FREGE CASES AND RATIONALIZING EXPLANATIONS[*]

Mahrad Almotahari
University of Edinburgh

Aidan Gray
University of Illinois, Chicago

Abstract

Russellians, Relationists, and Fregeans disagree about the nature of propositional-attitude content. We articulate a framework to characterize and evaluate this disagreement. The framework involves two claims: i) that we should individuate attitude content in whatever way fits best with the explanations that characteristically appeal to it, and ii) that we can understand those explanations by analogy with other 'higher-level' explanations. Using the framework, we argue for an under-appreciated form of Russellianism. Along the way we demonstrate that being more explicit about the framework in which debates about attitude content take place allows us to more precisely characterize the space of possible positions and the dialectic between them.

> "Indeed the essence of explanation lies precisely in the fact that a wide, possibly unsurveyable, manifold is governed by one or a few sentences." -Frege, 1880/81[1]

Suppose you're an *intentional realist*. That is, suppose you think intentional explanations often approximate the truth and thus reveal underlying patterns of nomic dependence between one's propositional attitudes and one's actions. How might you then theorize, in a systematic way, about the nature of propositional attitudes? In particular, how might you decide between competing conceptions of propositional-attitude content?

One promising thought is that the nature of propositional-attitude content is revealed by the role it plays in a certain class of explanations. Exactly how to characterize this class is a main theme below. But for now we can note that, paradigmatically anyway, intentional explanations explain a subject's behavior in terms of the attitudes that gave them a reason to perform it. Attitude content, one might think, is whatever it needs to be for these explanations to be true and explanatory.

---

[*] Forthcoming in *Noûs*.
[1] Taken from 'Boole's logical Calculus and the Concept-script', p. 36 of *Posthumous Writings*. The principle is part of Frege's argument for why his formal language is superior to Boole's.

We have three related goals in this paper. The overarching goal is to demonstrate that we can advance the ongoing debate about attitude content by being more explicit about the nature of intentional explanation. We'll do that, first, by developing a substantive picture of the form such explanations take. We do not claim that ours is the only viable picture. But we do claim that everyone needs some such picture if they want to implement the promising thought we began with. And our framework has two noteworthy advantages. Its core assumptions are widely shared (especially in the literature on Frege's Puzzle[2]) and independently plausible: they're largely motivated by our understanding of, and commitment to, intentional realism. Initially, then, the framework will seem compatible with (or even congenial to) competing conceptions of content.

Our second goal is to show how our framework allows us to characterize the motivation for *Semantic Relationism* about attitude content. This is a more recent contender as a theory of attitude content and, in our view, it has been liable to be misunderstood. Our framework allows us to clearly frame the position and succinctly state its argument against a certain form of *Fregeanism.*

But our third goal is to explain why our framework actually favors an austere form of *Russellianism*—even in light of Frege's Puzzle. So, contrary to what one might expect at the outset, fine-grained theories of content will need a different story about the nature of intentional explanation. But it's not clear what that story could be, or whether it will fit as nicely into a realist metaphysics of mind.

1. Explanationism

We're assuming that the best way to arrive at a theory of attitude content is to take the explanations in which attitudes canonically participate and hold that attitude content is whatever it has to be to play the right role in those explanations. Call this orientation *Explanationism.* As the paper goes on, we'll make this perspective more determinate and explore how it shapes debates about attitude content.

Our starting point is intentional realism. We assume that propositional attitudes and rational behavior participate in certain higher-level patterns. What an agent does, in a suitably broad sense of the verb 'to do', nomically depends on her propositional attitudes. In line with a standard conception of explanation (*i.e.*, the representation of a particular case as an instance of a robust regularity), we'll

---

[2] See (Peacocke, 1981), (Lewis, 1983), (Block, 1986), (Loar, 1988), (Devitt, 1989), (Fodor, 1991; 1994), (Braun, 2000; 2001*a*; 2001*b*; 2020), (Segal, 2000), and (Schneider, 2005; 2011).

assume that explanations of rational behavior require the subsumption of a token act under a lawlike generalization that describes the relevant higher-level pattern.

The patterns that constitute the domain of propositional-attitude psychology are derivative, not fundamental. We take this to mean, among other things, that they obtain in virtue of the goings-on at 'lower' levels. The intentional laws are implemented by patterns characterizable in functional, neurological, chemical, biological, or whatever, terms.

Here is a substantive (though, we hope, not too controversial) claim about situations of this sort: realism about a 'higher' level of explanation presupposes that this level is *perspicuous* with respect to some explanatory generalizations (Pylyshyn 1984). The idea here is that higher-level phenomena—genes, species, weather systems, economies, beliefs, *etc.*—must earn their keep by constituting patterns that are not as clearly visible in the goings-on of phenomena at lower levels. We might be led to realism about species, for example, rather than merely about genes, or about certain chemical structures, because there are explanatory patterns that species participate in such that, if the patterns are characterizable at all in terms of genes or chemical structure, they would be highly disjunctive and seemingly *ad hoc* (that is, not very pattern-like) at those lower levels.

Given Explanationism, this assumption of perspicuity means that we should individuate content as finely as, and no finer than, is needed to capture the generalizations that are characteristic of the intentional level of explanation. Intentional realism presupposes that there are explanatory joints characterizable in terms of attitude contents; Explanationism tells us to individuate attitude contents such that contents carve those joints.

Another substantive assumption (probably more controversial, but still relatively tame): intentional generalizations have *ceteris paribus* hedges (going forward, we'll call these '*cp*-hedges'). That is, we assume that their canonical form is something like: '*Ceteris paribus*, if p, then q'.

This is the structure we typically find when generalizations at one level are implemented by structures at lower levels. Consider (1a):

(1) (a) *Ceteris paribus*, if x is a raven, then x is black.

Suppose that (1a) states an explanatory generalization and, correlatively, that *raven* marks a biological joint.[3] What role is played by the *cp*-hedge?

The answer is, of course, complex. But here is the basic idea. (1a) is implemented by phenomena at lower levels: for example, by genes and the chemical processes that act on them in the conditions typical for raven development. There are different lower-level ways of instantiating ravenhood. Different ravens have different genetic make-ups. And some ways of genetically being a raven interrupt the connection between ravenhood and blackness. Some ravens that are genetically *abnormal*—albino ravens—are not black; their underlying genetic make-up is such as to not produce black pigment under conditions normal to raven development. We treat albino ravens as abnormal, that is, as falling under the *cp*-condition of (1a), and thus as not falsifying it. Call this an 'unthreatening exception' to a *cp*-generalization.

It would be natural to worry, at this point, that the *cp*-hedge, and the concomitant appeal to 'normality', vitiates ravenhood's claim to being a joint in nature. If the pattern we claim to have identified as involving ravenhood and blackness only holds some of the time, isn't that evidence that we haven't found the real joints?

In fact, the opposite is true. Qualifying a generalization with a *cp*-hedge, and acknowledging that there are unthreatening exceptions to it, can *reveal* joints by maximizing the "unifying power" of a system of generalizations (Nickel 2010, pp. 17-18). Sometimes it is only by rendering the generalizations of a higher-level domain exception-tolerant that we can adequately represent the fact that a number of different generalizations, which would look disconnected at a lower level, are bound together in an explanatory nexus.

Consider the following pairs of generalizations:

(1) (a) *Ceteris paribus*, if x is a raven, then x is black.
　　(b) If x has genetic features $g_1...g_n$, then x is black.

---

[3] Throughout the paper we use generalizations about species and phenotypes as an example from which we draw lessons about *cp*-explanations. We recognize that some of our claims—about species as natural kinds, about the relation between species and selection, *etc.*—aren't universally accepted, but neither are they completely out of step with contemporary ways of thinking among specialists (Coyne, 2009). We hope that none of the general lessons we draw depend on any specific claims about these matters. We simply want to represent important relationships between kinds, generalizations, and explanatory unity. We rely on this particular example because our understanding of these issues has been heavily influenced by (Nickel, 2010), in which the basic point is explained this way. For a similar picture, somewhat different in detail, see (Strevens, 2012).

(2) (a) *Ceteris paribus*, if x is a raven, then x has two wings.
     (b) If x has genetic features $g*_1...g*_n$, then x has two wings.

Suppose that $g_1...g_n$ are the genetic features that produce black pigment under conditions normal for raven development and that $g*_1...g*_n$ are the (at least partially distinct) genetic features that produce two wings under conditions normal for raven development. Imagine the examples continuing—each pair replacing a generalization about a raven phenotype with a generalization about the particular genetic basis of that phenotype.

    The (b)-generalizations are stricter than the (a) ones (though, notice that there really isn't any reason to suppose that they should not also be *cp*-hedged). If we replaced the (a)-generalizations with the (b) ones, we would be doing less *cp*-hedging. But this doesn't mean that *ravenhood* is not a joint in nature. The price of less hedging is that we lose sight of the fact that these generalizations are non-accidentally related to each other. It's not just that if we introduce *ravenhood* as a higher-level property, which can be instantiated by an open-ended, but not unconstrained, variety of related genetic make-ups, then we're able to express these generalizations in a more compact way (though, we can indeed do that). More importantly, representing the different patterns as intersecting generalizations about *ravens* allows us to see the interdependence between them.

    One reason to think that species are natural kinds—or, to put it another way, that generalizations involving species are perspicuous with respect to some real patterns—is that this level of explanation reveals the interconnectedness of distinct (a)/(b) generalizations (for example, the connection between (1a)/(1b) and (2a)/(2b). Suppose we ask: how is it that the different (a)/(b)-generalizations have come to hold; and what processes sustain them? The answer, ultimately, will appeal to selection on *ravens*.[4] Or, more neutrally, the process of natural selection that generates and sustains each generalization essentially involves the others; black feathers are adaptive, in part, because of the presence of other raven phenotypes (wings, eyes, *etc.*). The same is true for other raven phenotypes. If we see the

---

[4] It's very unlikely that selection accounts for *all* explanatory generalizations about raven traits. There are significant non-selectionist explanations, in terms of drift, spandrels, *etc.* This wouldn't threaten the idea that there's an underlying unity in the mechanisms that generate and hold in place the relevant patterns.

different phenotypic generalizations as intersecting on *ravens*, we can capture the unity of the processes that bring them about and sustain them.

Going further, what's true about raven-generalizations is true about generalizations concerning other species too. So, in general, species talk allows us to see higher-order patterns in which different species participate, and even different domains where selection operates.

To return to the attitudes, we'll assume that intentional explanation has this structure. Intentional regularities are implemented by regularities at lower levels. Those lower-level regularities might be stricter, but they don't perspicuously represent the forms of dependence that are characteristic of intentional explanation. This assumption is, to be sure, not mandatory. But it isn't implausible. And it will open up space to bring new considerations to bear on debates about attitude content.

It will turn out—in Section 5—that this initial characterization of the Explanationist framework is underspecified in an important way. But what's already on the table is enough to advance the debate about attitude content.

2. Russellian Exceptionalism

Russellianism about attitude content is the view that the content of an attitude is exhausted by its referential features; that is, once we specify how an attitude represents properties and relations as distributed over objects, we have fully specified its content. The first thing we'll do with our framework is to show what the most plausible form of Russellianism about attitude content looks like in it.

(The view we develop in this section is inspired by Fodor (1994). But the details, and the argument we provide for it later in the paper, diverge substantially from Fodor's discussion. Space doesn't allow for extensive comparison, but along the way we'll mention a few important points of contact with Fodor and his critics.)

Explanationism requires that we characterize competing theories of content in terms of their corresponding regimentations of intentional explanations. To see how this should go, it's helpful to have particular cases to work with.

Suppose Frederica believes (as she would put it) that 'Stefani Germanotta' (her eccentric neighbor—aka Lady Gaga) is down in the dumps. She wants 'Stefani' to be happy. So Frederica phones her to try to cheer her up. Call this Case 1.

Frederica's behavior is perfectly intelligible. The Russellian Explanationist will represent it as an instance of a certain generalization—perhaps something like (R).

(R)   *Ceteris paribus*, if x has a desire with the referential content that y is happy, and x has a belief with the referential content that y is unhappy, x will try to cheer y up.

The *referential content* of an attitude is the way that it represents the distribution of properties and relations over objects. We'll say that two attitudes are *referentially equivalent* when they share their referential content.

Russellianism is the thesis that propositional-attitude content is exhausted by referential content.[5] Given Explanationism, this thesis predicts that intentional explanation ought to take a particular shape when properly regimented. It should mention nothing other than subjects, attitude-types (belief, desire, etc.), and referential contents. The Russellian Explanationist's statement of the *cp*-law by means of which Frederica's behavior is made intelligible is specified in (R). We obtain the relevant instance of (R) by fixing values for the variables:

(R₁)   *Ceteris paribus*, if Frederica has a desire with the referential content that Stefani/Gaga is happy, and Frederica has a belief with the referential content that Stefani/Gaga is unhappy, Frederica will try to cheer Stefani/Gaga up.[6]

---

[5] This thesis is compatible with a wide range of views about the semantics of attitude ascription. Going forward, though, we largely bracket questions about natural-language semantics in order to streamline our discussion. Some readers might think this decision is misguided, either because the semantical and metaphysical issues are inextricably linked or because there's some dialectical advantage to be had by focusing on language first. Our view is that strong opinions about the right way to proceed here are unwarranted, given how difficult the problems are. Far better to embrace a spirit of tolerance and see where different paths lead. But our decision to separate the issues in philosophy of mind from the issues in philosophy of language isn't random. It's based on the conviction that everyone needs a story about the mechanics of folk-psychological explanation, whether one is theorizing about the semantics of 'believes' or the nature of belief. After all, judgments about the explanatoriness of certain folk-psychological claims are crucial for both the semantics of attitude ascription and the nature of propositional attitudes. See, in particular, Heck (2014*a*). It's necessary, then, to have some theoretical basis for accepting some such claims and not others. We think this basis should ultimately derive from the principles that justify belief in propositional attitudes—hence our starting point. For an illuminating discussion that's friendly to a referentialist conception of attitude content, but which focuses on the language side of things, we recommend (Goodman & Lederman, 2021). We're grateful to Harvey Lederman for discussion of this issue.

[6] The slash notation indicates that a certain position is referentially transparent. We don't assume that the subject of the attitude ascription (in this case, Frederica) is aware of the corresponding

As we understand them, Russellian intentional explanations ignore a great deal of machinery that Russellians posit in their broader theory of mind. Russellians hold that attitudes are three-place relations between subjects, referential contents, and some non-intentional third thing. Terms for this 'third thing' vary—some call it a *guise*; others call it a *notion* or *concept*.[7] Call it what you will. What's crucial for our brand of Russellianism is that none of these candidate 'third things' are explicitly mentioned in canonical intentional explanations (the force of 'canonical' here will become clear in Section 5).

This structure should be familiar from our general discussion of joint-carving higher-level explanations in the previous section. The kind of Russellianism for which we'll argue sees the three-place analysis of attitudes as abstractly characterizing the machinery that *implements* the intentional regularities that are captured by *cp*-generalizations like (R). The Russellian holds that the joints that are unique to the level of intentional explanation are fully characterized in terms of referential content. To put it metaphorically, intentional generalizations don't 'see' any of the implementing details; they only see referential content.

This is crucial for Russellianism as we understand it, because it guides the Russellian's response to the challenge posed by Frege cases. In brief, Russellians should hold that Frege cases are like albino ravens. That is, they're unthreatening exceptions to intentional generalizations.

Albino ravens are ravens that instantiate ravenhood in a genetically abnormal way. And they're unthreatening exceptions to (1a). Those two facts are connected. It is normal, given some higher-level generalizations, for unthreatening exceptions to be characterizable only in terms of some lower level vocabulary (Nickel 2010). (1a) is a species-level generalization; one way that a particular raven can fall under the *cp*-clause, and thus be an unthreatening exception to it, is for it to be abnormal with respect to the genetic implementation of ravenhood.

So if Russellianism is correct, it would be no surprise that an individual might be an unthreatening exception to generalizations like (R) in virtue of bearing the attitudes mentioned in its antecedent in a way that is only characterizable in terms of guises, concepts, or whatever.

---

identity. More generally, if it turns out that English 'that'-clauses do more than specify referential content, we ask that you treat the language of (R$_I$) as a technical language that just happens to resemble colloquial English.

[7] See (Salmon, 1986), among others.

To see how this works, consider a variation on our case. Suppose Frederica, instead of believing (as she would put it) that 'Stefani' is unhappy, believes (as she would put it) that 'Lady Gaga' is unhappy. As before, Frederica wants Stefani to be happy, but she's unfamiliar with popular culture, and so is unaware that Stefani is Lady Gaga. Frederica has no special interest in (as she would put it) 'Lady Gaga's' wellbeing, and from her point of view 'Stefani' is doing just fine. So she makes no attempt to cheer anybody up. Call this Case 2.

Case 2 appears to falsify (R_l) and thus (R). After all, Frederica has attitudes with referential contents that satisfy the antecedent of (R_l), and yet she doesn't attempt to cheer anyone up. So it looks like her behavior in Case 2 is a counterexample to Russellianism.

As we've said, the Russellian should resist this conclusion by saying that in Case 2, Frederica is an unthreatening exception to (R) rather than a counterexample.[8] In particular, the Russellian should say that she is unthreatening, because she possesses the attitudes mentioned in the antecedent in a way that is abnormal in relation to (R).

Abnormal in what way? To spell this out, we must introduce the idea of *coordination.* Two representations of the same object are coordinated if and only if the identity of their referents is "directly" encoded in the way they're represented. "Direct encoding" obviates the requirement for an explicit representation of identity in certain forms of intentional explanation.[9]

The nature of coordination is at the center of the debate between Russellians, Fregeans, and Relationists. The fundamental difference is over whether/how coordination is an aspect of content. The Russellian holds that it's not. But this isn't to say that she can't countenance coordination at all, nor even that she must take it to be irrelevant to intentional explanation. It's only to say that she should hold that the coordination of certain representations of the same object is a lower-level feature and is part of the *normality conditions* for possessing attitudes that instantiate the higher-level patterns identified by certain intentional generalizations.

Exactly how the Russellian should understand coordination will depend on the particulars of her view concerning attitude implementation. If she believes in a

---

[8] In addition to (Fodor, 1994), see (Soames, 1990), (Braun, 2000; 2001*a*; 2001*b*; 2020), (Schneider, 2011), and (Cappelen & Dever, 2013).

[9] Compare: the notion of *representation as the same* in (Fine, 2007) and the notion of *trading on identity* in (Campbell, 1987). Related ideas can be found in (Sainsbury, 2002) and (Taschek, 1995; 1998).

language of thought, she could hold that two representations of the same object are coordinated if and only if they're implemented by token "words" of the same language-of-thought type. We won't assume any particular Russellian account of coordination here. We'll only assume that the Russellian must appeal to some such account.

　　With all of this in view, we can state how our Russellian should respond to Frege's Puzzle. She should embrace

> (*Exceptionalism*)　　For all subjects, S, and intentional generalizations, G, if (*i*) G involves coreference at referential positions $r_1...r_n$, (*ii*) S satisfies the antecedent of G in virtue of attitudes that are not coordinated at $r_1...r_n$, and (*iii*) S fails to satisfy the consequent of G, then S isn't (thereby) a counterexample to G.

The 'positions' in an intentional generalization are just the particular aspects of referential content that they attribute to the relevant subjects. A generalization involves coreference at two positions when they're introduced by occurrences of the same variable (or, equivalently, when any subject who instantiates the generalization will have attitudes that are coreferential at those positions). So, for example, (R) involves coreference at the position that represents the person desired to be happy and the person believed to be unhappy.

　　Crucially, this principle does *not* imply that being in a Frege case is abnormal *as such*. An attitude state is normal or abnormal only *relative* to a particular intentional generalization. It's consistent with Exceptionalism that, in Case 2, Frederica satisfies many other intentional generalizations in virtue of her Stefani/Gaga-directed attitudes.

　　To avoid confusion, we should mention an apparent disanalogy between Frederica and an albino raven. We just said that Frederica's uncoordinated attitudes about Stefani/Gaga are not abnormal *as such*. They're only abnormal with respect to (R) (and perhaps other related generalizations). In contrast, we're generally happy to say that an albino raven is abnormal *period*, not just abnormal-with-respect-to-a-generalization. But this disanalogy is merely superficial; there's no important contrast here. In both cases, the fundamental thing is abnormality-with-respect-to-a-generalization. Albino ravens, notice, participate perfectly normally with respect to many other generalizations about raven phenotypes (for example,

the two-winged-ness of a black raven and the two-winged-ness of an albino raven are explained by the same feature: that they're both ravens). And Stefani's attitudes that are abnormal with respect to (R) can participate perfectly normally in other intentional explanations. We return to this point in Section 7.

We should note that this is one important respect in which our position might differ from the one offered in Fodor (1994). Fodor has been read as suggesting that subjects in Frege cases fall outside the bounds of rationalizing explanation all together.[10] (We say 'might differ' and 'has been read' because we aren't, ourselves, sure how to interpret Fodor's discussion.) We take this to be implausible in the extreme and nothing in our view entails it.

Let's call the position we've characterized in this section, *Russellian Exceptionalism*. So far, we haven't offered any argument for it. In particular, we haven't offered any argument that referential content is joint-carving for the intentional level of explanation. Nor have we offered any reason to take the Russellian's claim about the normal conditions for satisfying the antecedent of generalizations like (R) seriously. One of our goals in this paper is to show that Russellian Exceptionalism has more going for it than you might expect. But, for now, it's enough to see the general outline of the view and, in particular, to understand it as an instance of the general claims we made about higher-level explanation in the previous section. We'll defend the view in Section 7.

3. Fregeanisms

In the previous section, we showed how Explanationism allows us to characterize a version of Russellianism that's not otherwise in view. In this section, we'll show how it also allows us to distinguish different varieties of Fregeanism.

Fregeans treat Case 2 as a straightforward counterexample to (R) and thus as motivating a rejection of Russellianism about content. The moral of Case 2, according to them, is that attitude content must be individuated more finely than

---

[10] For example, here is (Arjo, 1996, p. 239) defending Oedipus (the protagonist of a well-known Frege case) from Fodor: "Oedipus's behaviour seems perfectly rational and understandable. He does not seem beyond the pale of intentional explanation in the manner of, say, someone who has ingested LSD, or suffered a blow to the back of the head with a 2x4." We agree with Arjo about Oedipus. But nothing in our view requires us to deny this. Note, for example, that there are many perfectly good rationalizing intentional generalizations that the Russellian can take Oedipus to instantiate. For example: *ceteris paribus*, if x wants to marry y, and x believes y wants to marry x, x will propose to y.

referential equivalence permits. For Fregeans, the joints revealed by intentional generalizations are finer grained than referential contents.

But what should Fregean generalizations look like? Fregeans posit a layer of content that determines reference but that isn't itself determined by reference. Call this additional layer of content *sense*. For the Fregean, every representation expresses a sense from which its referential content can be recovered. Senses are, therefore, *modes of presentation*. There's a long and torturous history of trying to characterize more precisely what a Fregean sense is.[11] Our aim is to skirt the issue as much as possible here. Whatever senses are, Frederica's '"Stefani"-sense' and her '"Gaga"-sense' had better be distinct. This is, presumably, why the Fregean holds that her attitudes in Case 2 have a different content from their counterparts in Case 1. In Case 2, Frederica's belief presents Stefani/Gaga via the 'Gaga'-sense while the belief in Case 1 presents her via the 'Stefani'-sense.

Should we replace (R) with a generalization that specifically mentions Frederica's 'Stefani'-sense? Presumably not. This sort of generalization would be much too narrow to be perspicuous. If some generalization in the vicinity of (R) reveals a pattern in rational behavior, it's not because that pattern has something in particular to do with thinking of Stefani/Gaga via the 'Stefani'-sense. If Frederica wanted (as she would put it) that 'Lady Gaga' be happy, and believed (again, as she would put it) that 'Lady Gaga' was unhappy, she would have tried to comfort her, other things being equal. The relevant difference between Case 1 and Case 2 is that, in Case 1, the two representations of Stefani/Gaga are coordinated, while, in Case 2, they aren't. And, for Fregeans, coordination is sameness of sense. Given all of this, it would be highly imperspicuous to replace (R) with indefinitely many structurally analogous Fregean generalizations, each making reference to a different sense (or different kinds of senses, or whatever).

The upshot, then, is this: if we embrace Fregeanism on the basis of Case 2, we'll replace (R) with (F).

---

[11] Actually, we believe the task is substantially more difficult than is widely recognized. Senses need to have a certain modal profile if the generalizations in which they figure are to satisfy a minimal requirement for being explanatory. Frege's Puzzle provides no support for thinking that they have this profile. We argue for these points at length elsewhere (Almotahari & Gray, 2021). In many ways, this paper picks up where that one leaves off.

(F)   *Ceteris paribus*, if x has a desire with the referential content that y is happy in which y is presented via sense s, and x has a belief with the referential content that y is unhappy in which y is presented via sense s, x will try to cheer y up.

Instantiating so that x = Frederica, y = Stefani/Gaga, and s = Frederica's 'Stefani'-sense, (F) subsumes the situation in Case 1. Furthermore, Case 2 doesn't pose a threat to the truth of (F), because there's no way of instantiating (F) to render its antecedent true relative to Case 2.

The thing to note about (F) is that it puts no constraint on the identity, or nature, of the sense to which it appeals. It only requires that the senses associated with the person believed to be unhappy and desired to be happy are the same.

We're now in a position to distinguish two varieties of Fregeanism. Let's call the brand of Fregeanism that motivates a fine-graining of content exclusively by appeal to the phenomenon exhibited in the relation between Case 1 and Case 2 *Relational* Fregeanism. For the Relational Fregean, (F) is characteristic of the way that senses are involved in intentional explanation. That is, Relational Fregeans only posit intentional generalizations that constrain patterns of sameness or difference of sense; they posit no intentional generalizations that make essential appeal to particular senses, senses of particular kinds, *etc.* To put it simply, Relational Fregeans hold that senses are relevant to intentional patterns only insofar as they are responsible for establishing (or failing to establish) the coordination that is required for certain forms of intentional explanation.

We don't claim that any self-professed Fregeans would self-identify as Relational Fregeans.[12] Rather, our claim is that failure to be explicit about Explanationism has obscured an important choice-point here. One of the goals of the paper is to clearly identify Relational Fregeanism and show how little it has going for it.

We'll call Fregeans who aren't Relational Fregeans *Substantive Fregeans*.[13] Substantive Fregeans posit intentional generalizations that make reference to features of sense other than their identity and distinctness. Given this, Substantive

---

[12] Though maybe (Devitt, 1989) and (Sainsbury, 2002) count. We're reluctant to come down one way or the other, mainly because of the general unclarity about these matters at the time this pioneering work was written.

[13] This position is identified in (Gray, 2022), but called simply 'Fregeanism' there.

Fregeanism must be motivated by something other than the phenomenon exhibited by Case 1 and Case 2, which only motivates generalizations like (F).

There are different possible motivations for Substantive Fregeanism. One might, for example, take 'essential indexicality' to motivate a fine-graining of content that doesn't fit the mold of (F). This would require arguing—contra recent skeptics (Cappelen and Dever 2013; Magidor 2015)—that this phenomenon is genuinely distinct from 'run-of-the-mill' Frege cases.[14] Alternatively, one might argue that we need a fine-graining of content to accommodate the sort of *a priori* knowledge revealed by conceptual analysis (Jackson 1998; Chalmers 2006). In either case, the argument for the fine-graining of content, and the theory of content developed, will be substantially different from Relational Fregeanism. Discussion of Substantive Fregeanism would take us off topic, so we'll set it to one side in this paper. (Though we'll mention that we take the kinds of considerations developed in Section 7 against Semantic Relationism to tell against some forms of Substantive Fregeanism.)

Relational Fregeanism is the kind of Fregeanism that is motivated by reflection on standard Frege cases. Explanationism allows us to frame it clearly. And, we will argue, it also allows us to see how little it has going for it.

4. From Relational Fregeanism to Semantic Relationism

In this section, we'll show how to fit Relationist approaches to content into our framework by arguing that, given our assumptions, Relational Fregeanism collapses into Relationism.

From an Explanationist point of view, Relational Fregeanism is an unstable position. Explanationism recommends that we let our conception of content be guided by the structure of explanatory intentional generalizations. If we find ourselves positing intentional generalizations that are only ever sensitive to sameness and difference of sense, and never to the natures of particular senses, it's clear that we aren't, in fact, entitled to think of senses as aspects of intentional content. And this is exactly the situation of the Relational Fregean. To see this, note that any attitude state, with the appropriate referential content, and any senses distributed in the right way will instantiate (F). As we have already noted, (F) places no constraint on the features of the senses that instantiate it. So, unless the particular nature of senses-qua-modes-of-presentation plays some *other* role in

---

[14] See (Gray, 2022) and (Almotahari & Gray, 2021) for arguments that it is.

intentional explanation, considerations of perspicuity guide us to explore whether we can frame our intentional generalizations *directly* in terms of coordination. And, recall, Relational Fregeanism is defined by its insistence that this is the only role for sense.

To put this point another way, perspicuity guides us to draw the line between the content of attitudes and the underlying facts about their implementation in a certain way: we should only admit into our ontology of content those features that are necessary to characterize intentional patterns. If a putative assignment of contents is more committal than is required to characterize intentional patterns, we should search for a more perspicuous alternative. The Relational Fregean runs afoul of this by insisting that senses are aspects of content while, at the same time, admitting that only relations of sameness and difference of sense are required to characterize the relevant patterns.

Therefore, we should explore the possibility of replacing (F) with something like (C).

> (C) *Ceteris paribus*, if x has a desire with the referential content that y is happy, a belief with the referential content that y is unhappy, and the representation of y in x's desire is semantically coordinated with the representation of y in x's belief, x will try to cheer y up.[15]

As we use the term, 'coordination' *sans* modifier denotes a relation that Russellians, Fregeans, and Relationists alike can appeal to. It's supposed to pick out a cognitively relevant relation that can hold between representations of the same object without a commitment to how that relation should be understood. 'Semantic coordination', on the other hand, is an element of Relationist ideology, which builds in a certain theoretical understanding of what coordination is.

Semantic coordination, according to the Relationist, is a *representationally primitive* intentional relation. This means that it's a relational feature of propositional content that's not determined by the distribution of any representational property. And it's precisely on this issue that Relationists and Fregeans differ. Fregeans hold that coordination is determined by sameness of sense, where sense is, itself, a non-relational feature of *content*. Relationism holds

---

[15] This kind of generalization is inspired by the explanations offered in (Heck, 2012). It differs from theirs insofar as it appeals to *semantic*, rather than *formal*, coordination. The significance of this distinction is discussed in (Almotahari, 2013) and (Heck, 2014*b*).

that coordination is a feature of content, but it isn't determined by the distribution of any non-relational features of content.

Semantic Relationists owe us an account of semantic coordination, just as Fregeans owe us an account of sense. Here we won't dive into the details of either approach, as our goal is to point out that the structural motivations for either, *vis-a-vis* Russellianism, are less robust than one might expect. But we should say a little about how the Relationist conceives of semantic coordination as this approach may be less familiar.[16]

Fine (2007, 2010) characterizes semantic coordination in two related ways. According to one, semantic coordination is understood in terms of a modest extension of the Russellian framework (2007, ch. 2-3). The Russellian holds that the intentional content of an attitude state is a set of Russellian propositions. Russellian propositions, themselves, are structured occurrences of objects, properties, and semantic operations. The Semantic Relationist posits extra structure at this level. Along with a set of Russellian propositions, there is a *coordination scheme*: an equivalence relation on occurrences of individuals in Russellian propositions. A coordination scheme encodes, at the level of intentional content, the transparent sameness of certain objects of thought. So Frederica's attitudes have the same Russellian content in Case 1 and Case 2, but different coordination schemes. In Case 1, but not in Case 2, the occurrences of Stefani/Gaga in the content of the belief that Stefani/Gaga is unhappy and in the desire that Stefani/Gaga be happy are related by the coordination scheme operative in Frederica's attitudes. The 'relationism' of this account lies in the fact that the coordination scheme of a subject's attitudes is not determined by any non-relational semantic feature of their attitudes (for example, it's not determined by which Russellian propositions they have attitudes towards). At the level of intentional content, coordination is irreducibly relational.

Fine sometimes talks as if this way of characterizing semantic coordination is merely an expository convenience (2010, pp. 73-74). He has a deeper commitment to the idea that the fundamental representational facts about an attitude state are to be characterized in terms of the notion of a 'semantic requirement'. The crucial fact about semantic requirements, in this context, is that it can be a semantic requirement on Frederica's attitudes that the relevant desire be about Stefanie/Gaga and a semantic requirement that relevant belief be about Stefanie/Gaga, while failing to be semantic requirement that these attitudes *be*

---

[16] For a fuller account, see (Fine, 2007; 2010) and (Gray, 2017).

about the same person (2007, p. 45ff). That extra, purely relational, semantic requirement is independent of the first two. It's present in Case 1 and not present in Case 2. On this way of thinking, talk of coordination schemes on the Russellian content of an agent's attitudes is a convenient shorthand for a description of the relational semantic requirements on their attitudes.

But the crucial idea here, again, is that there can be *relational* representational features of Frederica's attitudes that are not determined by any *non-relational* representational features. This allows the Relationist to more perspicuously represent the difference between Frederica's attitudes in Case 1 and Case 2; the difference, at the level of content, that makes a difference between those attitude states is purely relational. Relationism allows us to incorporate that relational difference directly, without having to posit non-relational features of intentional content (*i.e.* senses) whose only role is to ground the relevant relations.

Returning to Frederica, the Relationist line is straightforward: in Case 1, Frederica's situation is subsumed, and thus explained, by (C). In Case 2, Frederica doesn't satisfy the antecedent of (C), so she poses no threat to its truth.

Thus our framework enables a simple argument for Relationism:[17] Explanationism demands that, if the individual nature of senses makes no contribution to our understanding of rational agency via intentional generalizations, then senses aren't aspects of content. So Relational Fregeanism collapses into Semantic Relationism. One can see Relationism as a middle position between Russellianism and Fregeanism. The Relationist takes mental content to have a finer granularity than the Russellian takes it to have, but a coarser granularity than the Fregean.

Like the Russellian, the Semantic Relationist can allow that the full story about the implementation of intentional content will involve non-intentional entities. Just as the Russellian holds that an adequate metasemantics for thought might appeal to symbols in Mentalese, the Relationist might hold that an adequate metasemantics for semantic coordination might appeal to patterns of Mentalese symbol recurrence. What the Relationist insists on is that none of the underlying structure is directly relevant to intentional generalizations: the underlying structure *implements* the intentional patterns, but does not constitute them. And the takeaway from this discussion is that the Relational Fregean is in no position to

---

[17] This argument shows the influence of (Taschek, 1995; 1998), (Fine, 2007), (Pinillos, 2011; 2015), (Heck, 2012), and (Pryor, 2016*a*; 2016*b*).

disagree. So, to reiterate, Relational Fregeanism collapses into Semantic Relationism.

5. Inegalitarianism

We made explicit some assumptions about intentional explanation in Section 1 and put them to work in Sections 2-4. In this section, we'll introduce two further assumptions about intentional explanation that will have a role in our argument against Semantic Relationism.

The first is implicit in what we've already said; in particular, it's implicit in the idea that an explanatory scheme should be perspicuous with respect to real patterns at some level. The assumption is that not all true generalizations are explanatory. This should not be controversial. Compare (3a) and (3b).

> (3) (a) *Ceteris paribus*, metal conducts electricity.
>   (b) *Ceteris paribus*, pennies conduct electricity.

They are both true. But if we're interested in explaining the conductivity of some particular penny—or we're interested in the joints visible at the level of material science—(3a) is superior to (3b). Though it's true that pennies conduct electricity, they don't do so *because* they're pennies.

The concept *penny* is too fine-grained to characterize a joint in material science (though it's the right grain to characterize some joints in relation to economics, we presume). Accounts have been offered about how to tell when a true generalization has the right level of grain to count as an explanation (Yablo, 1997). And, in earlier work, we leveraged one such account to argue against Fregeanism (Almotahari & Gray, 2021).

Everyone should accept that not all of the true generalizations about some domain identify the joints in that domain. So the Russellian or Relationist might acknowledge the *truth* of generalizations like (F) but simply insist that (F) is roughly analogous to (3b), in that it contains more detail than is required to carve the intentional joints.

It might appear odd to suggest that the Russellian (or Relationist) could acknowledge the truth of (F). After all, (F) mentions senses. Can a Russellian really accept the existence of *senses*?

We've framed the point in an admittedly provocative way. If it's definitional of *sense* that senses are parts of intentional content, then the Russellian can't accept

that generalizations like (F) are true. But this is just terminology. The important point is that the Russellian can accept that there are true generalizations that operate at the level of granularity that the Fregean posits. She can resist the claim that the existence of these generalizations motivates Fregeanism by claiming that although the generalizations are true, they are not at the right level of grain to be explanatory.

This much has already been recognized in the literature on intentional content. What has *not* been recognized, as far as we know, is a further claim: not all *explanatory* generalizations that involve intentional phenomena can be taken as joint-carving with respect to intentional content. We must distinguish between (what we call) *canonical* and *non-canonical* intentional explanations.

As before, we'll work our way towards this by starting with discussion of explanations in other domains. We assumed above that *ravenhood* is a joint at a level of explanation that seeks to unify the connection between an open-textured range of genetic features and a class of phenotypic features characteristic of normal ravens. Recall (1).

(1) (a) *Ceteris paribus*, if x is a raven, then x is black.
    (b) If x has genetic features $g_1...g_n$, then x is black.

Our claim was that there's an important connection between the exception-tolerant nature of (1a)—that there are genetically abnormal ways of being a raven that are unthreatening exceptions to it—and the idea that *ravenhood* is a joint in relation to a particular level of explanation. The idea was that framing phenotypic generalizations at the level of *species* allows one to capture that a variety of different generalizations overlap and are mutually held in place by a process that connects them (*i.e.* natural selection).

We have thus been assuming a constitutive connection between the *natural kinds* at some level and the explanations that govern that level. Roughly, we've been assuming that the kinds are just those categories that figure in the relevant explanations.[18] In the case at hand, we're assuming that species are kinds at a level of explanation that's unified by the role that natural selection plays in generating and holding fixed clusters of phenotypes.

---

[18] "Natural kinds are the sets that one picks out in giving explanations. They are the sets corresponding to predicates that figure in our explanatory schemes" (Kitcher, 1984, p. 315, fn. 11). "Explanation guides us to what kinds there are" (Burge, 2010, p. 315).

Clearly, though, we need a nuanced picture of the relationship between explanations and kinds. Compare (1a) and (4):

(1) (a) *Ceteris paribus*, if x is a raven, then x is black.
(4) *Ceteris paribus*, if x is an albino raven, then x is white.

Raven is a species. And (1a) is an explanatory generalization. *Albino raven* is not a species. Are we then forced to say that (4), though true, is not an explanatory generalization? To perspicuously capture the pattern expressed in (4), must we replace talk of albino ravens with talk of genetic features?

That would be an unwelcome result. (4) possesses some of the features that make the exception-tolerant (1a) preferable to its lower-level alternative. Albino ravenhood, just like ravenhood *simpliciter*, comes in genetically different varieties. If we replaced (4) with a purely lower-level generalization, it would be massively disjunctive and would fail to register that, aside from their whiteness, albino ravens' phenotypes cluster in just the way that ravens' do. And the process by which albino ravens get their color is, in theoretically important respects, much like the process by which ravens do (it's not as if someone is going around painting albino ravens). We might say that albino ravens participate in the raven pattern in a parasitic way.[19]

So we should hold that (4) is an explanatory generalization. But, then, why is *albino raven* not a species? Why does its participation in an explanatory generalization analogous to (1a) not earn it that honor? Or, to put it another way, in what sense does it participate in the phenotypic patterns only in a parasitic way?

The answer here isn't surprising. (4) doesn't express a regularity of the right kind. (1a) expresses a regularity that's produced and explained by selection. (4) does not. The unifying principle that led us to identify a level of explanation, and corresponding kinds, in relation to (1a) and (2a) is the role that selection plays in sustaining and connecting those generalizations. The pattern identified in (4) isn't the result of selection in the way that (1a) and (2a) are (Nickel 2010). It doesn't stand in the right relation to the principles that unify the level of explanation exhibited by (1a), (2a), *etc.* We'll say that (4) is *non-canonical* with respect to the species-phenotype explanatory level.

---

[19] Here, again, we're following (Nickel, 2010). He notes that the truth of the generic, 'Albino ravens are white', is constitutively connected to the fact that the mechanism that generates the whiteness of albino ravens 'piggybacks' on the mechanism that generates the phenotypic properties of normal ravens (p. 21).

To sum up: not all explanatory generalizations at a certain level are on a par with respect to the kinds at that level. There can be privileged kinds—in this example, species—that are identified by their participation in canonical generalizations—in this example, generalizations that stand in the appropriate relation to selection.

There are explanatory generalizations about ravens; but there are also explanatory generalizations about *genetically abnormal* ravens. We'll call generalizations like (4), *mixed-level* generalizations. They are generalizations about what happens when kinds at one level are implemented in an abnormal way at a lower level. If we like, we can think of *albino raven* as a mixed-level kind (and thus not a species).

Returning to intentional explanation, we'll follow our practice of assuming that what goes for other high-level explanations goes for intentional explanation. So we'll assume:

(*Inegalitarianism*)   Only some explanatory intentional generalizations—the "canonical" ones—are content-individuating.

This is just an application of the lesson of this section to the case of intentional explanation. The interest will come in offering a story about what the canonical intentional explanations are and deciding what that means for theories of content.

6. Rationalizing Explanation

To carry out the Explanationist project, we need to decide *which* explanatory intentional generalizations are canonical. We should expect that some explanatory generalizations that mention intentional states will be mixed-level generalizations. Thus, some of the features mentioned in those explanations won't be intentional features. But what principle distinguishes canonical intentional explanations from mixed-level ones?

Our goal in this section is to identify a plausible answer to this question. We don't claim that it's the only possible answer. But it's pretty standard. And it's one that doesn't prejudge the debate between Russellians and their rivals. Consider the following passages:

[...] any actions for whose utility p is a necessary and sufficient condition might be called a belief that p (Ramsey 1927, p. 159).

The right assignment of content to Macbeth's states will be the one [...] that does best at assigning contents that rationalize behavior, according to the principles of common sense psychology (Lewis 1983, p. 375).

To believe that P is to be disposed to act in ways that would tend to satisfy one's desires, whatever they are, in a world in which P (together with one's other beliefs) were true (Stalnaker 1984, p. 15).

[...] psychological explanation is a matter of showing, roughly, that the subject thought and acted as *he ought to have done*, given the conditions prevailing (Campbell 1987, p. 291).

[...] the intentional stance [...] treats the system whose behavior is to be predicted as a rational agent; one attributes to the system the beliefs and desires it ought to have, given its place in the world and its purpose, and then predicts that it will act to further its goals in the light of its beliefs (Dennett 1988, p. 496).

[...] whatever is represented explains something about the person representing it. And what it explains is primarily the success of the actions that the person bases upon the representing (Blackburn 2005, p. 22).

We point to representations to explain how organisms and other systems manage to interact with their environment in useful and intelligent ways. The explanandum is a pattern of successful behaviour of a system in its environment (Shea 2018, p. 22).

[...] the correct interpretation of an agent is that which best rationalizes the agent's actions given their evidence (Williams 2020, p. 3).

[...] the correspondence of a representation to what it represents must be a cause of the usefulness of the representation. It must figure in a causal

explanation of success not just in a statistical explanation (Millikan 2023, p. 2).

Readers can, no doubt, supply additions to this list. But, even as it stands, the list shows the reach of an influential perspective, both in time and across very different philosophical orientations.

We don't claim that all of these passages express exactly the same idea. But they do express a fundamental sort of consensus: what's characteristic of the attitudes—or perhaps even of mental representations more broadly—is that they can participate in an explanation of a subject's actions that reveals why the agent did something that had a good chance of succeeding (at least relative to their goals). In a slogan: intentional explanation is *rationalizing* explanation. We'll assume, then, that the canonical intentional explanations are the rationalizing ones.[20]

Take two intentional generalizations that involve the same content:

(R)   *Ceteris paribus*, if x has a desire with the referential content that y is happy, and x has a belief with the referential content that y is unhappy, x will try to cheer y up.

(U)   *Ceteris paribus*, if x has a desire with the referential content that y resent x, and x has a belief with the referential content that y is unhappy, x will mock y's unhappiness.

Both are intentional patterns about (among other things) what subjects who believe someone to be unhappy do. We're assuming that these patterns are implemented in states with non-intentional features. But by framing the pattern at the level of intentional content, not only do we massively simplify the statement of the regularity, we capture the fact that they overlap in a way that's relevant to the processes that generate and sustain them. If the underlying implementation of

---

[20] This is another place where we're inspired by the spirit, though not the letter, of Fodor (1994). Fodor's argument for his brand of Russellian Exceptionalism crucially relies on the claim that "no remotely acceptable intentional psychology could count [the fact that rational behavior tends to succeed] as accidental" (p. 41). This isn't exactly the claim that we assume, and Fodor makes no mention of a distinction between canonical and non-canonical intentional generalizations. We'll also mention that we aren't sure, ourselves, exactly how Fodor's argument is supposed to go. To the extent that we understand it, we're sympathetic to the criticisms in (Arjo, 1996), (Segal, 1997), and (Aydede & Robbins, 2001).

believing that *x* is unhappy were such that, in conjunction with desiring that *x* be happy, it tended to generate behaviors that got the agent something they wanted, but that in conjunction with other desires, or other beliefs, the system generated behaviors that were random, or counter-productive, the relevant patterns would not be sustained. The patterns expressed by (R) and (U), along with indefinitely many other intentional patterns, are generated and sustained by the way they lead to agential success.

Several clarifications are in order. First, we're happy to work with a capacious sense of both 'did something' and 'success'. The quoted passages tend to focus on the role that attitudes play in explaining behaviors that are likely to bring about the satisfaction of needs or desires. But we're happy to think of the relevant 'actions' as including inferential transitions, emotional reactions, or other things besides. These are all things that a subject's attitudes can give them a reason to do, so we're happy to include them in the canonical explanations. Canonical intentional generalizations capture patterns in how rational agents behave so as to satisfy their needs and desires, acquire true beliefs on the basis of their evidence, form appropriate emotions based on their situation, *etc.* (If someone wants to work with a narrower sense of rationalizing explanation, we would be happy to go along. That would make our argument in the next few sections more straightforward.)

Second, the assumption that canonical explanations are rationalizing doesn't entail that canonical explanations can never explain a subject's failure. Take (R); that a subject's behavior is subsumed by (R) is no guarantee that they'll get something they want or need. After all, the relevant belief might be false, and thus there might be no cheering up to be done. Or the attempted cheering up might not come off. The claim that (R) is rationalizing is the claim that for someone whose behavior is subsumed by it, if their belief is true, it will be no accident if they get something they want or need.[21] Canonical explanations can explain failures, but for that to happen the failure must be subsumed under a generalization that explains success in good cases.

Finally, note that our assumption is importantly weaker than some related assumptions. Some theorists assume that the content of a belief is to be *identified* with the conditions under which actions based on the belief would be successful

---

[21] This is slightly too strong. A *cp*-generalization can have *deviant satisfiers* (for example: an albino raven that's painted black is subsumed by (1a) but its blackness is not explained by its ravenhood). For an individual that's subsumed by a *cp*-generalization to be such that their satisfaction of the antecedent explains their satisfaction of the consequent, they must satisfy the antecedent in a normal way.

(Millikan 1984, 1989). This is clearly a related idea, but is significantly stronger. It *presupposes* that the content of a belief is a truth-condition and so rules out non-Russellian approaches to content by *fiat*. We don't want to do that. We want to start with a weaker assumption—that canonical intentional explanations are rationalizing—and argue from there to the conclusion that non-Russellian approaches to content are on the wrong track.

On the face of it, there's nothing in our assumption that the Fregean or Relationist should balk at. In fact, the Campbell quote above comes from an explication of his brand of Fregeanism (and the Williams quote is from a work in which he endorses Fregeanism about attitude content). What Fregeans and Relationists should think, if they accept our elaboration of the Explanationist framework, is that generalizations like (R) are too coarse-grained *qua* rationalizing explanations; so intentional content is more fine-grained than referential content.

7. Relationism and Russellian Exceptionalism

In Section 4, we argued that Relational Fregeanism collapses into Relationism. In this section, we'll compare Relationism and Russellianism. We'll argue that, perhaps contrary to expectations, the Russellian does a better job carving the joints of canonical intentional generalizations. (We'll mention, in case you were unconvinced by the argument in Section 4, that the argument in this action applies to Relational Fregeanism too. To see how that argument would go, just replace talk of semantic coordination below with talk of sameness of sense).

Compare the vocabulary in which the Russellian and the Relationist can characterize intentional patterns: the Russellian can only appeal to referential content; the Relationist can, in addition, appeal to semantic coordination and its absence—call coreference without coordination *accidental coreference*. So our question is: does the extra grain in the Relationist vocabulary carve any rationalizing joints? Or is the Russellian vocabulary sufficient?

We'll take the question in two stages. First, do we need semantic coordination to carve any rationalizing joints? Recall (R) and (C).

(R)   *Ceteris paribus*, if x has a desire with the referential content that y is happy, and x has a belief with the referential content that y is unhappy, x will try to cheer y up.

(C)   *Ceteris paribus*, if x has a desire with the referential content that y is

25

happy, a belief with the referential content that y is unhappy, and the representation of y in x's desire is semantically coordinated with the representation of y in x's belief, x will try to cheer y up.

They exhibit the general relation between Russellian and Relationist principles. They are, ultimately, descriptions of the same intentional pattern. They differ with respect to whether coordination is understood as a background normality condition for possession of the coreferential attitudes specified in the antecedent, or as a relational feature of their content.

 This will be the general structure of the difference between Relationist and Russellian generalizations. For an intentional pattern that involves coreference at some positions—because its giving an agent a reason to do something constitutively depends on the attitudes' being about the same object—the Russellian will think that coordination is the normal way that an agent will come to possess those attitudes. We'll show that there is reason to think that the Russellian is right about the normality conditions here.

 Generalizations like (R) aim to capture the conditions under which the referential content of an agent's attitudes about an object rationalizes a particular action, partly in virtue of the fact that the attitudes are coreferential. The Russellian Exceptionalist claims that when an agent has attitudes like that, it's normally the case that the relevant representations are coordinated. Our argument for this claim has two parts: (*i*) having attitudes about an object o that rationalize a particular action, partly in virtue of being coreferential, is typically the result of o-directed *inquiry* (of a certain sort), and (*ii*) o-directed inquiry (of that sort) typically issues in coordinated attitudes about o.

 We'll begin by defending (*i*). Simplifying substantially, when one has attitudes about o that rationalize a particular action, partly in virtue of being coreferential, one has some beliefs about o, and some motivational states about o, that fit together in the right way. For example, if the relevant action is a bit of behavior, then the motivational state might be a desire with respect to o and the belief might be an instrumental belief about how o would have to change to satisfy the desire. If the relevant behavior is an inference, the motivational state might be a question about o and the beliefs might be information about o that's relevant to settling the question.

 Of all the attitudes that one *could* have about o, only a privileged subset would rationalize any particular action. Suppose that Frederica believes that Stefani

likes tulips and desires that Stefani become President. What should these attitudes lead Frederica to do? Nothing in particular. Getting in a position where one has attitudes whose referential content identifies a way to act that has a good chance of getting something one wants is no easy task. When you're in that position, it's not usually a matter of luck. It's usually a matter of inquiry.

We mean 'inquiry' here in a quite loose sense. Although we won't offer a definition, the idea should be familiar. One's thought about objects is structured into bundles of coordinated attitudes and connected recognitional capacities. Following an established tradition, we'll call such bundles *mental files*.[22] Frederica has a coordinated body of attitudes that she would express with 'Stefani' that's connected to various cognitive capacities (to recognize her neighbor by sight, to recognize her voice on the phone, etc). She has a distinct body of attitudes that she would express with 'Gaga' that's connected to other capacities (to recognize the singer in her elaborate costumes, to recognize her singing voice, and so on). Call these, respectively, her 'Stefani'-file and her 'Gaga'-file. Inquiry, as we understand it, is an activity in which an agent leverages the elements of an *o*-file to get themselves in a position to achieve one of their *o*-directed aims.

To see the idea, let's spell out Case 1 a bit more. How did Frederica acquire the attitudes in question? Here's a realistic possibility: because they were friends, Frederica had an antecedent interest in Stefani's happiness. And this interest gave her a reason to deploy the attitudes and recognitional capacities in her 'Stefani'-file to discover facts about Stefani's current state of mind (or, weaker even, to be sensitive to available evidence about her state of mind). She dropped by her neighbor's house and chatted with her. In doing this, she deployed many aspects of her 'Stefani'-file (knowledge of Stefani's address, knowledge of Stefani's typical affect and conversational style, knowledge of Stefani's history, etc). She thereby acquired the belief that Stefani is unhappy. She thus acquired a reason to act.

Contrast this with Case 2. We can imagine, again, that Frederica had an antecedent interest in Stefani's state of mind. But she learns about Gaga's unhappiness by hearing a report about the singer on the radio. So she now has a belief with referential content that, in conjunction with her desire that Stefani be happy, gives her a reason to act. Crucially, though, the acquisition of this belief was

---

[22] See (Evans, 1982), (Perry, 1980), (Recanati, 2012) and many others for this idea. Note that we are *not* endorsing the theoretical machinery deployed by some file theorists. For us, talk of files is just a convenient shorthand for referring to a body of coordinated attitudes along with some functionally integrated epistemic capacities. See (Goodman & Gray, 2022) for a defense of this way of unpacking the file metaphor.

cognitively unrelated to the attitudes and capacities associated with her 'Stefani'-file (the file which contains the relevant desire). In this case, discovering that Gaga was unhappy was, with respect to Frederica's antecedent interest in Stefani's happiness, purely serendipitous. It wasn't the result of inquiry that deploys the 'Stefani'-file (in our sense of 'inquiry').

Our claim is that Case 1 is the normal way of acquiring coreferential attitudes that rationalize an action partly in virtue of their coreferentiality and Case 2 is an abnormal way. Here we don't mean 'normal' in the statistical sense (though the claim is probably true in that sense), but rather in the normative sense: *the characteristic result of proper functioning.* It's part of the internal psychological momentum of 'files' that they develop rationalizing attitudes when such are available to be had. In contrast, if one has distinct files about o—that is, one has uncoordinated bodies of attitudes that are, unbeknownst to one, coreferential—there's no mechanism that generates beliefs in one file and desires in the other whose referential content jointly rationalizes any action. If one ends up with attitudes, distributed across distinct o-files, that jointly rationalize an o-directed behavior, this can only be an accident (assuming, of course, that there is no explicit judgment of identity linking the files).

Having argued in this way for (i), (ii) should be uncontroversial. It merely states that inquiry guided by a particular file typically produces attitudes that are coordinated with that file. When Frederica deploys her knowledge of Stefani's address, and her knowledge of Stefani's typical mannerisms, to go next door and discover that Stefani is unhappy, the result is a belief that is coordinated with the attitudes in the 'Stefani'-file (in particular, with her desire that Stefani be happy). If this were not how things typically went, there would be no point to inquiry, because its results would not typically lead an agent to perform the action that the content of their attitudes rationalized. In brief: *coordination begets coordination.* (It's also worth mentioning that, dialectically, Fregeans are not in a position to deny (ii). It's central to non-descriptive approaches to sense that attitudes that are the upshot of the kind of cognitive process we're describing are typically coordinated.[23])

We've now done what we did not do when we introduced Russellian Exceptionalism in Section 2—we've offered a reason to think it's correct in its claim about the normal conditions for possessing attitudes whose coreferentiality is, as

---

[23] See, among others, (Evans, 1982), (Campbell, 2002), and (Recanati, 2012).

such, relevant to rationalizing explanation. One might wonder exactly what that means about Relationist generalizations. We'll return to that question shortly.

First, though, we'll finish the comparison between Russellian and Relationist vocabularies. Recall that the Relationist can also appeal to coreference in the *absence* of coordination—that is, *accidental* coreference. Are there any rationalizing patterns that appeal to accidental coreference? Does accidental coreference carve any intentional joints?

Once we get this question clearly in view, the answer is obviously 'no'. An intentional joint that was characterized in terms of accidental coreference would involve a kind of action that was rationalized only when an agent had coreferential attitudes such that she was unaware of their coreferentiality. There are no forms of rationalization like this. If a subject's attitudes are coreferential but not coordinated, it is, from the perspective of rationalizing explanation, as if they weren't coreferential at all. More precisely: accidentally coreferential attitudes can rationalize action only in virtue of being subsumed by generalizations that also subsume non-coreferential attitudes.

To give an example, recall Case 2: Frederica believes that "Gaga" is unhappy and wants "Stefani" to be happy, but is unaware that Stefani is Gaga. Those accidentally coreferential attitudes *can* jointly rationalize some actions. Suppose that Frederica also believes that "Stefani" is mean-spirited, and likes to hear about celebrities being miserable. In that case, Frederica's attitudes about Stefani/Gaga give her a reason to call up Stefani to tell her about Gaga's unhappiness. Note, though, that the relevant generalization is something like (M).

> (M)   *Ceteris paribus*, if x has a desire with the referential content that y is happy, x has a belief with the referential content that y is mean-spirited, and x has a belief with the referential content that z is unhappy, x will tell y about z's unhappiness.

The crucial thing here is that this generalization does not involve coreference at the position of the person who is desired to be happy and the person who is believed to be unhappy (it *can* be instantiated by attitudes that are coreferential at those positions, but needn't be). Although Frederica instantiates this pattern in virtue of attitudes that are accidentally coreferential in those positions, that is irrelevant to the rationalizing role of those attitudes with respect to her action. She

would have instantiated the same pattern had she believed, instead, that Miley Cyrus was unhappy and told Stefani about it.

Returning to the larger question: does the Relationist use the extra structure in their vocabulary to identify any canonical intentional patterns that are not visible in the Russellian vocabulary? With respect to accidental coreference, the verdict is clear: we don't need it. With respect to semantic coordination, the situation is less straightforward. We have reason to think that coordination *is* the normal way of possessing coreferential attitudes that, as such, rationalize behaviors. So (C) foregrounds an element of the intentional explanation that ought to be backgrounded. We might say, then, that (C) makes something explicit that (R) leaves implicit and is, then, not a genuine competitor to it. And if we want to carve intentional patterns at the joints, thereby highlighting the differences that make a difference to intentional explanation, we would do better to characterize the patterns directly in terms of coreference. So the austerity of the Russellian vocabulary speaks in its favor: canonical intentional patterns can be characterized with referential content alone.[24]

8. Non-Russellian Explanations

We'll note one more important feature of Russellian Exceptionalism: if, as we've argued, canonical intentional explanations are rationalizing, and referential content is sufficient for characterizing rationalizing patterns, we should still expect the existence of mixed-level intentional explanations. These would be genuine explanations of what happens when subjects instantiate intentional states in ways that are abnormal with respect to some underlying level. Seeing that this is not only consistent with Russellianism, but predicted by it, allows us to see how the view avoids some problems it might be thought to founder upon.

---

[24] It would be natural to wonder whether the considerations we've adduced against fine-graining intentional content might be applied to our own preferred conception of such content (individuated in terms of the objects and properties that an attitude is about). Proponents of possible-worlds content will view our theory in much the same way that we view Fregeanism and Relationism: as illicitly incorporating features of the underlying implementation of intentional states into the contents of those states. Discussion of this issue would take us too far afield of the dialectic of this paper. Suffice to say, for now, that we would not, in principle, be hostile to an argument of this form. We doubt, though, that the proponent of possible-worlds content could incorporate much of our discussion. Our claims about the relationship between normal and abnormal implementations of intentional states make ineliminable reference to the referential properties of intentional content. We don't see how we could even state those claims in the idiom of possible-worlds content. This is much the same reason that (Heck, 2012) gives for appealing to Russellian content in their Relationist theory.

To see this, we'll rehearse a criticism that Heck (2012) levels against the Russellian.[25] Consider an extension of Case 2. Frederica wants "Stefani" to be happy. At time t_1, she hears on the radio, and so believes, that "Gaga" is unhappy. Those two attitudes don't lead her to do anything. Then, at time t_2, she goes next door and learns that "Stefani" is unhappy. She tries to cheer Stefani up. Call this Case 3.

Can we explain Frederica's behavior at t_2 by subsuming her under (R)? The answer appears to be less than straightforward. She instantiates it. And the attitudes in virtue of which she instantiates the antecedent give her a reason to do what she did. But, notice, she already instantiated the antecedent of (R) at t_1 and yet did nothing. So how could her instantiation of (R) explain her behavior at t_2 if she was in the same state at t_1 and did nothing?

Really, though, this isn't so puzzling from the perspective of Russellian Exceptionalism. Frederica went from satisfying the antecedent of (R) in an abnormal way to satisfying it in the normal way. That's what changed between t_1 and t_2. Given our commitments, it should come as no surprise that such cases exist.

Admittedly, the structure of the case means that questions about explanation are slightly delicate. If we ask: why did Frederica try to cheer up Stefani, we can simply point to the attitudes she possessed whose referential content gave her a reason to cheer up Stefani. In this sense, the explanation of why Frederica did what she did is the same in Case 1 and Case 3. (R) captures it.

But if we ask, why did Frederica cheer up Stefani at t_2 *rather* than at t_1? Or: what was the change at t_2 that led Frederica to cheer up Stefani *then*? The answer will be more complicated. We'll have to give the explanation by appealing both to the content of her attitudes and to facts about their implementation. That is, we'll need to subsume her under a mixed-level generalization, something like (R+).

(R+)   *Ceteris paribus*, if at t_1, x has a desire with the referential content that y is happy, and x has a belief with the referential content that y is unhappy, but those beliefs are uncoordinated, x will do nothing; but if at t_2, x acquires a belief with the referential content that y is unhappy that is coordinated with the desire, x will try to cheer y up at t_2.

---

[25] The view that Heck criticizes is not exactly the view that we're defending. Heck doesn't appear to endorse the specific Inegalitarian version of Explanationism that's guiding us, so it's not entirely clear what they would make of Russellian Exceptionalism. And there's a way in which their *Formal Relationism* resembles our Russellian Exceptionalism: coordination isn't an aspect of intentional content (which content, itself, can be characterized fully in referential terms). In any case, our view is significantly influenced by reflection on Heck's approach.

This generalization essentially involves subjects with attitudes that are accidentally coreferential (note the appeal to the absence of coordination).[26] But, consistent with the argument of the previous section, it's not a purely rationalizing explanation. It does more than simply explain why a subject does something by appealing to the attitudes that gave them a reason to do it. It also explains why a subject *doesn't* do something despite having attitudes with referential content such that, if it had been acquired via inquiry, would have led them to do it.[27]

(R+) is not in competition with (R) any more than generalizations about albino ravens are in competition with generalizations about ravens. And, just like that case, the explanatory power of (R+) is parasitic on the *rationalizing* pattern identified in (R) and characterized solely in terms of referential content.

9. Closing

Traditionally, intentional explanation has been thought to be unhappy ground for Russellian approaches to attitude content. It has seemed more or less obvious that referentially equivalent attitudes can play different roles in intentional explanation. There's a sense in which this is undeniable. But what that means for the individuation of content depends on assumptions about the nature of intentional explanation, and the role that content plays in it. If we make some plausible, though admittedly non-mandatory, assumptions, it looks like the Russellian can turn the

---

[26] It's worth noting that those who have replied to (Fodor, 1994) by offering intentional generalizations which the Russellian Exceptionalist can't capture have, in fact, pointed to generalizations about the conditions under which agents *do not* perform this or that action. This is the explicit theme of (Wakefield, 2002), but see also (Arjo, 1996, p. 236). From our perspective, this is telling. It amounts to admitting that non-Russellian intentional generalizations are non-canonical.

Wakefield offers an argument that explanations of non-action are a part of 'intentional psychology'. But the argument relies on the assumption that the domain of intentional laws is closed under entailment, which is independently implausible.

[27] As mentioned in footnote 5, our subject isn't the semantics of attitude ascription; it's the content of the attitudes. But, of course, the questions are related. The distinction we draw between canonical intentional explanations and mixed-level explanations fits nicely with a certain class of approaches to ascriptions. Many approaches hold that the truth-conditions of ⌜ x v's that p ⌝ are hybrid in an important sense: they're determined by the semantic content of p as well as by contextual information about non-intentional features of subjects' content-carrying states (Crimmins & Perry, 1989), (Fodor, 1990), (Richard, 1990), (Goodman & Lederman, 2021). This fits very nicely with our approach. We can hold that many attitude ascriptions are (at least implicitly) engaged in mixed-level explanation. The purely intentional aspects of the explanation are captured by the contribution of p to the truth-conditions of ⌜ x v's that p ⌝; the non-intentional aspects of the explanation are contributed by the contextually-determined constraints on non-intentional features of attitude states.

tables on her opponents. If the relation between intentional explanation and its lower-level implementation is analogous to what we find in other higher-level forms of explanation, there's good reason to think that the intentional joints can be carved using referential content alone.[28]

References

Almotahari, M. (2013). Flaws of Formal Relationism. *Thought* 2: 367-376.

Almotahari, M., & Gray, A. (2021). Frege Cases and Bad Psychological Laws. *Mind* 130: 1253-1280.

Arjo, D. (1996). Sticking up for Oedipus: Fodor on Intentional Generalizations and Broad Content. *Mind & Language* 11: 231-245.

Aydede, M. & Robbins, P. (2001). Are Frege Cases Exceptions to Intentional Generalizations? *Canadian Journal of Philosophy* 31: 1-22.

Blackburn, S. (2005). Success Semantics. In H. Lillehammer & D. H. Mellor (Eds.), *Ramsey's Legacy* (pp. 22-36). Oxford University Press.

Block, N. (1986). Advertisement for a Semantics for Psychology' *Midwest Studies in Philosophy* 10: 615-678.

Braun, D. (2000). Russellianism and Psychological Generalizations. *Noûs* 34: 203–236.

Braun, D. (2001*a*). Russellianism and Explanation. *Noûs* 35: 253–289.

Braun, D. (2001*b*). Russellianism and Prediction. *Philosophical Studies* 105: 59-105.

Braun, D. (2020). Still For Direct Reference. In A. Bianchi (Ed.), *Language from a Naturalistic Point of View: Themes from Michael Devitt* (pp. 193-215). Springer International Publishing.

Burge, T. 2010. *Origins of Objectivity*. Oxford University Press.

Campbell, J. (1987). Is Sense Transparent? *Proceedings of the Aristotelian Society* 88: 273-292.

Campbell, J. (2002). *Reference and Consciousness*. Clarendon.

Cappelen, H. & Dever, J. (2013). *The Inessential Indexical: On the Philosophical Insignificance of Perspective and the First Person*. Oxford University Press.

Chalmers, D. J. (2006). The Foundations of Two-Dimensional Semantics. In M. García-Carpintero & J. Macià (Eds.), *Two-Dimensional Semantics* (pp. 55-140). Oxford University Press.

Coyne, J. (2009). *Why Evolution is True*. Oxford University Press.

Crimmins, M. & Perry, J. (1989). The Prince and the Phone Booth: Reporting Puzzling Beliefs. *Journal of Philosophy* 89: 685-711

Dennett, D. (1988). Précis of *The Intentional Stance*. *Behavioral and Brain Sciences* 11: 495-546.

Devitt, M. (1989). Against Direct Reference. *Midwest Studies in Philosophy* 14: 206-240.

Evans, G. (1982). *The Varieties of Reference*. Clarendon.

Fine, K. (2007). *Semantic Relationism*. Wiley-Blackwell.

Fine, K. 2010. 'Semantic Necessity'. In B. Hale and A. Hoffmann, eds., *Modality: Metaphysics, Logic, and Epistemology*. Oxford University Press.

Fodor, J. A. (1990). A *Theory of Content and Other Essays*. MIT Press.

Fodor, J. A. (1991). You Can Fool Some of the People All of the Time, Everything Else Being Equal: Hedged Laws and Psychological Explanation. *Mind* 100: 19-34.

Fodor, J. A. (1994). *The Elm and the Expert: Mentalese and its Semantics*. MIT Press.

Goodman, J. & Lederman, H. (2021). Perspectivism. *Noûs* 55: 623-648.

Goodman, R., & Gray, A. (2022). Mental Filing. *Noûs* 56: 204-226.

Gray, A. (2017). Relational Approaches to Frege's Puzzle. *Philosophy Compass* 12: e12429.

Gray, A. (2022). Minimal Fregeanism. *Mind* 131: 429-458.

Heck, R. K. (2012). Solving Frege's Puzzle. *Journal of Philosophy* 109: 132–174.
    (Originally published under the name "Richard G. Heck, Jr.")

Heck, R. K.. (2014*a*). Intuition and the Substitution Argument. *Analytic Philosophy* 55: 1-30.
    (Originally published under the name "Richard G. Heck, Jr.")

Heck, R. K. (2014*b*). In Defense of Formal Relationism. *Thought* 3: 243-250.
    (Originally published under the name "Richard G. Heck, Jr.")

Jackson, F. (1998). *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. Oxford
    University Press.

Kitcher, P. (1984). Species. *Philosophy of Science* 51: 308-333.

Lewis, D. (1983). Individuation by Acquaintance and by Stipulation. *Philosophical Review* 92:
    3-32. Reprinted in D. Lewis, *Papers in Metaphysics and Epistemology*. Cambridge
    University Press. Page references to the reprinted version.

Loar, B. (1988). Social Content and Psychological Content. In R. H. Grimm & D. D. Merrill
    (Eds.), *Contents of Thought* (pp. 99-139). Arizona University Press.

Magidor, O. (2015). The Myth of the De Se. *Philosophical Perspectives* 29: 249-283.

Millikan, R. (1984). *Language, Thought, and Other Biological Categories*. MIT Press.

Millikan, R. (1989). Biosemantics. *Journal of Philosophy* 86: 281-297.

Millikan, R. G. (2024). Teleosemantics and the frogs. *Mind & Language*, 39(1), 52-60.

Nickel, B. (2010). *Ceteris Paribus* Laws: Generics & Natural Kinds. *Philosophers'
    Imprint* 10.

Peacocke, C. (1981). Demonstrative Thought and Psychological Explanation.
    Synthese 49: 187-217.

Perry, J. (1977). Frege on Demonstratives. *Philosophical Review* 86: 474-497.

Perry, J. (1980). A Problem about Continued Belief. *Pacific Philosophical Quarterly* 61:
    317-332.

Perry, J. (1990). Self Notions. *Logos* 11: 17-31.

Pinillos, N. Á. (2011). Coreference and Meaning. *Philosophical Studies* 154: 301–324.

Pinillos, N. Á. (2015). Millianism, Relationism, and Attitude Ascriptions. In A. Bianchi (Ed.),
    On Reference (pp. 322-334). Oxford.

Pryor, J. (2016*a*). De Jure Codesignation. In B. Hale, C. Wright, & A. Miller (Eds.), A
    Companion to the Philosophy of Language (pp. 1033-1079), second edition.

Pryor, J. (2016*b*). Mental Graphs. *Review of Philosophy and Psychology* 7: 309–341.

Pylyshyn, Z. (1984). *Computation and Cognition*. MIT Press.

Ramsey, F. (1927). Facts and Propositions. *Proceedings of the Aristotelian Society,
    Supplementary Volumes* 7: 153-170.

Richard, M. (1990). *Propositional Attitudes: An Essay on Thoughts and How we Ascribe Them*.
    Cambridge University Press.

Recanati, F. (2012). *Mental Files*. Oxford University Press.

Sainsbury, R.M. (2002). *Departing From Frege: Essays in the Philosophy of Language*.
    Routledge.

Salmon, N. (1986). *Frege's Puzzle*. Ridgeview Publishing Company.

Schneider, S. (2005). Direct Reference, Psychological Explanation, and Frege Cases. *Mind &
    Language* 20: 423-447.

Schneider, S. (2011). *The Language of Thought: A New Philosophical Direction*. MIT Press.

Segal, G. (1997). Content and Computation: Chasing the Arrows A Critical Notice of Jerry

Fodor's *The Elm and the Expert*. *Mind & Language* 12: 490-501.

Segal, G. (2000). *A Slim Book about Narrow Content*. MIT Press.

Shea, N. (2018). *Representation in Cognitive Science*. Oxford University Press.

Soames, S. (1990). 'Belief and Mental Representation'. In P. P. Hanson, ed., *Information, Language and Cognition*. University of British Columbia Press.

Stalnaker, R. (1984). *Inquiry*. MIT Press.

Strevens, M. (2012). Ceteris Paribus Hedges: Causal Voodoo That Works. *Journal of Philosophy* 109: 652-675.

Taschek, W. W. (1995). Belief, Substitution, and Logical Structure. *Noûs* 29: 71–95.

Taschek, W. W. (1998). On Ascribing Beliefs: Content in Context. *Journal of Philosophy* 95: 323–353.

Wakefield, J. C. (2002). Broad versus Narrow Content in the Explanation of Action: Fodor on Frege Cases. *Philosophical Psychology*, 15, 119-133.

Williams, J. R. G. (2020). *The Metaphysics of Representation*. Oxford University Press.

Yablo, S. (1997). Wide Causation. *Philosophical Perspectives* 11: 251–281.