what conditions are r_t and s_t members of the same file-stage?

CoStaging-8 r_t and s_t are costaged iff r_t and s_t are de jure coreferential

However, as noted, there is a live debate about how to define and characterise de jure coreference. Evaluating CoStaging-8 requires a clearer understanding of what de jure coreference is.

I started by considering Fine's (2007) account of de jure coreference. I identified two characterisations of de jure coreference, RAS-coreference and RR-coreference, and concluded that neither could be used to answer the costaging question.

One problem for using RR-coreference in an answer to the costaging question was that two records may be representationally required to corefer, even though it is not explicit to the thinker that they corefer.[1] I suggested that this highlighted a decision point in accounts of de jure coreference: whether de jure coreference is explicitly guaranteed coreference, or merely guaranteed coreference. Given the constraints of the core account of files,[2] to have any hope of stating the conditions on costaging in terms of de jure coreference, de jure coreference must be explicitly guaranteed coreference.

In this chapter, I evaluate two further accounts of de jure coreference, both on their own merits and for their potential to answer the costaging question. Both accounts treat de jure coreference as explicitly guaranteed coreference. The first (AP-coreference) is that presented by Pinillos (2011). The second (IK-coreference) is my own characterisation, developed out of my discussion of AP-coreference.

There are good reasons to focus on Pinillos's (2011) account of de jure coreference. Distinctively, it offers a thorough definition of de jure coreference which

[1] See 3.8.
[2] See 2.1.3.

does not rely on any theoretical commitments as to how de jure coreference is to be explained.[3] As such, Pinillos's definition of de jure coreference has informed subsequent discussions of the topic (e.g. Recanati 2012). Moreover, to the best of my knowledge, Pinillos is the first to suggest that de jure coreference is non-transitive in the *intra*personal case. If intrapersonal de jure coreference is non-transitive, then we should not expect de jure coreference to answer the costaging question,[4] so it is important to consider evidence that de jure coreference is non-transitive.

4.2 AP-coreference

4.2.1 Introducing AP-Coreference

Call de jure coreference, as Pinillos's characterises it, 'AP-coreference'. Pinillos introduces AP-coreference as a relation holding between NP-occurrences that not only externally-corefer, but do so whilst representing as the same (Pinillos 2011, p303).[5]

He provides several examples of utterances containing AP-coreferential NP-occurrences (Pinillos 2011, p303).

(46) The Prime Minister personally invited Smith, but he didn't show up.

(47) The Prime Minister personally invited Smith, but Smith didn't show up.

(48) The Prime Minister personally invited Smith, but the inconsiderate jerk didn't show up.

(49) The Prime Minister personally invited Smith, but that inconsiderate jerk didn't show up.

[3] Contrast with Fine (2007) who gives thorough definitions of SR- and RR-coreference, but only in terms of a semantic relationist explanation of the phenomena.

[4] See 4.2.2.

[5] See (Fine 2007, p40), and 3.5.

(Pinillos 2011) considers a broader range of examples of de jure coreference than (Fine 2007), including anaphoric epithets such as those in (48) and (49). I will continue to assume that paradigm examples of de jure coreference are anaphoric pronouns and their antecedents, as well as occurrences in a single use of a name.[6]

Pinillos identifies three characteristics of de jure coreferential NP-occurrences (2011, pp303-305), which can be illustrated using an utterance of (50).

(i) *Aprioricity*: a rational agent, who fully grasps an utterance of (50) and of (50_e) knows that the utterance of (50) entails the utterance of (50_e).[7]

(50) Jack wants to be an actor, but he works in a coffee shop.

(50_e) Someone who wants to be an actor works in a coffee shop.

(ii) *Attitude closure*: If an utterance of (50) is embedded in the 'that-clause' of 'thinks' to form (50_t), an utterance of (50_t) entails an utterance of (50_{te}).[8]

(50_t) Mary thinks that Jack wants to be an actor but he works in a coffee shop.

(50_{te}) Mary thinks that someone who wants to be an actor works in a coffee shop.

(iii) *Knowledge of conditional coreference*: A competent language user who fully understands an utterance of (50) can deduce a priori just in virtue of understanding the utterance, that if both the 'Jack'-occurrence and the 'he'-occurrence refer, then they externally-corefer.

[6] See 3.4.

[7] Throughout this chapter, I grant Pinillos the idea that we have clear intuitions about what is involved in 'fully grasping' utterances. I also follow Pinillos in discussing entailments between utterances (or 'sentence uses' in Pinillos's terminology). The idea of utterances entailing other utterances may seem unfamiliar, until it is remembered that utterances are simply sentences indexed to a context.

[8] The idea is that (50_t) attributes the same thought to Mary as would be expressed by Mary producing the embedded utterance of (50) (Pinillos 2011, p304n).

Pinillos claims that these characteristics are not merely typical of de jure coreference, but are in fact definitive of it (Pinillos 2011, pp305-306).

Where D is a truth-evaluable sentence utterance or sequence of sentence utterances, D does not contain 'x', and A and B are externally-coreferential NP-occurrences in D:

> 'A' and 'B' are [*AP-coreferential*] if and only if (i) any rational agent who fully grasps D and '$\exists x(\ldots x \ldots x \ldots)$' is in a position to see that the latter follows from the former; (ii) 'S thinks D' entails 'S thinks $\exists x(\ldots x \ldots x \ldots)$'; and (iii) any rational agent who fully grasps D will know of 'A' and 'B' that: if they both refer, they refer to the same object.
>
> (Pinillos 2011, p306, adapted to my terminology)

AP-coreference is defined as a relation between NP-occurrences in a truth-evaluable construction, made by a single language-user in a single discourse. Pinillos mentions that there is a way of understanding AP coreference, or "something close to it, as possibly holding across distinct discourses (and also across participants)" (Pinillos 2011, p315n). This means that AP-coreference is of limited use as a characterisation of de jure coreference in general. It is unable, for example, to characterise de jure coreference in utterances produced by different speakers,[9] or in non truth-evaluable constructions, such as utterances of (51).

(51) Check on Panya, and make sure she isn't stuck in the cat-flap.

Test (iii) seems readily applicable to cross-speaker and non truth-evaluable cases. Presumably, AP-coreference is not defined for such cases because tests (i) and (ii) cannot be adapted for use in these types of case.

AP-coreference is defined as a relation between externally-coreferential NP-occurrences. In 3.6.3, I suggested that there is a relation of de jure pseudo-coreference holding between some referential NP-occurrences that do not externally corefer, as in utterances of (42).

[9] E.g. IX, see p77.

(42) The best thing about Father Christmas is that when Father Christmas visits he brings presents.

As Pinillos points out, nothing in the three tests for AP-coreference blocks some non externally-coreferential NP-occurrences from passing these tests (2011, p318). Therefore, I assume that there is a relation of AP-pseudo-coreference, holding between referential but not externally-coreferential NP-occurrences iff they pass all three tests for AP-coreference.

Pinillos also provides an explanation of AP-coreference. Like Fine (2007), Pinillos gives a semantic relationist explanation of AP-coreference. However, rather than give that explanation in terms of semantic requirements of coreference, Pinillos proposes that AP-coreferential NP-occurrences (and AP-pseudo-coreferential NP-occurences) instatiate a newly posited semantic-relation: *p-linking* (2011, p318).[10] Some non-referential NP-occurrences are also p-linked. For example, on one reading of (25), the underlined NP-occurrences do not refer.

(25) Every cat caught a mouse and killed it.

But even on this reading of (25), Pinillos claims the underlined NP-occurrences are p-linked (2011, p319).

My concern is largely with Pinillos's characterisation of de jure coreference as AP-coreference rather than with its explanation. Nevertheless, because AP-coreferential NP-occurrences are supposedly p-linked, identifying the characteristics of AP-coreference will enable me to identify some of the characteristics of this putative new semantic relation.

[10]Lawlor (2010) employs a similar strategy, arguing for a semantic-primitive linking both de jure coreferential and de jure pseudo-coreferential NP-occurrences.

4.2.2 The non-transitivity of AP-coreference

AP-coreference is presumed to be reflexive and symmetric (Pinillos 2011, p322). Notably, however, Pinillos argues that AP-coreference is non-transitive. This claim is interesting in itself. AP-coreference is a characterisation of de jure coreference, and de jure coreference is typically assumed to be a transitive relation.[11] And if de jure coreference turns out to be non-transitive, this will place limits on our explanations of the phenomenon. For example, many explanations of de jure coreference appeal to "third objects" — for example, syntactic types, indices, discourse referents or mental files. The idea is that a thinker who fully understands the discourse will associate the de jure coreferential NP-occurrences with the same third object.[12] However, if de jure coreference is non-transitive, such explanations are disrupted.

But the claim that de jure coreference is non-transitive should be of interest to anyone hoping to use de jure coreference in answering the costaging question. Suppose we adapt CoStaging-8 to move beyond the simplifying assumption and to allow for empty mental files. This gives CoStaging-8*:

CoStaging-8* r_t and s_t are costaged iff c_{r_t} is part of r_t and c_{s_t} is part of s_t, and either c_{r_t} de jure corefers with c_{s_t} or c_{r_t} de jure pseudo-corefers with c_{s_t}.

In 2.2.1, I pointed out that on non-container pictures of files, *being members of the same mental file* is not be an equivalence relation. But nonetheless the picture is that records are costaged in virtue of containing some component: a particular node, or tokens of some particular type. This means that if r_t and s_t are costaged in virtue of component-stages c_{r_t} and c_{s_t} respectively, and s_t and u_t are costaged in virtue of c_{s_t} and c_{u_t} respectively, then r_t and u_t should be costaged in virtue of

[11] Fiengo and May imply that 'grammatically determined' coreference is an equivalence relation (2006, p37), Taylor implies that 'explicit' coreference is an equivalence relation (2003, p2), and in his doctoral dissertation Wasow proposed a Transitivity Condition on anaphoric relations (see Lasnik 1976, p13). As I discussed in 3.6.2, Fine considers failures of transitivity only in *inter*personal cases.

[12] See Pinillos (2011, pp312-316), and 2.2.2.

component-stages c_{r_t} and c_{u_t}. This means, if r_t and s_t are costaged in virtue of some relation obtaining between c_{r_t} and c_{s_t}, that relation should be transitive.

Hence, if AP-coreference is non-transitive, and AP-coreference correctly characterises de jure coreference, then de jure coreference is non-transitive and we should reject CoStaging-8*.

Pinillos gives two sets of examples in his attempt to demonstrate that AP-coreference is non-transitive. In the remainder of this section, I discuss Pinillos's first set of examples, and argue that it is not straightforward that these examples demonstrate the non-transitivity of AP-coreference. Rather, whether the relevant NP-occurrences pass the tests for AP-coreference depends on our account of the reference of 'confused' NP-occurrences. In 4.2.3, I discuss Pinillos's second set of examples and argue that they don't demonstrate the non-transitivity of AP-coreference.

Pinillos's first set of examples of the supposed non-transitivity of AP-coreference are utterances of (52).[13]

(52) a. We were debating whether to investigate both Hesperus$_A$ and Phosphorus$_B$; but when we got evidence of their true identity, we immediately sent probes there$_C$.

b. As a matter of fact, my neighbour John$_A$ is Professor Smith$_B$, you will get to meet (the real) John Smith$_C$ tonight.

c. Hesperus$_A$ is Phosphorus$_B$ after all, so Hesperus-slash-Phosphorus$_C$ must be a very rich planet.

In each case, Pinillos claims that the C-occurrence is AP-coreferential with the A- and B-occurrences, but that the A- and B-occurrences are not AP-coreferential

[13]These examples are taken in full from Pinillos (2011, p315). I have added subscript labels to significant NPs to facilitate discussion.

with one another. He claims that these examples therefore demonstrate that AP-coreference is non-transitive.

> Consider [(52a)]. Anyone who fully understands it will know of 'there' and 'Hesperus' that they refer to the same thing if they refer at all. The same goes for 'there' and 'Phosphorus'. However, people who fully understand [(52a)] don't have to know that 'Hesperus' and 'Phosphorus' refer to the same thing if they refer at all.
>
> (Pinillos 2011, p315, adjusted sentence numbering)

I agree with Pinillos that the A- and B-occurrences are not AP-coreferential. However, to demonstrate the non-transitivity of AP-coreference, the C-occurrence must pass the tests for AP-coreference with both the A- and B-occurrences. But whether or not it passes these tests depends on what account of confused reference is given.

To illustrate, consider cases that are structurally similar to (52) but involving false identity claims.

(53) a. We were debating whether to holiday in both Amsterdam$_A$ and Paris$_B$; but when we got evidence of their true identity, we immediately headed straight there$_C$.

b. Amsterdam$_A$ is Paris$_B$ after all, so Amsterdam-slash-Paris$_C$ must be a very lovely city.

c. Bacon$_A$ was Shakespeare$_B$ after all, so he$_C$ was even more talented than many people realise.

In these cases, it is clear that the A- and B-occurrences refer, and what they refer to. However, it is less clear what to say about the reference of the occurrence of the *confused* C-occurrence. There are three options worth considering:[14]

[14] A fourth view is possible: utterances of, for example, (53a) express two propositions, in one the 'there'-occurrence contributes Amsterdam, in the other it contributes Paris. Considering such a view would require rethinking Pinillos's definition of AP-coreference, which rests on the idea of *utterances* being true.

No-Reference: The confused C-occurrence fails to refer.

Amalgam: The confused C-occurrence refers to an amalgam or mereological sum of the referents of the A-occurrence and B-occurrence.

Single Reference: The C-occurrence is not confused. It externally-corefers with either the A-occurrence or B-occurence.

No-Reference is the natural position if confusion undermines reference.[15] The Amalgam view is suggested by Fine (2007, pp126-127).[16]

Intuitive appeal for the Single Reference view comes from the fact that in certain contexts, it is easy to 'hear' the C-occurrence as referring to just one of the antecedents. For example, if the topic of the discourse preceding an utterance of (53a) had been Amsterdam, and Paris had not yet been mentioned, then it is easy to hear the occurrence of 'there' referring to just Amsterdam. The suggestion is that in many discourses it is possible to be clear as to which antecedent the C-occurrence externally-corefers with.

Further motivation for accepting Single Reference comes from considering 'good' cases, where the identity claim is true, but where it is not widely known, e.g. utterances of (54).

(54) Rigil Kentaurus is Toliman after all, so it must be very important for navigation.

In such cases, it is easy to hear the 'it'-occurence as anaphoric on, and hence de jure coreferring with just the occurrence of 'Rigil Kentaurus' rather than on both 'Rigil Kentaurus' and 'Toliman'.

However, there are also circumstances where it is less easy to hear the C-occurrence as anaphoric on one particular antecedent, for example in 'slash'-cases

[15] This suggestion is made by Recanati (2012, p132), among others.

[16] Though Fine's suggestion is not made in response to this kind of example.

like (53b). Then we might implement Single Reference by suggesting that the C-occurrence externally-corefers with either the A- or B-occurrence, but it is vague which one.[17]

The point of setting up the confused examples listed as (53) is that a rational agent, on encountering an utterance of (52), does not know a priori whether its identity claim is correct. So for all the rational agent knows a priori, she might be in a confused situation, like (53) rather than (52). Hence, what happens if the C-occurrence is confused is very important for whether the NP-occurrences pass test (iii), i.e. whether a rational agent can know a priori of both the A- and B-occurrences, that if it and the C-occurrence both refer, then it and the C-occurrence externally-corefer. With the options for confused reference set up, we can see how each affects the argument for the non-transitivity of AP-coreference.

Follow Pinillos and assume that the utterances of (52) don't occur in a context where it is clear that the C-occurrences externally-corefer with a particular one of the A- and B-occurrences. And also assume (as Pinillos must) that a rational agent knows which of the accounts of confused NP-occurrences is correct. If we don't make this assumption, then none of the occurrences can be AP-coreferential with one another, as they would automatically fail test (iii).

Whatever account of confused reference is given, the A- and C-, and B- and C-occurrences of utterances of (52b) and (52c) clearly pass tests (i) and (ii) for AP-coreference, in virtue of the A- and B-occurrences being part of an identity-claim of the form $a = b$ (I leave readers to demonstrate this for themselves). Matters are less straightforward for utterances of (52a) because the identity is not explicitly claimed. But to demonstrate the non-transitivity of AP-coreference, only one of these examples needs to work. So the fact that the A- and C-, and B- and C-

[17]Or we might accept a hybrid view, where Single Reference is correct for some cases, and Amalgam or No Reference is correct for others. For ease of exposition, I assume that hybrid views are incorrect.

occurrences of utterances of (52b) and (52c) clearly pass tests (i) and (ii) for AP-coreference means we should focus our attention on whether they also pass test (iii). If they pass test (iii), then these examples demonstrate that AP-coreference is non-transitive.

We can run test (iii) under the different assumptions about the reference of the confused C-occurrence. If *No-Reference* is correct, a rational agent who fully grasps an utterance of one of the sentences of (52) knows a priori that if both the A-occurrence and the C-occurrence refer, then they externally-corefer. This is because she knows a priori that either the identity claim is true, in which case the C-occurrence refers, and externally-corefers with the A-occurrence; or the identity claim is false, in which case the C-occurrence fails to refer, and hence it is not the case that the A-occurrence and the C-occurrence both refer. *Mutatis mutandis*, the same is true of the B-occurrence and C-occurrence. Hence, if No-Reference is true, the A-occurrence and B-occurrence AP-corefer with the C-occurrence but not with one another, and utterances of (52) show that AP-coreference is non-transitive.

If *Amalgam* is correct, a rational agent who fully grasps an utterance of one of the sentences of (52) knows a priori that if the identity claim is true, then the A- and C-occurrence both refer and externally-corefer. And she knows a priori that if it is false, then the A- and C-occurrences both refer but do not externally-corefer. The C-occurrence refers to an amalgam of the referents of the A- and B-occurrences. Hence, the rational agent does not know a priori that if both the A-occurrence and C-occurrence refer, then they externally-corefer. *Mutatis mutandis*, the same is true of the B-occurrence and C-occurrence. Therefore, if Amalgam is true, the A-, B- and C-occurrences do not AP-corefer with one another, and AP-coreference is not shown to be a non-transitive relation.

Suppose *Single Reference* is correct. We already know we are in a context where it is unclear which antecedent occurrence the C-occurrence externally-corefers with

in an utterance of (52). In this situation the rational agent knows a priori that either the A- and C-occurrences, or the B- and C-occurrences externally-corefer if they refer at all, but the rational agent can't know which of these is true. Hence, in such contexts, none of the A-, B- or C-occurrences AP-corefer with one another, and utterances of (52) are not evidence that AP-coreference is non-transitive.

Note that if the 'there'-occurrence in (52a) is anaphoric on both the 'Hesperus'- and 'Phosphorus'-occurrences, and either Amalgam and Single Reference is correct, then in utterances of (52a), the 'there'-occurrence is not AP-coreferential with any antecedent occurrence. So if Amalgam or Single Reference were correct, then in those unusual cases where a singular pronoun is used with multiple non de jure coreferential antecedents, there can be referring anaphoric pronouns that are not de jure coreferential with any antecedent.

To summarise: Pinillos argues that AP-coreference is a non-transitive relation. But I have shown that it is far from clear that his first set of examples demonstrate the non-transitivity of AP-coreference. Rather, he only demonstrates the non-transitivity of AP-coreference if we assume that No-Reference is the correct account of confused reference.

4.2.3 Further evidence for the non-transitivity of AP-coreference?

Pinillos offers a second set of examples, which are also supposed to show the non-transitivity of AP-coreference. [18]

(55) a. Smith$_A$ is wearing a costume, and (as a result) Sally thinks he$_B$ is someone other than Smith$_C$.

[18]The examples are taken in full from (Pinillos 2011, p315), with subscript labels added to facilitate discussion.

b. He$_A$ was in drag, and (as a result) Sally thought that Smith$_B$ wasn't Smith$_C$.

In utterances of (55), Pinillos claims that the B- and C-occurrences both AP-corefer with the A-occurrence, but do not AP-corefer with one another, further demonstrating the non-transitivity of AP-coreference (Pinillos 2011, p316). His argument uses an utterance of (55a) as an example, and goes as follows: if the occurrences of B- and C-occurrences were AP-coreferential, then by test (i),[19] a rational agent would know a priori that an utterance of (55a) entails an utterance of (55a$_e$).

(55a$_e$) $\exists x$(Smith is wearing a costume, and Sally thinks x is someone other than x)

But Pinillos argues that an utterance of (55a$_e$) cannot be entailed by (55a), because it ascribes Sally an incoherent belief and so cannot be true (Pinillos 2011, p316). Therefore, the occurrences of 'he' and 'Smith' inside the that-clause do not AP-corefer. Similarly, Pinillos argues that the B- and C-occurrences fail test (ii) for AP-coreference (Pinillos 2011, p316). For the B- and C-occurrences to pass test (ii), an utterance of (55a$_t$) must entail an utterance of (55a$_{te}$).

(55a$_t$) Jack thinks that Smith is wearing a costume and (as a result) Sally thinks he is someone other than Smith.

(55a$_{te}$) Jack thinks that $\exists x$(Smith is wearing a costume and (as a result) Sally thinks x is other than x)

Pinillos suggests that (55a$_{te}$) ascribes Jack a belief he wouldn't have (presumably because it involves Jack ascribing Sally an incoherent belief). Hence, (55a$_{te}$) is not entailed by (55$_e$), and the B- and C-occurrences also fail test (ii) for AP-coreference.

[19] See p114.

However, these examples don't demonstrate the non-transitivity of AP-coreference, because (55a$_e$) does not ascribe Sally an incoherent belief and hence there is no reason to suppose that (55a$_t$) does not entail (55a$_{te}$). Moreover, it is inappropriate for Pinillos's tests (i) and (ii) for AP-coreference to be applied to NP-occurrences in the complement clause of a non-factive attitude-report. I will explain both these claims in greater detail below.

The B- and C-occurrences have no problem passing test (iii) for AP-coreference. But according to Pinillos, (55a$_e$) ascribes Sally an incoherent belief, and this alleged incoherence is what makes the B- and C-occurrences fail tests (i) and (ii) for AP-coreference. Hence to establish that AP-coreference is non-transitive, it is crucial to establish that (55a$_e$) does in fact ascribe Sally an incoherent belief. However, Pinillos does not succeed in doing this.

It is a familiar idea, from outside the coreference literature, that we don't always ascribe beliefs to others in ways that accurately reflect the way in which the subject of the attitude-report would herself report the beliefs. Suppose a friend, Sylvia, reports that she found someone she met objectionable. She never found out the person's name, but it is clear who the person is from her description. I might report her belief to another friend as 'Sylvia thinks Bobby is objectionable'. The attitude-report is not faithful to the way Sylvia is able to express her beliefs, but is perfectly felicitous. Even in contexts where that attitude report would be misleading and therefore infelicitous, it is far from obvious that the report would be false.

But we also find this phenomenon in the area of de jure coreference. Contra Pinillos, it can be felicitous to ascribe someone beliefs using anaphoric phrases, even when they themselves would not use such phrases to report their own beliefs. For example, we might report Peter's 'Paderewski'-beliefs using (56).[20]

[20] See p13.

(56) Peter thinks that Paderewski is a brilliant musician and that he is a hopeless musician!

The occurrence of 'he' is clearly anaphoric on just the occurrence of 'Paderewski', and such uses of anaphoric pronouns are paradigm cases of de jure coreference. However, it is equally clear that Peter would not report his own beliefs using such an expression. Nonetheless, utterances of (56) are felicitous.

Utterances of sentences like (56) demonstrate that not all belief reports capture how the subject of the belief report would herself report the belief.[21] Apply test (i) to the NP-occurrences in (56): an utterance of (56) might be thought to entail an utterance of (56_e).

(56_e) $\exists x$(Peter thinks that x is a brilliant musician and x is a hopeless musician)

However, so long as an utterance of (56_e) does not reflect the way Peter would report his own belief, there is nothing incoherent about ascribing Peter a belief using (56_e). An utterance of (56_i) would ascribe Peter an incoherent belief:

(56_i) Peter thinks that $\exists x$(x is a brilliant musician and that x is a hopeless musician)

But so long as (56_e) does not reflect the way Peter would report his own belief, then it does not entail (56_i), and no incoherent belief is ascribed.

Pinillos claims, and is right to claim, that sometimes when we report another's beliefs it matters that we get right the way that the believer would herself report those beliefs (Pinillos 2011, p316). But it does not follow that it *always* matters. We must allow for felicitous belief reports that don't accurately capture the way the believer reports her beliefs. And this is exactly the kind of situation we have in utterances of (55) and (56).

[21] See Fine (2007, pp102-104) for a related discussion of the ways felicitous belief reports can deviate from how the subject of the reports would herself report the beliefs. Recanati (2012, pp107-108) makes a similar point.

An utterance of ($55a_e$) only ascribes Sally an incoherent belief if it accurately reflects how Sally would report the belief. But there is no such requirement, and so there is no reason to suppose ($55a_e$) is incoherent. Therefore, the B- and C-occurrences in utterances of (55a) are not shown to fail tests (i) and (ii) for AP-coreference.

Independent of this objection to Pinillos's line of reasoning, it may be that Pinillos's tests (i) and (ii) should not have been applied to the B- and C-occurrences of utterances of (14), because it seems that the tests should not be applied to NP-occurrences in the 'that'-clauses of non-factive attitude reports.[22]

For the occurrences of 'Obama' and 'he' in an utterance of (57) to pass test (i) for AP-coreference, a rational agent must know a priori that an utterance of (57) entails an utterance of (57_e). For them to pass test (ii), an utterance of (57_t) must entail (57_{te}).

(57) Jack thinks Obama is president and that he is tall.

(57_e) $\exists x$(Jack thinks x is president, and that x is tall)

(57_t) Mary thinks that Jack thinks Obama is president and that he is tall.

(57_{te}) Mary thinks $\exists x$(Jack thinks x is president, and that x is tall)

However, on the face of it, a rational agent can fully grasp the content of an attitude-report without thereby a priori knowing whether the NP-occurrences in the that-clause refer. And if this is true, then even if an utterance of (57) entails an utterance of (57_e), a rational agent can grasp an utterance of (57) without thereby a priori knowing that it entails an utterance of (57_e), because as far as the rational agent knows a priori, the NP-occurrences may fail to refer, in which case (57_e) is not entailed.

[22]The same points apply to NP-occurrences in the "that"-clauses of negations of factive attitude reports, such as "Jack doesn't know that…".

Similar problems occur when applying test (ii) to NP-occurrences in that-clauses of non-factive attitude-reports and negations of factive attitude-reports. It is possible that an utterance of (57_t) is true, but that Mary erroneously thinks that the occurrences of 'Obama' fail to refer, and so an utterance of (57_{te}) is not entailed by an utterance of (57_t).

These examples demonstrate either that in utterances of (57), the 'Obama'-occurrence and 'he'-occurrence are not AP-coreferential, or that (i) and (ii) generate false negatives when applied to non-factive attitude reports. The first option seems unlikely. The NP-occurrences in utterances of (58) are paradigm examples of de jure coreference, and hence paradigm examples of AP-coreference.

(58) Obama is president, and he is tall.

It would be extremely surprising if such constructions were to stop being de jure coreferential when embedded in non-factive attitude reports. Instead, we should accept that the range of cases for which AP-coreference is defined is even more limited than previously thought. Not only is AP-coreference undefined for NP-occurrences in non truth-evaluable constructions, it is also undefined for NP-occurrences in the "that"-clauses of non-factive attitude reports and negations of factive attitude reports. Therefore tests (i) and (ii) shouldn't have been used to test for the de jure coreference of the B- and C-occurrences in utterances of (55).

Pinillos's arguments that the B- and C-occurrences of utterances of (55) fail tests (i) and (ii) for AP-coreference are unsuccessful. Furthermore, there is reason to think that tests (i) and (ii) should not have been applied to the B- and C-occurrences in the first place, because AP-coreference is not defined for NP-occurrences occurring in non-factive attitude reports. As the B- and C-occurrences pass test (iii) for AP-coreference, utterances of (55) give no reason to suppose that AP-coreference is non-transitive.

4.3 Concerns with AP-coreference

Pinillos's argument that de jure coreference is non-transitive depends in part on his characterisation of de jure coreference as AP-coreference. But there are several difficulties with AP-coreference as a characterisation of de jure coreference. The first two can be answered by adjusting the definition of AP-coreference. The third is more pressing.

My first concern is with test (ii), as it is currently presented. The semantics of attitude ascriptions are far from straightforward, so there are sensitive issues as to what thought-ascriptions entail what other thought-ascriptions. Test (ii) introduces these into the account of de jure coreference. For example, once we acknowledge that a belief report need not reflect the way the subject of the belief report would report her own belief, then test (ii) is unworkable as it stands. Running test (ii) on an utterance of (50) requires seeing whether an utterance of (50_t) entails an utterance of (50_{te}).[23] But if (50_t) does not reflect how Mary would report her own beliefs, then it does not entail (50_{te}).

So test (ii) must be revised to require that the initial belief report, here the utterance of (50_t), accurately reflects the subject's own way of reporting her belief.

My second concern is that it is sometimes unclear what a 'rational agent' who 'fully grasps' an utterance knows a priori. Hence, it can be unclear whether tests (i) and (iii) are passed, as both turn on what the rational agent knows a priori just by grasping the utterance.

To illustrate, run tests (i) and (iii) on occurrences of the underlined NPs in an utterance of (59).

(59) <u>The truth value of P</u> is TRUE, and <u>the truth value of $(\neg(P \leftrightarrow R) \wedge (R \rightarrow \neg R)) \vee (P \wedge R)$</u> is Mary's favourite truth value.

For the NP-occurrences to pass test (iii), a rational agent who grasps an utterance

[23] See p114.

of (59) must thereby know a priori that the NP-occurrences externally-corefer if they refer at all. But although we would expect that the rational agent could work out that the symbolic phrases are logically equivalent just in virtue of her full grasp of the utterance, we might expect her to require a little time to work this out (we have no reason to suppose that 'rational' means 'immediately omniscient of a priori truths'). And so we expect that it might take the rational agent a little time to work out that if the NP-occurrences refer, they externally-corefer. Similarly, we would expect a rational agent to take a little time to work out that (59) implies (59_e).

(59_e) $\exists x(x$ is TRUE, and x is Mary's favourite truth value)

So to work out if the NP-occurrences pass tests (i) and (iii) for AP-coreference, we must ask: does the fact that it takes a rational agent a little time to work out that P mean that the rational agent doesn't a priori know that P?

There are conflicting temptations here. On the one hand, the a priori knowledge of the rational agent is supposed to be a matter of *knowledge*. So if a rational agent has not worked out that P, she should not count as knowing P even if she could work out that P given the time. On the other hand, there is a long tradition of talking about the 'a priori knowledge' of a rational agent, when it might be better to talk of the agent's *potential* a priori knowledge. On this tradition, a rational agent can count as knowing a priori that P in virtue of being able to work out (a priori) that P.

On the second understanding of a priori knowledge, the NP-occurrences pass tests (i) and (iii) for AP-coreference. But on the first understanding of a priori knowledge, the NP-occurrences do not clearly pass tests (i) and (iii), as merely grasping the utterance is not sufficient for the rational agent to have worked out that the phrases are logically equivalent.

Whilst utterances like (59) highlight the lack of clarity in the notion of what a rational agent knows a priori, the problem reoccurs for utterances like (50). If we

expect the rational agent to take a little time working out that (59) entails (59_e), we can also expect the rational agent to take a little time working out that (50) entails (50_e). And it is this that shows how we resolve this problem. To ensure that the NP-occurrences in (50) pass the tests for AP-coreference, we must interpret the test as involving the rational agent's *potential* a priori knowledge, rather than what the rational agent has in fact got around to working out. Hence, we can suppose that the NP-occurrences in (59) do pass tests (i) and (iii) for AP-coreference.[24]

Therefore, these first concerns with AP-coreference can be resolved by slight adjustments to the definition of AP-coreference, and are not fatal to the AP-coreference account. My third concern is less easily resolved.

As AP-coreference is defined, it can only give an account of de jure coreference in a very limited group of cases: between referring NP-occurrences uttered by a single speaker in a single truth-evaluable discourse, and only between referring NP-occurrences that are not in the that-clause of a non-factive attitude report.[25] As such, it is unable to shed any light on the de jure coreference of NP-occurrences in example IX,[26] or utterances of (51).[27] And so long as tests (i) and (ii) are included in the definition of AP-coreference, there is no obvious way to define AP-coreference for a wider range of cases.

[24]Though as I show in 4.4.1, they do not pass test (ii), so are not AP-coreferential.
[25]See 4.2.3.
[26]See p77.
[27]See p115.

4.4 An alternative account: IK-coreference

4.4.1 Is de jure coreference *explicitly* guaranteed coreference?

Given the problems tests (i) and (ii) cause for the AP-coreference characterisation of de jure coreference, why not drop them and define de jure coreference just in terms of test (iii)?

In contrast to tests (i) and (ii), test (iii) is readily applicable to NP-occurrences in non truth-evaluable discourses, to NP-occurrences in non-factive attitude reports, and to NP-occurrences uttered by different speakers. And the fact that NP-occurrences pass test (iii) is often an explanation of why they pass tests (i) and (ii).[28] So perhaps it is unsurprising that some (e.g. Recanati 2012) just use a version of test (iii) in defining de jure coreference.

But whether test (iii) is sufficient depends on what stance we take on the *explicitness* of de jure coreference. In 3.10, I suggested that we might treat de jure coreference as explicitly guaranteed coreference, or merely as guaranteed coreference.

The significance of this decision point is emphasised by examples such as (59).[29] At first glance, we might not expect occurrences of the underlined NPs to be de jure coreferential. The paradigm examples of de jure coreference are anaphoric pronouns and occurrences in a single use of a name, cases where it is immediately obvious to a competent language user that there is a guarantee of coreference. In contrast, even someone competent at manipulating the symbols in (59) might take a little time to work out that the symbolic phrases are logically equivalent.

One response is to stick to the claim that de jure coreference is guaranteed

[28] For example, we might explain why occurrences of the underlined NPs in an utterance of (46) pass tests (i) and (ii) by pointing to the fact that they pass test (iii). However, this is not the only possible explanation of why NP-occurrences pass tests (i) and (ii). For example, the occurrences of the underlined NPs in an utterance of "Hesperus is Phosphorus" pass tests (i) and (ii) because this utterance, like any identity statement, entails "$\exists x(x \text{ is } x)$".

[29] See also the Oswald/Euphemia example in 3.8.

coreference, and to say that these examples merely illustrate that de jure coreference may not always be explicit to the language user. It appears that Fine (2007) would take this response.[30] Another response is that these examples show that guaranteed coreference isn't enough for de jure coreference, and the paradigm examples show that de jure coreference isn't just guaranteed coreference, but guaranteed coreference where that guarantee is explicit to the language user.

This latter response seems to be the sort favoured by Pinillos, who highlights the role of de jure coreference in reasoning. In particular he associates de jure coreference with the existence of an "*easy* inference" (Pinillos 2011, p306, my italics) from utterances of e.g. (50) to (50_e). Moreover, Pinillos's test (ii) blocks the NP-occurrences in an utterance of (59) from counting as AP-coreferential. In test (ii), there is no requirement that the person to whom the beliefs are attributed has any particular logical abilities, or has taken the time to work out the equivalences. Hence, there is no reason to suppose that an utterance of (59_t) entails an utterance of (59_{te}).

(59_t) Jack thinks the truth value of P is TRUE, and the truth value of $(\neg(P \leftrightarrow R) \wedge (R \rightarrow \neg R)) \vee (P \wedge R)$ is Mary's favourite truth value.

(59_{te}) Jack thinks $\exists x(x$ is TRUE, and x is Mary's favourite truth value)

Pinillos's test (ii) puts him on the side of those who think that de jure coreference is explicitly guaranteed coreference. This explains why we cannot just rely on (iii) in defining de jure coreference. Test (iii) only gets guaranteed coreference, not *explicitly* guaranteed coreference.

How we characterise de jure coreference depends on whether we take the narrower understanding of de jure coreference as *explicitly* guaranteed coreference, or take the broader understanding of de jure coreference as merely guaranteed coreference. Of

[30] See 3.8.

course, 'de jure coreference' is a term of art, and to an extent we can define terms of art as we wish. But 'de jure coreference' is supposed to identify an interesting relation holding between some externally-coreferential NP-occurrences and not others. Deciding how to characterise de jure coreference means taking a stance on which understanding of de jure coreference is more worthy of attention.

I suggest we take the narrower understanding of de jure coreference. This allows the idea of de jure coreference to capture in full what is distinctive about the paradigm cases of de jure coreference: not only that the NP-occurrences are guaranteed to corefer, but that this is immediately obvious to a language user who fully grasps the utterances. This position is also convenient for the purposes of this thesis. As I noted in 3.10, we can only hope for an answer to the costaging question from de jure coreference if de jure coreference is explicitly guaranteed coreference.

4.4.2 IK-coreference

Pinillos (2011) introduces *explicitness* into his account of de jure coreference through test (ii). But test (ii) introduces difficulties for Pinillos's characterisation of de jure coreference (see 4.3). I suggest using what has been learnt from studying AP-coreference to develop a new characterisation of de jure coreference avoiding these difficulties. Call this new characterisation 'IK-coreference'.

Where O_1 and O_2 are referring NP-occurrences uttered in a single discourse E:

> O_1 and O_2 *IK-corefer* iff any competent language user who fully grasps E will, just in virtue of being a competent language user who fully grasps E, immediately know a priori of O_1 and O_2 that: if O_1 and O_2 both refer, then they corefer.

Elements of this definition require further clarification. The notion of a priori knowledge at work in the definition of IK-coreference is different from that in the definition of AP-coreference. On this notion, in order to immediately a priori know

that P, the thinker must have immediately worked out that P (it is not enough that they would be able to work out that P). However, this working out may be tacit. After all, most thinkers don't form explicit beliefs about the NP-occurrences they understand, let alone about their coreference.

The definition relies on a fairly intuitive sense of what counts as immediate a priori knowledge. In this intuitive sense, immediate knowledge is instantly acquired on fully grasping the NP-occurrences in question, and is not the product of conscious reasoning, though it may be grounded in other knowledge, for example knowledge of the language's rules.

The definition of IK-coreference uses the idea of competent language users rather than rational agents, to avoid anything being packed into the idea of rational agents that would grant them immediate knowledge not granted to ordinary language users. The focus on the immediate knowledge of competent language users ensures that IK-coreference captures the explicitness of de jure coreference. The NP-occurrences in utterances of (59) would not be immediately known to be coreferential (if they refer at all) by an ordinary competent language user. Hence, they are not IK-coreferential.

My suggestion is that IK-coreference, although imprecise, is preferable to AP-coreference as a characterisation of de jure coreference. Like the AP-coreference characterisation, it is able to capture the explicitness of jure coreference. But because the definition of IK-coreference contains no equivalent of tests (i) and (ii) for AP-coreference, the relation can be defined for NP-occurrences in a much wider range of cases than AP-coreference is defined for, including non-factive attitude reports, non truth-evaluable constructions, and multiple constructions or constructions uttered by different speakers. And because the definition of IK-coreference contains no equivalent of test (ii), it avoids difficulties with the semantics of thought-ascriptions.

Starting from the IK-coreference characterisation of de jure coreference, I suggest

that we characterise de jure pseudo-coreference as IK-pseudo-coreference. Two referential NP-occurrences are IK-pseudo-coreferential just when they meet the necessary and sufficient conditions on IK-coreference, except that they are non-referring.

4.4.3 Is IK-coreference non-transitive?

Suppose that de jure coreference is characterised as IK-coreference. On this understanding of de jure coreference, is de jure coreference transitive?

Given the similarities between test (iii) for AP-coreference and the definition of IK-coreference, and that the A- and C-occurrences, and B- and C-occurrences, but not the A- and B-occurrences of (52) pass test (iii) for AP-coreference so long as we assume No-Reference, we might expect that IK-coreference is non-transitive so long as we assume No-Reference. However, things arent so simple. IK-coreference is defined in terms of the knowledge of an ordinary competent language user, rather than an idealised rational agent. In 4.2.2, I allowed that the rational agent would know which of No-Reference, Amalgam and Single Reference is correct. However, there is no reason to assume this of competent language users. If one has to know which account of confused reference is correct to count as a competent language user, then competent language users would be a rare species indeed.

So we can assume that ordinary competent language users don't know which account of confused reference is correct. As a result, it is not the case that any competent language user knows (a priori and immediately), of the A- and C-occurrences and of the B- and C-occurrences in utterances of (52), that if they both refer then they corefer. That knowledge would only be available to any competent language user if No-Reference is correct, and one has to know that No-Reference is correct to count as a competent language user. Therefore, none of the NP-occurrences in an utterance of (52) IK-corefer with one another. On the IK-coreference characterisation of de jure coreference, utterances of (52) provide no evidence for the non-transitivity

of de jure coreference.[31]

If we accept that IK-coreference is the best possible characterisation of de jure coreference, then Pinillos's examples have not shown that de jure coreference is non-transitive. However, the C-occurrence in utterances of (52a) is a pronominal anaphora, and so should be referentially dependent on an antecedent. And this means that Pinillos's first set of examples have still shown an interesting result: in unusual cases where there are multiple antecedents for a singular anaphor, that anaphor is not always de jure coreferential with its antecedents even if derives its reference from them.

4.5 De jure coreference and costaging

My purpose in investigating de jure coreference was to evaluate CoStaging-8 as an answer to the costaging question.[32] With a new, clearer understanding of de jure coreference, it is possible to evaluate CoStaging-8.

4.5.1 AP-coreference and costaging

If we follow Pinillos in using AP-coreference to characterise de jure coreference, we may be tempted to assume we can identify AP-coreference as a relation between component-stages of record-stages as well as NP-occurrences, and give CoStaging-10* as an answer to the costaging question.

CoStaging-10* r_t and s_t are costaged iff c_{r_t} is part of r_t and c_{s_t} is part of s_t, and either c_{r_t} AP-corefers with c_{s_t} or c_{r_t} AP-pseudo-corefers with c_{s_t}.

The prospect looks better for CoStaging-10* than attempts to answer the costaging question in terms of RR-coreference. As I showed in 4.4.1, test (ii) ensures that

[31] Pinillos's second set of examples, utterances of (55), also give no evidence for the non-transitivity of IK-coreference. The A-, B- and C-occurences in utterances of (55) all IK-corefer (as readers may confirm for themselves).

[32] See p79. See also p117 for CoStaging-8*.

AP-coreference is a kind of explicitly guaranteed coreference. We therefore expect no repeat of the difficulties with explicitness that were raised for CoStaging-9* (see 3.8).

However, CoStaging-10* has its own difficulties. First, it is far from clear we can work out a satisfying definition of AP-coreference as a relation between component-stages of record-stages. We may hope to find a mental substitute for 'understanding a discourse' to use in giving a mental version of a linguistic relation defined in terms of a language-user understanding a discourse.[33] However, it is extremely difficult to see how we could find a mental substitute for test (ii), as test (ii) involves embedding utterances into attitude reports. And, we cannot simply dispense with test (ii) as test (ii) is what ensures that AP-coreference is *explicitly* guaranteed coreference.

We might think that this is a sign that I have looked in the wrong place for an answer to the costaging question. I should have considered p-linking, the semantic relation that Pinillos proposes to explain AP-coreference and AP-pseudo-coreference.[34] We can suppose that p-linking can obtain between component-stages of record-stages, just as it can obtain between referring NP-occurrences or bound variables and their antecedents. This proposal gives us CoStaging-11*:

CoStaging-11* r_t and s_t are costaged iff c_{r_t} is part of r_t and c_{s_t} is part of s_t, and c_{r_t} is p-linked with c_{s_t}.

However, problems remain. Given how p-linking is introduced, we know little about it as a relation between component-stages of record-stages. But we know of two features of p-linking as a relation between NP-occurrences which should deter us from accepting CoStaging-11*.

First, whilst I have argued that the evidence for the non-transitivity of p-linking is not as strong as Pinillos suggests, the fact remains that if we assume No-Reference,

[33] I attempt this when giving a definition of IK-coreference as a relation between record-component-stages, see 4.5.2.

[34] See 4.2.1.

Pinillos has demonstrated that p-linking is non-transitive. But we cannot answer the costaging question with a non-transitive relation (see 4.2.2). We should be wary of CoStaging-11* at least until we have further investigated the alleged non-transitivity of p-linking. *Mutatis mutandis*, the same objection can be used against CoStaging-10*.

Second, Pinillos's examples may not have definitively proved that p-linking is a non-transitive relation. But they have demonstrated that p-linking does not guarantee shared referential-value. If No-Reference is correct, then in utterances of (53)[35], the A- and C-occurrences, and B- and C-occurrences have different referential-values — the A- and B-occurrences refer, and the C-occurrence fails to refer. But as AP-coreference and AP-pseudo-coreference only requires that if the occurrences both refer then they externally corefer, the occurrences can be AP-pseudo-coreferential (and hence p-linked) despite having different referential-values.

However, the core account of files says that file-stages determine the referential-value of the component-stages in virtue of which record-stages are members of that file-stage.[36] So we cannot use p-linking, or AP-coreference and AP-pseudo-coreference, to answer the costaging question, because these relations would allow files to be costaged in virtue of component-stages with different referential values.

4.5.2 IK-coreference and costaging

We might hope for more success from CoStaging-12*:

CoStaging-12* r_t and s_t are costaged iff c_{r_t} is part of r_t and c_{s_t} is part of s_t, and either c_{r_t} IK-corefers with c_{s_t} or c_{r_t} IK-pseudo-corefers with c_{s_t}.

The first step in investigating CoStaging-12* is to define IK-coreference and IK-pseudo-coreference between component-stages of record-stages.

[35] See p119.
[36] See 2.1.3.

Where c_{r_t} and c_{s_t} are referring component-stages of record stages r_t and $s_{t'}$ respectively, and r_t and $s_{t'}$ belong to a thinker T:

> c_{r_t} and c_{s_t} *IK-corefer* iff T, just in virtue of meeting the conditions on being a competent thinker, immediately knows a priori of c_{r_t} and c_{s_t} that: if c_{r_t} and c_{s_t} both refer, they corefer.

And c_{r_t} and c_{s_t} IK-pseudo-corefer just if they meet the necessary and sufficient conditions on IK-coreference, but are non-referring referential component-stages.

These definitions of IK-coreference and IK-pseudo-coreference for component-stages substitute understanding a discourse with the thinker T simply possessing record-stages. Because linguistic IK-coreference is defined merely in terms of competent speakers rather than idealised rational agents, there is no need to suppose that T is ideally rational, merely that she is sufficiently competent to qualify as a competent thinker. I leave the boundary between competent and not-competent unspecified, relying on an intuitive grasp of the distinction. However, I do require that the knowledge of conditional coreference which is definitive of IK-coreference is the result merely of those abilities that mean that T qualifies as a competent thinker, and not from any additional or special reasoning abilities she may have.

Like CoStaging-10* and CoStaging-11*, an advantage of CoStaging-12* over answers to the costaging question in terms of RR-coreference is that it answers the costaging question in terms of an explicit relation. Over CoStaging-10*, CoStaging-12* has the advantage of being stated in terms of a relation defined as a relation between record-component-stages. And a further advantage of CoStaging-12* over CoStaging-10* and CoStaging-11* is that there is no evidence that IK-coreference is non-transitive.

But there are two points of caution. The first is that nothing in the definition of IK-coreference blocks NP-occurrences or record-component-stages with different referential-values from being IK-pseudo-coreferential. So far, there is no evidence

that IK-pseudo-coreference does obtain between tokens with different referential-values. This might be enough, but ideally we should supplement CoStaging-12* with an argument that IK-pseudo-coreference is such that it cannot obtain between tokens with different referential-values. The argument might take as its starting point some claim of the following sort: IK-coreferential tokens must be in sufficiently simple constructions to allow immediate knowledge of conditional coreference. But only complex constructions result in tokens that do not share referential-value, but would externally-corefer if they referred. If we were to leave CoStaging-12* as our final answer to the costaging question, we would need to explore these issues further.

However, CoStaging-12* is not my final answer to the costaging question, because a concern remains. Most of the time, language-users only form implicit beliefs about NP-occurrences. So only a highly idealised account of what a language-user knows a priori will make IK-coreference widespread enough to work as an characterisation of de jure coreference. Similarly, most thinkers don't form non-dispositional beliefs about their records and component-stages. So if CoStaging-12* is to work as an answer to the co-staging question, we need a highly idealised account of what a thinker knows a priori, in order to allow for IK-coreferential component-stages to be common-place.

In the case of language we have some understanding of when to attribute the language-user this idealised knowledge. We attribute it to her when she realises it in her use of the language, both in her own utterances and in her responses to the utterances of others. In outline, a thinker realises the immediate a priori knowledge that NP-occurrences O_1 and O_2 are coreferential when she immediately treats O_1 and O_2 as if it's guaranteed that if they refer then they corefer, and the rules of her language guarantee that if O_1 and O_2 refer then they corefer.

But we also need some account of the circumstances in which we can attribute the thinker the idealised knowledge required for IK-coreference. The natural place

to look for this is in how the thinker realises this knowledge in her reasoning. But as yet, we don't know what kind of reasoning realises this knowledge. Merely treating component-stages as about the same isn't enough for a priori knowledge of coreference (see 2.1.3).

Even if we do identify how the thinker realises this idealised knowledge, we might think that this realisation provides a better answer to the costaging question than CoStaging-12*. In part, this is because describing the thinker's reasoning will not rely on making highly idealised claims about a thinker's a priori knowledge, claims which seem suspect when we remember that we have reason to attribute mental files to infants and animals as well as to mature humans.[37]

4.6 Final remarks

In this chapter, I considered two further proposed characterisations of de jure coreference, which both treat de jure coreference as explicitly guaranteed coreference. I considered Pinillos's suggestion that de jure coreference is characterised as AP-coreference, and raised concerns with both his characterisation of de jure coreference and his arguments that de jure coreference is non-transitive. I then argued that we should not use de jure coreference to answer the costaging question if de jure coreference is characterised as AP-coreference.

I offered IK-coreference as an alternative characterisation of de jure coreference, and showed that there is no evidence that IK-coreference is non-transitive. I considered the suggestion that we use an IK-coreference characterisation of de jure coreference to answer the costaging question, and have shown that the objections which were fatal to earlier proposals don't threaten this proposal. However, we should not yet be satisfied. We need some account of how a thinker realises the idealised knowledge attributed to her. And so we should turn away from considering

[37] See 2.2.2.

the thinker's knowledge, and consider her reasoning.

Chapter 5

Trading On Coreference

5.1 Introduction

There is a popular argument that when thinkers reason, they must sometimes *trade on the coreference* of their records. My objective in this chapter is to give an adequate account of trading on coreference, so that in 5.4 I can use trading on coreference to answer the costaging question.

I start by explaining what trading on coreference is, and exploring some overlooked features of the popular argument that thinkers trade on coreference. I then consider apparent evidence that thinkers can equivocate when trading on the coreference of cotemporal records. This evidence challenges a *prima facie* obvious picture of what kind of mistakes in reasoning a rational thinker might make, as well as counting against using trading on coreference to answer the costaging question. I show that we reach a stalemate of conflicting intuitions, which cannot be resolved except by taking some stance on how we explain trading on coreference. I argue that the core account of files gives us reason to conclude that thinkers cannot equivocate when trading on the coreference of cotemporal record-stages, and I explore how to explain intuitions that conflict with this conclusion. I finish the chapter by briefly considering the relationship between trading on coreference and IK-coreference, and arguing that we can answer the costaging question in terms of trading on coreference.

In this chapter, I start giving the contents of records directly, rather than associating them with the utterance the speaker would use to express the record. As before, I give the content of the record using an utterance of an English sentence. The utterance has the same coarse-grained content as the record, and indicates how the thinker would report that content if she could, but carries no commitment that the thinker can or would report that content with an utterance.

I also slightly adapt my notation system. It will be necessary to use many different records in the examples, and with a limited set of letters we risk losing

track. When discussing the records of a specific thinker, I indicate who the record belongs to, as well as giving the record's label.

For example, Thales assents to (1) and (2′)[1] in virtue of holding belief-records Th(r) and Th(s).

Th(r) Hesperus is bright

Th(s) It is not the case that Phosphorus is bright.

As before, stages of records are indicated with subscript times. So the t_1 stage of Th(r) is ${}^{Th}(\mathrm{r})_{t_1}$.

5.2 The argument for trading on coreference

5.2.1 Trading on coreference

I take it that there is a primitive category of reasoning: *reasoning as if record-components are coreferential*. *Trading on coreference* is one species of this.

Reasoning as if record-components are coreferential can be roughly characterised as reasoning that turns on record-components being about the same thing. It can best be introduced by examples. T reasons as if components of belief-records r and s are about the same thing when she infers a belief-record u just from r and s.[2]

r Fa

s Gb

u $\exists x(Fx\&Gx)$

However, if T does not infer u, this is not conclusive evidence that she isn't reasoning as if components of r and s are coreferential. Suppose T holds r and acquires s. If she also holds (for example) v, T may reject the belief recorded as r.

[1] See 1.1.1.

[2] Though as I point out below, if a thinker infers u from r, s and some other record, she may not be reasoning as if components of r and s are coreferential.

v $\neg\exists x(Fx\&Gx)$

But in rejecting r because of v, T is reasoning as if components of r and s are coreferential.

A thinker may reason as if record-components are coreferential when she reasons using a belief in an identity. So if T infers u from r, s and either w or x, she is reasoning as if components of r and s are coreferential.

w $a = b$

x 'a' corefers with 'b'

Similarly, if w or x are not belief-records, but merely suppositions or pretences for the sake of the argument, and the thinker supposes or pretends that $\exists x(Fx\&Gx)$, on the basis of r and s, with w or x, then she is reasoning as if components of r and s are coreferential.

A thinker can also reason as if record-components are coreferential without thinking that the records are about anything at all. Suppose that a thinker has belief-records y and z which she would express with utterances of (60) and (61) respectively.

(60) Only Rumpelstiltskin could spin this straw into gold.

(61) Rumpelstiltskin does not exist.

The thinker then infers the belief-record r_2 she would express (62).

(62) No one can spin this straw into gold.

In inferring r_2 from y and z, she is reasoning as if components of y and z are coreferential.

Cases where a thinker reasons as if record-components are coreferential contrast with cases where a thinker does not reason as if record-components are coreferential. For example, suppose T infers u from r, s and s_2

s_2 a & b are qualitatively identical

In this case, T is not reasoning as if components in r and s are coreferential. Her reasoning would not be invalid if record-components in r and s were coreferential, but her reasoning does not turn on the coreference of r and s's components.

As I have said, trading on coreference is one type of reasoning as if record-components are externally-coreferential. In particular:

> A thinker *trades on coreference* of record-components c_r and c_s iff she reasons as if c_r and c_s are coreferential without employing an additional representation implying that c_r and c_s are coreferential.

So if a thinker reasons as if c_r and c_s are coreferential in virtue of any kind of additional identity representation, whether this is a record in a mental file or a link between files, she is not trading on the coreference of c_r and c_s.

5.2.2 The argument for trading on coreference

This characterisation of trading on coreference will be made clearer by giving the popular and compelling argument that thinkers must be able to trade on coreference.[3]

In outline, the argument goes as follows: if we imagine a thinker who cannot trade on the coreference of her record-components, we find that she cannot reason as if her record-components are coreferential; rather she will be launched into a vicious regress. So any thinker who can reason as if her record-components are coreferential must be able to trade on the coreference of her record-components.

The first step is showing that how the regress is launched. Suppose Chris infers ${}^{C}(\mathrm{u})$ from ${}^{C}(\mathrm{r})$ and ${}^{C}(\mathrm{s})$.

${}^{C}(\mathrm{r})$ Hesperus is bright.

[3]This argument appears to originate with Campbell (1987-1988), but it is widely used elsewhere (e.g. Campbell 1995; Sainsbury 2002 [1997]; Brown 2004; Fine 2007; Schroeter 2008; Recanati 2012).

^{C}(s) Hesperus is large.

^{C}(u) $\exists x$(x is large & x is bright)

The fact that ^{C}(r) and ^{C}(s) contain externally-coreferential record-components does not make this inference warranted. Rather, for the inference to be warranted, it must be available to Chris that ^{C}(r) and ^{C}(s) contain externally-coreferential record-components. Suppose Thales inferred a belief he'd express using (63) from just the belief-records he'd express (1) and (5).[4]

(1) Hesperus is bright.

(5) Phosphorus is large.

(63) Something is large and bright.

Thales' inference is not warranted, even though the beliefs expressed (1) and (5) are externally-coreferential, just because it is not available to Thales that those beliefs are externally-coreferential.

Suppose that the only way it could be available that any two record-components are about the same thing is via an identity representation of the form $a = b$, 'bridging the gap' between the records. This means that to infer ^{C}(u) from ^{C}(r) and ^{C}(s), Chris would need a record ^{C}(v).

^{C}(v) Hesperus is Hesperus.

But ^{C}(v) can only bridge the gap between ^{C}(r) and ^{C}(s) if it is available to Chris that both ^{C}(r) and ^{C}(v), and ^{C}(s) and ^{C}(v), contain externally-coreferential record-components. But on the assumption this availability can only come from a record of the form $a = b$, Chris must have additional bridging-premises, one to bridge the gap between ^{C}(r) and ^{C}(v), and another between ^{C}(s) and ^{C}(v). But now the source of the regress should be clear: each new bridging-premise does not resolve the difficulty but simply adds to the problem, creating new gaps that need to be bridged.

[4]Examples from 1.1.1.

One might hope to avoid this difficulty by claiming that the bridging-records are enthymematic, suppressed from conscious awareness. But this won't help. The regress is a threat to the warrant of coreferential reasoning, and whether the relevant records are conscious is irrelevant to this threat.

Faced with this threatened regress, the next stage of the argument is to say that there is an alternative to supposing that the only way a thinker can reason as if record-components are coreferential is via a bridging-premise of the form $a = b$. This alternative is to allow that a thinker sometimes trades on the coreference of her record-components; that is, she sometimes reasons as if they are coreferential without employing any additional representation implying they are coreferential. If we allow that thinkers sometimes trade on coreference, no regress is threatened.

5.2.3 Reasoning as if components are coreferential is ubiquitous

5.2.2 presents the standard argument for trading on coreference. But important features of this argument and of trading on coreference are often overlooked. In 5.2.4 and 5.2.5, I will fill in additional details which are needed to make the argument, as it is usually given, successful. Before that, it is worth highlighting that reasoning as if record-components are coreferential is ubiquitous across kinds of mental representation.

First, thinkers 'reason as if coreferential' across different kinds of attitude. Suppose Chris has a belief-record ${}^{C}(\mathrm{w})$ and a desire-record ${}^{C}(\mathrm{x})$.

${}^{C}(\mathrm{w})$ If I go into the garden, I will see Hesperus.

${}^{C}(\mathrm{x})$ I see Hesperus.

For Chris to be motivated by ${}^{C}(\mathrm{w})$ and ${}^{C}(\mathrm{x})$ to go into the garden, he must be able to reason as if ${}^{C}(\mathrm{w})$ and ${}^{C}(\mathrm{x})$ contain coreferential components.

Second, reasoning as if record-components are coreferential is not just a feature of reflective conscious thought. It is also a feature of reasoning that is generally subpersonal or not subject to reflection. Millikan writes:

> [S]uppose that I perceive that α is orange and that β is round and that γ smells sweet and that δ is fist-sized and that ϵ is within reach. Why does it matter whether $\alpha = \beta$, or whether $\delta = \epsilon$, and so forth? Because if $\alpha = \beta = \gamma = \delta = \epsilon$, but only then, probably this is a reachable orange... Only by using these various bits of information *together* can this understanding be reached.
>
> (Millikan 2000, pp141-142)

Whenever we build up understanding of an object from records gathered at different points of time or from different sense-modalities, we are reasoning as if components of those records are coreferential. So reasoning as if record-components are coreferential is partly responsible for our conception of the world as full of persisting, multi-propertied objects. This means that animals and infant humans must also reason as if record-components are coreferential.

Third, reasoning as if record-components are coreferential is just part of a wider phenomenon of *reasoning as if record-components are co-valued*. Suppose Mary infers ${}^{M}(u)$ from ${}^{M}(\mathrm{r})$ and ${}^{M}(s)$.

${}^{M}(\mathrm{r})$ Jack is a fool.

${}^{M}(s)$ Alice is a fool.

${}^{M}(u)$ At least two things are fools.

This inference relies on Mary reasoning as if components in ${}^{M}(\mathrm{r})$ and ${}^{M}(s)$ share semantic value.

My focus is on thought about objects, so I focus on reasoning as if record-components are coreferential rather than on the broader phenomenon of reasoning as if record-components are co-valued.

Thinkers reason as if non-referring referential record-components are coreferential, they reason as if record-components are coreferential across attitudes, and they reason as if record-components are coreferential whether or not they are mature human thinkers. And reasoning as if record-components are coreferential is just part of a wider phenomenon of reasoning as if record-components are co-valued. If the standard argument for trading on coreference is successful, it can be reused for these examples too — showing that regress threatens unless we allow that animals and infants can trade on coreference, that thinkers can trade on coreference across attitudes, and that thinkers can trade on co-value as well as co-reference.

5.2.4 A metarepresentational alternative

Understanding the ubiquity of reasoning as if record-components are coreferential allows us to fill in one gap in the argument of 5.2.2.

The standard argument for trading on coreference assumes that the only alternative to trading on coreference is supplying a bridging-record of the form $a = b$. But there is another alternative: a metarepresentational bridging-record, of the form '*a*' *corefers with* '*b*'. And it does not lead to regress to suppose that thinkers cannot trade on coreference but instead use this form of bridging-record.

Suppose Chris cannot trade on coreference, but he has a metarepresentational record ^{C}(y) bridging the gap between ^{C}(r) and ^{C}(s). The referential components in ^{C}(y) do not refer directly to the planet Venus. Rather, they refer to referential components in ^{C}(r) and ^{C}(s).

^{C}(y) 'Hesperus$_{\text{r}}$' corefers with 'Hesperus$_{\text{s}}$'

Components of ^{C}(y) refer to components of ^{C}(r) and ^{C}(s), so Chris needs only ^{C}(y) for it to be available to Chris that ^{C}(r) and ^{C}(s) are coreferential. No further bridging-premises are needed for this inference, and although Chris doesn't trade on coreference, regress is not threatened.

For the standard argument that thinkers trade on coreference to go through, we need to rule out the suggestion that whenever thinkers reason as if record-components are coreferential, they always do so via a metarepresentational bridging-premise of some sort. The threat of regress is not enough to rule out this suggestion, so some other considerations must be introduced.

Brown (2004) is unusual in considering the metarepresentational proposal as an alternative to trading on coreference. However, she dismisses it by simply claiming that it is implausible that thinkers cannot make a "groundfloor inference... without invoking a premise that involves the concept of a concept" (2004, p182), and does not explain why the proposal is implausible.

One problem might be that the proposal is introspectively implausible: we are rarely aware of metarepresentational records like ^{C}(y). But once we allow that bridging-premises might be enthymematic, the introspective evidence is inconclusive. We cannot exclude the possibility that thinkers have these bridging-records, merely suppressed from conscious awareness.

The second problem is more decisive. In 5.2.3, I explained that reasoning as if coreferential does not just happen in conscious mature human thought, rather, it occurs in subpersonal thought, and in infant and animal thought. However, we have little reason to suppose that animals and infants have metarepresentational abilities, or that metarepresentation occurs in subpersonal thought. Suggesting that animals and infants do not have metarepresenational bridging-premises, but instead trade on coreference, is a more parsimonious explanation of their ability to reason as if record-components are about the same thing, requiring no attribution of metarepresentational abilities.

Once we allow that animals and infants can trade on coreference, we would need good reason to suppose that mature humans cannot trade on coreference and instead must deploy metarepresentational bridging representations. Absent any such reason,

and with the inconclusive introspective evidence in favour of the claim that mature humans trade on coreference, we should conclude that mature humans can also trade on coreference.

5.2.5 When do thinkers trade?

There is another gap in the standard way of discussing trading on coreference. The argument that thinkers must be able to trade on coreference is usually given to defend the claim that thinkers can trade on the coreference of atomic records like ${}^{C}(\mathrm{r})$ and ${}^{C}(\mathrm{s})$. But the argument does not prove that thinkers can trade on the coreference of components of atomic records, only that they can, at some point, trade on coreference. One can still suggest that thinkers can't trade on the coreference of components of atomic records like ${}^{C}(\mathrm{r})$ and ${}^{C}(\mathrm{s})$, but instead can only trade on the coreference of components of atomic records with components of identity premises. So Chris can trade on the coreference of components of ${}^{C}(\mathrm{r})$ and ${}^{C}(\mathrm{v})$, and ${}^{C}(\mathrm{s})$ and ${}^{C}(\mathrm{v})$, but not ${}^{C}(\mathrm{r})$ and ${}^{C}(\mathrm{s})$.

Again, the introspective evidence is against this proposal. The introspective evidence is that we often do trade on coreference of atomic records, without supplying an additional identity premise. But again, the introspective evidence is not conclusive because of the possibility that bridging-records may be enthymematic.

However, in this case, I suggest we should accept the introspective evidence. Although the argument for trading on coreference does not get us to the claim that thinkers can trade on the coreference of components of atomic records like ${}^{C}(\mathrm{r})$ and ${}^{C}(\mathrm{s})$, there is no reason to suppose thinkers cannot trade on coreference in this way. Absent any reason to reject the introspective evidence, we should conclude that thinkers can trade on coreference in the way standardly proposed.

To summarise, I considered a popular argument given that thinkers can sometimes trade on the coreference of their record-components — that is, they can reason as if those components are coreferential without supplying any additional representation implying that coreference. I discussed some features of the argument that are often neglected, filled in gaps in the reasoning, and highlighted the fact that reasoning as if record-components are coreferential, and hence trading on coreference, is ubiquitous across forms of reasoning.

5.3 Trading on Coreference and Transparency

5.3.1 The problem

Suppose Mary has belief-records ^{M}(r) and ^{M}(v), and trades on their coreference to infer ^{M}(w).

^{M}(r) Jack is a fool.

^{M}(v) Jack wants to be an actor.

^{M}(w) Someone who is a fool wants to be an actor.

If ^{M}(r) and ^{M}(v) were about different people, then Mary's reasoning would equivocate, that is, she would fallaciously treat record-components as having the same semantic-value when they in fact have different semantic-values. This would render invalid the inference to ^{M}(w).[5]

Alternatively, suppose Mary infers ^{M}(w) from ^{M}(r), ^{M}(v) and ^{M}(y).

^{M}(y) Jack is Jack.

In this case, Mary trades on the coreference of record-components in ^{M}(y) with record-components in ^{M}(r) and ^{M}(v). In this case, so long as the record-components

[5]I suppose that if two record-components must share referential-value, but it is indeterminate what that referential-value is, then a thinker who treats those components as having the same referential-value does not equivocate.

whose coreference is traded upon are about the same thing, Mary does not equivocate and her reasoning using ${}^{M}(\mathrm{w})$ is valid, even if ${}^{M}(\mathrm{r})$ and ${}^{M}(\mathrm{v})$ are about different people and ${}^{M}(\mathrm{y})$ is false.

A pressing question about trading on coreference is whether it is possible to trade on the coreference of record-components with different referential-values. In 5.2, I argued that trading on coreference underpins coreferential reasoning across different kinds of thought. If thinkers can trade on the coreference of components with different referential-values, then thinkers risk equivocation in their most basic coreferential reasoning. But we do not commonly suppose that there is danger of equivocation when trading on coreference.

We can state this natural position as a principle of transparency for trading on coreference (TC).

TC If a thinker trades on the coreference of record-components $\mathrm{c_r}$ and $\mathrm{c_s}$, then $\mathrm{c_r}$ and $\mathrm{c_s}$ share referential-value.

TC is less demanding than the related theses of *transparency of sameness*:

> If two of a thinker's token thoughts possess the same content, then the thinker must be able to know *a priori* that they do.
>
> (Boghossian 1994, p36)

and *transparency of difference*:

> If two of a thinker's token thoughts possess distinct contents, then the thinker must be able to know *a priori* that they do.
>
> (Boghossian 1994, p36)

If TC is correct, then a sufficiently reflective thinker might know that if she can trade on the coreference of record-components, they share referential value. But this does not mean that the thinker can trade on the coreference of any components that share referential-value. And so for any pair of components, a reflective thinker

cannot tell whether or not they share referential-value just by seeing whether or not she can trade on their coreference.

Problematically, although TC may be both intuitively appealing and widely accepted, there are putative counterexamples to TC, supposedly showing that rationally blameless thinkers can equivocate when trading on coreference.

These supposed counterexamples to TC would be of interest to anyone discussing trading on coreference, but are particularly troubling to anyone hoping to use trading on coreference to answer the costaging question. As I have emphasised, the core account of files requires that costaged record-stages contain a component-stage with a referential-value fixed by the file-stage, and this component-stage supplies the same referential-value to each record-stage in the file-stage.[6] If we allow that thinkers might equivocate when trading on the coreference of component-stages of cotemporal record-stages, this will undermine hopes of answering the costaging question in terms of trading on coreference.

Given my interest in answering the costaging question, I will focus on TCS, rather than TC.

TCS If a thinker trades on the coreference of cotemporal component-stages $\mathrm{c}_{\mathrm{r}_t}$ and $\mathrm{c}_{\mathrm{s}_t}$, then $\mathrm{c}_{\mathrm{r}_t}$ and $\mathrm{c}_{\mathrm{s}_t}$ share referential-value.

The challenge to TCS comes from various puzzle-cases which are supposed counterexamples to TCS. These puzzle-cases are commonly discussed in terms of externalism's challenge to first-person knowledge of mental states,[7] with connections to trading on coreference left aside.[8] However, the same cases that challenge the common-sense picture that thinkers have privileged access to their mental states challenge the intuitively plausible principle TCS, and the issues are closely connected. I limit the scope of my discussion just to the challenge to TCS.

[6] See 2.1.3 and 2.2.1.

[7] Key readings include Boghossian (1989, 1992, 1994); Ludlow and Martin (1998) (particularly Burge (1998)); and Brown (2004).

[8] Exceptions include Campbell (1987-1988, 1995); Lawlor (2001); Recanati (2012).

There are three principal types of case that challenge TCS: tracking failures, cross-modal identification failures, and reidentification failures. To illustrate the challenge to TCS, I give one example of each.

Each example involves the thinker reasoning as if component-stages in record-stages identifiable as the premises in each piece of reasoning are coreferential. The component-stages treated as coreferential are the A- and B-component-stages. Where I have sketched the content of each record, I have labelled the English expressions corresponding to the component-stages treated as coreferential.

At this stage, I make no claims about the reference the A- and B-component-stages. I indicate this by placing the relevant term in brackets. The English sentences containing bracketed terms are therefore not complete specifications of the record's content, merely indicators of it.

> **Example XI:** *Tracking failure*: Luc is watching a gull. He tracks it as at t_1 it takes flight from the beach and flies towards him. At t_2, the original bird (Gull_1) is imperceptibly switched with a superficially indistinguishable bird (Gull_2). At t_3, Gull_2 is close enough that Luc can judge that it is a yearling. At t_4, Luc has record-stages ${}^{L}(\text{r})_{t_4}$ and ${}^{L}(\text{s})_{t_4}$, and at t_4, Luc infers that *something which is a yearling was on the beach.*
>
> ${}^{L}(\text{r})_{t_4}$ $[\text{That}]_A$ was on the beach.
>
> ${}^{L}(\text{s})_{t_4}$ $[\text{That}]_B$ is a yearling.

It is important to remember that trading on coreference is a mental rather than linguistic phenomenon. If, at t_1, Luc had said "that is on the beach", then we cannot suppose that the 'that'-occurrence's referent changes once the switch occurs at t_2. But matters are less straightforward for thought. It is at least open to suppose that ${}^{L}(\text{r})_{t_4}$ is a stage of a record that switched its referential-value between its formation and t_4.

Example XII: *Cross-modal identification failure*: Jess is looking for a cold, clean glass. She sees a group of glasses, some of which are dirty, and some of which are warm from washing. She touches various clean glasses to find a cold one, thinking she is guiding her touch by sight. Eventually, at t_1, she moves to pick one of the glasses up. We explain this by claiming Jess has record-stage ${}^{J}(\mathrm{r})_{t_1}$ from sight, and record-stage ${}^{J}(\mathrm{s})_{t_1}$ from touch, and from these Jess has inferred that some particular glass is both cold and clean.

${}^{J}(\mathrm{r})_{t_1}$ $[\mathrm{That}]_A$ is clean.

${}^{J}(\mathrm{s})_{t_1}$ $[\mathrm{That}]_B$ is cold.

Unfortunately, Jess is in an experimental set-up where she is not looking at what she is touching. Rather, she is looking and touching a different set of glasses, and a visual illusion is used to make Jess suppose that she is touching the glasses she is looking at.

Example XIII: *Reidentification failure*: Antoine lives normally on Earth for twenty years. Shortly before his twentieth birthday, he visits a lake where he is excited to see Pavarotti swimming. On his twentieth birthday, he is switched to Twin-Earth, but does not realise he has been switched. Antoine lives a further thirty years on Twin-Earth. He enjoys seeing Twin-Pavarotti sing several times, and when he is about fifty, he attends a talk in which Twin-Pavarotti mentions that he has never swum. On hearing this (at t_1), Antoine concludes that he is being lied to. We can explain this conclusion as an inference from his records ${}^{A}(\mathrm{r})_{t_1}$ and ${}^{A}(\mathrm{s})_{t_1}$.

${}^{A}(\mathrm{r})_{t_1}$ $[\mathrm{Pavarotti}]_A$ was swimming.

$^{A}(\mathrm{s})_{t_1}$ [Pavarotti]$_B$ claims he has never swum.

The problem in each case is the same. It is plausible in each case to say that the thinker is trading on the coreference of the A- and B-component-stages in the premise record-stages. But it is also plausible to claim that the A- and B-component-stages in each of the record-stages do not share referential-value, and rather each is about the object which was the causal source of the record-stage. And both these claims cannot be true if TCS is true. If TCS is correct, a thinker cannot trade on the coreference of component-stages of record-stages unless they share referential-value.[9]

5.3.2 Possible responses

The problem can be presented as an inconsistent triad, each component of which is independently tempting: (i) TCS, (ii) that in the puzzle-cases (examples XI-XIII) the thinker trades on the coreference of the A- and B-component-stages, and (iii) that in the puzzle-cases the A- and B-component-stages have different referential-values.

Clearly, one of these three must be abandoned. The difficulty is choosing which ones should be kept, and which one rejected.

Trading on coreference Of the three claims, the one least often challenged is that the thinker trades on the coreference of the A- and B-component-stages.[10]

In 5.2.5, I pointed out that, problematically, the argument for trading on coreference does not tell us at what stage of reasoning we trade on coreference, only that at some point we must. It is compatible with the standard argument that thinkers

[9]One might object that these puzzle-cases are fantastic, and therefore we should not be concerned about them. But even if the examples were only fantastical, they should still be of concern to anyone who claims it is not *possible* to trade on the coreference of component-stages with different referential-values. Moreover, although fantastical examples are both vivid and commonly used, there is no need to employ them. Stalnaker (2008, pp123-130) provides a good range of non-fantastical examples that challenge TCS.

[10]This claim is generally taken for granted rather than defended, perhaps an artifact of the typical focus on self-knowledge rather than trading on coreference.

always deploy one bridging-premise of the form $a = b$ whenever reasoning with the beliefs that Fa and Gb.

We might think that these puzzle-cases give us a good reason to think that we always do deploy at least one bridging-premise of the form $a = b$. If we accept this, we can affirm both TCS and that the A- and B-component-stages have different referential-values, without being forced to conclude that thinkers equivocate in their reasoning.

Nonetheless, we should reject the suggestion that the thinkers do not trade on the coreference of the A- and B-component-stages. The cost of keeping the other two claims is misrepresenting the most basic forms of reasoning. Examples XI and XII in particular involve very basic kinds of informational integration — information acquired from different moments in an incident of visual tracking and information from different sense-modalities used to build a fuller picture of an object's properties. These are basic functions of reasoning, relied on by infants and animals as well as mature humans. Suggesting additional representations, functioning as premises that each piece of information is about the same thing, over-intellectualises a basic part of reasoning. Campbell makes this point clearly:

> [I]t would be wrong to think of the hypothesis that ordinarily vision and touch give one information concerning the same things, as if it were on a par with the hypothesis, which might be used in establishing the identity of the Evening Star and the Morning Star... The integration of touch and vision plays a far more fundamental role in our cognitive lives than that.
>
> (Campbell 1987-1988, p282)

Once we accept that thinkers do sometimes trade on the coreference of component-stages of independently formed atomic records of the form Fa and Gb,[11]

[11] By 'independently formed records, I mean records like ${}^{L}(\mathrm{r})_{t_4}$ and ${}^{L}(\mathrm{s})_{t_4}$ which have independent causal paths back to the source of the record. Contrast these with records with content *Panya is a cat* and *Panya is vicious*, where the record *Panya is vicious* was formed as a result of inference from *Panya is a cat* and *all cats are vicious*. In this case, one record includes the other in its causal path back to the source of the record.

then the only way of challenging the claim that thinkers trade on the coreference of the A- and B-component-stages is by coming up with arguments as to why the thinker, in that particular example, does not trade on coreference. But this strategy will not help us deal with the problem. So long as we think that there is sometimes trading on the coreference of component-stages of independently formed atomic record-stages, there is room to come up with cases where the record-stages causally originate in different objects, and therefore where it appears that a thinker trades on the coreference of component-stage with different referential-values. Arguing over the details of an individual case will not help us avoid apparent counterexamples to TCS.

On this basis, I suggest we should accept that in each of the puzzle-cases the thinker trades on the coreference of the A- and B-component-stages, and look instead at the other two members of the triad. However, as I will discuss, simply considering each member on its own merits does not give us a clear answer as to which one we should reject and which one we should keep.

TCS As I mentioned when it was introduced, TCS is an intuitively plausible principle. We are usually relatively untroubled by the suggestion that someone is equivocating. It is a familiar idea that a sloppy reasoner may, in the course of a single argument, use concepts with slightly different extensions as if they have the same extension, undermining the validity of the argument.

Nonetheless, counterexamples to TCS are peculiarly troubling. One reason for this is that in these examples it seems that the thinker equivocates through simple empirical bad luck, despite following good rational practice. Whereas in more familiar cases of equivocation, we suppose the thinker could, with encouragement, fix her mistake a priori. A second reason is that in familiar cases of equivocation, the equivocation involves concepts with similar extensions, for example, the different extensions associated with different uses of terms like 'innocent' or 'alive'. It is a

very different matter to say that thinkers equivocate in virtue of treating thoughts about completely different individuals as if they were thoughts about the same individual. The former kind of mistake seems unfortunate, but the second makes it hard to treat that thinker as meeting the most minimal conditions on rationality.

Boghossian writes:

> [L]et's say that being *minimally* rational is a matter of being able to *avoid* obvious violations of the principles of logic, given enough time to reflect on the matter and so on.
>
> (Boghossian 1994, p42)

Denying TCS disrupts our understanding of the minimal conditions on rationality in a way that accepting thinkers equivocate (in a priori fixable ways) when using terms like 'innocent' or 'alive' does not.

However, this is not itself a conclusive argument for keeping TCS, as our understanding of the conditions on rationality may be wrong. As Brown (2004, p184) points out, there is evidence of thinkers making invalid inferences in even simple pieces of reasoning.[12] We would demonstrate that our understanding of the conditions on rationality is wrong if we could find convincing arguments that the A- and B-component-stages in the puzzle-cases have different referential-values. And even if we cannot conclusively demonstrate anything about the referential-values of the A- and B-component-stages, this alone does confirm that our our understanding of the conditions on rationality is correct. So TCS is appealing, but has not been shown to be correct.

Different referential-values When we consider the record-stages containing the A- and B-component-stages, and focus on how each record-stage was formed rather than its role in reasoning, it is natural to say that the record-stage is about the object

[12] A famous example is the Wason selection task, demonstrating that many adults fail to apply modus tollens correctly, even after the logical structure of the problem has been drawn to their attention (see e.g. Wason and Shapiro 1971).

interaction with which caused that record to be formed, and hence the A- and B-component-stages refer to different objects. For example, when we consider ${}^{J}(\mathrm{r})_{t_1}$ and ${}^{J}(\mathrm{s})_{t_1}$ in turn, it is natural to say ${}^{J}(\mathrm{r})_{t_1}$ contains a component-stage referring to the seen glass, and ${}^{J}(\mathrm{s})_{t_1}$ contains a component-stage referring to the touched glass.

However, we might hope for more argument than simply pointing towards what is natural to think.

The first argument I will consider originates in the discussion of cases like example XIII. The considerations relate to memory, so are relevant to example XI as well.

Boghossian offers the following "platitude":

> [I]f S knows that p at t1, and if at (some later time) t2, S remembers everything S knew at t1, then S knows that p at t2.
>
> (Boghossian 1989, p23)

This "platitude" is intuitively plausible.[13] Moreover, it provides a reason to think that ${}^{A}(\mathrm{r})_{t_1}$ is about Pavarotti. Before the switch to Twin-Earth, we can suppose that Antoine knew that *Earth-Pavarotti was swimming*, and that after the switch to Twin-Earth, Antoine forgets nothing about the incident. Therefore, assuming the "platitude" is correct, at t_1 Antoine still knows that Earth-Pavarotti was swimming. On the reasonable assumption this is non-dispositional knowledge, he must have some record with content *Earth-Pavarotti was swimming*, and we can suppose that this is the record he uses in his reasoning to the conclusion that he is being lied to, and hence that ${}^{A}(\mathrm{r})_{t_1}$ is about Earth-Pavarotti rather than Twin-Pavarotti. Once we argue that ${}^{A}(\mathrm{r})_{t_1}$ is about Earth-Pavarotti, we can simply add the intuitively plausible claim that ${}^{A}(\mathrm{s})_{t_1}$ is not straightforwardly about Earth-Pavarotti, to get the result that the A- and B-component-stages have different referential-values.

[13] A thinker may lose knowledge that P by acquiring a belief conflicting with P or with the thinker's grounds for believing P. Boghossian clearly abstracts away from this kind of complication, and I do the same.

However, I suggest that although Boghossian's "platitude" is plausible, it isn't correct.[14]

If we distinguish between content and its vehicle, we can distinguish two cases where we may be tempted to say a thinker remembers something: retaining a content and retaining a vehicle.[15] Some kinds of externalist semantics allow that a thinker's *current* environment has a role in determining the contents of the thinker's mental representations, howsoever they were originally formed (call this *current-externalist semantics*).[16] Current-externalist semantics suggest that thinkers can retain a vehicle whilst not retaining the vehicle's original content. So if one counts as remembering everything known at t_1 simply through retaining vehicles from t_1, and current-externalist semantics are correct, then Boghossian's "platitude" is false.

If current-externalist semantics are incorrect and a thinker cannot retain a vehicle without retaining its original content, Boghossian's "platitude" is correct. But Boghossian's platitude is in itself no argument for supposing that current-externalist semantics is incorrect. Rather, it is merely a statement that it is incorrect (see also Ludlow 1995). So it is no platitude, rather it affirms a not uncontroversial semantic position.

The following principle is plausible: beliefs resulting from perceiving some object are about the object perceived. We can use this principle in an alterna-

[14]If 'remember' is factive (cf. Williamson 2000) then Boghossian's platitude is trivially true, but we can still wonder whether it is legitimate to suppose that after the switch Antoine remembers everything from before the switch, or whether we should suppose that Antoine may have only pseudo-memories. Here, I assume that 'remember' is sometimes used non-factively.

[15]In practice, we say someone 'remembers' something in many different circumstances. We allow for remembering-how and remembering-which as well as remembering-that, and we sometimes allow for remembering-that when neither content nor vehicle is retained. Suppose someone is discussing details of their early childhood, and claims to remember that P whilst acknowledging this memory may not be fully accurate. In this case we can suppose the thinker might not fully retain either vehicle or content, but still might attribute them the memory that P. This example demonstrates an additional problem with Boghossian's "platitude": it is seems context sensitive whether or not it is felicitous to claim *S remembers that P*. We should be cautious about putting too much argumentative weight on such claims without considering contextual factors.

[16]This includes file-theorists who give externalist accounts of how files fix reference, and allow that files change their reference. See 2.2.2.

tive argument that the A- and B-component-stages in example XIII have different referential-values. Because ${}^{A}(\mathrm{r})_{t_1}$ records a belief that is the result of perceiving Earth-Pavarotti, and ${}^{A}(\mathrm{s})_{t_1}$ records a belief that is the result of perceiving Twin-Pavarotti, ${}^{A}(\mathrm{r})_{t_1}$ is about Earth-Pavarotti, and ${}^{A}(\mathrm{s})_{t_1}$ about Twin-Pavarotti.

However, although this principle is intuitively plausible, there are counterexamples. Consider an example of mistaken identity:

> **Example XIV:** *Mistaken identity*: Martha sees Frank on the other side of the street, but mistakes him for Humphrey. She is struck by the red shoes the man is wearing. Later, she is able to recall the incident and even picture what 'Humphrey' was wearing that day. When recalling the incident, she sincerely utters (64).
>
> (64) Humphrey was wearing red shoes.

On a very natural reading of XIV, when Martha 'remembers' the incident she entertains (false) beliefs about Humphrey rather than (true) beliefs about Frank. The corresponding belief-records are based on the memory of a perceptual experience, but are not about the object experienced, rather they are about the object Martha mistook the actual object for.[17]

Were Martha to doubt that she had correctly identified the man, she could retreat to purely demonstrative beliefs, for example the belief expressed using (65).

(65) That man was wearing red shoes.

It is more natural to suppose that these beliefs are about Frank rather than Humphrey. However, the fact that Martha can retreat to these 'safe' beliefs does not mean she does, and so does not neutralise the counterexample. Unless Mary has reason to, she will not entertain *that man* beliefs, but rather uses *Humphrey* beliefs.

[17]The point would still be made if Martha's beliefs failed to be about anything, or were about an amalgam of Humphrey and Frank. So long as the beliefs are not about Humphrey, case XIV serves as a counterexample to the principle in question. Sainsbury and Tye (2012, p94) use this kind of example to make a similar point.

Another strategy for defending the claim that the A- and B-component-stages have different referential-values is to show that there are problems with the proposals raised for what shared referential-value the A- and B-component-stages have. For example, a large part of Schroeter's (2007) argument that the A- and B-component-stages have different referential-values is an extended criticism of Burge's (1998) account of the referential value of the A- and B-component-stages in example XIII.

I will not consider this strategy here. Burge's account of the referential-value of the A- and B-component-stages may, like several other accounts of their referential-value, be flawed. However, there are many ways we could account for the referential-value of the A- and B-component-stages, and arguing against one or two of these does not demonstrate that the A- and B-component-stages have different referential-values.[18]

Absent convincing arguments that the A- and B-component-stages have different referential-values, we might just focus on the intuitions we have that the A- and B-component-stages have different referential-values.

These intuitions are relatively easy to induce. One way is to emphasise the causal origin in discussions of the record-stages, and to suppress discussion of their role in reasoning. Another is to ask readers to focus on the sensory aspects of the belief. For example, discussing memories of Pavarotti swimming in an example like XIII, Boghossian asks us to focus on the thinker's "vivid and accurate representations of the scene" (1994, p39).

However, as Schroeter (2007, p609) acknowledges, our intuitions around these cases are unstable and liable to be influenced by which aspects of the case are highlighted. This can be illustrated obliquely by considering example XIV. If we just focus on the fact that when Martha recalls the incident, she produces utterances

[18]In 5.3.4, I outline my own proposal about the referential-value of the A- and B-component-stages.

of sentences like (64), it is natural to suppose that her memories of the incident are beliefs about Humphrey rather than beliefs about Frank. But if we start by focussing on Martha's ability to recall the scene and picture what the man was wearing, it is natural to say that her beliefs are about Frank. So if belief-records in confused scenarios have determinate referential-values, we don't get reliable information about those referential values simply by inducing intuitions.

5.3.3 Breaking the stalemate

We seem to have reached a stalemate. On the one hand, TCS is an appealing principle. On the other, it seems that there are counterexamples to TCS. We are simply contrasting intuitions about referential-value in the puzzle-cases with intuitions about the value of TCS, and getting no further towards working out which intuitions are correct.

We need to change approach, and step back from intuitions about TCS and the puzzle-cases. We should instead see whether we can come to some conclusion about TCS based on other theoretical commitments. In this section, I argue that if we subscribe to the core account of files presented in chapter 2, then we should accept TCS.

In outline, my line of reasoning is this: for component-stages c_{r_t} and c_{s_t} of records r_t and s_t respectively (i) if a thinker can (or does) trade on the coreference of c_{r_t} and c_{s_t}, then r_t and s_t are costaged in virtue of c_{r_t} and c_{s_t};[19] (ii) if r_t and s_t are costaged in virtue of c_{r_t} and c_{s_t}, then c_{r_t} and c_{s_t} share referential value; therefore (iii) TCS is correct.

Step (ii) is simply part of the account of files presented in chapter 2.[20]

[19] I assume that we are giving a global version of the mental file account, according to which records can only be associated with a file by being contained within it (see 2.2.1). *Mutatis mutandis* the arguments can be run for a conception view, in terms of r_t and s_t being associated with the same file-stage, rather than being costaged.

[20] See, in particular, p45.

Step (i) is a little more complex. Start by allowing that when a thinker can (or does) trade on the coreference of component-stages, she treats those component-stages as about the same object. Clearly, when a thinker does trade on the coreference of component-stages, she is treating them as about the same object. But it is also normal to count a thinker as treating records as about the same object when she is merely disposed to reason as if they are about the same thing. For example, I can be counted as treating my belief-records about Frege as about the same person, even if I am not currently thinking about Frege. This means that whenever a thinker can (or does) trade on the coreference of component-stages c_{r_t} and c_{s_t}, she is treating r_t and s_t as about the same object in virtue of c_{r_t} and c_{s_t}.

According to the account of files presented in chapter 2, if record-stages r_t and s_t are treated as if there's some object that both r_t and s_t are about, then either (a) r_t and s_t are co-staged, or (b) r_t and s_t are in linked file-stages. I have said that when a thinker can trade on the coreference of c_{r_t} and c_{s_t}, then in virtue of c_{r_t} and c_{s_t} she is treating r_t and s_t as about the same thing. This means that when a thinker can (or does) trade on the coreference of c_{r_t} and c_{s_t}, then in virtue of c_{r_t} and c_{s_t} either (a) r_t and s_t are co-staged, or (b) r_t and s_t are in linked file-stages.

But if the thinker can (or does) trade on the coreference of c_{r_t} and c_{s_t}, we can rule out scenario (b). If r_t and s_t are in linked file-stages in virtue of c_{r_t} and c_{s_t}, then the thinker treats c_{r_t} and c_{s_t} as about the same thing in virtue of the *additional* linking representation. This disqualifies her from counting as trading on the coreference of c_{r_t} and c_{s_t}.

Therefore, if the thinker can (or does) trade on the coreference of c_{r_t} and c_{s_t}, she must be in scenario (a), and so r_t and s_t are costaged in virtue of c_{r_t} and c_{s_t}.

Putting this together with the claim that if r_t and s_t are costaged in virtue of c_{r_t} and c_{s_t}, then c_{r_t} and c_{s_t} share referential-value, we get the claim that thinkers can (and do) only trade on the coreference of record-stages that share referential-value,

and hence TCS is correct. Therefore the account of files presented in chapter 2 gives reason to accept TCS.

I have been allowing that files may have their reference fixed by a definite description, perhaps even a very complicated description.[21] I have also allowed that thinkers might equivocate (in a priori fixable ways) when using terms like predicates like 'innocent' or 'alive'.

Now suppose a thinker has a mental file F that refers to a via some long reference-fixing description 'the $F\ldots$'. Now suppose that the thinker equivocates in her use of 'F', such that the definite description 'the $F\ldots$' is sometimes used in a way that picks out a, and sometimes in a way that picks out b, where $a \neq b$. It now seems that the file F itself equivocates between referring to a and referring to b.

Suppose the thinker trades on the coreference of $\mathrm{c}_{\mathrm{r}_t}$ and $\mathrm{c}_{\mathrm{s}_t}$, and r_t and s_t are members of F_t in virtue of $\mathrm{c}_{\mathrm{r}_t}$ and $\mathrm{c}_{\mathrm{s}_t}$. If the file F equivocates, we might worry that this will generate a counterexample to TCS. However, there is only a counterexample to TCS if the file equivocates in such a way that $\mathrm{c}_{\mathrm{r}_t}$ and $\mathrm{c}_{\mathrm{s}_t}$ may have different referential-values. But files are reference-fixing and so costaged records share referential-value. If file F equivocates in virtue of the thinker equivocating in her use of 'F', then the different stages of F may have different referential-values. But this does not mean that a single file-stage can deliver distinct referential-values to its members, and so this does not mean that co-temporal component-stages $\mathrm{c}_{\mathrm{r}_t}$ and $\mathrm{c}_{\mathrm{s}_t}$ may have different referential-values. So we can suppose that F equivocates, without denying TCS.

[21] See 3.8.

5.3.4 Referential-value in the puzzle-cases

If a mental file theorist is committed to TCS, she is committed to claiming that contra some intuitions, in examples XI, XII and XIII[22] the A- and B-component-stages share a referential-value.

However the file-theorist is not committed to any particular claim about what that referential-value is, or even to the claim that there is some determinate referential-value. There is no canonical 'mental files' account of what determines the reference of a mental file, or of referential-value in confused cases. In fact, file theory is compatible with saying that there are different types of files with different systems for reference fixing. For example, we might imagine that files associated with long-term representations of individuals (as in example XIII) have their reference fixed by the dominant causal source of the records in the file, but shorter-term files collecting records from soon-to-be-forgotten sensory perceptions (as in example XII) are about the unique source of the visual information in the file (if there is one).

As a place-holder, suppose that files have their reference fixed by the dominant causal source of records in the file, and that where there is no dominant causal source, the file fails to refer to anything. In example XIII, we can suppose that Antoine started with a file that was about Earth-Pavarotti, so the belief he would report using (66) is about Earth-Pavarotti.

(66) Pavarotti was swimming.

But after the switch to Twin-Earth, more information about Twin-Pavarotti gets into the file, so first the belief Antoine reports using (66) stops having a referent, then eventually becomes about Twin-Pavarotti. In example XII, we could suppose that visual information dominates touch information, and so Jess's A- and B-component-stages are about the seen-glass not the touched-glass. And in example XI, we could suppose that the current perceptions dominate memories, and so Luc's A- and B-

[22] See p158.

component-stages are about Gull_2 rather than Gull_1.

If the record-stages in the puzzle-cases whose coreference is traded upon share referential-value, what can be said about the intuition that they do not? A first point is that we might think that these intuitions are not very troubling. In 5.3.2 I highlighted that our intuitions in these cases are relatively malleable, and so do not form conclusive data against TCS.

A second point is that these intuitions are, at least to some extent, explicable by remembering that if the thinker had come to suspect that she was confused, she would be able to retreat to 'safe' unconfused beliefs. For example, had Jess (example XII) come to suspect that she were looking at and touching different glasses, she could retreat to thinking about *that-seen glass* and *that-touched glass*. Or Luc (example XI) could retreat to thinking *that-gull-I-saw* and *that-gull-I-see*. The fact that the thinker could retreat to these 'safe' beliefs makes it tempting to suppose that the thinker has records with these contents. However, it is not evidence that the thinker does in fact have such records, nor that if she did have such records then she would trade on the coreference of their components (or component-stages). Rather, we can suppose that the thinker would not trade on the coreference of components in the retreated-to records. The point of the retreat was to avoid treating records with different causal origins as about the same thing, and this point would be lost if the thinker traded on the coreference of components of those records.

Switching to making the simplifying assumption that all records are single-reference,[23] I can now tie up a loose end from 3.5.3 and from my discussion of CoStaging-4 in 2.3.2. In 2.3.2 and 3.5.3, I observed that it seems possible to doubt that costaged records were about the same thing — the thinker just has to suppose she is a victim of confusion. Suppose in example XI, Luc were to suspect that the gull

[23] See p19.

had been switched. Then it seems possible for Luc to question whether his memory of the gull and his current perception of the gull are about the same thing. However, as I pointed out in 2.3.2, this is in tension with the claim that: if the corresponding record-stages are genuinely costaged, they must share referential-value.

We can resolve this tension using the idea that thinkers can retreat to 'safe' beliefs. Suppose at t_5, Luc comes to suspect that the gull was switched. He will retreat to safe record-stages ${}^{L}(\mathrm{r})_{t_5}$ and ${}^{L}(\mathrm{s})_{t_5}$.

${}^{L}(\mathrm{r})_{t_5}$ That-gull-I-saw was on the beach.

${}^{L}(\mathrm{s})_{t_5}$ That-gull-I-see is a yearling.

He will abandon the confused mental file, replacing it with two safe files (and where he isn't sure where a record belongs, he keeps it separate from the safe files). If he makes the split in the right place, he can create a separate file on Gull_1 and another on Gull_2. Even if he makes the split in the wrong place, he will generate two files with possibly distinct referential-values.

Looking back on his past mental states, Luc may think of ${}^{L}(\mathrm{r})_{t_4}$ and ${}^{L}(\mathrm{r})_{t_4}$ as having different referential-values, even if it is not possible they had different referential-values.[24] One explanation of this is that now he suspects ${}^{L}(\mathrm{r})_{t_4}$ and ${}^{L}(\mathrm{r})_{t_4}$ have their causal origin in different gulls, he maximises the accuracy of the mental states he attributes himself by attributing himself record-stages about the causal origins of his records, rather attributing record-stages which accurately reflect the contents of his record-stages at the time. Another explanation is that once Luc comes to suspect that the gull was switched, he no longer has the mental file he deployed at t_4 (this file has been 'split'). So at t_5 it is natural for him to interpret his records at t_4 through the next best thing — the files he has at t_5. At t_5 he has distinct files which potentially have distinct referential-values, and so at t_5, it

[24] In practice, Luc is more likely to think of somewhat longer time-slices of records than individual record-stages. Talking in terms of record-stages is a convenient idealisation.

appears to Luc that at t_4 he had record-stages with potentially distinct referential-values.

More generally: at t_2 we can doubt that costaged record-stages from t_1 were about the same thing. But this does not mean that at t_1 it was a live possibility that those record-stages were about different things.

5.4 Trading on coreference and costaging

My suggestion was that we might find an answer to the costaging question in terms of the ability to trade on coreference.[25]

CoStaging-13* T's record-stages r_t and s_t are costaged iff r_t and s_t contain component-stages c_{r_t} and c_{s_t} respectively, and T can trade on the coreference of c_{r_t} and c_{s_t}

There are various points in favour of CoStaging-13*. One is that the ability to trade on the coreference of two mental representations is commonly *explained* as being the result of those representations sharing a MOP (e.g Campbell 1987-1988; Dickie and Rattan 2010). If files play the MOP role, then this role may not simply be meeting the Frege-constraint,[26] but also explaining the ability to trading on coreference.

Another point in favour of CoStaging-13* is that it does not succumb to the objections raised against alternative answers to the costaging question. Unlike assumed-coreference or AP-coreference,[27] trading on coreference is clearly a mental phenomenon involving records. And as CoStaging-13* is defined in terms of what

[25] I assume a global version. On a conception version, CoStaging-13* would just give conditions on being associated with the same file. For a genuine answer to the co-staging question, we need an extra clause limiting co-staging to just belief-records and records of belief-like information.

[26] See p12.

[27] See 3.3.2 and 4.5.1.

the thinker *can* do, there is no risk that r_t and s_t count as costaged by CoStaging-13*, but the thinker is not able to treat them as about the same object. Contrast this with attempts to answer the costaging question in terms of RR-coreference, where it may be that c_{r_t} and c_{s_t} are RR-coreferential even if it is not explicit to the thinker that c_{r_t} and c_{s_t} are about the same thing.[28]

I was cautious about answering the costaging question just in terms of IK-coreference because IK-coreference is defined in terms of some highly idealised understanding of what a competent thinker immediately knows a priori.[29] In contrast, CoStaging-13* requires no idealisation of a thinker's knowledge of her own mental states, only an account of how she is able to reason. Given the ubiquity of trading on coreference,[30] we can use CoStaging-13* in an account of the mental files of infants and animals, as well as mature humans.

Moreover, in 4.5.2, I pointed out that when thinking about component-stages IK-coreferring, we must give a highly idealised account of what a thinker knows a priori. But I observed that we didn't yet have any understanding of when we could attribute the (idealised) knowledge required for IK-coreference. However, it is tempting to claim that we can attribute a thinker the knowledge required for c_{r_t} and c_{s_t} to IK-corefer (the immediate a priori knowledge that if c_{r_t} and c_{s_t} both refer, then they corefer), just when a thinker can trade on the coreference of c_{r_t} and c_{s_t}. I suggested we might attribute a language-user the knowledge required for NP-occurrences O_1 and O_2 to IK-corefer when she immediately treats O_1 and O_2 as if it's guaranteed that if they refer then they corefer, and the rules of her language guarantee that if O_1 and O_2 refer then they corefer. And if a thinker can trade on the coreference of c_{r_t} and c_{s_t} she can immediately treat them as if it's guaranteed that if they refer then they corefer, because she isn't supplying an additional representation

[28] See 3.8.
[29] See 4.5.2.
[30] See 5.2.3.

implying that coreference. And if a thinker trades on the coreference of c_{r_t} and c_{s_t} then, by TSC, if c_{r_t} and c_{s_t} refer they corefer.

However, we might want an argument for CoStaging-13* beyond observing that it does not fall victim to the same problems as other proposed answers to the costaging question. In 5.3.3, I argued that if a thinker can trade on the coreference of c_{r_t} and c_{s_t}, this is sufficient for r_t and s_t being costaged in virtue of c_{r_t} and c_{s_t}. It is also possible to show that if r_t and s_t are costaged in virtue of c_{r_t} and c_{s_t}, this is sufficient for the thinker being able to trade on the coreference of c_{r_t} and c_{s_t}. This means that the thinker's ability to trade on the coreference of c_{r_t} and c_{s_t} is necessary and sufficient for r_t and s_t being costaged in virtue of c_{r_t} and c_{s_t}.[31]

To see that if r_t and s_t are costaged in virtue of c_{r_t} and c_{s_t}, this is sufficient for the thinker being able to trade on the coreference of c_{r_t} and c_{s_t}, turn again to the account of mental files developed in chapter 2.[32] If r_s and r_t are costaged in virtue of c_{r_t} and c_{s_t}, this is sufficient for the thinker (at t) treating c_{r_t} and c_{s_t} as about the same thing. This means that the thinker can treat c_{r_t} and c_{s_t} as about the same thing without supplying any additional representation implying they are about the same thing,[33] and so the thinker can trade on the coreference of c_{r_t} and c_{s_t}.

Reasons to suppose that the ability to trade on the coreference of c_{r_t} and c_{s_t} is necessary and sufficient for r_t and s_t being costaged in virtue of c_{r_t} and c_{s_t} provide a further reason to accept CoStaging-13* as answer to the costaging question.

Having an independent argument for CoStaging-13* allows us to claim that if a thinker can trade on the coreference of c_{r_t} and c_{s_t} , then c_{r_t} and c_{s_t} possess the characteristics we would expect of component-stages in virtue of which record-stages

[31] Again, *mutatis mutandis*, the same arguments can be made in terms of 'associated with the same file-stage' to suit the conception versions of file pictures.

[32] See especially p33 and p45.

[33] Indeed any additional record implying c_{r_t} and c_{s_t} are about the same thing will be tautological, and redundant to reasoning.

are costaged.

I have already argued that a thinker can only trade on the coreference of component-stages which share referential-value. Similarly, we can suppose that the ability to trade on the coreference of c_{r_t} and c_{s_t} is a transitive relation — so if a thinker can trade on the coreference of c_{r_t} and c_{s_t} , and on c_{s_t} and c_{u_t}, then the thinker can trade on the coreference of c_{r_t} and c_{u_t}. This is because a thinker can trade on the coreference c_{r_t} and c_{s_t} iff r_s and r_t are costaged in virtue of c_{r_t} and c_{s_t}. This means that either c_{r_t} and c_{s_t} are stages of a shared node, or stages of tokens of the same type. And being a shared node or tokens of the same type is a transitive relation. This point is supported by the absence of counterexamples to the claim that this relation is transitive.

5.5 Final remarks

In this chapter, I have given a detailed account of trading on coreference, a crucial form of reasoning required for all kinds of coreferential reasoning. I have filled in lacunae in the standard arguments for trading on coreference, and have investigated whether thinkers can trade on the coreference of cotemporal component-stages with different referential-values.

I demonstrated that we have conflicting intuitions, and that there is no path to resolving this conflict simply by focussing on these intuitions. My solution was to step back from the intuitions, and show that if one accepts file theory as presented in chapter 2, then one should accept that thinkers can only trade on the coreference of cotemporal component-stages with shared referential value.

This conclusion, and some of the argument used to reach it, set me up to give CoStaging-13* as an answer to the costaging question. I argued that CoStaging-13* avoids the difficulties encountered by previous answers to the costaging question, and I have given independent reason to think that CoStaging-13* gives a correct

account of the necessary and sufficient conditions on costaging.

Chapter 6

Costaging

6.1 Introduction

In 5.4, I finally identified a satisfactory answer to the costaging question, that is: under what conditions are record-stages r_t and s_t members of the same file-stage?

CoStaging-13* T's record-stages r_t and s_t are costaged iff c_{r_t} is part of r_t and c_{s_t} is part of s_t, and T can trade on the coreference of c_{r_t} and c_{s_t}

I suggested that not only does CoStaging-13* avoid difficulties raised with previous answers to the costaging question, but there are also independent reasons for accepting CoStaging-13*.

In chapter 2.1.2, I suggested that before we could evaluate the claim that thinkers use mental files in their thought about individuals, we first needed a theory of mental files to evaluate. Having an answer to the costaging question answers one of the two questions about the individuation of mental files that I posed in 2.3.1, getting us at least part-way towards a theory of mental files. So this is a good time to consider the mental file theory developed so far. I will outline the theory of files, and see what CoStaging-13* allows us to add to it and what work is still to be done. I will compare my account of the individuation of file-stages with that presented by Recanati (2012), who gives the most detailed theory of mental files yet published. And I will respond to the pressing objection that the theory is circular.

6.2 A partial theory of files

6.2.1 The theory

Mental files are collections of records treated as about the same object, though not all records treated as about the same object are members of the same mental file. According to the node picture, records partially consist in referential nodes, and according to the token picture, records partially consist in referential tokens. On

the node picture, records r and s are members of the same mental file in virtue of sharing a referential node, and on the token picture, records r and s are costaged in virtue of containing referential tokens of the same type. If a record contains more than one referential node (or referential tokens of more than one type) that record is in more than one mental file.

Records have their content determined, in part, by which mental files they are in. In particular, if a record r containing component c_r is a member of file F in virtue of c_r, then F determines the referential value of c_r. Because of this, record-stages in the same file-stage contain component-stages sharing at least one referential-value. But files can change their referential-value across time, so it may be that a record which is in a file F at t_1 but not t_2 does not share any referential-value with another record which is in F at t_2 but not t_1.

CoStaging-13* gives the necessary and sufficient conditions on two record-stages being members of the same file-stage. In particular, T's record-sages r_t and s_t are costaged iff r_t and s_t contain components c_{r_t} and c_{s_t} respectively, and T can trade on the coreference of c_{r_t} and c_{s_t}. This also gives synchronic individuation conditions for nodes, and tells us what it is to be tokens of the same type. Two stages of record-components are the same node, or tokens of the same type, iff the thinker can trade on the coreference of those record-components.

Thinkers form identity judgements when they start with two mental files, and form a judgement implying that those mental files are about the same object. The thinker may link the mental files, so she can treat the records in each file as about the same thing but only via the additional linking representation. Alternatively, those files may come to be merged.

CoStaging-13* allows us to determine when identity judgements result in files being merged rather than linked. Cotemporal record-stages r_t and s_t, containing component-stages c_{r_t} and c_{s_t} respectively, are (in virtue of c_{r_t} and c_{s_t}) in linked

mental files when the thinker treats c_{r_t} and c_{s_t} as about the same object, but does so via an additional representation. r_t and s_t are costaged in virtue of c_{r_t} and c_{s_t} when the thinker can trade on the coreference of c_{r_t} and c_{s_t}.

Between merging and linking, there's a third scenario. Suppose Lois starts off with a *Superman*-file and a *Clark*-file. She learns that *Superman is Clark*, and so links these files. The link allows her to treat the records in each file as about the same thing. So she is able to treat records ${}^{Lo}(r)$ and ${}^{Lo}(s)$ as about the same thing as records ${}^{Lo}(u)$ and ${}^{Lo}(v)$:

${}^{Lo}(r)$ Superman is brave

${}^{Lo}(s)$ Superman is fast

${}^{Lo}(u)$ Clark is a journalist

${}^{Lo}(v)$ Clark wears glasses

We might imagine that in virtue of treating these records as about the same thing, it becomes natural for Lois to think *Clark is brave*. In this case, we might suppose that in virtue of the link between the *Superman*-file and *Clark*-file, Lois is able to form a new record ${}^{Lo}(w)$ in the *Clark*-file.

(67) ${}^{Lo}(w)$Clark is brave

Lois can trade on the coreference of a component of ${}^{Lo}(w)$ with components of ${}^{Lo}(u)$ and ${}^{Lo}(v)$. ${}^{Lo}(w)$ is partially constituted by a node or token, in virtue of which it is a member of Lois's *Clark*-file. But the content of ${}^{Lo}(w)$ means ${}^{Lo}(w)$ should contain only one referential-component, so it cannot also be in the *Superman*-file, because if it were in both the *Clark*-file and the *Superman*-file, it would have to contain two nodes, or tokens of two different types. So ${}^{Lo}(r) \neq {}^{Lo}(w)$. In this kind of scenario, linking files F and G leads to the thinker adding replicas of some records in F to G. These replicas are numerically distinct from their originals, and may have different content.

So described, files seem appropriate to perform many of the roles outlined for them in 2.2.2. There is no requirement that records in a single file originate in the same object, so files can be used in accounting for confusion. Thinkers can trade on the coreference of record-components even when they do not suppose that those record-components refer, so we can allow for a thinker to have mental files on objects she knows not to exist. Trading on coreference is an ubiquitous phenomenon (see 5.2.3), so the necessary and sufficient conditions for costaging are met in subpersonal, animal and infant thought. And I have suggested that we can attribute the kind of knowledge required for IK-coreference just when a thinker can trade on the coreference of record-stages (see 5.4), so files can be used to explain de jure coreference (correctly understood).

Crucially, using CoStaging-13* enables file-stages to meet the Frege-constraint and so file-stages can play the MOP role.[1] Making the simplifying assumption that all records are single-reference:[2] if a thinker is in a Frege-case involving cotemporal record-stages, then she has two cotemporal record-stages about the same object, but does not treat them as about the same thing. If a thinker has cotemporal record-stages but does not treat those record-stages as about the same thing, then the thinker cannot trade on the coreference of those record-stages. So if a thinker is in a Frege-case, she cannot trade on the coreference of record-stages, and by CoStaging-13* must have different mental files.

Moreover, file-stages are individuated in terms of the ability to trade on coreference, so we cannot suppose that a thinker has a single file-stage whilst being unable to reason as if she has a single file-stage. This rules out scenarios where a thinker has just one file-stage on an object but ends up in a Frege-case because she reasons as if she has two file-stages.

However, because CoStaging-13* only gives an account of the synchronic indi-

[1] See 1.1.3.
[2] See p19.

viduation conditions on singular thought, we are no closer to identifying whether mental files can be used in an explanation of what it is to have a continued belief. And CoStaging-13* is not in itself useful for determining whether files constitute singular thought or are used in language understanding. We need further investigation of these phenomena before we can ascertain what role mental files play.[3]

In 3.8, I mentioned that a thinker T may have a belief that P even if it is not explicit to T that P. Suppose at t_1, Alexandra cannot bring to mind her belief about where Oliver Cromwell was born. Several hours later (at t_2) she is able to bring to mind that *Cromwell was born in Huntingdon* without intermediate prompting. It is natural to say that she believed that *Cromwell was born in Huntingdon* at t_1, even though at t_1 this was not explicit to her.

Assuming that Alexandra has only ever had a single file on Oliver Cromwell, if she believed at t_1 that *Cromwell was born in Huntingdon*, then record-stage ${}^{Al}(\mathrm{r})_{t_1}$ should have been in the t_1 stage of her Cromwell-file.

${}^{Al}(\mathrm{r})_{t_1}$ Cromwell was born in Huntingdon.

However, we might think that at t_1 she would have been unable to trade on the coreference of any component-stage of ${}^{Al}(\mathrm{r})_{t_1}$ with any component-stages from record-stages recording explicit beliefs, for example ${}^{Al}(\mathrm{s})_{t_1}$.

${}^{Al}(\mathrm{s})_{t_1}$ Cromwell was Lord Protector.

However, by CoStaging-13*, if she cannot trade on the coreference of component-stages in ${}^{Al}(\mathrm{r})_{t_1}$ and ${}^{Al}(\mathrm{s})_{t_1}$, then ${}^{Al}(\mathrm{r})_{t_1}$ and ${}^{Al}(\mathrm{s})_{t_1}$ are not costaged.

Rather than supposing that something has gone wrong in the account of mental files, I suggest we allow that a thinker can, in some suitably idealised sense, trade on the coreference of component-stages belonging to costaged record-stages, even when

[3]Though in the case of singular thought, there is room for simply stipulating that singular thought is constituted by mental files (see 1.1.2).

those record-stages are not immediately available to the thinker to use in reasoning. What matters is what the thinker would be able to do *were* those record-stages immediately available for use in reasoning. Whilst Alexandra may need a little time to recall that *Cromwell was born in Huntingdon*, once she has brought this to mind she needs no extra time or additional representations to realise that this belief is about the same person as her other Cromwell beliefs. Were ${}^{Al}(\mathrm{r})_{t_1}$ to be available for use in reasoning at t_1, she could trade on the coreference of ${}^{Al}(\mathrm{r})_{t_1}$ and ${}^{Al}(\mathrm{s})_{t_1}$. Hence, at t_1, Alexandra can trade on the coreference of ${}^{Al}(\mathrm{r})_{t_1}$ and ${}^{Al}(\mathrm{s})_{t_1}$.

If Alexandra can't reason using ${}^{Al}(\mathrm{r})_{t_1}$, she can't treat it as if it is not about the same object as other record-stages in the t_1 stage of her *Cromwell*-file. She may fail to make certain inferences, for example, at t_1 she is in some sense not disposed to infer that *the first Lord Protector was born in Huntingdon*, but we do not explain her mistake in terms of her failing to realise that $a = b$. Rather we explain her mistake as failing to access one of her beliefs about Cromwell. So allowing that a thinker can have files containing records that aren't immediately available for use in reasoning will not lead to Frege-cases, and so does not undermine files playing the MOP role.

6.2.2 Remaining issues

Although giving an answer to the costaging question has allowed me to fill in some details of file theory, there are still undecided questions.

One of these is the choice between the node and token picture of mental files. Nothing I have said in answering the costaging question helps us choose between these pictures.

In contrast, we might hope that adopting CoStaging-13* will help us chose between conception and global versions of mental file theory, that is between theories

which say files only contain belief-records and records of belief-like information, and theories which say records contain all types of record.[4] In 5.2.3, I observed that the argument that thinkers trade upon coreference applies to all attitudes, not just belief. We might think that CoStaging-13* gives reason to claim that files contain all kinds of attitudes about an object, so global versions are correct.

However, nothing in the argument for CoStaging-13* allows us to draw this conclusion. Suppose that there is something privileged about a thinker's beliefs and belief-like information which means only these records count as being contained in a mental file (perhaps because of a special reference-fixing role). Then we can stipulate that only these records are contained in a mental file. Other types of record are merely associated with the file.[5]

In this case, we must adjust the argument originally given for CoStaging-13* in 5.4 to argue that: r_t and s_t are associated with the same file-stage iff r_t and s_t contain c_{r_t} and c_{s_t} respectively, and the thinker can trade on the coreference of c_{r_t} and c_{s_t}. We can then add an extra clause to CoStaging-13* to give an answer to the costaging question compatible with conception versions of file accounts.

CoStaging-14* r_t and s_t are costaged iff r_t and s_t contain c_{r_t} and c_{s_t} respectively, and the thinker can trade on the coreference of c_{r_t} and c_{s_t}, and r_t and s_t are belief-records or records of belief-like information.

However, as I pointed out in 2.2.1, in the node and token pictures of files, there is only one way for a record to be associated with a node, so there's no difference in implementation between the global and conception versions. At the moment it appears that only terminological issues are at stake between the global and conception versions.

[4] See 2.2.1.
[5] See 2.2.1.

A gap in the file theory developed above is that it only gives a synchronic individuation condition on mental files, disregarding the fact that files are supposed to be persisting mental particulars. In chapter 7, I explore how this gap can be filled. But first, I complete my discussion of the theory developed so far by comparing my answer to the costaging question to Recanati's (2012), and responding to a pressing objection to file-theory as it has been developed so far.

6.3 Recanati's answer to the costaging question

In 2.1.2, I noted that although appeals to mental files are commonplace, mental file accounts are generally only sketched in outline. Recanati's *Mental Files* (2012) is a notable exception. His book-length treatment of the topic describes an 'indexical model' of files, defends it from accusations of circularity, and then uses it to respond to various long-standing puzzles in the philosophy of language.

There is not space to give critical attention to Recanati's entire account. However, I will orient my position in the files' literature by discussing what Recanati says about costaging and showing that difficulties with Recanati's account of costaging force us back towards CoStaging-13*, my own answer to the costaging question.

It is important for Recanati to have an answer to the costaging question for the same reasons it is important for any theory of mental files to answer the costaging question.[6] However, Recanati does not give a clear statement of either the synchronic or diachronic individuation conditions on mental files. Instead, he makes various comments from which it is is possible to extract possible answers to the costaging question.

The first of these is:

> To say that there are two distinct mental files is to say that information

[6] See 2.3.1.

> in one file is insulated from information in the other file. Files are a matter of information clustering. Clustering takes place when all the information derives from the same source, through the same ER relation...[7]
>
> (Recanati 2012, p42)

This suggests two possible answers to the costaging question, CoStaging-15 and CoStaging-16. Making the simplifying assumption (as Recanati generally does):

CoStaging-15 r_t and s_t are costaged iff r_t and s_t causally originate in the same object and enter through the same acquaintance relation

CoStaging-16 Record-stages r_t and s_t are costaged iff r_t and s_t are not insulated from one another

CoStaging-15 cannot be successful. First, as Recanati goes on to say, it is merely a norm that the information in a file derives from a single source. In practice, information in a mental file may derive from multiple sources or from no object at all (2012, p63). Moreover, if we accept CoStaging-15, then files cannot play the MOP role. In the Illustrious case,[8] Tony is in a Frege-case, so must have two file-stages on the Illustrious. But each file-stage contains records originating in the same object through the same acquaintance relation (visual attention), so CoStaging-15 only allows Tony to have one file-stage on the Illustrious. Hence CoStaging-15 is inadequate.

And CoStaging-16 is at best a provisional answer. When a thinker forms an identity judgement, she links the relevant mental files, and linking overcomes the files' informational isolation (2012, p43), allowing information to flow between files (2012, p94).

After incorporating linking into his theory, Recanati gives another clue as to how he might answer the costaging question:

[7] 'ER relations' are 'epistemically rewarding relations to objects', i.e. acquaintance relations.
[8] See p13.

> Two pieces of information go into the same file if they are taken to concern the same object.
>
> (Recanati 2012, p101)

As stated, this is only a sufficient condition on costaging, but this is the closest Recanati gives to a final answer. Switching from my talk of 'records' to Recanati's terminology of information, this suggests that for information-pieces i and j belonging to thinker T:

CoStaging-17 i and j are costaged at t iff at t T takes i and j to be about the same thing.

Recanati doesn't explore the consequences of CoStaging-17. However, exploring these consequences shows that CoStaging-17 gives an implausible account of costaging.

In 2.3.2, I considered CoStaging-5 which is very similar to CoStaging-17. I rejected CoStaging-5 on the basis that it implies an ideal of a 1:1 mapping from files to objects, and I argue in 2.1.3 that files are not maintained in accordance with a 1:1 mapping ideal from files to objects.

Recanati also rejects a 1:1 mapping ideal in favour of incorporating linking into his file theory, but he gives different reasons. His explanation for rejecting this ideal is that:

> It is of the essence of modes of presentation that there can be a multiplicity of modes of presentation for the same object.
>
> (Recanati 2012, p45)

The idea is that if a thinker hears a bird and simultaneously sees a bird, and comes to judge *that-heard bird = that-seen bird*, the bird is still presented in two distinct ways, and so she retains two MOPs, i.e. two (linked) files.

But *prima facie* this is incompatible with CoStaging-17. If pieces of information are costaged iff they are taken to be about the same thing, then it appears that a

thinker who has judged (by t) that *that-heard bird = that-seen bird* should (at t) only have one mental file on the bird, because (at t) she takes all the information in the heard-bird file to be about the same thing as the information in the seen-bird file.

Recanati's only option is to claim that there are two files, one based on the auditory ER-relation, one on the visual, and that all the information about the bird is simultaneously in both files. The very same piece of information must be in both files at once (rather than simply be duplicated in each file), because if a thinker taking cotemporal information to be about the same thing results in that information being costaged, the only way a thinker can retain two separate files of duplicate information is if she does not take the information to be about the same thing.

One of the difficulties in understanding this proposal is that Recanati does not give a gloss for what a *piece of information* is. But clearly, pieces of information cannot be fine-grained contents, if fine-grained contents are partially constituted by MOPs. Using the simplifying assumption: if information is fine-grained content, then information-piece i cannot simultaneously be associated with distinct file-stages F_t and G_t, because in virtue of being associated with F_t, i is fine-grained content P, and in virtue of being associated with G_t, i is fine-grained content Q, where $P \neq Q$. But then i is not self-identical.

It is more tempting to suppose that Recanati's "pieces of information" correspond to my "records" — mental representations that the world is a certain way. On this view of information it is hard to see how a single-reference information-piece can simultaneousy be in two different files. According to the node and token pictures of files, a record in two files must contain two referential-components. But if the content of a record is composed from its components, it isn't clear how a record containing two referential-components can still count as a single-reference record.

So it is difficult to see how Recanati can make CoStaging-17 compatible with his rejection of a 1:1 mapping ideal from files to objects. Moreover, CoStaging-17 has consequences which undermine our ability to suppose that files can be reference-fixing MOPs. In particular, using Recanati's understanding of de jure coreference, CoStaging-17 implies that not all information in a file is de jure coreferential. But if files play the MOP role, all information in the file should be de jure coreferential.

Recanati characterises de jure coreference in linguistic terms, using something similar to Pinillos's test (iii) for AP-coreference.[9]

> I characterize *de jure* co-reference in terms of a priori knowledge of (conditional) co-reference: two terms are *de jure* co-referential just in case anyone who understands the utterance in which they occur knows that they co-refer if they refer at all.
>
> (Recanati 2012, p110)

Recanati also discusses the 'de jure coreference' of information (e.g. 2012, p95), though he gives no separate definition of what it is for information to be de jure coreferential. I suggest FR-coreference as a reconstruction of what Recanati would call 'de jure coreference' for information.

Making the simplifying assumption, for information-pieces i and j belonging to T:

> i and j are *FR-coreferential* iff T knows a priori that i and j are externally-coreferential (if they refer at all)

This is a rough definition, relying on the simplifying assumption. Nonetheless, it is adequate for the purposes at hand, and I assume that where Recanati discusses the 'de jure coreference' of information, he is discussing FR-coreference.

It is not the case that: if I take it to be the case that i and j are about the same thing, then i and j are FR-coreferential. If I start with information i about *that-heard bird* and information j about *that-seen bird*, and by t have come to judge

[9] See 4.2.1.

(on empirical grounds) that *that-heard bird = that-seen bird*, I do not thereby have at t a priori knowledge that i and j externally-corefer (if they refer at all). So long as I can think independently about the heard-bird and the seen-bird, i and j may be about different things. Nonetheless, by CoStaging-17, i and j are costaged at t. So according to CoStaging-17, not all costaged information is FR-coreferential. Indeed, Recanati claims that if i and j are in the same file due to information flowing between two linked files, they are not FR-coreferential. Whereas i and j are FR-coreferential iff they occur in the same file without a prior linking operation (2012, pp94-95).

Because Recanati claims that i and j are FR-coreferential iff they occur in the same file without a prior linking operation, he must suppose that FR-coreference is widespread. Given how widespread FR-coreference must be, it is difficult to see what else can be required for FR-coreference beyond (i) the thinker treating the information as about the same object, and (ii) the information being guaranteed to be about the same thing (if about anything at all). Any other requirements threaten to make FR-coreference so demanding that few information-pieces will count as FR-coreferential.

But if file-stages are reference-fixing MOPs, all costaged information must be FR-coreferential. The Frege-constraint on MOPs means that if i and j are costaged, the thinker must treat i and j as about the same thing.[10] Otherwise, i and j could be treated as about different things, resulting in a Frege-case involving costaged records. And because files are reference-fixing, if i and j are in the same file-stage then they are about the same thing (if they are about anything at all). So all costaged information meets the requirement for being FR-coreferential

Because Recanati's file-stages are reference-fixing MOPs (Recanati 2012, pVIII), all information in a file-stage must be FR-coreferential. But by CoStaging-17, it is

[10] See 1.1.3.

not the case that all information in a file-stage is FR-coreferential. So CoStaging-17 is unsuccessful, because it has a consequence incompatible with file-stages being reference-fixing MOPs.

Although CoStaging-17 is inadequate, it might be possible to construct alternative answers to the costaging question from Recanati's materials. There are two ways a thinker T may come treat information-pieces i and j as about the same thing. In one, i and j start in distinct mental files, but T makes a judgement (or supposition) implying those files are about the same thing. In the other, T simply treats i and j as about the same thing from the start, not making any judgement (or supposition) that i and j are about the same thing. In the latter sort of case, Recanati says there is a "presumption of identity" between i and j, and he equates presumptions of identity with Campbell's 'trading on identity'.[11] Recanati suggests that the distinction between presumptions and judgements of identity

> gives us a criterion for telling apart the cases in which there is a single file and the cases in which there are two. If the subject 'trades upon identity' and proceeds to integrate various pieces of information directly, without appealing to a further identity premise, that means that there is a single mode of presentation.
>
> (Recanati 2012, p83)

This suggests a new answer to the costaging question. Again using the simplifying assumption:

CoStaging-18 i and j are costaged at t iff there is a presumption of identity between i and j at t.

But CoStaging-18 is not acceptable as it stands. 'Presumption of identity' is ambiguous. According to the *information-gathering* understanding, there is a presumption of identity between i and j just in case T treats them as if they are about the

[11] I use Campbell's earlier terminology of 'trading on coreference' when discussing trading on identity. See Recanati (2012, pp47-50); Campbell (1987-1988, 1995).

same thing without ever having formed an identity judgement implying that they are about the same thing and without having ever called into question whether they are about the same thing. According to the *current-reasoning* understanding, there is a presumption of identity between i and j just in case T can reason as if i and j are about the same object without using an additional premise implying this. In other words, there is a *current-reasoning* presumption of identity between i and j iff T can trade on the coreference of i and j (as I define trading on coreference, see 5.2.1).

These disambiguations come apart in at least one direction. If we allow that identity judgements can become so embedded that a thinker can come to trade on the coreference of information originally associated with different files, then there can be a current-reasoning presumption of identity between i and j even if there is not an information-gathering presumption of identity between i and j. However, it isn't clear which disambiguation Recanati had in mind. His focus on how information is gathered and whether a thinker has ever made a relevant judgement of identity seems to suggest the *information-gathering* understanding. But Campbell's argument for trading on identity (reiterated by Recanati) treats trading on identity as a type of reasoning, suggesting the *current-reasoning* understanding. So CoStaging-18 could be disambiguated in two different ways.

CoStaging-19 i and j are costaged at t iff there is a information-gathering presumption of identity between i and j at t.

CoStaging-20 i and j are costaged at t iff there is a current-reasoning presumption of identity between i and j at t.

CoStaging-19 is incompatible with Recanati's own claims that judgements of identity lead to linking, and that linking results in information flowing between the files (e.g. 2012, p94). But we might not think this a particular problem in itself.

After all, Recanati does not defend his claim that information can flow between linked mental files.

However, we should still reject CoStaging-19. Mental files are supposed to explain a thinker's current reasoning, not to give a history of how that information was accumulated. Consider a version of Davidson's swampman (Davidson 2001): Thales has beliefs he expresses (1) and (2′).[12]

(1) Hesperus is bright.

(2′) It is not the case that Phosphorus is bright.

We explain this in terms of Thales having two unlinked mental files, one associated with the word "Hesperus", one with the word "Phosphorus". Call Thales' t-stage 'Thales$_t$'. A storm at t generates a duplicate of Thales$_t$, sharing all properties intrinsic to Thales$_t$. Call this duplicate 'Swamp-Thales$_t$'. Like Thales$_t$, Swamp-Thales$_t$ is disposed to assent sincerely to (1) and (2′). It may be that Swamp-Thales$_t$ is not related to the world in the right way to have contentful mental representations. But assuming that the cognitive processes of reasoning (even if not the content of the thoughts involved in reasoning) supervene on the intrinsic qualities of a thinker, we expect that Swamp-Thales$_t$ will be disposed to reason just as Thales$_t$ is disposed to reason. And if mental files track dispositions to reason (as they must to resolve Frege-cases), Swamp-Thales$_t$ must have just the same file-stages as Thales$_t$.[13]

However, according to CoStaging-19, Swamp-Thales$_t$ has no file-stages, because he has gathered no information using presumptions of identity. So CoStaging-19 is unsuccessful, because it incorrectly attributes Swamp-Thales$_2$ no file-stages.

This example highlights that what matters for costaging is what mental representations a thinker has, rather than how she came to acquire those representations. This is further motivation for CoStaging-20. But CoStaging-20 is just CoStaging-

[12] See 1.1.1.

[13] I go onto argue in 6.4.1 that we must suppose that the facts about trading on coreference are prior to facts about content, further supporting this line of reasoning.

13*, adjusted take account of Recanati's terminology and the simplifying assumption. So it turns out that investigating Recanati's options for an alternative answer to the costaging question leads back to 5.4, my preferred answer to the costaging question.

6.4 Circularity objections

A common objection to mental file accounts is that they are circular. In this section, I explore two versions of this objection. The first concerns the coherence of my answer to the costaging question, the second concerns the explanatory value of file theory.

6.4.1 First objection

Mental files are supposed to determine the referential-values of the records they contain, so the project of identifying the contents of records is not independent of identifying which record-stages are costaged. This is clearest in 5.3.3, where I used the fact that certain record-stages are costaged to resolve questions about their referential-value.

We might therefore be wary of any attempts to answer the costaging question by making claims about what a thinker knows or accepts (examples include CoStaging-9* [14] and CoStaging-12*[15]) What a thinker knows or accepts is partially constituted by the facts about costaging, so it is incoherent to suppose that the facts about costaging are themselves constituted by what a thinker knows or believes.[16]

Fortunately, CoStaging-13* does not fall victim to this objection, so long as records and record-components are understood as the vehicles of content. Vehicles

[14] See p96.
[15] See p139.
[16] This issue was brought to my attention by Cian Dorr.

of content are individuated by their non-semantic properties, and those non-semantic properties determine how the record or record-component is used in reasoning.[17] If records and record-components are the vehicles of content, we can allow that facts about trading on coreference are explanatorily prior to facts about the contents of records, so no circularity threatens.

In practice, our understanding of the mind and brain is such that we think about records by description, for example "the representation that realises T's belief that *P*". Whilst our practice of attributing contentful attitudes may not be without problems, it is better established than our practice of attributing non-semantic properties to mental representations. But the fact that the content of record r is *epistemically* prior to r's non-semantic features does not undermine the claim that r's non-semantic features are *explanatorily* prior to its semantic features. So we can allow that facts about trading on coreference are explanatorily prior to facts about the content of records.

We might worry that these considerations threaten the practice of using Frege-cases as data in answering the costaging question. Files fix referential-value, so even the coarse-grained contents of a thinker's thoughts are dependent on which records are stored in which files. So one might think that Frege-cases cannot be used as evidence for an account of costaging, because we need an account of costaging to work out whether a thinker is in a Frege-case.

But there is no problem here, because of how Frege-cases are used as data in answering the costaging question. There is no need to presuppose an account of which cases are Frege-cases. Instead, we simply need to know whether record-stages r_t and s_t are treated as about same object, and we can suppose that this is determined by pre-semantic facts about r_t and s_t. We then consider whether the proposed answer to the costaging question implies that r_t and s_t are costaged.

[17] See e.g. Sedivy (2004, p154).

If there are cases where r_t and s_t are treated as about different objects, but the proposed answer to the costaging question implies that r_t and s_t are costaged, then we know that r_t and s_t should share referential-value and that we are in a Frege-case despite only having one file-stage. This is enough to show that the proposed answer to the costaging question is unsuccessful.

6.4.2 Second objection

The second circularity objection targets the explanatory value of mental files.

It is sometimes objected that file theories are circular. If we try to answer the costaging question in terms of record-stages sharing nodes, but have no better understanding of what a node is than that it is the referential component shared by all costaged record-stages, then this objection seems well-founded. But I have been trying to give a more illuminating answer to the costaging question than this.

However, one might object as follows: you give CoStaging-13* to explain what it is to have a file-stage, claiming record-stages r_t and s_t are costaged iff the thinker can trade on the coreference of c_{r_t} and c_{s_t}. But trading on coreference is itself a crucial kind of reasoning that is itself in need of explanation. And the most plausible explanation is that a thinker can trade on the coreference of record-stages because they share a MOP.[18] The MOP role is played by files, so you are effectively saying that records are costaged because a thinker can trade on the coreference of c_{r_t} and c_{s_t}, and that you can trade on the coreference of c_{r_t} and c_{s_t} because r_t and s_t are costaged. But this is circular.

There are several possible variations on this objection. One might worry that part of what it is to have records in the same file is for them to be treated as about the same thing, so we cannot say that a thinker treats records as about the same object *because* they are members of the same file. More generally, the concern is

[18] See e.g. Campbell (1987-1988, 1995); Dickie and Rattan (2010).

that we explain what files are in terms of the very phenomena they are supposed to explain.[19]

Rather than discussing all variations on this circularity objection, I will focus on the first version I gave: that we explain what it is to have a mental file in terms of trading on coreference, but also claim that files explain trading on coreference. *Mutatis mutandis* , my response can be adapted for the other versions. My response to the objection is that I have said nothing yet about the status of CoStaging-13*. But there are several ways of understanding the status of CoStaging-13*, and some avoid this objection.

6.4.3 The status of CoStaging-13*

There are various ways of understanding the status of CoStaging-13*, each corresponding to a different way of responding to the circularity objection.

Misguided We might think it was misguided to attempt to answer the costaging question, because 'mental files' are no more than a place-holder for a proper theory of MOPs. An analogous case would be talk of 'belief boxes'. Philosophers sometimes say: "having a non-dispositional belief that P involves tokening a representation with content P in your belief-box, and having a desire that P involves tokening a record with content P in your desire-box". Those who appeal to belief-boxes commit themselves at least to claiming that having a non-dispositional attitude that P requires having a mental representation that P, but beyond this, no one takes talk of 'belief-boxes' terribly seriously. No one expects a theory of belief-boxes, or to find interesting similarities between non-mental boxes and whatever it is in the cognitive system that makes a record that P a belief-record rather than a desire-record. Talk of 'belief-boxes' is just a picturesque place-holder theory, freeing

[19]Versions of this circularity objection are given by Lawlor (2001, p80) and Fine (2007, p68). Recanati (2012, ch8) gives an alternative way of responding to such objections.

us to discuss other issues.

If we think of file-talk as analogous to belief-box talk, then in trying to answer the costaging question, I have been missing the point. We aren't supposed to be looking for an adequate theory of MOPs in terms of files, so it should be no surprise if attempts to give such a theory are explanatorily unsatisfying.

Tracing the history of file-talk, it is tempting to think that early file-talk would have been thought analogous to talk of belief boxes. Strawson (1974) uses a simple file model to discuss identity judgements, alongside a 'dot' model. He does not claim one model has advantages over the other, but writes:

> [S]uch models as these have various imperfections... Equally certainly they could be improved. But it would be idle to improve them beyond the point at which they fulfill their purpose. Their purpose is to help us to escape from a typically, philosophically obsessive way of looking at our question. Once we are clear, we can relax.
>
> (Strawson 1974, p56)

A definition We might think of CoStaging-13* as giving a definition of what it is for records to be costaged. The idea is that we have no adequate grasp on the idea of costaging except via a claim like CoStaging-13*. An analogous case is talk of 'Pica'. Abbreviating somewhat, Pica is defined as the persistent eating of nonnutritive substances (American Psychiatric Association 2000). Plausibly we have no grasp on what Pica is beyond this definition. We do not, for example, suppose that 'Pica' labels an underlying pathology causing the symptom of eating nonnutritive substances (causes of Pica are as diverse as childhood conditioning and pregnancy).

If we think of CoStaging-13* as giving a definition of costaging, then an utterance of (68) is non-explanatory, just as an utterance of (69) is non-explanatory:

(68) r_t and s_t are costaged because T can trade on the coreference of $\mathrm{c}_{\mathrm{r}_t}$ and $\mathrm{c}_{\mathrm{s}_t}$.

(69) Olivia has Pica because she persistently eats nonnutritive substances.

At best, (68) and (69) explain terminology. They cannot explain the phenomena under discussion.

If we think of CoStaging-13* as giving a definition of costaging, talk of costaging is merely short-hand for trading on coreference. Attributions of costaged records are no more interesting than claims that thinkers can trade on the coreference of components of those records. And CoStaging-13* has no explanatory significance of its own, merely clarifying terminology without advancing our understanding of mental representation.

Understanding CoStaging-13* as miguided or as a definition is to take a pessimistic view of the value of file-theory. In contrast, much contemporary writing on mental files appears more optimistic. 'Files' are not treated as one metaphor or shorthand among others. Rather, we are presented with detailed typologies of mental files (e.g. Perry 2001; Recanati 2012). Theories about what constitutes singular thought (e.g. Jeshion 2010) and what fixes the reference of singular thoughts (e.g. Dickie 2010) are developed that turn on claims about the nature and structure of mental files. And talk of files is attributed "empirical bite" (Recanati 2012, pVIII).

This points towards two more optimistic ways of understanding the status of CoStaging-13*.

A statement of identifying characteristics To understand this proposal about the status of CoStaging-13*, it is best to start with an analogous case: gold. We use certain characteristics to identify things that are made of gold. For the sake of the argument, suppose we use yellowness and density. But we are also prone to say things like (70):

(70) This ring is yellow and dense because it is gold.

One might think that if we use yellowness and density to identify gold, then "gold" just means "dense and yellow", and so (70) is no more explanatory than (69). But

there's a good case that this gets the semantics of natural kind terms wrong.[20] Instead:

> we use 'gold' as a term for a certain *kind* of thing... The kind of thing is *thought* to have certain identifying marks.
>
> (Kripke 1980 [1972], p118)

'Gold' does not mean 'yellow and dense'. Instead, those characteristics are used to identify a certain *kind* of thing: gold. Being a member of the gold-kind is explanatorily prior to being yellow and dense. So utterances of (70) give a genuine explanation: the characteristics of the ring are caused by its being made of gold. As things stand, it isn't a very illuminating explanation. We hope for empirical advances that tell us more about the kind and account for why gold things have these characteristics. Nonetheless, even without these empirical advances, we can use (70) to give an explanation of the ring's characteristics.

This gives us a way of thinking about the status of CoStaging-13*. We can suppose CoStaging-13* states the identifying characteristics of some kind of relation: costaging. r_t and s_t can be identified as instantiating the costaging relation when T can trade on the coreference of $\mathrm{c}_{\mathrm{r}_t}$ and $\mathrm{c}_{\mathrm{s}_t}$. Costaging is explanatorily prior to trading on coreference, so utterances of (71) are true, and utterances of (68) are false.

(71) T can trade on the coreference of $\mathrm{c}_{\mathrm{r}_t}$ and $\mathrm{c}_{\mathrm{s}_t}$ because r_t and s_t are costaged.

We should acknowledge that utterances of (71) are not fully illuminating, and hope for empirical advances to explain why costaged records have these identifying characteristics.

Understanding CoStaging-13* as giving identifying characteristics of costaging rescues file theory from the circularity objection made in 6.4.2. The objection in 6.4.2 misrepresents the status of CoStaging-13*. CoStaging-13* is not an expla-

[20] See Kripke (1980 [1972]), also Leibniz (1996 [1764], book III),

nation of costaging in terms of trading on coreference, rather, it is a statement of costaging's identifying characteristics. We can explain incidences of those identifying characteristics by claiming that record-stages are costaged.

However, if CoStaging-13* has this status, then we should expect CoStaging-13* to be false. CoStaging-13* claims that it is necessary and sufficient for costaging that costaged records possess the identifying characteristics of costaging. But we do not suppose that it is necessary and sufficient for something to be gold that it is yellow and dense. Instead, we allow that with improved understanding of the kind underlying the identifying characteristics of gold, we may recognise that some yellow and dense things are not members of that kind (for example, fool's gold). And we may discover things which are members of that kind but do not display the identifying characteristics. For example, if we identify tigers as orange, stripey, long-toothed creatures, we can still allow for white, unstriped, toothless tigers. We might even find that there is more than one kind picked out by the identifying characteristcs. For example, two different kinds of mineral both share the identifying characteristics of jade. Rather than saying that one kind is true jade and the other fool's jade, we say that there are two kinds of jade (Putnam 1975). So if we think CoStaging-13* just gives identifying characteristics, we should expect that CoStaging-13* itself is false. If CoStaging-13* is false, it cannot itself be used in explanations. Nonetheless, getting to the point where we claimed CoStaging-13* counts as an improvement in our understanding of the mental representation of objects, because it allowed us to identify costaging's identifying characteristics.

As it happens there have been empirical advances in our understanding of gold. We now know that something is gold iff it has atomic number 79. This suggests a fourth way of understanding the status of CoStaging-13*.

A necessary truth There are two different analogies which can help us think about how CoStaging-13* might be a necessary truth. Utterances of (72) state a necessary a posteriori truth.

(72) Something is water iff it is H_2O.

And suppose, for the sake of simplicity, that utterances of (73) state a necessary a priori truth.

(73) T knows that P iff T has a justified true belief that P.

We can think that CoStaging-13* is analogous with one of these necessary truths.

It is worth highlighting a difference between *knowledge* and *Pica*.

(74) T has Pica iff T persistently eats nonnutritive substances.

Utterances of (74) and (73) are both necessary a priori truths, and both can be used as definitions. As definitions they will inform someone who doesn't already have a grasp on the meaning of 'Pica' or 'know'. But most language users have a grasp on what it is to know something, independent of conceptually analysing 'knowledge'. So utterances of (73) state a non-trivial fact about knowledge, just as utterances of (72) state a non-trivial fact about water. In contrast, utterances of (74) cannot be used to state a non-trivial fact about Pica, because there is nothing more to grasping what it is to have Pica than grasping this definition.

If we accept that we have an adequate independent grasp on costaging prior to answering the costaging question, we can claim that CoStaging-13* has the status of a non-trivial necessary truth, analogous to utterances of either (72) or, as seems more likely (73).

CoStaging-13* has potential explanatory significance. For example, we can use CoStaging-13* in an explanation of why a thinker can only trade on the coreference of components which share referential-value.[21] Similarly, we can use utterances of

[21] See 5.3.3.

(72) and (73) in explanations, as in utterances of (75) and (76).

(75) It doesn't matter if Olivia drinks that H_2O, because something is water iff it's H_2O.

(76) Knowledge is valuable because S knows that P iff T has a justified true belief that P.

However, we might wonder whether standalone utterances of (68) and (71) can count as explanations of either costaging or trading on coreference. Consider utterances of (77) and (78).

(77) S knows that P because S has a justified true belief that P.

(78) S has a justified true belief that P because S knows that P.

Utterances of (77) and (78) may be used to explain terms, or even to explain how we know that S knows that P, or how we know that S has a justified true belief that P. But neither should be read as stating that one phenomenon is explanatorily prior to the other. Rather, they give an identity, and if $a = b$, then a cannot be explanatorily prior to b.

So if we suppose CoStaging-13* has the explanatory status of an a priori truth, then CoStaging-13* is informative and true, and so can itself potentially be deployed in explanation. However, if we attempted to make claims about either costaging or trading on coreference being explanatorily prior, circularity might threaten, because neither costaging nor trading on coreference is explanatorily prior to the other.

Bringing these options together, we see that how we respond to the circularity objection presented in 6.4.2 depends on what status we give CoStaging-13*. If we suppose that CoStaging-13* is misguided or a definition, then mental file explanations are circular, and moreover CoStaging-13* does not add to our understanding of mental representation. If CoStaging-13* states costaging's identifying character-

istics, then the circularity objection does not hold, but strictly speaking CoStaging-13* is false (even if working towards CoStaging-13* advanced our understanding of mental representation). And if CoStaging-13* has the status of a necessary truth, CoStaging-13* is informative and potentially useful in further explanations. However, we cannot (without risking circularity) claim either that trading on coreference is explanatorily prior to costaging, or that costaging is explanatorily prior to trading on coreference.

Before moving on, there are several points worth mentioning. First my concern is not with *why* record-stages come to be costaged, merely what can be explained given that record-stages are costaged. Whatever the status of CoStaging-13*, it is likely this other question will be answered by enumerating ways a thinker may come to recognise objects, and so the most interesting answers will come from psychology and psycholinguistics rather than from philosophy.

Second, one might wonder whether opening the possibility that CoStaging-13* gives only costaging's identifying characteristics undermines my earlier criticisms of potential answers to the costaging question. For example, in 3.8 I argued against using RR-coreference to answer the costaging question, because it was possible that T might not treat RR-coreferential record-stages as externally-coreferential. But acknowledging that not all costaged records exhibit the identifying characteristics of costaging does not mean that we shouldn't look for the best available account of those characteristics. We have an a priori argument that there is a mismatch between record-stages being RR-coreferential and behaving as we expect costaged records to behave. Given we have an alternative where there is no a priori argument for any mismatch,[22] we should prefer this alternative.

Third, these options reveal a further decision point for a theory of mental files, this time about the status of the theory's claims. The options oulined above are all

[22] And less threat of circularity. See 6.4.1.

plausible. I will therefore remain neutral between them, and show how subsequent issues interact with these different ways of understanding file-talk.

Fourth, we might well think that these options correspond to a decision point for *any* theory of mental representation. Whenever functionally characterised mental representations are posited, one can ask about the explanatory relationship between the mental representation and the functions in terms of which it is characterised.[23]

6.5 Final remarks

In this chapter, I have presented my partial theory of mental files, and highlighted some remaining decision points. I have contrasted my answer to the costaging question with Recanati's (2012), and shown that although Recanati takes a different approach to giving the synchronic individuation conditions on mental files, a careful exploration of the consequences of his position leads back to my preferred answer to the costaging question.

I have considered circularity objections to my account of files. I have argued we avoid one circularity objection by understanding records as vehicles of content, that our response to the second circularity objection depends on what decision we make about the status of claims like CoStaging-13*, and that there are plausible positions which are not threatened by the second circularity objection.

[23]These specific issues about files are comparable to more general debates between analytic and scientific functionalists, and role and realizer functionalists (see e.g. Lewis 2002 [1972]; Block 1980 [1978]).

Chapter 7

Cofiling

7.1 Introduction

I have focussed on the costaging question, putting aside the cofiling question until I had developed a satisfactory account of costaging. But now that I have defended an answer to the costaging question, it remains to be seen how this answer may be supplemented with an account of files' diachronic individuation conditions. I will focus on the cofiling question, that is:

Under what conditions are file-stages F_t and $G_{t'}$ stages of the same file?

Following the route I took in discussing the costaging question, I start by looking for necessary and sufficient conditions on cofiling. In 6.4.3, I considered various ways of thinking about the status of my answer to the costaging question: CoStaging-13*. One option is that CoStaging-13* gives costaging's identifying characteristics. On this understanding of CoStaging-13*, we naturally expect to give identifying characteristics for cofiling rather than necessary and sufficient conditions. So we might worry that it is misguided to look for necessary and sufficient conditions. Nonetheless, necessary and sufficient conditions are still a good place to start. We would have a strong account of the identifying characteristics on cofiling if we had an account of these characteristics such that there is no a priori argument that cofiled file-stages fail to meet these characteristics.

I start by introducing considerations that any answer to the cofiling question will have to take account of. I discuss various options for answering the cofiling question, and conclude that no available answer works across all accounts of mental files. And I briefly discuss why this conclusion should be troubling to anyone positing persisting MOPs.

7.2 Cofiling considerations

7.2.1 Records, referential-components and files

The first constraint on an answer to the cofiling question is that the files picture indicates that the diachronic individuation conditions on records, referential-components and files are closely related. Records are partially constituted by referential-components, i.e. nodes or tokens. Therefore, part of what it is for a record to persist is to have a persisting node or token.[1] Files are bundles of records containing the same node (or tokens of the same type). So having a persisting file is also partly a matter of having a persisting node (or token).

At the moment, we have no account of the diachronic individuation conditions on files, persisting nodes (or tokens), or persisting records. We can hope that these diachronic individuation conditions are sufficiently closely related that getting an understanding of one will get us an understanding of the others.[2] But the fact that the conditions are closely related also means that we cannot give an illuminating account of the diachronic individuation conditions on one of these by presupposing an account of the individuation conditions on one of the others.

7.2.2 Are there cross-temporal examples of Frege-cases?

One approach to answering the cofiling question would be to give a direct account of the conditions on two file-stages being stages of the same persisting file. An alternative approach would be to make the simplifying assumption that all records are single-reference,[3] and ask what it is for non-cotemporal record-stages to be members of the same file. A third approach is to give an account of what it is for a referential-component (a node, or a token) of a record to persist. Whichever

[1] Though having a persisting record is not a prerequisite of having a persisting node, or having tokens of the same type at different times.

[2] See 2.3.1.

[3] See p19.

approach we take, we need an account of what it is for two non-cotemporal stages to be stages of the same thing.

Answering the costaging question, I frequently used the claim that the MOP role is played by files. I introduced MOPs in minimal terms, claiming only that they are subject to the Frege-constraint.[4] The Frege-constraint requires that any thinker in a Frege-case has at least two MOPs for the same object, so I rejected answers to the costaging question which implied that thinkers could be in Frege-cases with costaged records.

Answering the cofiling question involves giving an account of what it is for non-cotemporal stages to be stages of the same thing, so using the Frege-constraint in answering the cofiling question requires *cross-temporal examples of Frege-cases*, that is, cases where describing non-cotemporal attitude-stages as a binary relation between a coarse-grained proposition and a thinker leads to ascribing the thinker *prima facie* irrationality, and where we explain the thinker's mistake by claiming "she does not realise that $a = b$."[5]

But there are no cross-temporal examples of Frege-cases. Frege-cases require attitudes with temporal overlap. We allow that thinkers change their minds or forget things, so we do not explain the *prima facie* irrationality of believing P and $\neg P$ by claiming "she doesn't realise that $a = b$" unless the beliefs that P and $\neg P$ temporally overlap.[6] Similarly, if there is no temporal overlap in her beliefs that Fa and Ga, there is no *prima facie* irrationality and so no Frege-case if a thinker believes that Fa, and then believes that Ga, and is not immediately being able to infer that $\exists x(Fx\&Gx)$.[7]

[4] See p12.

[5] See p11.

[6] If a thinker changes her mind very frequently, or for the wrong reasons, she may be *prima facie* irrational. But this is not a Frege-case, because we explain her mistake in terms of her changing her mind rather than failing to realise that $a = b$.

[7] This is not to say that Frege-cases do not play out over time. Suppose T believes P, and learns that $P \rightarrow Q$, and but does not immediately drop her belief in P or acquire the ability to infer that Q. The circumstances that make this a Frege-case have played out over time, but the *prima facie*

Alone, this doesn't make it obvious that there are no cross-temporal examples of Frege-cases. Suppose between t_1 and t_2, T is in a Frege-case because she has beliefs with coarse-grained content P and $\neg P$. We might think that this is a cross-temporal example of a Frege-case, involving the t_1 stage of T's belief that P and the t_2 stage of her belief that $\neg P$. Nonetheless, this is not a genuine cross-temporal example. It is *prima facie* irrational for T to have a belief-stage at t_1 that P and a belief-stage at t_2 that $\neg P$ only if we suppose that at t_2, T retains the belief that P she had at t_1. If the beliefs that P and $\neg P$ did not temporally overlap, there would be no *prima facie* irrationality. The impression that there might be *prima facie* irrationality still relies on cotemporal stages of those attitudes, and so we do not have a genuine cross-temporal example of a Frege-case.

If generating the impression of *prima facie* irrationality in Frege-cases always relies on holding attitudes simultaneously, we might attempt to gain traction on the cofiling question by using counterfactual elements to develop a cross-temporal example of a Frege-case (or something very much like one). We might ask whether, if the thinker had a t_1 stage of an attitude at t_2, she would reason in a *prima facie* irrational manner.[8] However, this won't work either. It isn't enough to suppose that at t_2 the thinker has the same attitude towards the same coarse-grained content. We need to find out how the thinker would reason if she had this attitude. But how the thinker would reason at t_2 would depend on what t_2 MOP the t_1 belief was associated with, and this isn't captured just by the coarse-grained content and the type of attitude. To say anything more, we need to know what it is to use the same MOP at t_2 as was used at t_1, but this presupposes an answer to the cofiling question.

And we cannot ask whether a thinker would reason in a *prima facie* irrational way

irrationality comes from her simultaneously holding beliefs that P and $P \to Q$, and not being able immediately to infer that Q.

[8]Clearly, a thinker cannot in fact have a t_1 stage of anything at t_2. But we can imagine what would happen if a t_1 stage is duplicated at another time and retains its intrinsic properties.

were she to have a t_1 record-stage r_{t_1} at t_2. Whether or not the thinker's reasoning is *prima facie* irrational depends in part on the contents of her representations. But according to the mental files account, the content of r_{t_1} is determined by the file-stage which r_{t_1} is a member of.[9] So the content of r_{t_1}, if it were held at t_2, would depend on what file-stage r_{t_1} were a member of at t_2. And as yet, we have no way of working this out. We cannot just say that r_{t_1} would be a member of the same file it was a member of at t_1, because this also presupposes an answer to the cofiling question.

These difficulties mean that we cannot generate cross-temporal examples of Frege-cases,[10] and so we cannot use the Frege-constraint to investigate the diachronic individuation criteria on MOPs.[11]

7.2.3 Why allow for persisting files?

MOPs were given a minimal introduction in terms of Frege-cases. So if there are no cross-temporal examples of Frege-cases, we might worry that there is no motivation for positing persisting MOPs. And as files were introduced originally as an account of what MOPs are, we might worry that this means there is no motivation for positing persisting files. Nonetheless, there are reasons to posit persisting files.

Cross-temporal rationality judgements Whilst there are no cross-temporal Frege-cases, we do have some cross-temporal rationality judgements to take account of. Thales doesn't realise that *Hesperus is Phosphorus*. At t_1 he holds record-stages

[9] See also 6.4.1.

[10] At least, not prior to answering the cofiling question.

[11] This point is made explicitly by Evans (1985 [1981], p308). It is acknowledged elsewhere, e.g. Longworth (forthcoming) gives a criterion of MOP-individuation using beliefs indexed to a single time. However, acknowledging the issue rarely leads to discussion of diachronic individuation conditions for MOPs.

Similarly, there are no interpersonal Frege-cases. I have been assuming that MOPs are mental particulars and hence not sharable. But if we were to seek an account of some MOP-like phenomenon, sharable across individuals and perhaps playing an explanatory role in accounts of successful communication, Frege-cases would be no help in giving their individuation conditions.

${}^{Th}(\mathrm{r})_{t_1}$ and ${}^{Th}(\mathrm{u})_{t_1}$.

${}^{Th}(\mathrm{r})_{t_1}$ Hesperus is bright

${}^{Th}(\mathrm{u})_{t_1}$ $\forall x(\mathrm{x}$ is bright $\rightarrow x$ is large)

Suppose that he uses ${}^{Th}(\mathrm{r})_{t_1}$ and ${}^{Th}(\mathrm{u})_{t_1}$ in an inference to a record-stage he holds at t_2. In one scenario, he infers ${}^{Th}(\mathrm{v})_{t_2}$, in the other he infers ${}^{Th}(\mathrm{w})_{t_2}$.

${}^{Th}(\mathrm{v})_{t_2}$ Hesperus is bright

${}^{Th}(\mathrm{w})_{t_2}$ Phosphorus is bright

Although ${}^{Th}(\mathrm{v})_{t_2}$ and ${}^{Th}(\mathrm{w})_{t_2}$ have the same content, *prima facie* it is irrational for Thales to infer ${}^{Th}(\mathrm{w})_{t_2}$ but it is not *prima facie* irrational for Thales to infer ${}^{Th}(\mathrm{v})_{t_2}$.[12]

Given that MOPs are in our explanatory toolkit, the most natural description of the difference is that a thinker reasons irrationally when she infers a record-stage u_{t_2} from record-stages r_{t_1} and s_{t_1}, and u_{t_2} deploys different MOPs from those deployed in r_{t_1} and s_{t_1}. To give this description, we need to suppose MOPs persist. If files play the MOP role, then we should also suppose that files persist.

Building up information about an object The ability to treat information acquired at different times as being about the same object is vital. Without it, thinkers would have an extremely impoverished understanding of objects. Whether or not we accept file theory, we need some account of how thinkers build up increasingly complex understandings of objects over time.

Given that an important part of the synchronic role of files is that record-stages associated with the same file-stage are treated as about the same thing, it is natural to say that representations gathered at different times are part of a persisting file.

[12]The scenario where Thales infers ${}^{Th}(\mathrm{w})_{t_2}$ is not a Frege-case. If we describe his t_2 beliefs as relations to coarse-grained propositions, we generate no *prima facie* irrationality, even if he has t_2 stages of ${}^{Th}(\mathrm{r})$ and ${}^{Th}(\mathrm{u})$.

And the standard file picture is that files persist and play a role in building up information about objects. When a thinker encounters an object o as if for the first time, she opens a new file F on o, and adds to F all subsequent representations she takes to be about o.

Persisting representations A thinker's ability to build up an increasingly complex understanding of an object over time relies on the fact that if T believes P at t_1, then *ceteris paribus* T believes P at t_2. If we think just in terms of thinkers having stages of representations and attitudes, this seems unlikely good fortune. The simplest explanation of why a thinker who believes P at t_1 will generally believe P at t_2 is that attitudes and the records realising them persist. Hence, if a thinker has a belief-record that P at t_1, then *all things being equal* that record will still exist at t_2, and will still be a belief-record with content P. Without thinking that there are continued beliefs and persisting records, the similarities between a thinker's mental states from one instant to the next become extremely mysterious,[13] as does the awareness that our beliefs persist, and the awareness that we are thinking of the same thing again.

Reference-fixing gives another reason to posit persisting records. If we suppose that at least sometimes reference is determined by causal relations between records and external objects, we must allow for persisting records so that we can have records that are causally related to previously encountered objects.

Mental files accounts must suppose that having persisting records requires having persisting nodes or persisting tokens (see 7.2.1). A file is a bundle of records containing the same node (or tokens of the same type), so if we have persisting nodes (or persisting tokens) and persisting records containing these nodes (or tokens), then we have persisting files. So the file picture suggests that once we allow for persisting

[13]See also Evans (1985 [1981], p309).

records, we must allow for persisting files.[14]

7.2.4 Shifts in referential-value

In 2.2.2, I claimed that accounts of reference-fixing indicate that files can change their referential-value across time. This point can be made vivid by an example:[15]

> **Example XV:** [16] Sam holds his newborn child (child_A). He has decided to call his child 'Amy' and forms a belief he expresses (at t_1) with an utterance of (79).
>
> (79) Amy looks like Gollum.
>
> A few hours later, child_A is substituted with child_B. Sam raises child_B as his own, interacting with her daily and never encountering child_A. Ten years later (at t_2) Sam forms a belief he expresses with an utterance of (80):
>
> (80) Amy needs to finish her homework.

As files are used to track objects across time, the beliefs expressed with Sam's utterances of (79) and (80) should be recorded in the same file F.

In 2.2.2, I sketched accounts of reference-fixing favoured by mental file theorists, noting that file theorists generally suppose that the causal source of at least some records in the file (at least in part) determines what the file is about. At t_1, we can presume that child_A is the only causal source of information in F, and at t_2 child_B

[14] Perry (1980) approaches much the same position by a different route. He suggests we can use persisting files in an account of what it is to have a continued belief (see 2.2.2).

[15] We might hope for an account of the diachronic individuation conditions on files before claiming that persisting files can undergo referential-shift. However, I use the fact that files can undergo referential-shift to constrain answers to the cofiling question. My hope is that we can use relatively familiar ideas of records, files and MOPs persisting over time to demonstrate that an adequate account of their diachronic individuation conditions must take account of referential-shift.

[16] A crucial difference between this example and those discussed in 5.3.1 is that XV focuses on beliefs held at *different* times, rather than cotemporal belief-stages.

is the dominant causal source of information in F. Whilst file-theory is compatible with different ways of spelling out the suggestion that the referential-value of a file is determined by the causal source of the records in that file, accounts using the idea of the 'dominant causal source' of records will almost certainly give the result that F_{t_1} refers to child_A and F_{t_2} refers to child_B. Even if we suppose that confusion can never lead to full reference-switch and instead leads to reference-failure, reference to an amalgam, or results in an indeterminate referential-value, then F_{t_2} will have a different referential value to F_{t_1}.[17]

The alternatives to allowing that F changes referential-value are unattractive. One is that Sam records the beliefs expressed with the utterances of (79) and (80) in F, and F retains its original referential-value throughout. This suggests that Sam does not have beliefs about child_B, even though he has interacted with her daily for ten years. The other alternative is that the utterances of (79) and (80) express beliefs recorded in different mental files. This would allow us to acknowledge that the beliefs concern different children without supposing that files shift referential-value. However, Sam does not suspect his child has been switched, so the cost of taking this position is losing important parts of the file picture. We no longer can think of files capturing the thinker's understanding of the sameness and difference of objects, and must abandon the idea that records treated as about the same thing are entered in the same file.[18]

I have previously suggested that if T has costaged records r_t and s_t containing component-stages $\mathrm{c}_{\mathrm{r}_t}$ and $\mathrm{c}_{\mathrm{s}_t}$ respectively, then we might say that (in some suitably idealised way) T knows that: if $\mathrm{c}_{\mathrm{r}_t}$ and $\mathrm{c}_{\mathrm{s}_t}$ refer, they externally-corefer.[19] But if non-cotemporal record-stages r_{t_1} and s_{t_2} are members of the same file and contain

[17] Stepping aside from theoretical commitments, the natural response to XV is that there is full reference-switch: at t_1 Sam expresses a belief about child_A, and at t_2 he expresses a belief about child_B.

[18] See 2.1.3.

[19] See 5.4.

$c_{r_{t1}}$ and $c_{s_{t2}}$ respectively, we should not attribute T a priori knowledge that: if $c_{r_{t1}}$ and $c_{s_{t2}}$ refer, they externally-corefer. There is no guarantee that $c_{r_{t1}}$ and $c_{s_{t2}}$ share referential-value, and hence no guarantee that if they refer, they externally-corefer.

7.3 Cofiling options

With these considerations in place, the next task is identifying necessary and sufficient conditions on cofiling. For ease of exposition, unless otherwise stated I make the simplifying assumption that all records are single-reference.

7.3.1 Shared records

Descriptions of files as persisting generally start with the idea that files are bundles of records. As the file evolves, more records are added and old records are lost, but the bundle survives these changes.

This might suggest that we account for cofiling just in terms of sharing records.

CoFiling-1 File-stages F_{t_1} and G_{t_2} are cofiled iff there is a record r such that F_{t_1} contains r and G_{t_2} contains r

However, intuitively files may persist through a complete change in their records. But this is ruled out by CoFiling-1. To allow for this, we can use the ancestral of record sharing.

CoFiling-2 File-stages F_{t_1} and G_{t_2} are cofiled iff F_{t_1} S* G_{t_2}

x S* y iff x S y or there is a chain of file-stages F_1, $F_2 \ldots F_k$ (k >0) such that

x S F_1, F_1 S $F_2 \ldots F_k$ S y

x S y iff x shares a record with y

But CoFiling-2 is unsatisfactory. First, if we drop the simplifying assumption, we can at best use record-sharing as a necessary condition on co-filing. But even as a

necessary condition on co-filing, it is unilluminating. It relies on the idea of records persisting through time. But we have no better grip on what it is for a record to persist through time than we have on what it is for a file to persist through time.[20]

7.3.2 Causal connectedness

Persisting files are supposed to be used by a thinker to build an increasingly complex picture of an object. It is difficult to see how this would work if the t_1 file-stage of file F does not causally influence how F is at t_2. This suggests CoFiling-3.

CoFiling-3 File-stages F_{t_1} and G_{t_2} are cofiled iff F_{t_1} and G_{t_2} are suitably causally connected

The difficulty with CoFiling-3 is spelling out what the "suitable causal connection" is. If I believe that Felix and Oscar are qualitatively identical, then my t_1 *Felix*-file will causally influence the t_2 stage of both my *Oscar*-file and t_2 *Felix*-file. It seems unlikely that we will be able to find any better statement of this "suitable causal connection" than giving CoFiling-4.

CoFiling-4 File-stages F_{t_1} and G_{t_2} are cofiled iff F_{t_1} and G_{t_2} are causally connected in the way appropriate for cofiled file-stages

Assuming some causal connection is required for cofiling, CoFiling-4 may be correct. But it is unilluminating.

7.3.3 De jure coreference

We might hope to find an answer to the cofiling question in the accounts of de jure coreference discussed in chapters 3 and 4. Constraints on answers to the cofiling question are different from constraints on answers to the costaging question,[21] so a

[20] See 7.2.1.

[21] For example, there is no requirement that cofiled file-stages share a referential-value.

relation that failed to provide an answer to the costaging question may still be able to answer the cofiling question. Moreover, accounts of de jure coreference are rooted in linguistic coreference, and linguistic coreference relations tend to hold between non-cotemporal occurrences.

AP- and IK-coreference Making the simplifying assumption, one option is CoFiling-5, which uses characterisations of de jure coreference in terms of knowledge of conditional coreference.[22]

CoFiling-5 File-stages F_{t_1} and G_{t_2} are cofiled iff F_{t_1} contains a record-stage r_{t_1}, and G_{t_2} contains a record-stage s_{t_2}, and r_{t_1} and s_{t_2} [AP-/IK-corefer] or [AP-/IK-pseudo-corefer]

A first worry is that CoFiling-5 is victim to a version of the first circularity objection raised against mental file theory in 6.4.1: facts about a thinker's attitudes are partly constituted by which records are costaged, so we cannot use attitudes in an answer to the costaging question. However, this concern does not obviously carry through to co-*filing*. If a single attitude-stage is enough for knowledge that if r_{t_1} and s_{t_2} both refer then they corefer, thinkers can meet the knowledge requirement for AP- or IK-coreference just in virtue of facts about costaging. So CoFiling-5 may presuppose an answer to the costaging question, but does not presuppose an answer to the cofiling question.

But CoFiling-5 is still problematic, even if we allow for the highly idealised attributions of knowledge required to suppose that AP- or IK-coreference is widespread.[23] For two record-stages to AP- or IK-corefer, a thinker must be able to know that if the relata refer, then they externally corefer. This requirement of conditional coreference means that AP- and IK-coreference cannot serve as answers to the cofiling question if cases of confusion like example XV result in full reference-swiching or

[22] See chapter 4.
[23] See 4.5.2.

reference to an amalgam. If, as is extremely plausible, confusion results in full reference-switching or reference to an amalgam, then cofiled file-stages F_{t_1} and G_{t_2} may contain record-stages r_{t_1} and s_{t_2}, which refer but do not externally-corefer, and so cannot AP- or IK-corefer.

RR-coreference My objection to using RR-coreference in answering the costaging question was that the core account requires that thinkers treat *costaged* records as about the same thing, and if we use RR-coreference to answer the costaging question, there could be costaged records that are not treated as about the same thing.[24] But RR-coreference might be more helpful now we are discussing *non-cotemporal* record-stages. Making the simplifying assumption:

CoFiling-6 File-stages F_{t_1} and G_{t_2} are cofiled iff F_{t_1} contains a record-stage r_{t_1}, and G_{t_2} contains a record-stage s_{t_2}, and r_{t_1} and s_{t_2} either RR-corefer or putatively RR-corefer

In 3.6.3, I observed that Fine did not give a complete account of the back-up semantics of SR- or RR-coreference, but one can be reconstructed. If my reconstruction is satisfactory, then Fine's characterisation of linguistic de jure coreference might be thought to handle examples like XV better than Pinillos's (2011) or my own. In example XV, it is tempting to say that the occurrence of 'Amy' in the utterance of (79) refers to child_A, and the occurrence of 'Amy' in the utterance of (80) refers to child_B. Assuming this is correct, these occurrences cannot AP- or IK-corefer. In contrast, there is nothing to rule out allowing that the occurrences of 'Amy' are putatively SR-coreferential, whilst referring to different objects. However, I have only given a definition for putative SR-coreference *within* a discourse,[25] so further work is needed to fully develop this argument.

[24] See 3.8. In this section, I make the tendentious assumption that there is some account of manifest consequence adequate for allowing us to claim that some records are RR-coreferential (see 3.9).

[25] See 3.6.3.

Nonetheless, CoFiling-6 is unsatisfactory, because there is no account of what it is for r_{t_1} and s_{t_2} to be putatively RR-coreferential adequate for use in answering the cofiling question.

Two relata are putatively RR-coreferential when a thinker must accept for the purposes of (at least part of) her reasoning that they are RR-coreferential to count as rational.[26] Clearly, the kind of 'acceptance that they are RR-coreferential' involved is highly idealised, as thinkers rarely have non-dispositional attitudes about their records. I suggested that this idealised acceptance might be realised in how a thinker treats her records when reasoning with them both. However, whilst this may be satisfactory when thinking about cotemporal record-stages, it is not satisfactory for non-cotemporal record-stages. By t_2, r_{t_1} no longer exists. So we cannot suppose that a thinker treats r_{t_1} and s_{t_2} as RR-coreferential in how she reasons with both r_{t_1} and s_{t_2}, because the thinker doesn't have both r_{t_1} and s_{t_2} to reason with at once.

One option is to say that taking r_{t_1} and s_{t_2} as RR-corefrential is a matter of some minimal awareness that one is thinking about the same thing again, or that one has not changed what one is thinking about. But this requires some kind of metarepresentational ability. There is a presumption in favour of the view that files will be useful in accounting not just for mature conscious human thought, but also subpersonal thought and animal and infant thought (see 2.2.2). CoStaging-13* supports this presumption by giving necessary and sufficient conditions on costaging that are met in animal, infant and subpersonal thought.[27] However, if we interpret putative RR-coreference as requiring some kind of metarepresentational ability, we cannot use CoFiling-6 to answer the cofiling question for mental files belonging to animals or infants, or for subpersonal mental files, because we cannot rely on metarepresentational abilities in accounting for cofiling across all these kinds of thought.

[26] See 3.7.2.
[27] See 5.4.

So we have severe difficulties developing an account of putative RR-coreference adequate for use in answering the cofiling question. Therefore, the prospects look dim for using RR-coreference in an account of the necessary and sufficient conditions on cofiling.

These same difficulties arise for any account of cofiling which require taking an attitude towards non-simultaneous record-stages or file-stages, including CoFiling-5. There is no way of giving an idealised account of what it is to take an attitude towards two non-simultaneous record-stages or file-stages which can give a satisfactory account of cofiling in animal, infant and subpersonal thought.

7.3.4 Trading on coreference

We might look for an answer to the cofiling question in the same place we found an answer to the costaging question: in a thinker's ability to trade on coreference. Still making the simplifying assumption, this suggests CoFiling-7.

CoFiling-7 T's file-stages F_{t_1} and G_{t_2} are cofiled iff F_{t_1} contains a record-stage r_{t_1}, and G_{t_2} contains a record-stage s_{t_2}, and T can trade on the coreference of r_{t_1} and s_{t_2}.

We might wonder how a thinker can trade on the coreference of non-cotemporal record-stages. For T to be able to reason with a record-stage, that record-stage must exist. So at t_2, T cannot trade on the coreference of s_{t_2} with r_{t_1}, because r_{t_1} doesn't exist at t_2.

For CoFiling-7 to be successful, we need a "highly counterfactual" account of what it is for a thinker to be able to trade on the coreference of non-cotemporal records.[28]

[28]Fodor (1994) considers similar issues, and gives his own "highly counterfactual" (p108) response.

> T can trade on the coreference of r_{t_1} and s_{t_2} iff were she to hold r_{t_1} at t_2, she could trade on the coreference of r_{t_1} and s_{t_2}, and were she to hold s_{t_2} at t_1, she could trade on the coreference of r_{t_1} and s_{t_2}.

With this understanding of trading on coreference between non-cotemporal records, CoFiling-7 is at least viable as an answer to the costaging question.

In 7.2.2, I argued that I cannot use a similar strategy to generate cross-temporal Frege-cases, because working out whether a thinker's reasoning is *prima facie* irrational requires working out what file-stage r_{t_1} would be in at t_2. There is no corresponding problem with this "highly counterfactual" understanding of trading on coreference, just so long as we make the tendentious assumption that record-stages have the non-semantic properties which enable trading on coreference intrinsically.[29] If these properties are intrinsic to the record-stage, then r_{t_1} would still have them if moved to t_2, so we can investigate which record-stages the thinker can trade on the coreference of r_{t_1} with, without working out which t_2 file-stage r_{t_1} belongs to.

The principal appeal of CoFiling-7 is that it treats cofiling as an extension of the better understood relation of costaging. The intuitively plausible suggestion is that if non-cotemporal record-stages meet the conditions on costaging, then they count as members of a persisting mental file.

CoFiling-7 might also explain why we expect cofiling to result in reflective thinkers having the impression that they are continuing to have beliefs about the same thing. If a thinker can be aware of some past record-stage, then she can implicitly imagine how it would interact with current record-stages. If it seems to the thinker that it would interact as if it were costaged with the current record-stages, then she will suppose that it is about the same thing as her current record-stages.

And trading on coreference can be used to answer the cofiling question even if

[29] I argue in 6.4.1 that CoStaging-13* is viable as an answer to the costaging question if we assume that facts about trading on coreference are prior to semantic facts.

we abandon the simplifying assumption.

Suppose that component-stages $c_{r_{t1}}$ and $c_{s_{t2}}$ are components of r_{t_1} and s_{t_2} respectively.

> T can trade on the coreference of $c_{r_{t1}}$ and $c_{s_{t2}}$ iff were T to hold r_{t_1} at t_2, she could trade on the coreference of $c_{r_{t1}}$ and $c_{s_{t2}}$, and were T to hold s_{t_2} at t_1, she could trade on the coreference of $c_{r_{t1}}$ and $c_{s_{t2}}$.

CoFiling-7* T's file-stages F_{t_1} and G_{t_2} are cofiled iff a record r_{t_1} is a member of F_{t_1} in virtue of containing $c_{r_{t1}}$ and a record s_{t_2} is a member of G_{t_2} in virtue of containing $c_{s_{t2}}$, and T can trade on the coreference of $c_{r_{t1}}$ and $c_{s_{t2}}$.

For ease of exposition, return to making the simplifying assumption. One might object that CoFiling-7 cannot account for files changing referential-value over time, because I have argued that thinkers can only trade on the coreference of record-stages sharing referential-value (see 5.3.3). However this argument turned on the fact that file-theorists are committed to claiming that costaged record-stages share referential-value. As file-theorists are not committed to non-cotemporal stages of the same record sharing a referent, the same argument cannot be made that if a thinker can trade on the coreference of non-cotemporal record-stages, then they share referential-value.

Moreover, there is no reason to think that if r_{t_1} existed at t_2 it would have the same referential-value it has at t_1. The referential-value of a record-stage is determined by which file-stage it is associated with. Suppose that at t_1, r_{t_1} refers to a in virtue of being associated with F_{t_1}. If file F changes its referential-value such that F_{t_2} refers to b, then if r_{t_1} were to exist at t_2, it would be associated with F_{t_2} and refer to b.

Another objection to this strategy is more compelling: it requires making very heavy-duty assumptions about records and what enables trading on coreference,

even though we know very little about what enables trading on coreference.

One assumption we need for CoFiling-7 to work is that the node picture of files is false. On the node picture, a record partially consists in a node, and records which share a node can have their coreference traded upon. Assuming the node picture, ask what would happen if r_{t_1} existed at t_2. One option is that we move the whole of r_{t_1} to t_2, including the node. But then the thinker would not be able to trade on the coreference of r_{t_1} with any of the records at t_2, because r_{t_1} would contain the node from t_1 and so would not share a node with any of the records existing at t_2. Alternatively, we could imagine what would happen if all of r_{t_1} except the node existed at t_2. But then r_{t_1} would be missing the component we are interested in. We might try to complete r_{t_1} with a node from t_2, but doing this would require already knowing what t_2 file-stage r_{t_1} would be a member of, presupposing just what we are trying to find out.

In contrast, adopting the token picture, CoFiling-7 seems viable. On the token picture, we can imagine that we are moving all of r_{t_1} to t_2, and still sensibly investigate which which t_2 record-stages can have their coreference traded upon with r_{t_1}.

However, even adopting the token picture, we need to add at least one further assumption for CoFiling-7 to work. To see this, continue making the simplifying assumption, and adopt the record-stage's colour as a placeholder for the intrinsic property of record-stages that enables their coreference to be traded upon. So T can trade on the coreference of two record-stages iff they are the same colour.[30] Suppose that non-cotemporal record-stages r_{t_1} and s_{t_2} are members of the same file F. r_{t_1} is the same colour as other record-stages in F_{t_1} at t_1, as s_{t_2} is the same colour as other record-stages in F_{t_2}. But this is compatible with the file changing colour between t_1 and t_2. Suppose at t_1, record-stages in F_{t_1} are green, but record-stages in F_{t_2} are

[30] Assume it's never indeterminate whether two records are the same colour.

yellow. If r_{t_1} existed at t_2, the thinker would not be able to trade on the coreference of r_{t_1} and s_{t_2}, because although *ex hypothesi* r_{t_1} and s_{t_2} are members of the same file, they are different colours.

So for CoFiling-7 we need to assume not only that record-stages intrinsically have the properties which enable their coreference to be traded upon, but also that these properties remain constant throughout the life of the file. If we allow for these properties to change (in my model, for the colour to change), CoFiling-7 will deliver the wrong results about co-filing.

Another consequence of accepting CoFiling-7 is that it allows for temporally gappy mental files. Suppose at t_1 and t_3, T has red record-stages, but T has no red record-stages at t_2. T will be able to trade on the coreference of the red t_1 record-stages with the red t_3 record-stages, and so those record-stages will count as members of the same file, even though that file did not exist at t_2. If considerations of causal connectedness mean that gappy files are unacceptable, we must assume that thinkers never reuse whatever intrinsic properties it is that enable trading on coreference (in this case, colours), and so this scenario never arises. But again, this is making a substantive assumption for which we have little independent evidence.

If the assumptions required for CoFiling-7 hold, then we can use CoFiling-7 to fill in details of our theory of files. For example, CoFiling-7 rules out models of linking and merging files where files retain their identity through merges and links. Suppose a thinker has two file-stages at t_1, B_{t_1} containing blue record-stages and R_{t_1} containing red record-stages. The thinker makes an identity judgement, and eventually by t_2 files B and R have been merged. The t_2 stage of this merged file is P_{t_2}.

Suppose that the original file-stages maintain their identity throughout this merging process, so P_{t_2} is a temporal part of both B_{t_1} and R_{t_1}. Then, by CoFiling-

7, the thinker should be able to trade on the coreference of record-stages from B_{t_1} and R_{t_1} with record-stages in P_{t_2}. But this means record-stages in B_{t_1} and P_{t_2} must be the same colour, as must record-stages R_{t_1} and P_{t_2}. But then record-stages in B_{t_1} and R_{t_1} would also be the same colour, and the thinker would be able to trade on the coreference of record-stages in B_{t_1} and R_{t_1}, and (by CoStaging-13*) B_{t_1} and R_{t_1} would not be distinct files.

This shows that P_{t_2} cannot be co-filed with both B_{t_1} and R_{t_1}, so by CoFiling-7, B_{t_1} and R_{t_1} cannot both survive the merging process.

However, some questions are left unanswered by adopting CoFiling-7 simply because we have a limited understanding of the properties of records that make it possible to trade on their coreference. Suppose we have files used to refer to times. We might wish to use the answer to the cofiling question to settle the question whether Mary's record-stage ${}^{M}(\mathrm{r})_{21}$ from 21^{st} December 2013 uses the same MOP as ${}^{M}(\mathrm{s})_{22}$ or ${}^{M}(\mathrm{u})_{22}$, which are both Mary's record-stages from twenty-four hours later.

${}^{M}(\mathrm{r})_{21}$ Today it is raining

${}^{M}(\mathrm{s})_{22}$ Yesterday it was raining

${}^{M}(\mathrm{u})_{22}$ Today it is raining

Here's one plausible model: sameness of colour is used to track which records are *today*-records. So each day, Mary's *today*-records are always red-coloured. Each morning, Mary's old *today*-records change colour, and she starts forming new red records about the new day.[31] Here's an alternative: samness of colour is used to track which day the record was formed on. So each day, Mary uses a new colour for her *today*-records. Those records retain that colour on subsequent days. So we can suppose that on 21^{st} December, Mary forms taupe *today*-records, and those

[31]This is the kind of picture suggested by the idea that thinkers have a single 'buffer'-file for collecting information about *here*, wherever *here* happens to be (Recanati 1997, pp122-123).

records will still be taupe on 22^{nd} December. On 22^{nd} December, T forms cerise *today*-records.

On the first model, ${}^{M}(\mathrm{r})_{21}$ and ${}^{M}(\mathrm{u})_{22}$ are both red. By CoFiling-7, ${}^{M}(\mathrm{r})_{21}$ and ${}^{M}(\mathrm{u})_{22}$ are members of the same file, and so assuming the MOP role is played by files it is ${}^{M}(\mathrm{r})_{21}$ and ${}^{M}(\mathrm{u})_{22}$ that share MOP.

On the second model, ${}^{M}(\mathrm{r})_{21}$ and ${}^{M}(\mathrm{s})_{22}$ are both taupe, and ${}^{M}(\mathrm{u})_{22}$ is cerise. The thinker trades on the coreference of ${}^{M}(\mathrm{r})_{21}$ and ${}^{M}(\mathrm{s})_{22}$, and so by CoFiling-7, it is ${}^{M}(\mathrm{r})_{21}$ and ${}^{M}(\mathrm{s})_{22}$ that are members of the same file, and so share MOP.

The difficulty is that at the moment, our understanding of records and what enables trading on coreference is so limited that we have no way of knowing which of these models best fits the facts, and little understanding of what would help us work out which model best fits the facts. More generally, this limited understanding means we can't judge whether the assumptions needed for CoFiling-7 to work are reasonable.

7.3.5 Identifying characteristics

CoFiling-7 gives an answer to the co-filing question, as long as we accept some heavy-duty assumptions about records. However, if we are cautious about accepting these assumptions, then we still don't have an adequate answer to the cofiling question. With no alternative accounts of the necessary and sufficient conditions on cofiling available, one might conclude that it is not possible to to give an adequate account of the diachronic individuation conditions on files.

How we respond to this conclusion depends on what kind of status we suppose an answer to the cofiling question would have, if it were found.[32] If we think that attempts to answer the costaging and cofiling question are misguided, then we would be untroubled by the failure to answer a question we should not have been asking in

[32]For details of these options, see 6.4.3.

the first place. If we expect the answer to have the status of a definition, then our failure to find an answer to the cofiling question should be troubling. We have no grasp on cofiling beyond giving a definition of cofiling, and we haven't been able to find a definition for cofiling. So talk of files as persisting appears to be meaningless.

If we think an answer to the cofiling question would be a necessary a priori truth, then again we should think our failure to find an answer calls into question the project of ascribing persisting files. But if we think an answer would be a necessary a posteriori truth, then we might think that the failure to answer the cofiling question is a symptom of using inadequate methodology to approach the question.

But suppose that an answer to the cofiling question just provides identifying characteristics for cofiling (i.e. we think of answers to the cofiling question given so far as analogous to claims like *something is gold iff it's yellow and dense*). Then we have an alternative way of answering the cofiling question: acknowledging that we cannot give necessary and sufficient conditions on cofiling, and instead giving ways to identify cofiling.

One route for identifying cofiling would be to start from an investigation of costaging, and identify cofiling as a matter of the persistence of whatever it is underlying costaging's identifying characteristics. So if we find out r_t and s_t are co-staged iff r_t and s_t are the same colour, then we suppose that co-filing is a matter of having non-cotemporal record-stages that are the same colour.

There are two difficulties with this. One is that it is not always straightforward what counts as the same thing. For example, we might think that whether the different isotopes of water count as the same kind depends on the purposes at hand. We need some understanding of cofiling's identifying characterics so that when all the empirical evidence is in, we can work out what counts as the same thing for the purposes at hand.

A second difficulty is that the suggestion succumbs to a variation on Millikan's (2000, pp133-135) objection to the suggestion that repeating a mental vehicle is sufficient for mental representations being under the same MOP. Her argument is that it is not enough to say that the vehicle is repeated. The thinker has to be able to use that repetition in reasoning as if the representations are under the same MOP.[33] It is not enough to show that non-cotemporal record-stages are the same colour, we have to also show that the thinker uses this persistence appropriately (after all, files might switch colour over time). This gives a further reason for needing an account of co-filing's identifying characteristics to check that the thinker is using the file-stages in a way appropriate to claiming that they are co-filed.

It has not been easy to give an account of the necessary and sufficient conditions on cofiling. However, if we are only looking for identifying characteristics, then there is more flexibility than I've been allowing — we can just give characteristics that we expect to be generally associated with cofiling. If it turns out that the persistence of whatever it is underlying costaging's identifying characteristics generally results in cofiling's identifying characteristics being displayed, then we have an excellent case that we have correctly identified cofiling. And if the persistence of the thing underlying costaging's identifying characteristics does not generally result in cofiling's identifying characteristics being displayed, then we have an independent route for identifying cofiling.

I suggest that there are several characteristics to use in identifying cofiling. The first is associated with the role file-stages have in building up information about an object.[34] Cofiled file-stages generally contain representations causally-originating in a single object. Thinkers are sometimes confused, so this characteristic is not universally displayed. But assuming that confusion is the exception rather than

[33] See 2.3.2.
[34] See 7.2.3.

the norm, we can allow that cofiled file-stages will generally contain representations causally-originating in a single object.

The second is associated with connections between cofiling, persisting representations and continued belief.[35] In reflective thinkers with sufficient metarepresentational abilities, non-cotemporal record-stages with the same content existing in cofiled file-stages will generally result in thinkers being aware that they are continuing to believe something, and that they have not changed their minds.

A third is that we should expect some causal connections between co-filed file-stages, so if F_{t_1} and G_{t_2} are co-filed, there is some causal connection between how F_{t_1} is and how G_{t_2} is. Files build up a body of information about an object. It is difficult to see how causally disconnected file-stages could play this role.

There are reasons to think that these characteristics can't provide an account of the necessary and sufficient conditions on cofiling. But if we accept that an answer to the cofiling question just gives cofiling's identifying characteristics, then we can suppose that these characteristics will be helpful in identifying cofiling.[36]

7.4 Final remarks

I have not given a definitive answer to the cofiling question. Rather, I have suggested two routes for answering the cofiling question. One relies on substantive assumptions about records, the other does not give necessary and sufficient conditions on cofiling. Neither is compatible with all plausible variations on mental file theory. And, so on some variations (for example, a node-picture of files, expecting a necessary a priori answer to the cofiling question) there is no satisfactory answer to the cofiling question, and no satisfactory account of what it is for a file to persist.

[35] See 7.2.3.

[36] A further complication would be if we thought these different characteristics corresponded to different purposes in identifying persisting file-stages, such that we may identify one group of file-stages as a persisting file for some purposes, but not for others.

One might think that difficulties answering the cofiling question are not a major concern, rather they are an idiosyncratic consequence of my breaking down the account of mental files into a costaging and cofiling component, and beginning my discussion with relationships between record-stages rather than records. To this, I respond that our understanding of files can be no better than our understanding of certain mental states and processes. If we suppose that the synchronic and diachronic characteristics of these states and processes are different, as file-theorists do, then we need to investigate these synchronic and diachronic characteristics independently. And hence we need to separate the costaging from the cofiling question.

Moreover, because mental files play the MOP role, and MOPs are introduced to respond to a problem in *synchronic* reasoning, we have a reasonable grasp on what it is to have *at a single instant* two different files. Conflating the synchronic and diachronic individuation conditions on mental files, as most file-theorists do, disguises that although it is normal to talk of persisting mental files, we have a relatively poor grasp on what it is to have a persisting file.

Another concern with my approach is that I am asking for too much. I am effectively asking for a functional definition of files. However, it is commonplace to suggest that mental states like 'being in pain' are functionally defined, even though there is little anxiety over failure to give an adequate functional definition of pain.

Moreover, I am asking for both synchronic and diachronic individuation conditions on files. But we often find it easier to give synchronic individuation conditions than diachronic individuation conditions. We may struggle to give diachronic individuation conditions on words or dances (Hawthorne and Lepore 2011, p482), but we still talk about words and dances.

However, there is a considerable difference between the cases. Mental states like *being in pain* are central to folk psychology. *Words* and *dances* are part of our folk ontology. In contrast, *mental files* are philosophers' posits. We might think that we

are stuck with 'pain'-talk and 'word'-talk, but 'files'-talk is optional. We should only use it if we can come up with an adequate account of what it is we are discussing.

Mutatis mutandis, the same things can be said for any theory positing persisting MOPs. MOPs may be better established, but like files they are a philosopher's posit rather than a core part of folk-psychology. And MOPs are introduced in response to a problem in synchronic reasoning. This means we have no clear understanding of what it is for a MOP to persist, or even of what considerations would inform an account of what it is for a MOP to persist. The difficulties we have answering the cofiling question should cause us look harder at the assumption that we can safely talk of MOPs persisting.

Chapter 8

A Cautionary Note

8.1 Introduction

I have used the bulk of this thesis to clarify the claim that thinkers use mental files. I have filled in the details of this claim as far as possible, and have set out remaining decision points for a theory of files. And I have defended the theory from some objections.

However, my remaining comments will be largely cautionary. I suggest we should separate the account of the mental representation of objects developed in this thesis from a thicker account of mental files, which suggests that the file metaphor should be taken seriously. I will suggest that the thicker account does not add to our understanding of mental representation, and I will argue that the terminology of 'files' may be misleading.

8.2 The thick and thin accounts

8.2.1 Beyond material objects

Before I argue that we should distinguish the thick and thin accounts of mental files, it is helpful to add a final detail to my account. Mental files are largely discussed by those concerned with singular thought. I have followed standard practice by focussing on examples involving reference to material objects. However, this gives a limited perspective, encouraging the view that there are special problems for the representation of material objects. But many of the phenomena discussed in this thesis are replicated beyond thought about material objects.

Suppose Ben is a competent user of the terms 'sofa' and 'settee', and can be accurately ascribed beliefs about sofas/settees. Nonetheless, Ben (falsely) supposes that 'sofa' and 'settee' are not cointensional. Ben sincerely assents to (81) and (82).

(81) Most sofas don't have stuffed cushions.

(82) Most settees have stuffed cushions.

It is natural to ascribe Ben beliefs that *most sofas don't have stuffed cushions* and *most settees have stuffed cushions* (Burge 2007 [1979]). But there are relevant similarities to Frege-cases: if we think of belief as a binary relation between a thinker and a coarse-grained proposition we would ascribe Ben *prima facie* irrational beliefs, and we explain Ben's mistake by saying "he doesn't realise that sofas are settees". As we explain Frege-cases by suggesting that thinkers have more than one mode of presentation for the same object, it is plausible to suggest we account for Ben's situation by ascribing him two modes of presentation for *sofas*.

In fact, as I discussed in 3.9, on many plausible semantic theories thinkers can have different takes on kinds like *cat* (Loar 1985) or *contract* (Burge 2007 [1979]), for distances like *one foot*, or times like *today* (Perry 1993 [1980]). Subsequently, thinkers might have different takes on properties like *being a cat* and *happening today*, or relations like *being in a contract with* or *being a foot from*.

Wherever a thinker might have different takes on a kind, property or relation, it is possible to get cases where treating a thinker T's attitudes as binary relations between T and coarse-grained contents results in attributing T *prima facie* irrational conflicting attitudes, or *prima facie* irrationality in her theoretical or practical reasoning. In these cases, we naturally describe T's mistake in terms like "she doesn't realise that F is the same as G". Call these cases, which don't qualify as full Frege-cases,[1] "quasi-Frege-cases". If we explain Frege-cases by saying the thinker has two MOPs on a single object, we might think we should explain quasi-Frege-cases as resulting from the thinker having two modes of presentation for the same property or relation. Conversely, if a thinker has cotemporal thoughts concerning F involving just one mode of presentation for F, the thinker should treat those thoughts as concerning the same thing.

[1] For the definition of Frege-cases, see p11.

In 5.2.3, I pointed out the regress argument for trading on coreference could be reused to argue that (for fear of regress), we must accept that thinkers can trade on co-value. Given that file theory suggests strong connections between MOPs and the ability to trade on coreference,[2] it is natural to associate the broader phenomena of modes of presentation with trading on co-value. The arguments connecting MOPs with the ability to trade on coreference used the assumption that files play the MOP role, so cannot be replicated without begging any questions. But if a thinker can trade on the co-value of non-referential components c_1 and c_2, then c_1 and c_2 meet a plausible necessary condition on being associated with the same mode of presentation: if a thinker can trade on the co-value of c_1 and c_2, we can expect that the thinker will not treat c_1 and c_2 as concerning different things, and so would not find herself in a quasi-Frege-case as a result of those components.

I have presented a file theory claiming that thinkers can trade on coreference in virtue of records sharing referential nodes (or containing referential tokens of the same type), and that if two records share a referential node (or contain referential tokens of the same type), then those records are associated with the same MOP. If we accept this, parsimony suggests we take the next natural step and say that: thinkers can trade on co-value in virtue of records sharing nodes (or containing tokens of the same type), and if two records share a node (or contain tokens of the same type) then they are associated with the same mode of presentation. If we don't take this step, we need to posit a new range of mental structures to do the same work as referential nodes (or referential tokens of the same type).

If we accept that there are nodes (or tokens) representing kinds, properties and relations, it is a further question whether to count records containing the same non-referential node (or non-referential tokens of the same type) as records in a single mental file. We might opt to restrict the term 'mental file' to bundles of records

[2]See 5.3.3 and 5.4.

containing shared referential nodes (or referential tokens of the same type), or even to restrict the term to bundles of records containing singularly referring nodes (or singularly referring tokens of the same type).

One non-arbitrary reason to restrict the term 'mental files' to a particular sort of node (or sort of token) is that we might associate mental files with a particular kind of reference-fixing, and suppose that the content of nodes (or tokens) representing kinds, properties and relations is determined differently to the content of nodes (or tokens) representing material objects.

Some of these differences are relatively subtle. For example, we might think that we have to handle the idea of representations 'causally originating' in something differently. Or we might think that outside singular thought about material objects, a thinker must have a more accurate understanding of something to count as thinking about it.

Another difference is potentially more significant. We are less concerned by the suggestion that thinkers sometimes equivocate and change the semantic-value of their thoughts when thinking about e.g. *being innocent* than we are by the suggestion that thinkers sometimes equivocate and change referential-value when thinking singularly about e.g. *Pavarotti*.[3] Assuming the node picture of files, we might take this as evidence that the node for *being innocent* doesn't deliver the same content to each cotemporal record-stage using that node, in the way we suppose the node for *Pavarotti* delivers the same referent to each cotemporal record-stage containing that node. Or we could say that 'acceptable' equivocation always involves complex concepts built up of several different nodes. 'Acceptable' equivocation occurs when a thinker uses a slightly different combination of nodes in cotemporal records and mistakenly treats those different combinations as the same. But as with referential nodes, a single node will deliver the same content to cotemporal

[3]See 5.3.2.

record-stages containing that node (or tokens of that type).

We might choose to extend the terminology of 'mental files' to a broader range of mental representations, rewriting the core account[4] to acknowledge relevant differences. Or we may restrict the terminology of 'mental files' to the original cases, and so preserve the core account. Whichever option we take, an important point remains: it appears that at some level of description, the mental structures realising singular thought about material objects are the same as those realising other kinds of thought. So *all* records can be thought of as a connected group of nodes or a string of tokens of several different types.[5] Some of these nodes or tokens may be referential, but others ascribe or represent relations.[6]

8.2.2 Distinguishing the thick and thin accounts

The suggestion that thinkers use mental files appears to make two distinctive contributions: (i) the claim that the 'file' role outlined in the core account of mental files is instantiated;[7] and (ii) the claim that, furthermore, there is some helpful analogy to be drawn between non-mental file systems and the mental representation of objects.

When evaluating the claim that thinkers use mental files, it is important to distinguish a thin account of files which endorses only (i), from a thick one which takes the file metaphor seriously and endorses both (i) and (ii). On the thin account, we need not take the file metaphor seriously, and the use of 'file'-talk is incidental. We can see this by stripping the language of 'files' from the core account.

Introduce a new term: *RMOP*. An RMOP is whatever mental representation

[4]See 2.1.3.

[5]So described, the token picture appears to be little more than a reinvention of the language of thought hypothesis Fodor (1975).

[6]One consequence of this conclusion is that attempts to define singular thought in terms of a special kind of mental representation (e.g. Jeshion 2010, see 1.1.2), rather than a special kind of content or form of reference-fixing, seem unsustainable. If we want to suppose there is a difference between thought involving mental files and thought involving other kinds of mental representation, this just is a difference in how the mental representation has its reference determined.

[7]See 2.1.3.

plays the MOP role. We can rewrite the slogans of the core account in terms of RMOPs.

I The MOP role is played by RMOPs

II If record-stages r_t and s_t are associated with the same RMOP, then r_t contains a component-stage c_{r_t} and s_t contains a component-stage c_{s_t}, and at t the thinker treats c_{r_t} and c_{s_t} as about the same thing.

III RMOPs are mental particulars.

IV If r_t and s_t are associated with the same RMOP F_t, then r_t and s_t contain component-stages c_{r_t} and c_{s_t} respectively, and the referential-value of c_{r_t} is the referential-value of c_{s_t}, and this referential-value is determined by F_t.

V If records r and s are treated as if there is some object that both r and s are about, then either there's some RMOP that both r and s are members associated with, or there are two linked RMOPs that r and s are associated with.

With the language of files removed, these slogans are merely a commitment to the claim that the MOP role is played by reference-determining mental representations,[8] along with some fairly commonplace claims about MOPs that would only be rejected outright by someone sceptical of standard ideas about MOPs.

8.2.3 The value of the file account

The thin account of files just endorses the claim that the 'file' role outlined in the core account is instantiated. Using this core account as a starting point has allowed me to develop the account of the mental representation of objects presented in this thesis.

[8]This view is by no means unique to file theory. See e.g. McGinn (1989, pp190-192), Fodor (1992).

The distinctive additional contribution of the thick account of files is the claim that there is a helpful analogy between non-mental filing systems and the mental representation of objects. However, considering the account of mental representation I have developed from the core account of files, there is little to suggest there is any helpful analogy between 'mental files' and the kind of non-mental filing system outlined in 2.1.1.

I quickly abandoned everything but the terminology of records being contained in files, treating a 'file' as just those records which contain the same referential node, or referential tokens of the same type. And I have since suggested that records consist of not only nodes or tokens representing objects, but also nodes or tokens representing properties, kinds and relations. If we understand mental representation in either of these ways, it is very difficult to see any helpful analogies between non-mental filing systems and the mental representation of objects.

So whilst the thin account of files may be a helpful way of thinking about the mental representation of objects, the thick account of files adds nothing distinctive of value. The value of the idea that thinkers use mental files comes from the thin account of files alone.

8.3 The dangers of 'files'-terminology

I have rejected the thick account of files. If we don't take the analogy between non-mental filing systems and mental representations seriously, talk of 'files' becomes a shorthand for the distinctive role described in the core account of mental files (or for whatever realises that role).

Even as shorthand, however, the terminology may be dangerous without further clarification.

One danger is that talk of 'files' often encourages us to lose track of the distinction between the phenomenon used as a model for thinking about mental representation

(non-mental files) and the phenomenon studied in terms of this model, because file theorists rarely talk about mental representation except in terms of 'files'. Modelling mental phenomena in terms of non-mental phenomena is helpful only insofar as we are alive to disanalogies as well as analogies. But using the same terminology for both model and system modelled makes it difficult to identify and describe disanalogies.

8.3.1 Status and the dangers of 'files'-terminology

Some further dangers depend on how we think of the claims about the individuation of files, such as CoStaging-13*.[9]

If we think of claims like CoStaging-13* as misguided, or as definitions, then we might think that the danger of 'files'-terminology is that it encourages us towards an overly optimistic picture of what mental files are. File theories frequently have the appearance of substantive theories of mental representation, and connections are drawn between mental files and the 'files' posited in psychology and linguistics.[10] Perhaps because such connections are available to be drawn, or because there is something *prima facie* plausible in thinking of the mind in terms of a filing system, the danger of files-terminology is that it is too suggestive, leading us to suppose that we are giving a substantive theory of mental files rather than using 'mental files' as a placeholder theory or as a shorthand for other mental processes.

And if we think of claims like CoStaging-13* as definitions, or as stating a priori necessary truths, then we might worry that a danger of the 'files'-terminology is that it disguises cases where talk of 'files' is not adequately defined, or we have not yet worked out the details of the theory. Using the familiar terminology of 'files' to discuss poorly understood phenomena of mental representation leads to us supposing we have a better understanding of what is being discussed using 'files'-terminology than we in fact have. This problem is exemplified by the fact that our familiarity

[9] See 6.4.3.
[10] See p201.

with the idea of non-mental files persisting disguises our poor understanding of the diachronic individuation conditions on files.[11]

If we think of claims like CoStaging-13* as giving identifying conditions, we might worry again that 'files'-terminology is too suggestive, encouraging us to suppose that whatever underlies the identifying characteristics of files has characteristics relevantly similar to a non-mental file. But we have no evidence of any such similarity.

Finally, if we think of claims like CoStaging-13* as giving identifying conditions or stating necessary a posteriori truths, then we suppose that files are something that can be investigated empirically. If we take this kind of view of files, it is natural to draw connections between mental files and the 'files' discussed by psychologists and linguists.[12] However, we should be extremely cautious because the shared terminology of 'files' risks disguising differences between the 'files' discussed by linguists and cognitive psychologists, and those discussed by philosophers.

8.3.2 Files in psychology and linguistics

There is much to be said about comparisons between mental files with 'files' discussed in psychology and linguistics. Here, I can only outline reasons to be cautious when drawing connections between different uses of 'files'-terminology.

In cognitive psychology, the claim that there are "object files" is a response to the *binding problem* in vision. The *binding problem* as investigated by psychologists concerns how humans integrate information across time and sensory modality . As studied in visual attention, the binding problem arises in particular because there is evidence that the distinct features (for example, colour and shape) of an object

[11] See chapter 7.

[12] E.g. Larson and Segal (1995); Recanati (2012). Jeshion deserves particular mention, for using the claim that "mental files' essential singularity is parasitic on . . . object files' essential singularity" (2010, p134n) to argue that files constitute singular thought.

are analysed by separate cognitive systems, leading to questions about how they are integrated (Treisman 1999). One response has been to suggest that there are "object files", binding together these features (Kahneman and Treisman 1984; Kahneman et al. 1992).

Whilst object files and mental files are supposed to play a similar role, integrating information about a single object, we should be cautious about assuming there are similarities between object and mental files. A first difficulty is that little is said about object files that suggests that we should expect similarities between object files and non-mental files. So if object files are supposed to be evidence that of the usefulness of non-mental files as a model for mental representation,[13] more work is needed to show that the 'file' metaphor is used in psychology for anything more than convenience.

A second difficulty is that, whilst object files and mental files both integrate information about a single object, they might be associated with different ways of understanding what it is to be a single object. Object files are:

> temporary "episodic representations of real world objects... separate from the representations stored in a long-term recognition network, which are used in identifying and classifying objects.
>
> (Kahneman et al. 1992, p176)

In contrast, while I do not rule out mental files playing other roles, mental files are primarily associated with the MOP role and so with reidentifying objects. Reidentifying an object is different from treating that object as a single object.

> Imagine watching a strange man approaching down the street. As he reaches you and stops to greet you he suddenly becomes recognizable as a familiar friend... Throughout the episode, there was no doubt that a single individual was present; he preserved his unity (in the sense that he remained the *same* individual)... Perception appears to define objects more by spatiotemporal constraints than by their sensory properties or

[13] Recanati (2012, ppVII-VIII) appears to make this suggestion.

by their labeled identity.

(Kahneman et al. 1992, pp176-177)

The scenario described by Kahneman et al. (1992) involves just one object file representing the man.[14] In contrast, file theory suggests that this scenario involves two mental files. Initially, you open a mental file F serving as a MOP for the unrecognised man. When you recognise the man, you deploy your pre-existing mental file G on the man, adding further information from the man to G, and either linking or merging F and G.

These issues are not evidence that there are no interesting connections between the representations described as 'mental files' and object files. But they give reason to think that identifying these connections requires careful argument, rather than simply pointing to the shared language of 'files'.

Conversely, the language of 'files' should not lead us to suppose that the only interesting connections between mental files and cognitive psychology concern object files. For example, we might think that mental file theory has much to learn from empirical work concerning proper names (e.g. Valentine et al. 1995; Semenza 2009), even though this work does not use the terminology of 'files'.

In linguistics, talk of 'files' is associated with dynamic theories of semantics. Dynamic theories of semantics attribute truth conditions to the body of information built up over the course of a discourse, rather than individual sentences. The meaning of a sentence is its ability to update that body of information. Some versions of dynamic semantics such as Heim's File Change Semantics (FCS)[15] and Kamp's Discourse Representation Theory (DRT)[16] present the meaning of sentences using representations of that body of information.[17] It is these representations that can

[14]In fact, object files were introduced as an alternative to claiming that perceiving an object as an object requires activating representations in the long term memory.

[15]Heim (See 1988 [1982], 1983).

[16]Kamp (See e.g. 2002 [1981]); Kamp and Reyle (See e.g. 1993).

[17]Others are explicitly non-representational (e.g. Groenendijk and Stokhof 1991).

be described in terms of files.

FCS and DRT are both significant semantic theories, with widespread implications for theories of meaning and truth. I cannot hope to do either justice here. However, giving an example of the representations posited by each theory will give a flavour of each theory, sufficient to indicate difficulties drawing connections between these theories and mental file theory.

FCS explicitly uses the 'file' metaphor, comparing the discourse representation to a file of file-cards, and the discourse participant's task to updating those file-cards. These file-cards play the same role as Karttunen's (1979 [1969]) discourse referents.[18] If (83) is uttered discourse-initial, the thinker's discourse-file when the utterance is complete can be represented as FCS-83.[19]

(83) A cat caught a mouse and ate it.

FCS-83

1
— *is a cat*
— *caught* 2
— *ate* 2

2
— *is a mouse*
— *was caught by* 1
— *was eaten by* 1

DRT does not use the imagery of files, but rather talks simply of discourse referents occurring in Discourse Representation Structures (DRSs). Nonetheless, Kamp's theory is associated with the terminology of "files" (Recanati 1995, p571; 2012, p205). At the end of the utterance of (83), the listener's DRS would be DRS-83.

DRS-83

s^1 s^2 u v w x
$cat(u)$
$mouse(v)$
$s^1 : caught(u, v)$
$u{=}w$ $\quad$ $v{=}x$
$s^2 : ate(w, x)$

s^1 and s^2 are discourse referents for states of affairs, u and w are discourse referents for the cat, v and x are discourse referents for the mouse.

[18] See 2.2.2.

[19] In this and what follows, I ignore tense issues.

These lingusitic 'files' again are like mental files in that they are used to integrate information about a single thing. And there are striking similarities between the imagery of Heim's file-cards and versions of the mental file picture. The imagery is the same as the label picture of mental files, and like some versions of file theory, Heim suggests that identity judgements lead to file-cards being merged (Heim 1988 [1982], pp318-320). But we should be cautious of drawing significant connections. Heim is neutral as to whether her discourse representations are eliminable heuristic devices. In contrast, file theorists appealing to empirical evidence for mental files are committed to the idea that files are mental representations.

In contrast to Heim, Kamp is explicitly committed to discourse representations being indispensable and mentally realised.[20] However, it is less easy to see connections between idea of *files* and the discourse referents of DRT. Whilst, with care it may be possible to draw connections between discourse referents and the nodes and tokens I have been discussing, there is no good reason to use files terminology to discuss nodes and tokens.

Even trying to identify connections between Kamp's discourse referents nodes or tokens, one has to acknowledge differences. Suppose (84) is uttered discourse initial. A listener's subsequent discourse representation would include DRS-84.

(84) Panya caught a mouse and then she ate it.

DRS-84

s^1 s^2 u v w x
Panya(u)
mouse(v)
s^1 : *caught(u, v)*
u=w v=x
s^2 : *ate(w, x)*

The utterance of (84) results in two discourse referents representing Panya, u and w, one resulting from the 'Panya'-occurrence, the other resulting from the 'she'-occurrence. But we can suppose that if the listener believed the utterance, she

[20] Kamp (especially 1990).

would form new records representing Panya with just one node, or tokens of just one type.

Kamp also uses DRSs to represent attitudes.[21] Suppose Mary has a referring belief she'd express with an utterance of (84). The representation of her attitudes would include DRS-84^M.

DRS-84^M $\left\{ \begin{array}{l} \langle \text{ANCH}, \boxed{\begin{array}{|c|}\hline u \\ \hline \textit{Panya}\ (u) \\ \textit{Cat}(u) \\ \textit{Fluffy}(u) \\ \hline \end{array}} \rangle \\ \\ \langle \text{BEL}, \boxed{\begin{array}{|c|}\hline u\ v \\ \hline \textit{mouse}(v) \\ s^1 : \textit{caught}(u,\ v) \\ s^2 : \textit{ate}(u,\ x) \\ \hline \end{array}} \rangle \end{array} \right\} \langle u, \mathbf{u} \rangle$

The attitude has three components. There is the belief Mary would express using (84) (indicated by 'BEL'), and an individual representation of Panya (indicated by 'ANCH'). This individual representation will have some content itself, but does not contain all Mary's beliefs about Panya. Finally, there is an anchoring relation between the individual representation of Panya and Panya herself (indicated as $\langle u, \mathbf{u} \rangle$). Again, it's not straightforward drawing connections between this account of attitudes and anything from file theory. We might think that individual representations correspond to 'files' containing just a privileged selection of the belief-records which a thinker treats as about a single object. But 'files', so understood, do not conform to the core account of mental files (see p35n).

Considering, even briefly, the details of 'files' in linguistics and psychology suggests we should be cautious drawing connections between these and the use of the 'file'-metaphor in philosophy.

Furthermore, these examples from linguistics and cognitive science point towards

[21] See Bende-Farkas and Kamp (2001).

a further danger from 'files'-terminology: it may encourage us to conflate distinct kinds of mental representation. 'Files'-terminology encourages us to suppose that the same kind of thing underlies the identifying characteristics of mental files throughout the cognitive system. Consider just trading on coreference of belief records and records of belief-like information: it may be that the mechanisms responsible for trading on coreference when early-vision guides action,[22] are very different from the mechanisms responsible for trading on coreference when a thinker interprets an anaphoric pronoun, and these are different again from mechanisms responsible for trading on coreference in conscious reasoning. The picture is even more complex when we add desire-records and the like into the mix. When using the terminology of 'files' we must be careful not to think too simplistically about the mind and presuppose a unified explanation of these phenomena.

8.4 Final remarks

I have argued that the thick account of files, which takes the file metaphor seriously, does not make a valuable distinctive contribution to our understanding of the mental representation of individuals. I have also argued for caution in using the terminology of 'files'. However, in this thesis I do more than draw negative conclusions.

I have investigated linguistic coreference phenomena, critiquing existing definitions of *de jure* coreference before offering my own. And I have developed the thin account of files, giving synchronic individuation conditions on the mental representations which play the MOP role. I presented this account using 'files'-terminology, but this terminology is incidental to the account and does not undermine its positive contribution, provided the cautionary comments of 8.3 are borne in mind. And I have also laid the groundwork for further work: identifying routes for giving these mental representations' diachronic individuation conditions; discussing the options

[22] See e.g. Pylyshyn (2003, p150).

for thinking about the status of statements of these mental representations' individuation conditions; and illustrating how thinking about this status can influence how the theory is developed.

Bibliography

American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders* Fourth Edition, Text Revision. American Psychiatric Association, Washington DC, 2000.

J. Azzouni. Singular Thoughts (Objects-directed Thought). *Proceedings of the Aristotelian Society*, Supplementary Volume LXXXV:45–61, 2011.

K. Bach. Default Reasoning: Jumping To Conclusions And Knowing When To Look Twice. *Philosophical Quarterly*, 65(1), 1984.

K. Bach. *Thought and Reference*. Clarendon Press, Oxford, 1994 [1984].

A. Bende-Farkas and H. Kamp. *Indefinites and Binding: from Specificity to Incoporation*. 13th European Summer School in Logic, Language and Information, Helsinki, revised version edition, August 2001.

N. Block. Troubles with Functionalism. In N. Block, editor, *Readings in Philosophy of Psychology*, volume One, pages 268–305. Harvard University Press, Cambridge MA, 1980 [1978].

P. A. Boghossian. Content and Self-Knowledge. *Philosophical Topics*, XVII(1), 1989.

P. A. Boghossian. Externalism and Inference. *Philosophical Issues*, 2:11–28, 1992.

P. A. Boghossian. The Transparency of Mental Content. *Philosophical Perspectives*, 8:33–50, 1994.

D. Braun. Understanding Belief Reports. *The Philosophical Review*, 107(4):555–595, 1998.

J. Brown. *Anti-Individualism and Knowledge*. MIT Press, Cambridge MA, 2004.

T. Burge. Memory and Self-Knowledge. In P. Ludlow and N. Martin, editors, *Externalism and Self-Knowledge*, pages 351–370. CSLI Publications, Stanford CA CA, 1998.

T. Burge. Individualism and the Mental. In *Foundations of Mind*, pages 101–150. Oxford University Press, 2007 [1979].

J. L. Camp, Jr. *Confusion: A Study in the Theory of Knowledge*. Harvard University Press, Cambridge MA, 2002.

J. Campbell. Is Sense Transparent? *Proceedings of the Aristotelian Society*, 88: 273–292, 1987-1988.

J. Campbell. *Past, Space and Self*. MIT Press, Cambridge MA, 1995.

S. Carey and F. Xu. Infants' knowledge of objects: beyond object files and object tracking. *Cognition*, 80(1-2):179–213, 2001.

T. Crane. The Singularity of Singular Thought. *Proceedings of the Aristotelian Society*, Supplementary Volume LXXXV:21–43, 2011.

M. Crimmins. *Talk About Beliefs*. MIT Press, Cambridge MA, 1992.

M. Crimmins and J. Perry. The Prince and the Phone Booth: Reporting Puzzling Beliefs. *The Journal of Philosophy*, 86(12):685–711, 1989.

D. Davidson. Knowing One's Own Mind. In *Subjective, Intersubjective, Objective*, pages 14–38. Oxford University Press, Oxford, 2001.

I. Dickie. We are Acquainted with Ordinary Things. In R. Jeshion, editor, *New Essays on Singular Thought*, pages 213–245. Oxford University Press, Oxford, 2010.

I. Dickie. How Proper Names Refer. *Proceedings of the Aristotelian Society*, 111: 43–78, 2011.

I. Dickie and G. Rattan. Sense, Communication, and Rational Engagement. *Dialectica*, 64(2):131–151, 2010.

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

G. Evans. The Causal Theory of Names. In *Collected Papers*, pages 1–24. Clarendon Press, Oxford, 1985 [1973].

G. Evans. Understanding Demonstratives. In *Collected Papers*, pages 291–321. Clarendon Press, Oxford, 1985 [1981].

R. Fiengo and R. May. *Indices and Identity*. MIT Press, Cambridge MA, 1994.

R. Fiengo and R. May. *De Lingua Belief*. MIT Press, Cambridge MA, 2006.

K. Fine. The Role of Variables. *The Journal of Philosophy*, 100(12):605–631, 2003.

K. Fine. *Semantic Relationism*. Blackwell, Oxford, 2007.

K. Fine. Reply to Lawlor's 'Varieties of Coreference'. *Philosophy and Phenomenological Research*, LXXXI(2):496–501, 2010.

J. A. Fodor. *The Language of Thought*. Harvard University Press, Cambridge MA, 1975.

J. A. Fodor. Substitution Arguments and The Individuation of Beliefs. In *A Theory of Content and Other Essays*. MIT Press, Cambridge MA, 1992.

J. A. Fodor. *The Elm and the Expert*. MIT Press, Cambridge MA, 1994.

J. A. Fodor. *LOT2: The Language of Thought Revisited*. Oxford University Press, Oxford, 2008.

J. A. Fodor and Z. W. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1):3–71, 1988.

G. Forbes. Cognitive Architecture and the Semantics of Belief. *Midwest Studies in Philosophy*, 14(1):84–100, 1989.

G. Forbes. The Indispensability of Sinn. *The Philosophical Review*, 99(4):535–563, 1990.

G. Frege. Sense and Reference. *The Philosophical Review*, 57(3):209–230, 1948 [1892].

G. Frege. The Thought: A Logical Inquiry. *Mind*, LXV(259):289–311, 1956 [1918].

G. Frege. Letter to Jourdain. In A. W. Moore, editor, *Meaning and Reference*, pages 43–45. Oxford University Press, Oxford, 1993.

T. Goodsell. Is De Jure Coreference Non-Transitive? *Philosophical Studies*, forthcoming.

H. P. Grice. Vacuous Names. In D. Davidson and J. Hintikka, editors, *Words and Objections*, pages 118–145. D. Reidel, Dordrecht, 1975 [1969].

J. Groenendijk and M. Stokhof. Dynamic Predicate Logic. *Linguistics and Philosophy*, 14(1):39–100, 1991.

J. Hawthorne and E. Lepore. On Words. *The Journal of Philosophy*, 108(9):447–485, 2011.

J. Hawthorne and D. Manley. *The Reference Book*. Oxford University Press, Oxford, 2012.

I. Heim. File Change Semantics and the Familiarity Theory of Definiteness. In R. Bäuerle, C. Schwarz, and A. von Stechow, editors, *Meaning, Use, and Interpretation of Language*, pages 164–190. Walter de Gruyter, 1983.

I. Heim. *The Semantics of Definite and Indefinite Noun Phrases*. Outstanding Dissertations in Linguistics. Garland Publishing, New York NY, 1988 [1982].

R. Jeshion. The Significance of Names. *Mind and Language*, 24(4):370–403, 2009.

R. Jeshion. Singular Thought: Acquaintance, Semantic Instrumentalism, and Cognitivism. In R. Jeshion, editor, *New Essays on Singular Thought*, pages 105–140. Oxford University Press, Oxford, 2010.

D. Kahneman and A. Treisman. Changing Views of Attention and Automaticity. In R. Parasuraman and D. R. Davies, editors, *Varieties of Attention*, pages 29–61. Academic Press, Orlando FL, 1984.

D. Kahneman, A. Treisman, and B. J. Gibbs. The Reviewing of Object Files: Object-Specific Integration of Information. *Cognitive Psychology*, 24(2):175–219, 1992.

H. Kamp. Prolegomena to a Structural Theory of Belief and Other Attitudes. In C. A. Anderson and J. Owens, editors, *Propositional Attitudes: The Role of Content in Logic, Langauge, and Mind*, Lecture Notes Number 20, pages 27–90. CSLI Publications, Stanford CA, 1990.

H. Kamp. A Theory of Truth and Semantic Representation. In P. Portner and B. H. Partee, editors, *Formal Semantics: The Essential Readings*, pages 189–222. Blackwell, Oxford, 2002 [1981].

H. Kamp and U. Reyle. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory.* Studies in Linguistics and Philosophy volume 42. Kluwer Academic Publishers, Dordrecht, student edition edition, 1993.

D. Kaplan. Words. *Proceedings of the Aristotelian Society Supplementary Volume*, 64:93–119, 1990.

D. Kaplan. Dthat. In A. P. Martinich, editor, *The Philosophy of Language*, pages 292–305. Oxford University Press, Oxford, 1996 [1975].

D. Kaplan. Words on Words. *The Journal of Philosophy*, 108(9):504–529, 2011.

L. Karttunen. Discourse Referents. *Syntax and Semantics*, 7:363–385, 1979 [1969].

S. Kripke. A Puzzle About Belief. In A. Margalit, editor, *Meaning and Use: Papers presented at the Second Jerusalem Philosophy Encounter*, pages 239–283. D. Reidel, Dordrecht, 1979.

S. Kripke. *Naming and Necessity.* Blackwell, Oxford, 1980 [1972].

R. K. Larson and G. Segal. *Knowledge of meaning: an introduction to semantic theory.* MIT Press, Cambridge MA, 1995.

H. Lasnik. Remarks on Coreference. *Linguistic Analysis*, 2(1):1–22, 1976.

K. Lawlor. *New Thoughts about Old Things: Cognitive Policies as the Ground of Singular Concepts.* Garland Publishing, New York NY, 2001.

K. Lawlor. Varieties of Coreference. *Philosophy and Phenomenological Research*, LXXXI(2):485–495, 2010.

G. W. Leibniz. *New Essays on Human Understanding.* Cambridge University Press, Cambridge UK, 2^{nd} edition, 1996 [1764].

D. Lewis. Psychophysical and Theoretical Identifications. In D. J. Chalmers, editor, *Philosophy of Mind: classical and contemporary readings*, pages 88–94. Oxford University Press, Oxford, 2002 [1972].

B. Loar. Social Content and Psychological Content. In R. H. Grimm and D. D. Merrill, editors, *Contents of Thought*, pages 99–110. University of Arizona Press, 1985.

M. Lockwood. Identity and Reference. In M. K. Munitz, editor, *Identity and Individuation*, pages 199–211. New York University Press, New York NY, 1971.

G. Longworth. Sharing Thoughts About Oneself. *Proceedings of the Aristotelian Society*, CXIII:1–18, forthcoming. http://www.aristoteliansociety.org.uk/pdf/longworth.pdf.

P. Ludlow. Social externalism, Self-knowledge, and Memory. *Analysis*, 55(3):157–159, 1995.

P. Ludlow and N. Martin, editors. *Externalism and Self-Knowledge*. CSLI Publications, Stanford CA, 1998.

C. McGinn. *Mental Content*. Blackwell, Oxford, 1989.

J. S. Mill. *A System of Logic: ratiocinative and inductive*. Longmans, Green and Co, London, 8^{th} edition, 1904 [1843].

R. G. Millikan. *On Clear and Confused Ideas: An Essay about Substance Concepts*. Cambridge University Press, Cambridge UK, 2000.

M. Murez. Mental Files and the Concept of Identity. Institut Jean Nicod, MS.

J. Perry. Frege on Demonstratives. *The Philosophical Review*, 86(4):474–497, 1977.

J. Perry. A Problem About Continued Belief. *Pacific Philosophical Quarterly*, 61 (4):317–332, 1980.

J. Perry. Cognitive Significance and New Theories of Reference. *Noûs*, 22(1):1–18, 1988.

J. Perry. Belief and Acceptance. In *The Problem of the Essential Indexical*, pages 53–63. Oxford University Press, Oxford, 1993 [1980].

J. Perry. *Reference and Reflexivity*. CSLI Publications, Stanford CA, 2001.

N. A. Pinillos. Coreference and Meaning. *Philosophical Studies*, 154(2):301–324, 2011.

H. Putnam. The meaning of 'meaning'. In *Mind, Language and Reality*, volume Philosophical Papers volume 2, pages 215–271. Cambridge University Press, Cambridge UK, 1975.

Z. W. Pylyshyn. *Seeing and Visualising*. MIT Press, Cambridge MA, 2003.

W. V. O. Quine. *Word and Object*. MIT Press, Cambridge MA, 1960.

G. Rattan. *Semantic Relationism*, by Kit Fine. *Mind*, 118(472):1124–1131, 2009.

F. Recanati. *Direct Reference: from language to thought*. Blackwell, Oxford, 1997.

F. Recanati. Deixis and Anaphora. In Z. G. Szábo, editor, *Semantics versus Pragmatics*, pages 286–316. Oxford University Press, Oxford, 2005.

F. Recanati. In Defence of Acquaintance. In *New Essays on Singular Thought*, pages 141–189. Oxford University Press, Oxford, 2010.

F. Recanati. *Mental Files*. Oxford University Press, Oxford, 2012.

T. Reinhart. Center and periphery in the grammar of anaphora. In B. Lust, editor, *Studies in the Acquisition of Anaphora*, volume 1, pages 123–150. D. Reidel, Dordrecht, 1986.

M. Richard. *Propositional Attituds: an essay on thoughts and how we ascribe thems*. Cambridge University Press, Cambridge UK, 1990.

B. Russell. *Principles of Mathematics*. Routledge, London, 2010 [1903].

R. M. Sainsbury. Fregean Sense. In R. M. Sainsbury, editor, *Departing from Frege*, pages 125–136. Routledge, London, 2002 [1997].

R. M. Sainsbury and M. Tye. *Seven Puzzles of Thought*. Oxford University Press, Oxford, 2012.

N. U. Salmon. *Frege's Puzzle*. MIT Press, Cambridge MA, 1986.

S. Schellenberg. Sameness of Fregean Sense. *Synthese*, forthcoming.

L. Schroeter. The Illusion of Transparency. *Australasian Journal of Philosophy*, 85 (4):597–618, 2007.

L. Schroeter. Why Be an Anti-Individualist? *Philosophy and Phenomenological Research*, LXXVII(1):105–141, 2008.

L. Schroeter. Bootstrapping our way to samesaying. *Synthese*, Forthcoming.

J. R. Searle. Proper Names. *Mind*, LXVII(266):166–173, 1958.

S. Sedivy. Minds: contents without vehicles. *Philosophical Psychology*, 17(2):149–180, 2004.

G. Segal. Two Theories of Names. *Mind & Language*, 16(5):547–563, 2001.

C. Semenza. The Neuropsychology of Proper Names. *Mind and Language*, 24(4): 347–369, 2009.

T. Sider. Four-Dimensionalism. *The Philosophical Review*, 106(2):197–231, 1997.

S. Soames. Presupposition. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume Volume IV: Topics in the Philosophy of Language, pages 553–616. D. Reidel, Dordrecht, 1989.

R. Stalnaker. Propositions. In A. F. MacKay and D. D. Merrill, editors, *Issues in the Philosophy of Language*, pages 79–91. Yale University Press, New Haven CT, 1976.

R. Stalnaker. *Inquiry*. MIT Press, Cambridge MA, 1984.

R. Stalnaker. Common Ground. *Linguistics and Philosophy*, 25(5/6):701–721, 2002.

R. Stalnaker. *Our Knowledge of the Internal World*. Oxford University Press, Oxford, 2008.

P. F. Strawson. *Subject and Predicate in Logic and Grammar*. Methuen & Co., London, 1974.

K. A. Taylor. *Reference and the Rational Mind*. CSLI Publications, Stanford CA, 2003.

A. Treisman. Solutions to the Binding Problem: Progress through Controversy and Convergence. *Neuron*, 24:105–110, 1999.

T. Valentine, V. Moore, and S. Brédart. Priming Production of People's Names. *The Quarterley Journal of Experimental Psychology*, 48A(3):515–535, 1995.

T. van Gelder. Compositionality: A Connectionist Variation on a Classical Theme. *Cognitive Science*, 14(3):355–384, 1990.

K. von Heusinger. Reference and Representation of Pronouns. In H. J. Simon and H. Wiese, editors, *Pronouns — Grammar and Representation*, Linguistik Aktuell/Linguistics Today volume 50, pages 109–135. John Benjamins Publishing Company, Amsterdam, 2002.

P. C. Wason and D. Shapiro. Natural and contrived experience in a reasoning problem. *Quarterly Journal of Experimental Psychology*, 23:63–71, 1971.

T. Williamson. *Identity and Discrimination*. Blackwell, Oxford, 1990.

T. Williamson. *Knowledge and its Limits*. Oxford University Press, Oxford, 2000.

A. Woodfield. Conceptions. *Mind*, C(4):547–572, 1991.