ABSTRACT

Title of Dissertation: ART, FICTION, AND EXPLANATION

Moonyoung Song, Doctor of Philosophy, 2019

Dissertation directed by: Professor Jerrold Levinson
Department of Philosophy

This dissertation consists of four stand-alone chapters that address topics at the intersection of art, fiction, and explanation.

Chapter 1, “the nature of the interaction between moral and artistic value,” aims to elucidate what it means to say that a work’s moral virtue or defect is an artistic virtue or defect. I address this question by showing that the following two strategies commonly used to establish such a claim are not successful: (1) appealing to the counterfactual dependence of the work’s artistic value on its moral virtue or defect; and (2) arguing that the work is artistically valuable (or defective) and morally valuable (or defective) for the same reasons.

Chapter 2, “aesthetic explanation,” argues for the psychological account of aesthetic explanation (i.e., the explanation of the aesthetic by the non-aesthetic), according to which the presence of certain non-aesthetic properties explains the presence of a certain aesthetic property just when the observer’s experiences of the non-aesthetic properties cause their experience of the aesthetic property. I

demonstrate how this account illuminates the selectivity of aesthetic explanation—the phenomenon of aesthetic explanation citing only some of the non-aesthetic properties on which an aesthetic property supervenes—, drawing an analogy between the selectivity of aesthetic explanation and causal explanation.

Chapter 3, "the fictionality puzzle, fictional truth, and explanation," proposes that what is true in fiction is determined by inference to the best explanation. I show that this account of fictional truth provides a novel solution to the fictionality puzzle, which concerns why certain kinds of deviant claims, such as deviant moral claims (e.g., female infanticide is permissible), are difficult to make true in fiction, whereas other kinds of deviant claims, such as deviant scientific claims (e.g., time travel is possible), are regularly true in fiction.

Chapter 4, "aptness of fiction-directed emotions," argues that the criteria governing the epistemic appropriateness of emotions directed towards fictional entities are analogous to the criteria governing the epistemic appropriateness of emotions directed towards real entities. In both cases, an emotion is epistemically appropriate if and only if it is fitting, justified, and salience-tracking, and these notions are understood in analogous ways.

ART, FICTION, AND EXPLANATION

by

Moonyoung Song

Dissertation submitted to the Faculty of Graduate School of the
University of Maryland, College Park in partial fulfillment
Of the requirements of the degree of
Doctor of Philosophy
2019

Advisory Committee:

Professor Jerrold Levinson, Chair
Professor Harjit Bhogal
Professor Peter Carruthers
Professor Patricia Greenspan
Professor Steven Mansbach
Professor Dan Moller

© Copyright by
Moonyoung Song
2019

Dedication

For my parents.

Acknowledgements

I am deeply indebted to my supervisor, Jerrold Levinson, as well as to the rest of my committee members, Harjit Bhogal, Peter Carruthers, Patricia Greenspan, and Dan Moller. Many thanks are also owed to my colleagues, especially Julius Schönherr and Evan Westra, for their feedback on my work. I am also grateful for comments from anonymous reviewers and editors from the *Journal of Aesthetics and Art Criticism*.

Table of Contents

Dedication……...ii

Acknowledgements……iii

Introduction…… 1

1. The Nature of the Interaction between Moral and Artistic Value……. 8

2. Aesthetic Explanation……35

3. The Fictionality Puzzle, Fictional Truth, and Explanation……80

4. Aptness of Fiction-Directed Emotions……117

Bibliography……....143

Introduction

This dissertation consists of four papers about how we appreciate and interpret artworks. Each paper is concerned with a distinct aspect of the question: Chapter 1 addresses how the moral value of artwork influences its artistic value; Chapter 2 elucidates in what sense an artwork having certain non-aesthetic properties explains it having a certain aesthetic property; Chapter 3 proposes a theory on how fiction's narrative content is interpreted; Chapter 4 discusses what makes an emotional response to a fiction apt or inapt. These four chapters are independent of each other and can be read in any order. However, they revolve around the following two broad projects.

The first project is to elucidate the relationship between an artwork's artistic or aesthetic features and its non-artistic or non-aesthetic features. When we attribute an artistic or aesthetic feature to a work of art, judging that it is artistically good, beautiful, elegant, etc. (or artistically bad, ugly, gaudy, etc.), we think the work has these features *in virtue of* or *because of* its other features. Picasso's *Les Demoiselles d'Avignon*, for instance, is artistically good partly in virtue of the original way in which it represents three-dimensional objects seen from multiple viewpoints on a two-dimensional canvas. And Jackson Pollock's *Summertime* is dynamic because of its composition of lines and colors. But exactly in what sense does the originality of *Les Demoiselles d'Avignon* make the work artistically good, and does the composition of *Summertime* make it dynamic? In other words, what is the relationship between *Les Demoiselles*

d'Avignon's artistic value and originality, and between *Summertime*'s dynamism and composition, such that the works have the former feature in virtue of the latter feature?

This question can be divided into two: one about an artwork's *overall artistic or aesthetic value* and the other about its *aesthetic properties*. Aesthetic properties are properties such as dynamism, elegance, unity, sublimity, garishness, and gaudiness. Unlike completely descriptive properties, such as being red or having curved lines, aesthetic properties are evaluative. Elegance, for instance, is positive while garishness is negative. But aesthetic properties also have descriptive aspects that make them more specific than overall artistic or aesthetic goodness or badness. Being elegant, for instance, is more specific than being artistically/aesthetically good, and being garish is more specific than being artistically/aesthetically bad. Given this distinction between overall artistic/aesthetic value and aesthetic properties, we need to consider the following two questions separately: (1) in what sense an artwork is *artistically/aesthetically good or bad* in virtue of its other features and (2) in what sense an artwork has its *aesthetic properties* in virtue of its other properties. The first question is addressed in Chapter 1, and the second question in Chapter 2.

More specifically, Chapter 1 focuses on the question of whether an artwork can be artistically good or bad in virtue of its *moral* virtue or defect. The current literature on this issue is mostly concerned with defending one of the following three particular positions: an artwork can never be artistically good or bad in virtue of its moral virtue or defect; an artwork can be artistically good or

bad in virtue of its moral virtue or defect, and when it is so, the interaction is always in the same valence (i.e., a moral virtue always positively contributes to artistic value, and a moral defect always negatively contribute to artistic value); a moral virtue can sometimes negatively contribute to artistic value, and a moral defect can sometimes positively contribute to artistic value. In Chapter 1, I suggest we step back from this debate among the three positions, to reflect on what it means to say that an artwork is artistically good or bad in virtue of its moral virtue or defect. My answer to this question is mostly negative—I focus on showing what the claim does *not* mean rather than what it means. I argue that the claim does not mean that the work's artistic value counterfactually depends on the moral virtue or defect (i.e., the work would be more, or less, artistically valuable if it did not have the moral virtue or defect), or that the work's artistic value and moral value share the same basis. However, I briefly offer a positive suggestion as well: to say that an artwork is artistically good or bad in virtue of its moral virtue or defect is to say that the moral virtue or defect partly *grounds* or *explains* the work's artistic value.

Chapter 2 focuses on aesthetic properties rather than overall artistic/aesthetic value, discussing in what sense an artwork's s non-aesthetic properties explain its aesthetic properties. It is a commonplace in aesthetics that an object having certain non-aesthetic properties (e.g., having a certain shape) explains it having a certain aesthetic property (e.g., elegance). In other words, the object has an aesthetic property in virtue of or because of its non-aesthetic properties. This chapter aims to elucidate the nature of such explanations. I argue

that the fact that an object has certain non-aesthetic properties explains the fact that it has a certain aesthetic property just when under the right circumstances, the right observer's experiences of the non-aesthetic properties causes her experience of the aesthetic property. According this account, which I call the psychological account of aesthetic explanation, aesthetic explanation is a kind of causal explanation broadly construed—although there is no causal relation between an object having the aesthetic and non-aesthetic properties, the explanatory relation between them is based on the causal relation between corresponding psychological events. After arguing for the psychological account of aesthetic explanation, I then point out that aesthetic explanation is selective in the sense that the explanation of an object having a certain aesthetic property does not include all of the non-aesthetic properties on which the aesthetic property supervenes, and illustrate how the existing literature on the selectivity of causal explanation sheds light on the selectivity of aesthetic explanation.

Chapters 3 and 4 pursue a different project, which is to elucidate the continuities and discontinuities between fiction and reality. The world of fiction is certainly different from the real world. After all, the characters in most fictions do not exist in the real world, and we are perfectly aware of this fictionality when engaging with fiction. However, the world of fiction is continuous with the real world in some other respects. For instance, gravity exists in the world of *Anna Karenina* simply because it does in the real world, even though Tolstoy never mentions that. Furthermore, we often emotionally respond to fiction as if the characters and events were real. Hence a question arises as to in what ways fiction

is continuous and discontinuous with reality. I suggest that the discontinuity crucially lies in the way in which *what is true* in a fiction is determined. Truth in fiction is determined by our interpretation while truth in reality is independent of what we think is true. However, I suggest that when it comes to *our responses* to what is true in fiction and reality, there are significant continuities. In particular, I suggest in Chapter 4 that our emotional responses to fiction are governed by epistemic norms analogous to those that govern our emotional responses to reality.

In Chapter 3, I articulate the exact way in which what is true in fiction is determined interpretively. I propose that what is true in a fiction is determined through inference to the best explanation. According to this proposal, the text of a fiction amounts to data to explain, and the set of propositions that together best explain the text is the set of propositions that are true in the fiction. In this respect, truth in fiction is different from truth in reality. In reality, what is true is not *determined* through any form of inference, although inference, when it is good, may *track* what is true.

After defending this account of fictional truth, I also demonstrate in Chapter 3 how this way of understanding fictional truth solves the so-called fictionality puzzle, which is about why certain kinds of deviant claims, such as deviant moral claims (e.g., female infanticide is morally permissible), are difficult to make true in fiction, whereas other kinds of deviant claims, such as deviant scientific claims (e.g., time travel is possible), are regularly true in fiction. The reason, I argue, is that the former kinds of deviant claims are usually not part of

the best explanation of a fiction's text, while the latter kinds of deviant claims usually are.

While Chapter 3 focuses on a discontinuity between fiction and reality, Chapter 4 argues for a continuity between them. I argue that the criteria governing the epistemic appropriateness of emotions directed towards fictional entities, such as characters and events in fiction, are analogous to the criteria governing the epistemic appropriateness of emotions directed towards real entities. The criteria in the two domains, of course, are not completely identical, since fictional characters and events are merely fictional. However, the criteria in the two domains are analogous (that is, they are structurally identical) in the following sense: in both cases, an emotion is epistemically appropriate if and only if it is fitting, justified, and salience-tracking, and these notions are understood in in the same way across reality- and fiction-directed emotions, except that fictional truth and salience in the context of our engagement with fiction are determined in a different way from how truth and salience are determined in reality. To defend this claim, I consider and reject two arguments that suggest otherwise: (1) contrary emotions are epistemically appropriate towards the same object depending on whether the object is fictional or real; (2) different kinds of considerations count as justificatory reasons for fiction-directed and reality-directed emotions. I demonstrate that the seeming disanalogies to which these arguments appeal stem from the failure to distinguish between emotions directed towards the fictions' contents and emotions directed towards the fictions themselves, or from other sources of differences that do not count as differences

in the criteria for epistemic appropriateness *per se*, such as the different ways in which truth and fictional truth, and salience in fictional and real domains, are determined.

Chapter 1: The Nature of the Interaction between Moral and Artistic Value[1]

1. Introduction

Over the last couple of decades, there has been extensive discussion on the interaction between the moral and the artistic value of an artwork, in particular, how an artwork's moral value affects its artistic value.[2] Various theoretical positions have been proposed, including the following three: autonomism, the view that an artwork's moral virtue or defect is never an artistic virtue or defect; moralism, which argues that, when relevant to artistic value, a moral virtue is always an artistic virtue and a moral defect is always an artistic defect; and immoralism (also called contextualism or the anti-theoretical view), which argues that a moral defect (virtue) can sometimes be an artistic virtue (defect).[3]

The aim of this essay is not to defend a particular position in this debate. I instead aim to advance our understanding of the nature of value interaction itself, by addressing questions as to what it means that an artwork's moral virtue or defect is an artistic virtue or defect, and how we can prove or disprove such a claim.

---

[1] This chapter was originally published in the *Journal of Aesthetics and Art Criticism* (Song, 2018). It has been reprinted here with permission from Wiley.

[2] Interaction in the other direction—the effect of artistic value on moral value— is not the focus of this essay. Also note that this essay concerns the interaction between moral and artistic, not aesthetic, value (by artistic value I mean the value of artwork *qua art)*. However, I will sometimes have to use the terms 'artistic' and 'aesthetic' interchangeably due to the fact that the distinction is often blurred in the value interaction literature.

[3] This is just one rough way to divide up the literature. For different taxonomic suggestions, see Carroll (1996) and Giovannelli (2007). For autonomism, see Anderson and Dean (1998) and Harold (2011). For moralism, see, for instance, Carroll (1996; 1998; 2000), Gaut (1998; 2007), Giovannelli (2013), and Stecker (2005; 2008). Note that although Carroll is viewed by many as committed to the moralist thesis, he himself considers his claim to be weaker—a moral defect (virtue) is sometimes an artistic defect (virtue)—, which is compatible with immoralism (2000). For immoralism, see, for example, Jacobson (1997; 2005), Kieran (2003; 2005), and Eaton (2012; 2013).

I approach these questions in the following way. First, in section II, I distinguish between two kinds of value interaction: (1) a moral virtue or defect's *intrinsic* or *pro tanto* artistic valence that does not vary with context, and (2) a moral virtue or defect's contribution to a work's overall artistic value *in the context of that particular work*.

The rest of the essay then focuses on the second kind of value interaction, namely, interaction in a particular context. To better understand the nature of this interaction, I examine two strategies commonly used in the literature to establish such interaction. The first strategy assumes that whether an artwork's moral virtue or defect contributes to its artistic value is a matter of whether the artistic value counterfactually depends on the moral virtue or defect. The second strategy assumes that if a work is artistically valuable (defective) and morally valuable (defective) for the same reasons, then its moral virtue (defect) is an artistic virtue (defect). I will discuss each of these two strategies in sections III and IV, and show that they both fail to establish value interaction. Lastly, in section V, I discuss the implications of these failures, namely, whether they should incline us toward any particular position in the value interaction debate, and how we should approach value interaction claims given the failures of those two strategies.

2. Two kinds of value interaction

Claims about value interaction have been made in various forms, but I will begin my discussion with the following formulation (I may later introduce other formulations where necessary): a moral defect or virtue is an artistic defect or virtue.

I choose this formulation because it is probably the most commonly used, and is general enough to cover both kinds of value interaction I discuss below. This formulation also obviates a problem with the following comparative formulations: a moral defect or virtue can *increase* or *decrease* a work's artistic value; or a moral defect or virtue can make a work *better* or *worse* as art. These expressions are very common, but they are unclear as to what the work in question is being compared with.[4] Although the nature of this comparison raises an interesting issue, one might question whether value interaction should be understood in comparative terms in the first place. So, I will start my discussion with the non-comparative formulation.

What does it mean to say that a moral defect or virtue is an artistic defect or virtue? I believe there are two ways to understand this claim: (1) in terms of a moral defect or virtue's intrinsic artistic valence that does not vary with context, and (2) in terms of a moral defect or virtue's contribution to the work's overall artistic value due to the ways in which it interacts with other features of the work in the particular context. For simplicity, I will call the former kind of value interaction *intrinsic* value interaction and the latter kind *contextual* value interaction.[5] Note that a moral defect or virtue's having intrinsic artistic valence does not necessarily mean that the presence of a moral defect or virtue guarantees that the work is artistically good

[4] For discussion of this issue, see Stecker (2008, pp. 148-149).

[5] In principle, one of these two kinds of interaction neither entails nor excludes the other. (I add "in principle" because this issue may be affected by what specific positive account of each of the two interactions we should adopt, which is a question I do not pursue in this essay.) That is, a moral defect or virtue's having positive, negative, or neutral intrinsic artistic valence is in principle compatible with any of the following three possibilities: it (1) positively or (2) negatively contributes to the work's overall artistic value in the particular context of the work; or (3) it makes no such contribution at all. Thanks to an anonymous referee for this journal for pressing me to clarify the relation between the two kinds of value interaction.

or bad. A more plausible way to understand intrinsic value interaction is through the notion of *pro tanto* merit or demerit. Although a moral defect or virtue does not guarantee that the work is artistically good (or bad), it is a *pro tanto* artistic merit (or demerit) in that the work is artistically meritorious (or flawed) to that extent or in that aspect.

This distinction between intrinsic and contextual value interaction is not new. Berys Gaut has made it nicely. Gaut formulates his view, called ethicism, in terms of *pro tanto* principles, which concern intrinsic value interaction as I characterize it (2007, pp. 57-66). According to the *pro tanto* principle he endorses, a work is aesthetically meritorious (flawed) *insofar as* it has a moral virtue (defect) that is aesthetically relevant.[6] This principle invariably holds, he argues, regardless of how the moral virtue or defect affects the overall artistic value of the particular work in which it is present. Even when a moral virtue (defect) of a work negatively (positively) contributes to the work's overall artistic value in the particular context, the work still has some artistic merit (demerit) insofar as it has that feature.

Consider, for instance, *Triumph of the Will*, a Nazi propaganda film of artistic excellence. Gaut would say that the film is artistically defective insofar as it is morally defective, even if, as some immoralists argue, the moral defect positively contributes to the film's overall artistic value due to some features of the particular film, such as the inextricable connection between its beauty and its moral defect.

[6] Although Gaut uses the term 'aesthetic,' he claims that the aesthetic value of art is artistic value (2007, pp. 34-41).

This distinction between intrinsic and contextual value interaction is often blurred in the literature, and I believe we have good reasons to draw it more clearly. To begin with, as the above example suggests, a theory about intrinsic value interaction and a theory about contextual value interaction do not in principle compete with each other.[7] For instance, as Gaut himself notes, his ethicism, which concerns intrinsic value interaction, is compatible with immoralism (or contextualism as he calls it) about contextual value interaction (2007, p. 64). In fact, to my knowledge, no one has defended immoralism about intrinsic value interaction—perhaps because it is counterintuitive that a moral defect has intrinsic positive artistic valence. What immoralists claim is that *in some particular cases*, a work's moral defect positively contributes to its overall artistic value. There is no inconsistency in holding this view along with any view on intrinsic value interaction.

Furthermore, defending the two kinds of value interaction requires different kinds of arguments. To defend a claim about intrinsic value interaction, one would need to establish some kind of invariant connection between moral and artistic value independent of any particular context. This would require defending *pro tanto* principles in aesthetics against the objection that there are no such principles (This debate is directly related to the debate between generalism and particularism in aesthetics). In contrast, defending a claim about contextual value interaction

[7] I add 'in principle' because I admit that the debates on intrinsic and contextual value interaction might overlap at some point. For instance, it may turn out that the only way to make sense of *pro tanto* principles is to construe them in terms of contextual value interaction in all relevant cases. The possibility of such overlap, however, does not mean that we do not need to distinguish between the two kinds of value interaction in the first place.

requires appealing to the context of a particular work. For instance, Matthew Kieran defends his immoralism by arguing that *in the case of some artworks*, a moral defect contributes to the work's cognitive value, which in turn contributes to its artistic value (2003; 2005). He does not need to establish an *invariant* link between a moral defect and cognitive value; he only needs to show that a moral defect *sometimes* contributes to cognitive value due to the particular context of the work.

All these considerations suggest that there are two different debates on value interaction, concerning intrinsic and contextual value interaction. In the remainder of this essay, I focus only on contextual value interaction. The issue of intrinsic value interaction is no less important, but exploring it would take us too far afield, especially because we would need to delve into issues not traditionally considered part of the value interaction debate, such as the tenability of *pro tanto* principles.

3. Contextual value interaction and counterfactual dependence

My aim is now to advance our understanding of the nature of contextual value interaction, by examining common ways in which this interaction is construed in the literature. In particular, I will assess two specific strategies often used to defend contextual value interaction claims, and show that neither of them is successful. Although these two do not exhaust all the strategies available in the literature, carefully examining them and seeing why they fail will help us make some progress toward a better understanding of the nature of contextual value interaction. I will discuss the first strategy in this section, and the second strategy in the following section.

While discussing these two strategies, I will use the following formulations to indicate claims about contextual value interaction: a work's moral defect or virtue *positively (negatively) contributes* to its overall artistic value; or the work is artistically valuable (defective) *in virtue of* the moral virtue or defect (these two claims are used interchangeably in the literature, and I will follow that practice). Note that these claims exclude cases in which a work with a moral virtue or defect is artistically valuable (defective) *despite* or *independently of* it. In such a case the moral virtue or defect itself makes *no* positive (negative) contribution to the work's artistic value, and it is not the case that the work is artistically valuable (defective), even partly, in virtue of the moral virtue or defect.[8]

Let us now look at the first strategy commonly used to establish claims about contextual value interaction. This strategy draws the conclusion that a work's moral virtue or defect positively (negatively) contributes to its artistic value from the truth of the following counterfactual conditional: if the work did not have the moral virtue or defect, it would be less (or more) artistically valuable. Thus this strategy assumes that counterfactual dependence is sufficient for value interaction. This assumption can be stated as follows:

> **Counterfactual Dependence (Sufficient Condition)**: A work's feature *x* positively (negatively) contributes to its overall artistic value if the

[8] A partial contributor is still a genuine contributor, although I will often omit the word 'partial' or 'partly' for simplicity.

following conditional is true: if the work[9] did not have *x*, it would be less (more) artistically valuable.

This assumption is frequently made in the literature, but not all cases of seeming appeal to it involve using it as a crucial part of an argument. I will later discuss one example in which one's argument crucially relies on **Counterfactual Dependence (Sufficient Condition):** Ted Nannicelli's immoralism.

Despite its intuitive appeal, **Counterfactual Dependence (Sufficient Condition)** is false, for two reasons: (1) features of an artwork are often interconnected and thus co-vary with each other in such a way that a work's artistic value counterfactually depends on its feature *x* without *x* contributing the artistic value; (2) **Counterfactual Dependence (Sufficient Condition)** does not distinguish genuine *contributors* to a work's artistic value from mere *necessary factors*.

(1) Let us start with the first reason, which I will call the problem of interconnection. It is worth noting that this problem does not concern the difficulty of *conceiving* a relevant counterfactual situation. It is often not conceivable in any concrete way what a work would be like if its moral defect or virtue were removed, due to the defect or virtue's being tied to other features. For instance, it is difficult to imagine what *Triumph of the Will* would be like if it did not have its moral defect,

[9] I will ignore the issue of whether the work would still be the same work after the removal of *x*. For our purposes, nothing important turns on this issue because the intuitive appeal of **Counterfactual Dependence (Sufficient Condition)** is not affected by whether the resulting work is the *same* work or not. For discussions related to this issue, see Jacobson (1997) and Harold (2008, pp. 48-50).

since its moral defect is inextricably tied to other central features of the film, including its formal features. Accordingly, it is almost impossible to determine whether the altered work would be less artistically valuable than the original one. This is, however, just a reason to think that **Counterfactual Dependence (Sufficient Condition)** is often difficult to *apply*. The problem of interconnection that I want to raise goes deeper. **Counterfactual Dependence (Sufficient Condition)** is simply false, even if we can conceive the relevant counterfactual situation.

Here is a counterexample. Suppose that an artwork has a feature *x* that does not positively (negatively) contribute to the overall artistic value of the work. Feature *x*, however, may satisfy the conditional—if the work did not have *x*, it would be less (more) artistically valuable— if it is connected with another feature of the work *y* that co-varies with the work's artistic value.

It is not easy to find a concrete example that completely fits the descriptions above, but here is an example that at least comes close. Moreover, even if this specific example turns out to be problematic, the abstract counterexample may still stand. Suppose that the so-called Mozart effect—the alleged enhancing effect of Mozart's music on a listener's cognitive performance—actually exists, and that one of Mozart's sonatas has that effect (feature *x*). Intuitively, having the Mozart effect neither positively nor negatively contributes to the sonata's artistic value.[10] That is, it is intuitively not true that the sonata is artistically valuable or defective in virtue

[10] I am not claiming that cognitive value is never relevant to artistic value. My claim is just that the kind of cognitive value the Mozart effect represents is intuitively irrelevant to artistic value.

of having the Mozart effect. However, it may be the case that if the sonata did not have the effect, its artistic value would differ, simply because the presence of the effect is connected with other features of the work, such as its sonic properties (feature y). It is quite plausible that if the sonata did not have the Mozart effect, it would have different sonic properties, in which case the artistic value of the sonata would differ as well (after all, the sonata would sound different). In such a case, having the Mozart effect satisfies the conditional, if the sonata did not have the Mozart effect, it would be less (or more) artistically valuable, but merely because having the Mozart effect is connected with the sonata's sonic properties.

One might argue at this point that we have simply chosen the wrong interpretation of the phrase, "if the work did not have *x*." This phrase can be interpreted in two ways. The first is what we have just considered, which is to remove *x* from the work. In this interpretation, the removal of *x* may result in changes in other features connected to *x*, causing the problem of interconnection. The second interpretation ignores such interconnections and simply assumes that the work does not have *x*, without changing anything else about the work.[11] For instance, in the case of *Triumph of the Will,* we can simply imagine that Nazism is moral, rather than trying to imagine what the film would be like with its moral defect removed.[12]

[11] Alessandro Giovannelli discusses this interpretation, specifically with respect to *Triumph of the Will* (2013). But he does not use this interpretation to *support* his brand of moralism. His point is just that his moralism *entails* that "the work would be better artistically, other things being equal, if its perspective's ethical status were different, that is, if the perspective were not immoral" (p. 346).

[12] By imagining that Nazism is moral, I mean simply changing the moral status of Nazism, without changing anything else about Nazism itself.

This second interpretation is free from the interconnection problem; however, it has other problems. To begin with, some philosophers believe that it is not possible for moral truths to be otherwise[13]. Furthermore, even if there is a possible world in which Nazism is moral, it is unclear whether this is the right possible world to consider when evaluating the conditional. The truth of a counterfactual conditional is commonly understood to be determined by whether the consequent is true in the possible world(s) closest to our own in which the antecedent is true. And the world in which Nazism is moral is arguably more distant from the actual world than the world in which the moral defect of *Triumph of the Will* is removed and Nazism is immoral. Additionally, it is hard to see how artworks are to be evaluated in such a distant world, and thus hard to determine how artistically valuable *Triumph of the Will* would be there.[14]

Even if we set aside all the problems related to possible worlds, the second interpretation of "if the work did not have *x*" still has a problem: it cannot give us any insight about *contextual* value interaction, which is our concern. Contextual value interaction occurs due to the ways in which a work's moral virtue or defect interacts with other features of the work in a particular context, but the second interpretation detaches a moral virtue or defect from all the other features of the

[13] See, for instance, Zangwill (1995).

[14] One might object that I am presupposing here that moral and artistic value can interact with each other, which is exactly what is at issue in the value interaction debate. I admit that I am presupposing *some kind* of interaction, but it is *not* the kind of interaction at issue in the debate. What I am presupposing is that both artistic and moral facts partly depend on basic facts about human nature (e.g., our tendency to feel pleasure at certain things and pain at others), and thus the two values may co-vary with each other *across different possible worlds* with different basic facts about human nature. This kind of interaction is not what is at issue in the debate, which is whether there is a relationship between the two values *within our own world*. (Thanks to James Harold for signaling this difficulty to me.)

work. For instance, when we simply assume that Nazism is moral, and thus that *Triumph of the Will* does not have a moral defect without changing anything about the work itself, we are detaching the moral defect from the particular context of the work. Thus, the truth or falsity of the conditional, "if *Triumph of the Will* did not have the moral defect, it would be more artistically valuable," interpreted in this way has nothing to do with contextual value interaction. At best, it could tell us what intuitions we have about *intrinsic* value interaction, that is, whether a moral defect has positive or negative intrinsic artistic valence, if we have any clear intuitions about this matter at all. Therefore, the proponent of **Counterfactual Dependence (Sufficient Condition)** faces a dilemma: the first interpretation of "if the work did not have *x*" is susceptible to the interconnection problem, and the second interpretation is irrelevant to contextual value interaction.[15]

(2) So far, we have seen how the fact that features of artwork are interconnected with each other poses a problem to **Counterfactual Dependence (Sufficient Condition)**. Let us now turn to its second problem, which is that it does not distinguish between what genuinely *contributes* to a work's artistic value and mere *necessary* factors.[16] I will illustrate this problem with an actual example in which **Counterfactual Dependence (Sufficient Condition)** is used to defend a value interaction claim: Ted Nannicelli's immoralism (2014). He claims that an artwork is sometimes artistically valuable in virtue of its moral defect. His argument

[15] Thanks to an anonymous referee for this journal for suggesting formulating my objection to **Counterfactual Dependence (Sufficient Condition)** as a dilemma.
[16] In a sense, this problem is just a special case of the interconnection problem because it arises from the connection between a merely necessary factor and a genuine contributor. However, this problem is worthy of separate attention.

for his claim, in a nutshell, is as follows: (1) an artwork is sometimes funny partly in virtue of a moral defect pertaining to how it is produced (call this a *genetic* moral defect), and (2) funniness can in turn contribute to artistic value. He appeals to counterfactual dependence to support (1), so let us focus on that part of his argument.[17]

One of his examples of a work that is funny in virtue of its genetic moral defect is a video skit titled *The Landlord.* In it, the director's two-year-old daughter plays a drunk landlord swearing at her renter. According to Nannicelli, coaching a child to engage in such behavior is a genetic moral defect, but *The Landlord* is funny partly in virtue of the defect. This is because, he argues, "without the moral flaw, the work is significantly less funny" (p. 173). The incongruity of a child engaging in adult behavior that grounds the funniness could not have been achieved without the moral defect.

Contra Nannicelli, I maintain that even if this conditional, "without the moral flaw, the work is significantly less funny," is true, it does not follow that the work's funniness is in virtue of its genetic moral defect. This is because the conditional's truth entails only that the moral defect is *unavoidable* or *necessary* to achieve the funniness, and not all necessary factors are genuine *contributors*. If so, it does not follow from the truth of the conditional that the moral defect itself

[17] Premise (1) concerns the relation between a moral defect and *funniness*, not artistic value, but this difference is not important for our purposes. Whether *x* is artistic value or funniness, the same principle is at issue—*x* is in virtue of *y* if *x* counterfactually depends on *y*.

contributes to *The Landlord*'s funniness. The funniness is more plausibly in virtue of the incongruity alone.

To illustrate, consider the following case, which is structurally similar to *The Landlord*. Suppose that there is a certain rare pigment that is very expensive, but has an incredibly beautiful color. No other pigment has the same color. It is then unavoidable for a painter to pay lots of money for the pigment if she wants to use that color. Now, consider a picture painted with that pigment, which is beautiful in virtue of the color. Would we say that the picture is beautiful also partly in virtue of paying lots of money for the pigment? I think not. Even though it is true that without the money spent, the picture would not be as beautiful, intuitively, the beauty is not in virtue of the money spent, not even partly; it is in virtue of the color. Similarly, even if without its moral defect, *The Landlord* would not be as funny, it does not follow that the funniness is in virtue of the moral defect; rather, the funniness is in virtue of the incongruity.

Admittedly, there might be a sense in which the picture is beautiful in virtue of the money spent, in just the sense that the beauty counterfactually depends on it. However, if we allowed counterfactual dependence to be enough for an in-virtue-of relation to hold, then a work's funniness or artistic value would be in virtue of a wide range of features that we consider irrelevant. For instance, we may have to accept that *The Landlord* is funny partly in virtue of its being a talkie, since it would be less funny if it were not a talkie. Of course, again, there is a sense in which it is funny partly in virtue of being a talkie. But I do not think this loose sense of in-virtue-of is what philosophers have in mind when engaging in the value interaction

debate. Value interaction based on such a loose sense of in-virtue-of is not philosophically interesting.

So far, we have seen that counterfactual dependence is not *sufficient* to capture exactly what feature of a work makes a positive contribution to its value. But one might think that counterfactual dependence can at least serve as a *necessary* condition, adopting the following principle.

> **Counterfactual Dependence (Necessary Condition)**: A work's feature *x* positively (negatively) contributes to its overall artistic value *only* if the following conditional is true: if the work did not have *x*, it would be less (more) artistically valuable.

Daniel Jacobson seems to have something like this principle in his mind when he makes the following remark regarding Emily Dickinson's poem *Tell all the truth but tell it slant*: "the moral defects of the poem's ethical perspective can sensibly be deemed a blemish—that is, an aesthetic flaw—only if the poem would be improved, aesthetically, by its alteration" (1997, p. 183).[18]

However appealing it might look, **Counterfactual Dependence (Necessary Condition)** is also false. To illustrate, let me first make an analogy with a non-art case.[19] A group of friends is having a party. One of them, Mary, is a lot

[18] This claim of Jacobson was criticized by Gaut (2007, pp. 57-66). Gaut contends that a moral defect is an intrinsic aesthetic flaw even if its alteration would not overall improve the work aesthetically. Thus Gaut's criticism of **Counterfactual Dependence (Necessary Condition**) is that counterfactual dependence is not necessary for *intrinsic* value interaction. My criticism of **Counterfactual Dependence (Necessary Condition)**, on the other hand, is that counterfactual dependence is not necessary for *contextual* value interaction. To my knowledge, Gaut's response to Jacobson is the only earlier critical discussion of **Counterfactual Dependence**.

[19] The following example is based on Jonathan Dancy's example (2004, p. 55).

of fun. So, the party is enjoyable partly in virtue of her. But the nature of the party happens to be such that it would be even more enjoyable if Mary were absent, due to the group becoming smaller and more intimate. Despite this, it is still intuitively the case that Mary positively contributes to the party's enjoyableness.

A similar scenario is possible in art. Suppose that a film has a cliché in its narrative. Intuitively, that cliché negatively contributes to the film's overall artistic value unless there is a special overriding reason (e.g., the cliché is satirized later in the film). But the nature of the particular film might be such that it would be artistically worse if the cliché were altered, perhaps because the overall plot would become less intelligible.

At this point, one might argue that I simply chose the wrong possible world to consider. Instead of considering a world in which the film becomes less intelligible as a result of altering the cliché, we should consider a world in which the film maintains its intelligibility even when the cliché is altered. That is, when altering the cliché, we should hold fixed everything else about the film that may influence its artistic value. On this interpretation, the film would be better without the cliché.

This way of understanding the conditional, however, involves the same problem as the one involved in assuming that Nazism is moral when interpreting the phrase "if *Triumph of the Will* did not have a moral defect." As we saw earlier, the problem is that a conditional understood this way detaches the feature in question from the particular context, and thus has implications only for whether the

feature is *intrinsically* artistically good or bad. That the film would be better without the cliché, *other things being equal,* shows only that the cliché is an intrinsic artistic defect. It has no implications for *contextual* value interaction, that is, how the cliché affects the film's overall artistic value due to the ways in which it interacts with other features of the film, which is what is at issue in this section.

I have so far argued that counterfactual dependence is neither necessary nor sufficient for value interaction. Admittedly, there is something intuitively appealing about understanding value interaction in terms of counterfactual dependence. This faulty intuition might result from the fact that counterfactual dependence often works as a *rough heuristic* for tracking other kinds of interactions, such as everyday causal responsibility. Although counterfactual dependence is not an accurate analysis of even everyday causal responsibility, everyday causal responsibility is usually less vulnerable to the problem of interconnection compared to the case of artwork, which is more holistic by nature. Hence, counterfactual dependence often, though not always, tracks everyday causal responsibility. This perhaps explains the strong intuitive appeal of counterfactual dependence as an analysis of value interaction. As we have seen, however, counterfactual dependence fails to accurately capture value interaction in the context of art. Thus we should resist the temptation to use it as a strategy to prove or refute value interaction claims.

4. Contextual value interaction and common reasons

The second strategy often used to support claims about contextual value interaction is to show that a work's moral value and artistic value share *common reasons*. This strategy relies on the following assumption:

> **Common Reasons**: A work is artistically valuable (defective) in virtue of its moral defect (virtue) if what is immoral (moral) about the work is just what is artistically valuable (defective) about it, or in other words, if the work is immoral (moral) and artistically valuable (defective) for the same reasons.

A.W. Eaton's robust immoralism uses this strategy (2012, 2013). Eaton's main claim is that "an immoral feature of an artwork can make a significant positive aesthetic contribution *precisely in virtue of its immorality*" in the case of works featuring what she calls rough heroes (2012, p. 283, emphasis in original).[20] Rough heroes are characters such as Tony Soprano from *The Sopranos*, who have irredeemable moral flaws but are at the same time portrayed as sympathetic and likable. One of Eaton's arguments for her main claim is that such works are morally defective and artistically valuable for the same reason.[21] Thus, her argument relies on **Common Reasons**.

Let us look closely at Eaton's argument. She first contends that "works featuring rough heroes are morally flawed in virtue of their perspective that endorses moral depravity by rendering it sympathetic, likeable, praiseworthy, and

[20] Although she uses the term 'aesthetic,' what she means by it seems very close to artistic value.
[21] Eaton seems to offer another argument that does not appeal to **Common Reasons**. The argument is that works featuring rough heroes are *compelling* (because they elicit conflicting attitudes toward rough heroes), which is an aesthetic merit. This argument is not my target.

glamorous" (2012, p. 287). Eaton then argues that this sort of moral defect is at the same time an aesthetic achievement. She points out that it is challenging for an artwork to elicit conflicting attitudes towards rough heroes—the attitudes of both blaming and liking them—, since audiences feel imaginative resistance to adopting such morally dubious attitudes. To meet this challenge, a work should strike the right balance between portraying the immoral and the likable aspects of rough heroes. As Eaton puts it, a work should have "a delicate equilibrium" between the two aspects, which takes "considerable artistry and finesse" (2012, p. 287). In this respect, eliciting conflicting attitudes towards a rough hero is an artistic achievement.

For these reasons, Eaton contends that eliciting conflicting attitudes toward rough heroes is both an artistic achievement and a moral defect. That is, "the very same feature of a work is both a moral blemish and an aesthetic merit," or to put it another way, "some immoral artworks are morally flawed and aesthetically meritorious for the same reasons" (2012, pp. 288-290). I will grant, for the sake of argument, that this claim is true. Now, the question is whether Eaton's main claim that "an immoral feature of an artwork can make a significant positive aesthetic contribution *precisely in virtue of its immorality*" follows from this (2012, 283, emphasis in original). I think not. **Common Reasons** is false.

To illustrate why, let me make an analogy between a common cause and a common reason. In the case of a common cause, two effects share the same cause, but there may not be a causal relation between the two effects. For example, the firing of a gun may cause both a muzzle flash and the death of a

person in front of the gun. However, it is not the case that the muzzle flash causes the death. Thus, two effects sharing a common cause is not a reason to think that one of them causes the other. Now, note that the structure of a common cause is similar to that of an artwork being morally defective and artistically valuable for the same reason. Just as a common cause causes two effects, a common reason grounds two value features: a moral defect and an artistic merit. If so, just as two effects sharing a common cause is not a reason to think that one of them is caused by the other, two value features sharing a common reason is not a reason to think that one of them is in virtue of the other.

Let me elaborate on this analogy a bit further. One sign that there is no causal relation between two effects caused by a common cause is the lack of asymmetry between them. When *x* is caused by *y*, *y* is prior to *x*, in some sense. But neither of the two effects caused by a common cause is prior to the other (assuming that there is no other relation between them). Such lack of asymmetry is shown in Eaton's examples, too. In works featuring rough heroes, the moral defect is not prior to the artistic merit in any sense: all that Eaton has shown about their relationship is that they share a common reason, which is a symmetrical relation. This suggests that there is no in-virtue-of relation between the moral defect and the artistic merit because when *x* is in virtue of *y*, *y* must be prior to *x* in some sense – this is part of the nature of an in-virtue-of relation.

Another way to make the same point is as follows. If we accept **Common Reasons**, and thus conclude that the artistic merit of featuring a rough hero is in virtue of its moral defect on the grounds that the two features share the same

reason, then we have to accept that an in-virtue-of relation holds the other way around, too—that is, the moral defect is in virtue of the artistic merit as well. This is because there is no asymmetry in two features' sharing the same reason. In fact, it seems quite plausible that the moral defect of featuring rough heroes is in virtue of its artistic merit. In a sense, what makes the works featuring rough heroes morally problematic is the delicate equilibrium between portraying the immoral and likable aspects of their rough heroes, which Eaton considers to be an artistic achievement. In this respect, it makes sense to say that the works are morally defective in virtue of their artistic achievement. At least, we have no reason to think that this claim is less plausible than the claim that the works are artistically valuable in virtue of their moral defects. This is because the reason we have for the latter claim is just that the artistic merit and moral defect share the same ground. If this reason alone is enough to support the claim that the works are artistically valuable in virtue of their moral defects, it should also support the claim that the works are morally defective in virtue of their artistic achievement. This symmetrical relation between the moral defect and the artistic merit, again, suggests that no in-virtue-of relation holds between the two. If *x* is in virtue of *y*, it cannot be the case that *y* is in virtue of *x*, because when *x* is in virtue of *y*, *y* is prior to *x*, which implies that *x* is not prior to *y*.

In response to this objection, Eaton might argue that what makes works featuring rough heroes morally defective is their *attempt* to elicit conflicting responses toward rough heroes, rather than their *success* in doing so, while the artistic merit of such works lies in the success. Thus, the moral defect is, in some

sense, prior to the artistic merit. However, now the problem is that this seems like a case in which a work is artistically valuable *despite* its moral defect, rather than *in virtue of* it. The work is artistically valuable because it has successfully *overcome* the challenge posed by the moral defect, thanks to its aesthetic finesse. This aesthetic finesse, not the moral defect, is that in virtue of which the work is artistically valuable.

Let me now contrast Eaton's argument with another argument that seems similar, but is free from the problem afflicting Eaton's argument. The argument is Noël Carroll's (1996; 1998). Comparing his argument with Eaton's will further clarify my objection to **Common Reasons**.

Carroll's view, which he calls moderate moralism, claims that "in some instances a moral defect in an artwork can be an aesthetic defect, and that sometimes a moral virtue can count as an aesthetic virtue" (1998, p. 419).[22] One of his examples in which a moral defect is an aesthetic defect is an imaginary story of Heinrich Himmler, a leading member of the Nazi Party, winning the Nobel Peace Prize. This story is intended to elicit our admiration for Himmler, and thus is morally defective. In a nutshell, Carroll's argument for the claim that this moral defect is an aesthetic defect is as follows: morally sensitive audiences would fail to show admiration for Himmler due to the evilness of the work's

[22] Although Carroll uses the term "aesthetic," he seems to understand aesthetic value quite broadly, almost coextensively with artistic value at least in the case of the aesthetic value of artwork. See Carroll (2006).

perspective, and this is an aesthetic defect because it is a failure in what the work aims to achieve.

In some passages, Carroll seems to put this argument in terms of common reasons (so much so that some of his opponents called Carroll's argument "the argument from common reasons"[23]). For instance, Carroll says, "[the story about Himmler] fails aesthetically for the same reasons that it is morally defective: the perspective it advances is evil. The evilness of the perspective is what makes it morally bad and aesthetically bad" (1998, p. 421).

Despite this seeming appeal to **Common Reasons**, Carroll's argument does not in fact rely on it. To illustrate, first note that it does not make sense to say that the moral defect of the Himmler story is in virtue of the aesthetic defect—it is not because the work fails to elicit the response it aims to elicit that its perspective is immoral. It makes sense to say that the aesthetic defect is in virtue of the moral defect, but not the other way around. This is a sign that Carroll's argument does not suffer from the same problem as Eaton's, namely, the lack of asymmetry. Then where does the asymmetry in Carroll's argument come from? His argument partly involves a *causal relation*, which is asymmetric by nature, although he does not invoke the term causation himself. In Carroll's example, morally sensitive audiences' recognition of the moral defect *causes* them to resist the response the work aims to elicit. This then allegedly counts as an aesthetic defect.[24] Of course, we might still say that the moral and aesthetic

[23] See Anderson and Dean (1998).
[24] This claim is not about causation, so Carroll's argument only *partly* relies on causation.

defects share a common reason, since both of them eventually result from the problematic perspective of the work. However, there is more to the relationship between the moral and aesthetic defects than merely sharing a common reason—the relationship also includes a causal relation. This gives Carroll's argument the asymmetry that Eaton's argument lacks.[25]

5. Implications

If, as I have argued, **Counterfactual Dependence** and **Common Reasons** cannot be used to support value interaction claims, what does this imply for the value interaction debate? I doubt that my arguments directly support any of the existing positions. I criticized two immoralist views in this essay, but immoralism *per se* is not undermined by my criticism. There can be arguments for immoralism that do not rely on **Counterfactual Dependence** or **Common Reasons**. On the other hand, some arguments for moralism might be vulnerable to my criticism if they rely on the two strategies. One might think my argument points in favor of autonomism, which denies that a moral defect or virtue can be an artistic defect or virtue. However, all I have shown is that two particular strategies commonly used for defending value interaction fail. There might be better strategies.

What could those better strategies be? I already mentioned one argument that is immune to my criticism, Carroll's argument for moderate moralism. As I understand it, his argument partly involves causation, which is one kind of

[25] I remain neutral as to whether Carroll's argument is successful *overall*. For criticisms of Carroll's argument that are not relevant here, see, for instance, Jacobson (1997; 2005) and Stecker (2005).

contributive or in-virtue-of relation, broadly construed. Thus, assuming that we have an intuitive notion of causation that does not need to be analyzed in terms of counterfactual dependence, this line of argument would be free from my criticism.

But value interaction does not always involve causation, and even Carroll's argument only partly relies on causation. So, we need a more comprehensive approach that can cover value interaction in general. I cannot here attempt to develop such a comprehensive account. Rather, I will just propose some new directions for research that may facilitate the progress of the value interaction debate, and may eventually lead to a comprehensive account of value interaction.

Firstly, I propose we pay more attention to the notion of *grounding*, which is usually construed as non-causal determination, or the in-virtue-of relation in a non-causal sense. This is exactly what is at issue when we discuss whether an artwork can be artistically valuable in virtue of its moral virtue or defect. It is plausible to think that a work is artistically valuable in virtue of its moral virtue or defect if and only if it (at least partly) grounds the work's artistic value.

Another notion that might be useful for the value interaction debate is *explanation*. This notion has already been invoked in the value interaction literature. For instance, Carroll has argued that sometimes a work's moral defect is "an ineliminable factor" in explaining why the work is aesthetically defective, and in this case, the moral defect is an aesthetic defect (1998). Jacobson (2005) then uses this argument to defend immoralism. He argues that a work's moral

defect can be an ineliminable factor in explaining why the work is aesthetically valuable, and therefore a moral defect can be an aesthetic virtue. This line of approach seems worth pursuing, although there are several questions that need exploration, such as when exactly a feature of an artwork is an ineliminable factor in explaining its artistic value, and whether and how explanation about artistic value differs from scientific explanation, the most discussed kind of explanation.

Admittedly, the notion of grounding or explanation may not offer the kind of *quick test* for value interaction that **Counterfactual Dependence** or **Common Reasons** is (wrongly) expected to offer. This is because the notions of grounding and explanation themselves require thorough investigation. Furthermore, it is widely agreed that grounding is a primitive notion, which we cannot further analyze in other terms. However, approaching value interaction through these two notions may still be useful. The extensive literature on grounding and explanation provides us with rich theoretical tools, some of which may even come close to quick tests. For instance, my objection to **Common Reasons** makes use of a widely agreed characteristic of grounding, its being asymmetric. This offers one quick test for value interaction—if there is no asymmetry, there is no in-virtue-of relation. This, of course, is not a comprehensive test for value interaction. But it helps to make some progress.

The true nature of value interaction remains to be fully elucidated. But whatever it is, it is not merely counterfactual dependence or the sharing of common reasons.[26]

[26] An earlier version of this essay was presented at the American Society for Aesthetics Annual Meeting in Seattle, Washington, in 2016. I am grateful to the audience for their valuable feedback. I also owe special thanks to James Harold, Jerrold Levinson, Dan Moller, Julius Schoenherr, Evan Westra, and an editor and anonymous referees for this journal for their helpful comments and suggestions.

Chapter 2: Aesthetic Explanation

1. Introduction

It is a commonplace in aesthetics that an object having certain non-aesthetic properties *explains* it having a certain aesthetic property.[27] In other words, the object has the aesthetic property *because of* the non-aesthetic properties. Brancusi's sculpture *Bird in Space* (1928), for example, is elegant because of, among other things, its tapered shape and smooth surface texture. Its shape and texture partly explain its elegance.[28] Let us call such explanation of the aesthetic by the non-aesthetic: *aesthetic explanation*.

Note that aesthetic explanation defined this way is *not* an explanation of overall aesthetic or artistic goodness. Explanation of overall aesthetic or artistic goodness includes more than aesthetic explanation: it also requires the explanation of the object's overall aesthetic or artistic goodness by its aesthetic properties. For instance, in order to explain why *Bird in Space* is artistically valuable, we need an explanation of how its elegance, as well as its other aesthetic properties, contributes to or detracts from its artistic value, in addition to an explanation of its elegance (and other aesthetic properties) by its non-aesthetic properties. This essay concerns only the latter form of explanation.

---

[27] There is no widely-accepted agreement on what distinguishes aesthetic properties from non-aesthetic properties, but there is good agreement on the extension of aesthetic properties, which include elegance, garishness, unity, delicacy, sublimity, beauty (as a specific aesthetic property rather than aesthetic goodness in general), and so on. For discussion on the nature of aesthetic properties, see, e.g., Kivy (1975); Levinson (2005); Sibley (1965).

[28] This example is based on Bender (1987, p. 34).

Although it is widely agreed that the non-aesthetic explains the aesthetic, surprisingly little work has been done on the nature of such explanation. This essay aims to make progress toward a better understanding of aesthetic explanation by achieving the following two goals. First, I offer an account of aesthetic explanation that can elucidate in what sense the non-aesthetic explains the aesthetic. My suggestion is that an object having certain non-aesthetic properties explains it having a certain aesthetic property just when, and in the sense that, roughly speaking, the observer's experience of the aesthetic property is caused by their experience of the non-aesthetic properties. Let us call this account the psychological account of aesthetic explanation. This account construes aesthetic explanation as a kind of causal explanation, broadly construed: although there is no causal relation between the object having the non-aesthetic properties and it having the aesthetic property, the explanatory relation between them eventually derives from the causal relation between the observer's experiences of them.

The second goal of this essay is to illuminate one interesting feature of aesthetic explanation, namely its selectivity. Aesthetic explanation is selective in that the explanation of an object having a certain aesthetic property cites only some of the non-aesthetic properties that can make a difference to the aesthetic property, or to put it another way, the non-aesthetic properties on which the aesthetic property supervenes. For instance, although *Bird in Space* might not be elegant if it were much smaller (e.g., a few inches high) or larger (e.g., 10 feet high), the sculpture's size does not intuitively seem to explain its elegance. Hence

a question arises as to why aesthetic explanation is selective and what principle governs such selection. I suggest that the selectivity of aesthetic explanation comes from it being a kind of causal explanation broadly construed—causal explanation in general is selective. As for the selection principle for aesthetic explanation, I will suggest a principle that mirrors L.R. Franklin-Hall (2015)'s selection principle for causal explanation, according to which causal explanation selects those factors that maximize the ratio of delivery (the degree to which the factors cited in the explanation stabilizes what is explained) to cost (the amount of information an explanation contains).

The essay will proceed as follows: in Section 2, I articulate and argue for my psychological account of aesthetic explanation. In Section 3, I focus on the selectivity of aesthetic explanation, defending the selection principle mentioned above. And in Section 4, I conclude the essay by briefly discussing the implications of my account for adjacent issues in aesthetics, such as the debate between generalism and particularism.

2. The psychological account of aesthetic explanation

In this section, I articulate and defend my account of aesthetic explanation, which understands the explanation of the aesthetic by the non-aesthetic in terms of the causation of the observer's experience of the aesthetic by their experience of the non-aesthetic.

2.1. Clarifying the account

More specifically, my account can be put as follows:

The Psychological Account of Aesthetic Explanation

> The fact that an object has certain non-aesthetic properties explains the fact that it has a certain aesthetic property just when, and in the sense that, under the right conditions, the right observer's experience of the aesthetic property (A-psychological event or, more simply, A-event) is caused by their experiences of the non-aesthetic properties (NA-psychological events or NA-events).

For instance, in the case of *Bird in Space*, its tapered shape and smooth surface texture partly explain its elegance in the sense that the observer's experience of the elegance is caused by their experience of the shape and surface, among other things.

Let me now make a couple of clarifications of the account. First, although I mentioned earlier that aesthetic explanation is a kind of causal explanation, this is so only in a broad sense of causal explanation. I do not claim that there is a direct causal relation between the explanandum or what is explained (i.e., an object having a certain aesthetic property) and the explanans or what explains (i.e., the object having a certain non-aesthetic properties). There is no temporal difference between the explanandum and the explanans, so assuming that a cause must temporally precede its effect, there can be no causal relation between the two. Thus, strictly speaking, my account does not take aesthetic explanation to be causal explanation.

Yet, there is a loose sense in which my account understands aesthetic explanation as causal explanation. It construes aesthetic explanation in terms of causal relations between psychological events that happen in the observer's mind, which are responses to, or experiences of, the explanans and the explanandum. Note that I use the term "response" and "experience" broadly here. A response or experience, as I will use the terms, can be perceptual, cognitive, and/or affective, and can also be conscious or unconscious.

The second clarification I want to make is that my account is not committed to a particular metaphysical position on the nature of aesthetic properties. To begin with, I wish to leave open who counts as a relevant observer(s) and what counts as the right conditions. My account only requires that there are *some* observers—who could be humans, or sentient beings in general, rather than a particular group of people—and *some* conditions that are relevant to aesthetic properties. I also am not committed to the claim that there is an objective standard that determines who is the right observer. Thus, my account is compatible with relativism, by which I mean the view that an object has a certain aesthetic property for a group of observers but not for others (and that neither is incorrect).

One might presume that my account is committed to anti-realism about aesthetic properties, since it understands aesthetic explanation in terms of psychological events. But this is not the case. For instance, my account is compatible with Jerrold Levinson's realist account of aesthetic properties, according to which aesthetic properties are "higher-order ways of appearing"

(Levison, 2005, p. 218), or with Vid Simoniti's realist account, which argues that aesthetic properties are powers that bring about certain experiences and that such powers are real properties of an object (2017). These theories, as well as my account, are mind-dependent in some sense. However, not any sort of mind-dependence implies anti-realism. Anti-realism requires eliminativist reduction, but my account does not imply that facts about aesthetic properties are reducible, in an eliminativist rather than vindicative way, to facts about psychological events. Thus, my account is compatible with realism as well as anti-realism.[29]

To further illustrate, consider an analogy with colors. Both realists and anti-realists about colors can agree that an explanation as to why a particular object is red should cite how its molecular structure, in conjunction with light, (actually or potentially) brings about a certain experience for an observer under the right conditions. In fact, it is unclear whether there is any explanation of a particular object being red that does not refer to the experience of red in any way because the property of being red is multiply realizable at the lower-level—the only thing that ties the different bases together seems to be the experience of red (or the power to bring about such an experience). Adopting such a mind-dependent explanation about the particular object being red is not to commit to anti-realism about redness. The same point applies to my account of aesthetic

[29] I wish to remain agnostic in this essay about whether my account is compatible with *any kind* of realism. My account is certainly compatible with naturalistic realism. Even if it turns out to be incompatible with non-naturalistic realism, this is not a serious problem because few endorse non-naturalistic realism about aesthetic properties. In fact, if my account is incompatible with non-naturalistic realism, the plausibility of my account would provide an additional reason to reject it.

explanation. Explaining why a particular object is elegant by citing corresponding psychological events is, by itself, not to commit to anti-realism about elegance.[30]

2.2. Defending the account

I will now defend the causal account of aesthetic explanation. I will do so by considering three possible objections. I believe this negative defense is enough in the context of this essay, given the absence of any alternative account in the current literature.

(1) The first possible objection to consider is that not all A-psychological events are caused by NA-psychological events and, thus, the causal account fails as a general account of aesthetic explanation. It may seem that A-events can occur before, or even without, NA-events. When we find a piece of music beautiful, for instance, we often recognize the beauty first, before thinking about relevant non-aesthetic properties such as the piece's sound structure. Sometimes we do not even know why an object is beautiful, elegant, sublime, etc. It just strikes us as having those aesthetic properties.

These observations, however, are compatible with my claim that all A-events are caused by NA-events. NA-events may cause A-events, even when the observer is unaware of this fact. There is empirical evidence that we are often

[30] It is worth noting that my account does not claim that whether an object has a certain aesthetic property is a matter of whether the right observer attributes the property to it (for instance, whether they judge that the object is elegant). I only claim that there is a certain psychological event that corresponds to the object having the aesthetic property, and the psychological event does not need to involve the attribution of the property to the object (the event, for instance, could be just a certain kind of feeling or perception that does not involve the concept of elegance).

unaware of what causes our mental states.[31] In some cases, NA-events can even be completely unconscious. It is widely agreed by psychologists that perception can be unconscious, and unconscious perception can give rise to other mental states, such as affect.[32] Thus, an unconscious perception of an object's non-aesthetic properties may cause an A-event. Alternatively, even when the observer is conscious of an NA-event, they might be inattentive or lack the vocabulary to articulate their experience. Thus, there are several reasons some A-events may appear to be caused not by an NA-event even when, in fact, all A-events are caused by NA-events.

This is, of course, not a positive argument that all A-events are caused by NA-events. To offer a positive argument, let me first note that the kind of NA-events I have in mind are minimal—any mental event that registers or represents an object in the observer's mind counts as an NA-event. To experience a piece of music as poignant, for instance, you need to hear the sound of the piece. To experience a painting as elegant, you need to see it. Even when the object is imaginary (as when you hear a piece of music in your head and find it poignant) or abstract (as when you find a mathematical proof elegant), there is a representation of the object in your mind. Whatever the exact nature of such registrations or representations is, they are what I mean by NA-psychological events.

[31] See, for example, Nisbett & Wilson (1997).
[32] See, for instance, Winkielman et al. (2005).

Now, let me illustrate why all A-psychological events are causally preceded by NA-psychological events by considering different types of A-psychological events—affective, cognitive, and perceptual—in turn. First, some aesthetic properties seem to correspond to *affective* psychological events. Poignancy, melancholy, and sublimity may plausibly belong to this category.[33] Now, it should be uncontroversial, I believe, that such affective A-events are preceded by some NA-events that provide the object of affect, whether it is perceptual, cognitive, or of some other nature. Affect involves evaluation or appraisal, so it requires things to evaluate. This is, of course, not to say that we are first aware of what to evaluate and then of the evaluation. NA-and A-events can become conscious at the same time, or the NA-event may not even be conscious. But this is compatible with my claim that the NA-event causally precedes the A-event.

The same point applies to *cognitive* A-events, such as judgment, if there is any aesthetic property that corresponds to such an event. Like affect, judgment cannot be made without first registering the object of judgment, consciously or unconsciously, so cognitive A-events should be preceded by some NA-events that provide the object of judgment.

What about *perceptual* A-events? Aesthetic properties such as unity, balance, and gracefulness might correspond to perceptual events—that is, they might be perceived in some sense of perception. These constitute the most

[33] It is not important for my argument which specific aesthetic properties correspond to which type of A-event. My argument works insofar as the three types exhaust all A-events (or if the point I make about the three types could apply to the remaining A-events as well).

compelling opposing cases. An objector might ask why, if both events are perceptual, the perception of such aesthetic properties should be preceded by the perception of relevant non-aesthetic properties.

In response, let us look more closely at the claim that aesthetic properties can be perceived. Even those who endorse this claim do not believe that aesthetic properties are perceived in the same way as lower-level sensory properties, such as shapes and colors. Dustin Stokes, for instance, argues that we can see an object as having a certain aesthetic property in a similar way that we can see an ambiguous figure like the duck/rabbit figure as a duck (or a rabbit) (2018). In such a case, he states, "an organization or gestalt of basic features" (p. 30) is part of the perceptual content, and "to perceptually experience the relevant gestalt just is to perceptually experience the [aesthetic] property" (p. 40). If, as Stokes argues, perceiving an aesthetic property is perceiving an organization or gestalt of lower-level, non-aesthetic features, it is quite plausible that the perception of the aesthetic property is preceded by the registration of the non-aesthetic properties, whether this registration happens consciously or unconsciously.

Stokes is not the only philosopher who argues that we can perceive higher-level properties. Peter Carruthers, for instance, offers a more empirically rigorous suggestion as to how higher-level concepts, such as mental state concepts, can be part of perception (2015). Although he does not use aesthetic concepts for his examples, his theory seems applicable to aesthetic concepts as well. According to Carruthers, higher-level concepts can be represented in the content of perception when they become conscious by being integrated with sensory input and entering

working memory as a single unit. Note that this event happens *after* the concept in question is processed based on sensory input. Thus, sensory input precedes perception of the higher-level concept. If the perception of aesthetic concepts occurs via this mechanism, then the perception of non-aesthetic properties should precede it. Therefore, whether the aesthetic property in question is affective, cognitive, or perceptual, we have good reasons to think that all A-events should be preceded by NA-events.

(2) The second objection to my account that I will consider is that there are non-causal connections between the aesthetic and the non-aesthetic, such as conceptual connections or grounding relations, and thus aesthetic explanation should be understood in terms of such non-causal connections. My response to this objection, to put it simply, is that even if there are such non-causal connections between the aesthetic and the non-aesthetic, this is compatible with the fundamental metaphysical connection between them being the causal relation between NA- and A-psychological events. Let me elaborate.

Let us first consider conceptual connections between the aesthetic and the non-aesthetic. Some philosophers have argued that at least some non-aesthetic and aesthetic properties or terms are connected conceptually. Peter Kivy (1975), for instance, argues that some aesthetic terms are "condition-governed," in the sense that certain non-aesthetic descriptions are sufficient for the application of the aesthetic terms. According to him, such aesthetic terms are equivalent to terms like 'intelligence' in that respect—although there is no necessary and sufficient definition of being intelligent, some descriptions are sufficient for the application

of the term. For instance, if the person "can understand Gödel's proof, has a high I.Q., speaks and reads seven modern and four ancient languages, and has won the Nobel prize for physics", then "you are not logically free to deny that he is intelligent" (p. 205). Similarly, Kivy argues, the description of "Bach's *Prelude in B-flat minor* (from Book I of the *Well-tempered Clavier*) as a twenty-four-measure piece in which the main theme (or a variation of it), occurs in every measure" is sufficient for the application of the term unity to the piece (p. 205).

Frank Sibley (1965) is another philosopher who has argued that aesthetic terms are condition-governed. Unlike Kivy, though, Sibley believes that no non-aesthetic descriptions are *sufficient* for the application of any aesthetic terms. Hence, no aesthetic terms are *positively* condition-governed. Sibley, however, argues that some aesthetic terms are *negatively* condition-governed, in that there are non-aesthetic descriptions that *preclude* the applications of them, or to put it another way, the negations of those non-aesthetic descriptions are "logically necessary" for the applications of the aesthetic terms (p. 153). His reasoning is as follows:

> It seems clear that with only pale pastels and no instances of bright (or apparently bright) colors in existence, there could be no examples of gaudy or garish coloring, and it would be dubious whether anyone could have learned, or whether there would be, any such concepts. Similarly, if there were only bright and intense colors, there could be none describable aesthetically as delicate. If all lines and movements were either straight or sharply angular, never being or giving the impression of being curving or

> flowing, there might be no use for the term "graceful." Were such relationships merely contingent, it would be conceivable that we might find occasional exceptions, a graceful straight line or some garish pastels (p. 153).

There are different available responses to this objection. First, one might argue that the aesthetic terms they mention are not in fact condition-governed. For instance, one could claim that the description that the main theme occurs in every measure in a musical piece is not sufficient for the application of the term unity, because such a piece might simply be boring rather than unified, depending on its other features (if unity is an inherently positive term). Similarly, one could come up with actual examples in which pastels are gaudy/garish or bright colors are delicate. Alan Goldman (2006), for instance, says that Sibley would not have made the claim that pastels cannot be garish "if he had ever seen the hotel facades in South Miami Beach" (p. 308).

Another possible response to this objection is to grant that the terms are condition-governed in some sense, while denying that the condition-governing is of a *conceptual* sort. This is how Beardsley (1974) and Levinson (2011) respond to Kivy and Sibley. Beardsley argues that while a graceful straight line is difficult to imagine, "so is a world in which fires send out waves of chilly air" (p. 246).[34] According to him, what makes the imagining difficult is "not conceptual analysis,

[34] It is worth noting that there is a sense in which such an imagining is not difficult. If the imaginability at issue is not the *perceptual* imaginability of a graceful straight line or garish pastels but whether we can imagine *that* we occasionally find a straight line graceful or some pastels garish, such an imagining seems easy—we only need to suppose that our psychological make-up is different (perhaps due to a different evolutionary history).

but experience" (p. 246). In a similar vein, Levinson states, "the possibility of rationally defending ascriptions of garishness by appeal to lower-level perceivable features does not show that there must be a conceptual connection between them; rather, a wide range of empirical experience with manifestations of garishness on the part of informed perceivers would be a perfectly adequate basis for such defense" (2011, p. 154).

Although I find these two responses plausible, I will in fact put forward a different response to the objection. Let us grant, for the sake of argument, that some conceptual connections between the aesthetic and the non-aesthetic exist. This, however, does not undermine my account of aesthetic explanation, for the following reasons.

First, even Kivy and Sibley acknowledge that conceptual connections exist only with respect to a very small number of aesthetic properties or terms.[35] Thus, even if I grant that a few instances of aesthetic explanation (such as that of the unity of Bach's *Prelude in B-flat minor*) are of a conceptual nature, I can still claim that in the vast majority of cases, aesthetic explanation should be understood in terms of causal relations between NA- and A-psychological events.

Second, even in such rare cases in which aesthetic explanation is conceptual, there is another sense—namely, in the metaphysical sense—in which aesthetic explanation can still be a matter of causal relations between NA- and A-

[35] What is the reason that conceptual connections exist only with respect to a very small number of aesthetic properties? I believe it is because, as I will discuss below, most non-aesthetic properties that are responsible for the presence of an aesthetic property are determinate properties, many of which are not conceptually construable.

psychological events.[36] The existence of a conceptual explanation does not preclude the possibility that what metaphysically connects the aesthetic and the non-aesthetic, even in such a case, is the causal relation between corresponding A- and NA-psychological events. Suppose, for instance, that the description that the same theme occurs in every measure in a musical piece is sufficient for the application of the term unity to the piece, and that this has to do with our concept of unity. Then we have an explanation of the piece's unity in terms of *our concepts*. At the same time, it might still be the case that what makes the object unified, *metaphysically speaking*, is the fact that the observer's experience of the same theme occurring in every measure gives rise to a characteristic experience of unity. In fact, it is possible that the reason we have the concept of unity, according to which a musical piece with the same theme occurring in every measure must be unified, is our psychological make-up—the experience of the same theme occurring in every measure is bound to give rise to the experience of unity.

So far, I have shown that even if there exist conceptual connections between the aesthetic and the non-aesthetic, it may still be the case that what metaphysically connects the two is causal relations between NA- and A-psychological events. Conceptual connection, however, is not the only kind of non-causal connection that may exist between the aesthetic and the non-aesthetic. Another possible connection between the aesthetic and the non-aesthetic is grounding. Grounding is a non-causal in-virtue-of relation that holds between, for

[36] Note that here and elsewhere in this essay, my concern is explanations in the ontological sense (i.e., explanations as facts in the world), not explanations as communicative acts.

instance, the fact that an object is maroon and the fact that it is red, and the fact that Socrates exists and the fact that {Socrates} exists, and so on—the former grounds the latter. And it is plausible that such a grounding relation holds between aesthetic and non-aesthetic facts. Audi (2012), for instance, mentions a vase being beautiful in virtue of its shape, color, and texture as an example of grounding (p. 690). Then why not, an objector might ask, understand aesthetic explanation in terms of grounding instead of causal relations between A- and NA-psychological events?

In response, let me first point out that there is a difficulty with understanding aesthetic explanation in terms of aesthetic grounding (the grounding of the aesthetic by the non-aesthetic). Many philosophers agree that what grounds a fact necessitates it—that is, the grounded fact must hold wherever the grounding fact holds. This means that the grounding fact of an object having a certain aesthetic property must cite all the non-aesthetic properties on which the aesthetic property supervenes. This conflicts with the selectivity of aesthetic explanation. As I briefly mentioned in Section 1, aesthetic explanation does not usually cite all those non-aesthetic properties. Of course, there are some ways in which the grounding account of aesthetic explanation may accommodate the selectivity—for instance, by claiming that the selectivity is merely pragmatic, and thus what really explains the presence of the aesthetic property is the presence of all those non-aesthetic properties. But, as I will illustrate in the following section, such a pragmatic move is not convincing in the case of aesthetic explanation.

I do not deny that aesthetic grounding *per se* holds. Aesthetic grounding is compatible with my account of aesthetic explanation. In fact, the reason the non-aesthetic grounds the aesthetic might be that there exist causal relations between A- and NA-psychological events. I believe that there are different species of grounding, and they may hold for different reasons. In the case of aesthetic grounding, the grounding relation may hold because of causal relations between A- and NA-psychological events. In fact, claiming that the non-aesthetic explains the aesthetic, because the non-aesthetic grounds the aesthetic, is unilluminating—it almost repeats the same fact in two different terms. My account, by contrast, can illuminate why the explanatory and the grounding relations hold: they hold in virtue of causal relations between A- and NA-psychological events.

(3) The third possible objection to my account that I will consider is that even if aesthetic explanation should be understood in terms of the kind of causal relation I have in mind between NA- and A-psychological events, aesthetic explanation is not causal explanation, even in the broad sense in which I have been using the term. There are accounts of explanation other than a causal account that are compatible with there being a causal relation between the explanans and the explanandum. On such views, it is not because the explanans causes the explanandum that the former explains the latter; the explanatory force comes from something else. I will show below that those other accounts of explanation are inadequate for aesthetic explanation.

One such account is the deductive-nomological account of explanation, according to which a set of propositions explains a phenomenon just when the

following conditions are met: they logically entail the phenomenon, contain at least one law that is essential to the entailment, and are all true.[37]

This essay is not the place to adjudicate between different accounts of explanation in general, but it should be clear that the deductive-nomological account is not adequate, at least for aesthetic explanation. This is because the deductive-nomological account requires *exceptionless* laws—otherwise the explanans would not logically entail the explanandum. But very few, if any, exceptionless laws exist with respect to the relations between the aesthetic and the non-aesthetic. Thus, the deductive-nomological account fails to be a general account of aesthetic explanation.

A proponent of the deductive-nomological account might respond that there are in fact numerous exceptionless laws about the relations between aesthetic and non-aesthetic properties, because the supervenience base of an aesthetic property fixes the aesthetic property. Thus, there exist numerous laws with the following form: an object with [insert a set of the non-aesthetic properties that an aesthetic property X supervenes on] has X. And such laws are exceptionless at least in some sense of the term (that is, if we hold human sensibilities and other contextual factors fixed). The problem with such laws, however, is that each of them would be applicable to virtually only one case, namely, the object in question. It is widely accepted that the supervenience base of an aesthetic property is very wide because any minor change in almost any

[37] See, for instance, Hempel & Oppenheim (1948).

non-aesthetic property of an object may potentially change its aesthetic properties. Given this, exceptionless laws understood in the above way would cite a very long list of non-aesthetic properties, which would be possessed virtually only by the object in question. If each law is applicable to only one object, it seems pointless to say that aesthetic explanation is a matter of subsumption under laws.

At this point, one might suggest that aesthetic explanation should be understood in terms of *ceteris paribus* laws, rather than exceptionless laws. *Ceteris paribus* laws are laws that include, whether implicitly or explicitly, a *ceteris paribus* (i.e., other things being equal) clause, and thus allow for exceptions. Examples include the following: defeasibly, matches light when struck; *ceteris paribus*, lying is wrong; the lack of vitamin C causes scurvy; and so on. Most laws in special sciences are *ceteris paribus* laws, so it is plausible (though controversial) that *ceteris paribus* laws can play explanatory roles.

I do not believe that aesthetic explanation should be understood in terms of *ceteris paribus* laws. First, there are some reasons to think that non-vacuous, informative *ceteris paribus* laws that connect the aesthetic and the non-aesthetic do not exist.[38] Sean McKeever and Michael Ridge (2011), for instance, argue that while there are non-vacuous and informative *ceteris paribus* laws that connect the non-moral and the moral, their argument for this claim does not apply to the

[38] To be clear, I do not deny that explanation of *overall aesthetic or artistic goodness* might be a matter of subsumption under *ceteris paribus*, or even exceptionless, laws. As I noted in Section 1, aesthetic explanation as I characterize it does not include explanation of overall aesthetic or artistic goodness.

aesthetics case.[39] According to them, in the case of *ceteris paribus* laws regarding morality—such as the law that lying is morally wrong, other things being equal—it is in principle possible to replace the other-things-being-equal clause with a specification of the finite factors that explain why lying is not morally wrong in some cases. By contrast, Mckeever and Ridge argue, such replacement does not seem possible in the case of aesthetics because aesthetic evaluation is *global*: aesthetic evaluation is usually concerned with a token rather than a type, and it "depends upon taking in the token—with all of its determinate features—as a whole" (p. 278).[40] One way to illustrate this asymmetry is to note that, while certain features of an action are never directly relevant to its morality (e.g., what color jacket the agent was wearing when performing the action), any feature of an object can be directly relevant to its aesthetic character. For these reasons, McKeever and Ridge suggest, there are no non-vacuous and informative *ceteris paribus* laws that connect the aesthetic and the non-aesthetic.[41] Laws such as "an object having a tapered shape is elegant except when it is not" are true, but only vacuously so, since no principled explanation as to when it is not is possible.

[39] McKeever and Ridge are concerned with the relation between non-aesthetic properties and overall aesthetic goodness, but their argument can be applied to the relation between non-aesthetic and aesthetic properties as well, which is the focus of this essay.

[40] A determinate property means the more specific one of a pair of property types (and the less specific one is called the determinable property). For instance, being red is determinate of being colored, and being colored is determinable of being red. Being determinate is thus a relative notion—with respect to being scarlet, being red is determinable rather than determinate, for instance. This relativity, however, is often ignored, and "determinate" usually just means very specific.

[41] To put it more precisely, their claim is that their argument for the claim that there are non-vacuous and informative *ceteris paribus* laws in the domain of morality does not apply to the domain of aesthetics, which opens the possibility that some other argument supports such laws in aesthetics.

What counts as genuine *ceteris paribus* laws, however, is a complex issue. So that my account of explanation does not depend on such an issue, let me provide a simpler reason why aesthetic explanation cannot be understood in terms of *ceteris paribus* laws. Even if there are some genuine *ceteris paribus* laws that connect the aesthetic and the non-aesthetic, this cannot provide a general account of aesthetic explanation because it fails to accommodate many instances of aesthetic explanation. The reason it fails has to with the fact that an explanation of an object's aesthetic property very often invokes determinate properties. For instance, as Sibley rightly observes, "[a] thing is graceful in virtue of being curved in exactly the way it is, not just in virtue of being curved" (1974, p. 11). Similarly, *Bird in Space* is elegant because of its exact shape and surface texture rather than its having a tapered shape and smooth surface texture (although, for the sake of convenience, I have been using the latter expression). Even if what explains the elegance is not having the exact shape and texture, it should be something much more specific than having a tapered shape and smooth surface texture, such as having a very similar shape and texture to those of *Bird in Space*. Thus many, if not all, *ceteris paribus* laws that connect non-aesthetic properties to aesthetic properties (if there are such laws) would invoke very specific properties.

Now, the problem with laws invoking such properties is that they are rather useless for explanation because, given that determinate properties are very specific, such laws would be applicable to only a very small number of cases. For instance, very few objects other than *Bird in Space* itself have the exact shape and surface texture of (or shape and surface similar enough to) *Bird in Space*. Thus,

the law that an object with the specific shape and texture is, other things being equal, elegant is useless in explaining most other cases of elegance. In fact, the situation is worse, because the explanatory base for the elegance of *Bird in Space* plausibly includes several determinate properties other than the exact shape and texture. Thus, it may well be the case that the law that cites all those properties can explain elegance only in the case of *Bird in Space*. This point would easily apply to many instances of aesthetic explanation, since aesthetic explanation usually invokes several determinate properties. One might, of course, argue that each specific case of aesthetic explanation involves its own *ceteris paribus* law only applicable to the object in question. But, at this point, the claim that aesthetic explanation is a matter of subsuming under laws starts to seem quite pointless, as I already pointed out with respect to exceptionless laws.

Let me now turn to the unificationist account of explanation. According to this account, explanation is a matter of illuminating connections between apparently unrelated phenomena.[42] For instance, Newtonian mechanics is explanatory in that it shows how movements of different objects are governed by the same laws. Although this account may be plausible as a theory of explanation in other areas, it is not adequate for aesthetic explanation. The reason is, again, that aesthetic explanation usually invokes determinate properties. An explanation involving determinate properties would rarely reveal connections between different objects, since there are few objects that share the same determinate properties. Recall the case of *Bird in Space*. The explanation that it is elegant

[42] See, for instance, Friedman (1974).

because of its particular shape and texture would not illuminate connections between different objects, since few objects would have exactly the same shape or texture. In fact, the basic idea of the unificationist account—explanation is a matter of unifying different phenomena—seems to conflict with our intuitive understanding of aesthetic explanation. In most cases, aesthetic explanation is a matter of highlighting the particularity of the object in question. Aesthetic explanation usually foregrounds what makes the object special and different from others, rather than what it has in common with other objects. Given these considerations, the unificationist account of explanation is inadequate for aesthetic explanation. Therefore, none of the major non-causal accounts of explanation are plausible with respect to aesthetic explanation.

It is worth nothing at this point that all the alternative accounts of explanation that I have considered above are inadequate for aesthetic explanation for a roughly similar reason: aesthetic explanation is based on the totality of the object's many, if not all, determinate properties. This feature of aesthetic explanation cannot be accommodated easily by accounts of explanation that appeal to some kind of abstraction or generalization, whether laws or unification. The causal account of explanation, on the other hand, can accommodate that feature easily because it does not need to appeal to abstraction or generalization.

This concludes my defense of the psychological account of aesthetic explanation. In the following section, I focus on aesthetic explanation's selectivity.

3. The selectivity of aesthetic explanation

In this section, I first illustrate the selectivity of aesthetic explanation, and then discuss what makes aesthetic explanation selective and what selection principle governs aesthetic explanation.

3.1. Selectivity

To illustrate how aesthetic explanation is selective, let me contrast it with aesthetic supevenience, the supervenience of the aesthetic on the non-aesthetic. It is widely agreed that aesthetic properties supervene on non-aesthetic properties.[43] Aesthetic supervenience is commonly understood as follows:

> *Aesthetic Supervenience*
>
> Two objects (e.g., artworks) that differ aesthetically necessarily differ nonaesthetically (i.e., there could not be two objects that were aesthetically different yet nonaesthetically identical: fixing the nonaesthetic properties of an object fixes its aesthetic properties). (Levinson 2011, p. 135)

How is aesthetic supervenience different from aesthetic explanation? Supervenience and explanation are different notions, so there is an obvious sense in which aesthetic supervenience is different from aesthetic explanation. For instance, supervenience is a relation between properties, while explanation is

[43] Existing disagreements mostly concern the details, such as the kinds of non-aesthetic properties aesthetic properties supervene on and, relatedly, the modal strength of the supervenience claim. See, e.g., Currie (1990a); Hick, (2012).

usually taken to be a relation between facts. Furthermore, they exhibit different formal features: supervenience is reflexive (any property supervenes on itself) and *non*-symmetric (sometimes it holds symmetrically while other times it holds asymmetrically), and explanation is irreflexive and asymmetric. Also, a property that an object necessarily has or fails to have (e.g., being self-identical or not being self-identical) supervenes on anything (because such a property cannot differ, period), and a property supervenes on its complement. By contrast, it is not the case that necessary truths and falsities are explained by anything, or that a fact is explained by its negation. [44]

The difference between aesthetic supervenience and aesthetic explanation that I want to emphasize, however, is the following: only some of the non-aesthetic properties that an aesthetic property supervenes on are cited in the explanation of the object having the aesthetic property. This is what I mean by aesthetic explanation being selective. Let me elaborate.

It is a commonplace in aesthetics that a slight change in an object's non-aesthetic properties (such as a slight change in a painting's background color) can lead to a change in its aesthetic properties. All those non-aesthetic properties that can affect a change in the object's aesthetic properties should be included in their subvenient base, since the base must fix the object's aesthetic properties for the supervenience relation to hold. Thus, an aesthetic property's subvenient base is very wide—it usually includes all perceptual properties of the object as well as

[44] See Benovsky (2012) for discussion on formal differences between aesthetic supervenience and explanation. He uses the term grounding instead of explanation, but what he means by it seems very close to what I mean by explanation.

(although not completely uncontroversial) some non-perceptual properties such as its genre and history of production. Now, what is important is that not all those non-aesthetic properties are cited in the explanation of the object's having the aesthetic property. Rather, in most cases, only a proper subset of the subvenient base of an aesthetic property is intuitively the explanatory base of its presence. A painting's background color, for instance, is rarely cited in an explanation of the painting's aesthetic properties, even if it is part of the subvenient base.

As a more concrete example, recall Brancusi's *Bird in Space*. I earlier mentioned that its elegance is partly explained by its shape and texture. The explanatory base plausibly includes more than the shape and texture, but whatever the explanatory base exactly is, there are some non-aesthetic properties of the work that do not belong to the explanatory base, while being part of the subvenient base of the elegance. For instance, the approximate size of *Bird in Space*—about 60 inches high—is part of the subvenient base of its elegance, since if the sculpture were much smaller (e.g., a few inches high) or bigger (e.g., 10 feet high), it might not be elegant.[45] However, intuitively, the sculpture's size does not, even partly, *explain* its elegance. That is, it is not the case that the sculpture is elegant partly *because of* its size.[46] As this example demonstrates, our intuitive understanding of aesthetic explanation is more selective than aesthetic supervenience.

[45] What is included in the subvenient base is plausibly not the exact size, but its approximate size. But even this approximate size does not seem to be part of the explanatory base.
[46] One might think that certain kinds of non-aesthetic properties, such as the size of a work, can never be part of an explanatory base. This is not the case. For instance, Barnett Newman's "zip" paintings are sublime partly because they are large.

Other philosophers have noted the selectivity of aesthetic explanation. Bender (1987), for instance, states that "many more properties than those which intuition would accept as responsible for an object's aesthetic character must be included in the supervenience base" (p. 37). For instance, while the absence of the drum seems to be responsible for the gossamer texture of Haydn's *La Roxalane* Symphony (no. 63), the absence of the bagpipe does not, even though the bagpipe's presence would destroy the gossamer texture (p. 36).[47]

Similarly, Frank Sibley (1965) distinguishes between two senses in which a specific object's aesthetic character depends on its non-aesthetic characteristics: what he calls total specific dependence, which seems close to aesthetic supervenience, and notable specific dependence, which seems close to aesthetic explanation. Sibley describes total specific dependence as follows: "[f]eatures one would hardly think of singling out as notably contributing to [an object's] aesthetic character—say, background colors, hardly noticed brush strokes, and so on—nevertheless do contribute because, being as they are, they at least allow it to have the character it has, a character it conceivably might not have if they were altered" (p. 138). On the other hand, Sibley points out, a critic rarely cites all those features when explaining an object's aesthetic character. Rather, the critic is concerned with notable specific dependence, that is, "[h]e is usually interested in much more pointed explanations: he tries to select certain peculiarly important or salient features or details" (p. 139).

[47] In this essay, I will assume, for the sake of simplicity, that a negative property (i.e., the absence of a property) can be a subvenient base or cited in an explanation. This is controversial, but nothing important for the purposes of this paper turns on this issue.

It should be clear by now, I hope, that our intuitive understanding of aesthetic explanation is more selective than aesthetic supervenience. Thus, a successful account of aesthetic explanation should make aesthetic explanation more selective than aesthetic supervenience. It should also illuminate why aesthetic explanation is more selective and what is the basis for such selection. Neither Bender nor Sibley attempts to answer these questions or provide an account of aesthetic explanation (or aesthetic determination, or notable specific dependence, as they call it).[48] The task I will undertake in the remainder of the essay is to demonstrate how my psychological account of aesthetic explanation can illuminate the selectivity of aesthetic explanation.

3.2. Why is aesthetic explanation selective?

On my account of aesthetic explanation, according to which aesthetic explanation is a kind of causal explanation broadly construed, the selectivity of aesthetic explanation comes from it being causal explanation. Causal explanation in general is selective—it usually cites only some of the factors that are necessary for the effect to occur. This phenomenon is called causal selection. A car crash, for instance, might be jointly brought about by a drunk driver, a bald tire, a slippery road, an approaching car, a blind corner, etc., in such a way that the crash would

[48] Sibley in fact briefly makes a tentative suggestion: "often in a work there are some features that strike us as making the most outstanding contribution, usually those in which a small alteration would work a remarkable aesthetic change" (1965, p. 139). But this suggestion will not work because, first, it only provides a difference in degree while the explanatory vs subvenient base distinction is all-or-none and, second, a factor that would work a remarkable aesthetic change may still not be part of an explanatory base. For example, the presence of the bagpipe would destroy the gossamer texture of Haydn's *La Roxalane* Symphony, even more than the presence of the drum would, but the absence of the bagpipe is not part of the explanatory base whereas the absence of the drum is.

not occur if one of those factors were absent.[49] The car crash, then, is causally dependent on all those factors. But not all those factors would be cited in an explanation of the crash. Similarly, an explanation of a forest fire would cite, say, a dropping of a lighted cigarette, without mentioning other factors necessary for the fire to occur, such as the fact that oxygen was present, there were flammable materials around and they were not too damp, etc.[50]

Now, given that causal explanation in general is selective in this way, it is not surprising that aesthetic explanation, which is a kind of causal explanation broadly construed, is selective as well. That is, the selectivity of aesthetic explanation is just an instance of the broader phenomenon of causal selection. I believe it is a merit of my account of aesthetic explanation that it can easily answer why aesthetic explanation is selective—it is because aesthetic explanation is causal explanation in some sense and causal explanation is selective. Answering this question may not be as easy for other accounts. For instance, it is hard to see how the deductive-nomological account of explanation, according to which the explanans must entail the explanandum, can accommodate the selectivity of aesthetic explanation (or explain away our intuition about it). It may be possible to do so, but it will need some additional resources that are not part of the account itself. In contrast, my account does not need any additional resources to accommodate the selectivity of aesthetic explanation. The selectivity naturally follows from my account.

[49] The example is from Lewis (1986a, p. 214).
[50] The fire example appears frequently in the literature. To my knowledge, the first appearance was in Hart and Honoré (1959).

3.3. The Selection Principle

What selection principle, then, governs aesthetic explanation? That is, what distinguishes between those non-aesthetic properties that are cited in the explanation of an object having an aesthetic property and the rest of the aesthetic property's subvenient base? This question, too, is a specific instance of a broader question—what is the principle for causal selection in general?

It is worth noting at this point that what selection principle turns out to be correct for aesthetic explanation does not affect the plausibility of my psychological account of aesthetic explanation. My account of aesthetic explanation is compatible with any selection principle, or even with the view that there is no objective selection principle (because the selection is merely pragmatic). Thus, for the purpose of defending my psychological account of aesthetic explanation, I do not need to endorse a particular selection principle for aesthetic explanation. Yet, the question as to what selection principle, if any, governs aesthetic explanation is an interesting question on its own. In the remainder of this essay, I will attempt to answer this question by consulting the literature on causal selection in general.

Finding a selection principle for causal selection is notoriously difficult. In fact, many philosophers believe that there is no objective or metaphysical basis for the selectivity of causal explanation. David Lewis, for instance, argues that "to explain an event is to provide *some* information about its causal history", while

including in the causal history anything on which the event in question counterfactually depends (1986a, p. 217, emphasis added).

According to this position, the selectivity of causal explanation is pragmatic—it is merely based on conversational pragmatics or other practical considerations that vary with the context of inquiry, such as the inquirer's interest. Lewis, for instance, states that "there are ever so many reasons why it might be inappropriate to say something true. It might be irrelevant to the conversation, it might convey a false hint, it might be known already to all concerned, and so on" (2004, p. 101) and that "we may select the abnormal or extraordinary causes, or those under human control, or those we deem good or bad, or just those we want to talk about" (1986b, p. 162).

This essay is not the place to adjudicate between different positions on causal selection; it is only concerned with the selectivity of *aesthetic explanation*. That said, the pragmatic account is at least *prima facie* implausible when it comes to aesthetic explanation. For the success of the pragmatic account, it is crucial to offer specific pragmatic mechanisms that can explain our intuitions about the selectivity of explanation. While such specific pragmatic mechanisms are available at least in some cases of causal selection, it is unclear whether there are any plausible pragmatic mechanisms in the case of aesthetic explanation. The car crash, for example, could be explained due to the different interests of inquirers.[51]

[51] There are other cases of causal selection, such as the forest fire case, that are more difficult for the pragmatic view to account for. Jonathan Schaffer (2012), for instance, rightly argues that conversational pragmatics cannot fully explain the case because, first, saying that the presence of oxygen is the cause of the fire seems like not just saying something irrelevant but false; second, ordinary speakers would not just avoid asserting the claim but assert its negation (i.e., they would

A police officer, for instance, who is looking for legal violation, would cite the drunk driver in the explanation of the crash, whereas a car engineer might focus on problems with the car, such as the bald tire. Neither of these explanations appears objectively better than the other. They seem, metaphysically speaking, equal.

In the case of aesthetic explanation, however, it is hard to see what pragmatic mechanisms can explain our intuition about its selectivity, such as the intuition that the shape, but not the size, of *Bird in Space* explains its elegance.[52] To begin with, unlike the car crash case, there seems to be no context of inquiry in which citing the work's size is a good explanation. If someone happens to focus on its size, perhaps because she just saw much smaller sculptures or she just happens to be interested in an artwork's size, to claim that *Bird in Space* is elegant because of its size, we would probably respond that she gave the incorrect explanation because she was influenced by her contingent situation. Thus, appealing to context of inquiry does not seem to account for the selectivity of aesthetic explanation. Of course, if the context of *occurrence* were different (for instance, if most existing sculptures were much smaller than *Bird in Space*), this explanation could hold good. But the context of occurrence (i.e., the context in which the work is) is not a pragmatic consideration, although the context of

---

assert that the presence of oxygen did not cause the fire); and, third, the assertion is not cancellable by adding that the speaker did not mean it while a conversational implicature is cancellable that way. As I mention below, the same points are applicable to aesthetic explanation as well.

[52] I am not claiming that no pragmatic consideration is relevant to the selectivity of aesthetic explanation. After all, the fact that aesthetic explanation does not cite all the factors on which an aesthetic property supervenes on might have to do with limited human cognitive abilities. But which factors are cited, I believe, is not a completely pragmatic matter.

inquiry (the context in which the inquirer is) is. The context of occurrence is external to the object itself, but it is still an objective, not pragmatic, consideration.

Appealing to conversational pragmatics would not help either. Although intuition may vary, saying that *Bird in Space* is elegant because of its size, or that the gossamer texture of Haydn's *La Roxalane* Symphony is in virtue of the absence of the bagpipe, seems *false*, rather than saying something *irrelevant* (or some other case of violating conversational pragmatics). And ordinary speakers are willing not just to avoid making such assertations, but also to assert their negations, saying, for instance, that the absence of the bagpipe is not part of the reason the symphony has a gossamer texture. These suggest that the selectivity here is not just a matter of conversational pragmatics.

One might attempt to appeal to common knowledge, arguing, for instance, that our common knowledge that Haydn's symphonies usually lack the bagpipe (or to put it another way, it would be quite unusual for such a symphony to include the bagpipe) explains why we do not cite it as an explanation of its gossamer texture. But if that is the case, what really accounts for the selectivity seems to be the *fact* that Haydn's symphonies usually lack the bagpipe, not our knowledge of the fact. After all, a false common belief would not influence what factors should be cited in an explanation.[53]

[53] Recall that my concern in this essay is explanations as facts in the world, not as communicative acts.

Thus, it seems that no pragmatic mechanism that is currently available in the literature can account for the selectivity of aesthetic explanation. This is not, of course, a knock-down argument against the pragmatic account. But I believe it does shift the burden of proof. Given that our actual practice of aesthetic explanation is selective and that, as I will show below, there is a plausible non-pragmatic account of the selectivity, the burden lies on the proponents of the pragmatic account to provide a specific pragmatic mechanism that can explain away our intuition.

If, as I suggest, there is a non-pragmatic selection principle for aesthetic explanation, what exactly is it? My answer is based on Franklin-Hall's Causal Economy account of explanation (2015). According to this account, "good explanations—those deemed 'complete' and capable of providing understanding—are special for their *economy*: they 'cost less,' in virtue of being abstract, and 'deliver more,' in virtue of citing causal influences that make the event to be explained stable or robust" (p. 434, emphasis in original). The principle for causal selection that she proposes is thus that an event's narrow causes are those that are cited in the explanation(s) of the event that has the highest delivery-to-cost ratio among all the candidate explanations—thus she calls this principle the "biggest bang-for-your-buck" standard (p. 423).[54] The rest of the

[54] She uses the term "narrow causes" instead of "causes", saying that on many accounts of causation, any factors that are necessary for an event to occur count as its causes (p, 417). I will follow her terminology to remain agnostic about whether only those factors that are cited in the explanation of an event count as its causes. Strictly speaking, the distinction between causes and background conditions does not need to correspond to the distinction between causal influences that are cited and not cited in causal explanation. One might, for instance, argue that any causal influences are causes, while maintaining that there is some other metaphysical distinction between the causal influences that are cited and not cited in causal explanation. This essay is concerned

causal influences—that is, the factors necessary for the target event to occur—that are not cited in this explanation are background factors.[55] As we will see later, which explanation has the higher delivery-to-cost ratio is not completely determined by pragmatic considerations, making Franklin-Hall's account a non-pragmatic account of causal selection.

Why choose her account over other non-pragmatic accounts of causal selection? My aim in this essay is not to defend her account as the correct account of causal selection. My claim is just that an account mirroring hers is the correct account for the selectivity of aesthetic explanation. Yet her account does have some virtues that other non-pragmatic accounts lack. As Franklin-Hall herself points out, other accounts use some kind of relativization, based on "an actual population," "an explanatory framework", "a cause-and-effect-event-contrast," "an object of comparison", and so on (p. 421). On such accounts, the cigarette-dropping is cited in the explanation of the fire because, roughly speaking, the explanatory task at issue is to explain what caused the fire *given that oxygen was present*. But those accounts do not attempt to explain what grounds such relativization. As Franklin-Hall puts it, "they out-source the explanatory task by suggesting that just what is selected is a consequence of which actual populations, frameworks, etc., are included in explanatory requests. And on the selection of *these* they remain silent" (p. 421, emphasis in original). In this respect, those

---

with the selectivity of causal explanation rather than the causes vs. background conditions distinction.

[55] Her account is pluralistic in that there can be more than one explanation that has the same highest delivery-to-cost ratio. In such a case, she can allow that pragmatic considerations may play a role in deciding which explanation, among the ones that have the same highest delivery-to-coast ratio, is the best.

accounts offer only incomplete solutions to the problem of causal selection. As we will see below, Franklin-Hall's account does not involve such a relativization.

Let me first elaborate on her account by explaining what she means by delivery and by cost when she claims that an event's narrow causes are those that are cited in the explanation that has the highest delivery-to-cost ratio. I will then modify it to apply to aesthetic explanation.

By the *cost* of an explanation, Franklin-Hall means its "total content," which is measured by "the number of ways that the world might have been that the explanation rules out" (p. 426). Thus, an explanation that cites fewer factors costs less. For instance, an explanation of a forest fire that only cites a cigarette-dropping costs less than an explanation that also cites other necessary factors such as the presence of oxygen.

The *delivery* of an explanation, on the other hand, is understood in terms of the "stability boost" it provides for the target event—that is, how much the factors cited in the explanation increase the stability of the target event to be explained, with the stability of an event meaning the degree to which the event would have happened even if other things had been different (p. 426). Thus, the stability boost is the difference between the *baseline stability* of the target event (i.e., the stability of the event simpliciter) and what Franklin-Hall calls its

construct stability—the stability of the target event *given the occurrence of the factors cited in the explanation*.[56]

The baseline stability is measured by the number of nearby possible worlds in which the target event takes place. For instance, in the forest fire case, the baseline stability of the target event (i.e., the forest fire) is the number of nearby possible worlds in which a fire occurs in the forest (regardless of what causes it). The construct stability is measured by, first, supposing that the factors cited in the explanation (e.g., a cigarette-dropping) occur in all nearby possible worlds, even in those worlds in which the factors would otherwise be absent without this supposition; and, second, by counting the number of all nearby possible worlds in which the target event (e.g., a fire in the forest) would occur. Now, the stability boost of the explanation is the construct stability minus the baseline stability, which reflects the *increase* in the target event's stability in virtue of the factors cited in the explanation. Note that on this account, an explanation provides a high stability boost when the factors cited in it are "those that are themselves unstable simpliciter [that is, they are absent in many nearby possible worlds], but such that, given their occurrence, the event to be explained is very stable" (pp. 427-428). If the factors cited in the explanation are themselves stable simpliciter—that is, they are already present in many nearby possible worlds—the supposition that they are present in all nearby possible worlds will

[56] Here and elsewhere in this essay, I am glossing over details of Franklin-Hall's account that are not important for the purposes of this essay.

not increase the stability of target event much, which means the explanation's stability boost is not high.

According to this account, the cigarette-dropping is the narrow cause of the forest fire because the explanation that cites it has the highest delivery-to-cost ratio among all candidate explanations. It has a low cost since it cites only one factor. And it delivers a lot because it offers a high stability boost to the fire. This is because in several nearby possible worlds, the cigarette-dropping would not happen, but if we suppose that it happens in those worlds, a fire would occur in most of them (because background factors such as oxygen would already be present in most nearby possible worlds—a possible world in which oxygen is absent in the forest would be quite far from the actual world). On the other hand, adding the presence of oxygen to the explanation, while increasing the cost significantly, does not significantly boost the stability of a fire because in most nearby possible worlds, the oxygen would already be present anyway. By contrast, if a fire occurs in a plant where oxygen is supposed to be absent, the narrow cause is intuitively the presence of oxygen. Franklin-Hall's account explains this intuition as well—in this case, but not in the case of the forest fire, the presence of oxygen is unstable simpliciter and, thus, can provide a significant stability boost to the fire.

Now let me apply Franklin-Hall's account to aesthetic explanation. To do so, I need to modify it a bit. She understands stability boost in terms of possible worlds ordered by physical similarities, but such an ordering does not seem to be relevant to aesthetic explanation. For instance, in terms of physical similarities,

the possible worlds in which *Bird in Space* is of a different texture and the possible worlds in which the sculpture is of a different size seem to be almost equally distanced from the actual world. Thus, if we order possible worlds by physical similarities, we cannot differentiate the explanatory significance of the texture and the size. For this reason, I suggest understanding stability boost in terms of possible worlds ordered based on *artistic* similarities.[57] That is, we should imagine possible worlds where the actual artwork in question is replaced with other artwork that differs from the actual one to a varying degree and everything else is identical to the actual world, and the distance between a possible world and the actual world is determined by how the work in the possible world is similar to the actual work in terms of artistic considerations, such as genre, style, art historical period, and so on.[58] For instance, if the actual work in question is Brancusi's *Bird in Space*, we have in nearby possible worlds a sculpture of a similar style and genre that was made around the same time as *Bird in Space*.

Now, I propose that the reason the explanation of *Bird in Space*'s elegance cites only some of its subvenient base, such as its shape and texture, while leaving out the rest of the subvenient base, such as its size, is that the explanation that cites only the former factors maximizes the delivery-to-cost ratio. Citing the sculpture's shape and texture substantially boosts the stability of the target event

[57] I wish to leave open how aesthetic explanation in the case of non-art objects should be understood.
[58] I use the possible world talk to maintain the continuity with Franklin-Hall's account, but you can imagine a web of artworks, instead of possible worlds, if you wish.

(i.e., it being elegant). This is because, in many nearby possible worlds, the work would not have the same shape and texture as *Bird in Space* (hence, this explanans is unstable simpliciter), and supposing that the work has the exact shape and texture of *Bird in Space* would make it elegant in many of those worlds. On the other hand, citing the size in addition to the shape and texture increases the cost substantially but does not substantially boost the stability of the target event. This is because first, in the nearby possible worlds, the objects would be in the same approximate size range, and second, even in those worlds where the work is significantly smaller or larger than *Bird in Space*, making it similar to *Bird in Space* size-wise would not make it elegant in many of the worlds where the work has a different shape or texture from that of *Bird in Space*.

Consider another example: while the absence of the drum seems to be responsible for the gossamer texture of Haydn's *La Roxalane* Symphony, the absence of the bagpipe does not. My proposal can explain why. The absence of the drum is unstable simpliciter, since, as Bender notes, the drum is found in Haydn's other symphonies (as well as in other similar symphonies by other composers of the same period) (1987, p. 36). Thus, in many nearby possible worlds, the musical piece would include the drum. By contrast, the absence of the bagpipe is stable simpliciter because the bagpipe would be absent in the musical piece in many nearby possible worlds (a symphony that includes the bagpipe is quite distant, in terms of style, genre, period, etc., from Haydn's *La Roxalane* Symphony—at least further than a symphony including the drum). Therefore, while the absence of the drum substantially boosts the stability of the target event

(i.e., the work having a gossamer texture), the absence of the bagpipe does not. Thus, the explanation that cites the absence of the drum has a higher delivery-to-cost ratio than the explanation that cites the absence of the bagpipe. Accordingly, the absence of the drum, but not the absence of the bagpipe, is part of the explanation of the symphony's having a gossamer texture.

As these examples illustrate, non-aesthetic properties cited in the explanation of the presence of an aesthetic property usually possess the following two features: (1) they are absent in many other artistically similar works, and thus they tell us what is special about this particular work; and (2) if those similar works had the non-aesthetic properties, they would likely have the aesthetic property as well, which reflects the non-aesthetic properties' power to contribute to the aesthetic property. This way, my selection principle captures how aesthetic explanation is a function of not only the intrinsic properties of the work in question, but also its place in the larger artistic context.[59]

Let me end this section by considering a possible objection to my selection principle. It could be argued that my suggestion cannot handle cases in which the object in question has the aesthetic property at issue in every nearby possible world. Suppose, for instance, that we want to explain why a certain classic ballet performance is elegant. In this case, the performances in all nearby possible worlds, which belong to the same or a similar genre, are elegant as well (or let us just assume so for the sake of argument). If so, the objection goes, the explanation

[59] It is also worth noting that my selection principle is compatible with not just my psychological account but with other accounts of aesthetic explanation. I believe this is a merit of my selection principle.

of the performance being elegant that is intuitively correct (for instance, it is elegant because it emphasizes elongated lines) would not count as the correct explanation because its stability boost would be zero (since the performance in every nearby possible world is already elegant, the stability of the target event is already maximized). How, then, can my proposal explain why emphasizing elongated lines is responsible for the performance being elegant?

I have several responses to this objection. To begin with, in cases like this example, what we really want to explain is not why the performance is elegant, period, but why it is elegant to the extent that it is elegant (for instance, why it is exceptionally elegant). And this fine-grained aesthetic property is not present in the performance in every nearby possible world. Most importantly, the fact that the stability boost of the intuitively correct explanation is zero is not a problem by itself, insofar as it has the highest delivery-to-cost ratio among all candidate explanations. If the other candidate explanations have negative stability boosts, then the explanation in question would still possess the highest delivery-to-cost ratio. And it is quite plausible that intuitively incorrect explanations would have negative stability boosts. For instance, the explanation according to which the performance is elegant because the dancers are wearing tutus might well have a negative stability boost because in some nearby possible world, adding this feature to the performance might result in its no longer being elegant. Of course, not all intuitively incorrect explanations have negative stability boosts. There might be a feature that the performance in every nearby possible world happens to have but is intuitively irrelevant to elegance (for instance, having a narrative). But

such cases can be dealt with by expanding the set of relevant nearby possible worlds so that it will include ballet performances that do not have a narrative. With respect to this set of possible worlds, the explanation that cites elongated lines would have a higher delivery-to-cost ratio than the explanation that cites a narrative.

4. Conclusion

I have argued so far that the explanation of aesthetic properties by non-aesthetic properties is a kind of causal explanation broadly construed and that this causal nature of aesthetic explanation can illuminate the selectivity of aesthetic explanation. By way of a conclusion, and to prevent possible misunderstandings of my account, in this section I will consider some implications (or lack thereof) of the account for adjacent issues in aesthetics.

First, my claim that aesthetic explanation is causal explanation may appear to be in tension with the fact that we often can provide reasons for the judgment that an object has a certain aesthetic property.[60] This is not the case. My causal account is perfectly compatible with the fact that aesthetic judgments call for, and can be justified by, reasons. This is because the relations between a judgment and the reasons we provide for it can be based on causal relations. I might, for instance, judge that soda is unhealthy and cite its other properties (e.g., it contains a lot of sugar) as a reason for the judgment. This reason supports the judgment in virtue of the causal relation between ingesting large quantities of sugar and its

[60] For related discussion, see Bender (1997) and Shiner (1996).

harmful effects on the body. In a similar way, citing the shape and texture of *Bird in Space* supports the judgment that it is elegant in virtue of a causal relation between the corresponding NA- and A-psychological events. This is not to say that the judge needs to be aware that the explanatory relation holds in virtue of the causal relation between the NA- and A-events. It also is not necessary that the NA- and A-events in question are those of the judge themselves—the judge may not be the right observer.

Second, my psychological account has no direct implication for the generalism vs. particularism debate, which concerns whether there is any general principle in artistic evaluation. First of all, my account is not concerned with the relationship between a work's overall artistic or aesthetic value and its aesthetic properties, so my account simply is irrelevant to the generalism vs. particularism debate with respect to this relationship.

Even when it comes to the relationship between a work's non-aesthetic and aesthetic properties, which is the topic of this essay, my account does not directly support particularism or generalism. Although I earlier mentioned Mckeever and Ridge's argument against *ceteris paribus* laws in aesthetics, I wish to remain neutral as to whether their argument is successful (even if it is successful, it refutes only one particular argument for generalism).

My account of aesthetic explanation, however, has an implication for how the generalism vs. particularism debate should proceed. If my account is right, the debate boils down to whether there are general principles regarding causal

relations between NA-psychological events and A-psychological events. Thus, the debate is partly empirical—it is about the psychological tendencies of the right observer—but only partly so. We will need to explore, non-empirically, how to characterize the right observer and the right conditions and what psychological event corresponds to each aesthetic property. Furthermore, we will need to discuss what counts as a general principle in the case of complex causal relations like those between NA- and A-psychological events, in which there are usually several interconnected causes working together to bring about more than one effect at the same time.

Chapter 3. The Fictionality Puzzle, Fictional Truth, and Explanation

1. Introduction

This essay proposes a novel solution to the fictionality puzzle.[61] To illustrate the puzzle, consider the following short story:

> *Death*
>
> Jack and Jill were arguing again. This was not in itself unusual, but this time they were standing in the fast lane of I-95 having their argument. This was causing traffic to bank up a bit. It wasn't significantly worse than normally happened around Providence, not that you could have told that from the reactions of passing motorists. They were convinced that Jack and Jill, and not the volume of traffic, were the primary causes of the slowdown. They all forgot how bad traffic normally is along there. When Craig saw that the cause of the bankup [sic] had been Jack and Jill, he took his gun out of the glovebox and shot them. People then started driving over their bodies, and while the new speed hump caused some people to slow down a bit, mostly traffic returned to its normal speed. So Craig did the right thing, because Jack and Jill should have taken their argument somewhere else where they wouldn't get in anyone's way. (Weatherson, 2004, p. 1)

[61] The fictionality puzzle was initially discussed, along with several other puzzles, under the name of imaginative resistance. It was not until Weatherson (2004) that these puzzles were distinguished from one another (note that Weatherson uses a different term "alethic puzzle" to refer to the fictionality puzzle). This essay concerns only the fictionality puzzle. Although I do not deny that it may be connected to other puzzles of imaginative resistance, I will not discuss this issue here. For an argument for such a connection, see Stear (2015). For imaginative resistance in general, see, e.g., Walton (1994; 2006) and Gendler (2000; 2006).

Most philosophers agree that it is not fictional (i.e., not true in the fiction) that Craig did the right thing, despite the claim being explicitly asserted. This is puzzling considering that a range of outlandish claims (e.g., people travel in time, a man turns into an insect while maintaining his identity, people dine at a restaurant at the end of the universe, etc.) are easily made true in fiction, often by a simple assertion. Let us call cases like *Death*, in which an explicitly asserted deviant proposition fails to be fictional, puzzle cases.[62] Puzzle cases are not restricted to morally deviant ones like *Death*: some puzzle cases discussed in the literature concern non-moral propositions (e.g., a nonjoke like "A maple leaf fell from a tree" is extremely funny) (Walton, 1994, p. 44) or even non-evaluative propositions (e.g., something indistinguishable from a knife is a television) (Weatherson, 2004, p. 5). The fictionality puzzle concerns why explicitly asserted propositions fail to be fictional in such puzzle cases.

This essay focuses on one influential line of solution to the puzzle, first suggested by Kendall Walton and subsequently developed by Brian Weatherson and then Nils-Hennes Stear. While the solutions these philosophers propose are not the same, their solutions share an important commonality: they all focus on the relation between the higher-level proposition in question (e.g., Craig did the right thing) and the lower-level proposition that is supposed to ground it in some sense (e.g., Craig killed Jack and Jill because they were holding up traffic). The basic idea, roughly speaking, is that the fictionality puzzle arises due to some kind of

[62] I am following Stear's (2015) terminology here.

conflict between these two levels. I will call this line of solution the *interlevel relation solution*.[63]

I believe the interlevel relation solution is the most promising solution currently available in the literature. I will argue, however, that the solution in its current form is incomplete. My aim in this essay is to articulate why it is incomplete and then to fill in the gap, thereby providing a complete solution to the fictionality puzzle.

I begin in Section 2 by elucidating the sense in which the interlevel relation solution in its current from is incomplete. I will show that the solution leaves some crucial questions unanswered. Over the remainder of the essay, I argue that we can find plausible answers to those questions—and, accordingly, a complete solution to the fictionality puzzle— by understanding the generation of fictional truth in terms of inference to the best explanation. In Section 3, I first motivate the proposal that the generation mechanism of fictional truth is based on inference to the best explanation. I then demonstrate in Section 4 that this way of understanding fictional truth offers a complete solution to the fictionality puzzle.

2. The Interlevel Relation Solution

The interlevel relation solution was first suggested by Walton (1994). He proposes that the fictionality puzzle arises due to certain determination or dependence relations that "cannot easily be different in fictional worlds and in the real one," while acknowledging that "[w]hy this is so, and what kind of determination or

[63] For other solutions to the fictionality puzzle and criticisms of them, see Weatherson (2004).

dependence is involved, is still a mystery" (1994, p. 44). Walton's suggestion was subsequently developed by Weatherson and then Stear.[64]

Weatherson's (2004) solution to the fictionality puzzle has two main elements, what he calls the rule of *Virtue* and "That's all" clauses. First, *Virtue* goes as follows:

> *Virtue*
>
> If *p* is the kind of claim that, if true, must be true in virtue of lower-level facts, and if the story is about those lower-level facts, then it must be true in the story that there is some true proposition *r* which is about those lower-level facts such that *p* is true in virtue of *r* (Weatherson, 2004, p. 18).

Weatherson argues that deviant propositions in puzzle cases fail to be fictional because they violate *Virtue*. For instance, in *Death*, the proposition that Craig did the right thing is the kind of claim that, if true, must be true in virtue of lower-level facts (that is, the proposition is not about a primitive fact), and the story is about those lower-level facts (e.g., Craig killed Jack and Jill because they were holding up traffic). But *Death*'s lower-level facts conflict with the proposition that

[64] There are complexities regarding the exact nature of the interlevel relation at issue. Walton (1994) often uses the term *supervenience*, but it should be noted that his essay was written around the time when the notion of supervenience was used more broadly, to include what is now called *grounding*. Weatherson (2004), on the other hand, says that the kind of interlevel relation at issue is not supervenience, instead using the expression "in virtue of" (p. 16). Stear (2015) uses the term grounding, but he mentions that the kind of grounding relation relevant to the fictionality puzzle is not the same as what grounding is usually taken to mean by metaphysicians (p. 14, fn.32). I will not delve into these complexities because they do not affect the purposes of this essay. Instead, I will assume that Walton, Weatherson, and Stear all have in mind the same or a similar enough notion, and I will use the term interlevel relation to refer to it.

Craig did the right thing. It cannot be the case that Craig did the right thing in virtue of his killing Jack and Jill for holding up traffic. Therefore, Weatherson argues, the proposition that Craig did the right thing violates *Virtue* and, accordingly, fails to be fictional.

"That's all" clauses, on the other hand, are needed to exclude the possibility of some unwritten lower-level fact that would morally justify Craig's action (e.g., Jack and Jill were about to kill everyone in sight including Craig) being imported into the fiction. In such a case, the proposition that Craig did the right thing would not violate *Virtue*. Weatherson argues that fiction has an implicit "That's all" clause that prohibits importing such virtue-preserving lower-level facts.

Although I agree with Weatherson that interlevel conflicts are important for solving the fictionality puzzle, I believe his solution is incomplete for two reasons. First, as Stear (2015) rightly points out, Weatherson's "That's all" clause is too strong (we will revisit this point when discussing Stear's view). Second, there is an important question regarding *Virtue* to which Weatherson does not provide a satisfactory answer: why should in-virtue-of relations invoked in *Virtue* be the same as those in reality? That is, why can it not be the case that, *in the fiction*, Craig's killing Jack and Jill is morally right in virtue of the fact that he killed them for holding up traffic? After all, it is just a fiction.

Weatherson in fact mentions that in-virtue-of relations might differ between fiction and reality in some cases, although he does not elaborate on this

possibility (2004, p. 7). He maintains, however, that "there is a default assumption that these relations are imported into stories or imaginations [from reality], and it is not easy to overcome this assumption" (2004, p. 17). Weatherson provides a reason in support of this default assumption: although the author can determine which world we are discussing by dictating what lower-level facts are true in her fiction, "once we have locked onto the world being discussed, the author has no special authority to say which concepts, especially which higher-level concepts like RIGHT. . ., are instantiated there" (2004, p. 22).

This response, however, is not fully satisfactory. Weatherson may well be right that there is such a default assumption, but this cannot be the whole story given that there might be exceptions to it as he himself acknowledges. We still do not know exactly when and how the default assumption might be overridden, and, accordingly, interlevel relations might differ across fiction and reality. As long as an explanation of the exceptions is absent in Weatherson's account, his solution to the fictionality puzzle, in its current form, is incomplete.

Nils-Hennes Stear (2015) adopts Weatherson's basic ideas about interlevel relations but with significant modifications. Like Weatherson's solution, Stear's has two elements: the notion of exhaustivity, which replaces Weatherson's "That's all" clause, and the notion of adequacy, which replaces *Virtue*. Stear argues that puzzle cases arise when, "on the best interpretation(s), a work attempts to prescribe a claim that, fictionally, is *inadequately* and *exhaustively* grounded" (2015, p. 12, emphasis in original). Grounds for a claim are *inadequate* when they fail to make the claim true, and grounds are *exhaustive* when "there are no

additional grounds available (that is, true) that will ground [the claim]" (Stear, 2015, p. 12).

Stear's solution differs from Weatherson's in two important respects. First, while no importation of virtue-preserving lower-level facts is permissible according to Weatherson's "That's all" clause, Stear rightly points out that importing a virtue-preserving lower-level fact is sometimes permissible, as the following story demonstrates:

> *Denise*
>
> Denise went to bed complaining of a strong headache. Her friends and family wished her goodnight. Three days later they buried her in the grounds of the old church. In the circumstances, it was the proper thing to do. (Stear, 2015, p. 11)

Stear argues, and I agree, that, in this case, it is permissible to import the virtue-preserving lower-level fact that Denise died before the burial. As for when importing a virtue-preserving lower-level fact is permissible, Stear contends that it is just when, on the best interpretation(s) of the fiction, the higher-level claim in question is not exhaustively grounded. In *Denise*, the claim that burying Denise was the proper thing to do is not exhaustively grounded, allowing for the importation of the virtue-preserving lower-level fact that she died before the burial, whereas, in *Death*, the claim that Craig did the right thing is exhaustively grounded, so no importation of virtue-preserving lower-level facts is permissible.

The second important difference between Stear's and Weatherson's solutions is that Stear takes a more contextual approach to interlevel relations in fiction. Recall that in Weatherson's *Virtue* the default assumption is that interlevel relations remain the same across fiction and reality, although he admits, without further elaboration, that this assumption might be overridden in some cases. Stear, on the other hand, argues that interlevel relations in fiction can differ from those in reality, depending on a range of contextual factors including the genre of the fiction, other explicit claims the author makes in the story, relevant conventions, and so on. In such cases, a deviant higher-level claim can be, in the fiction, adequately grounded by a lower-level claim that, in reality, would not adequately ground the higher-level claim.

An important question, then, is exactly why, in a puzzle case like *Death,* the relevant interlevel relation remains the same across fiction and reality, making the deviant higher-level claim inadequately grounded in fiction. In other words, why does the explicit assertion that Craig did the right thing fail to establish that the relevant interlevel relation in *Death* differs from that in reality? Stear's answer is that, given contextual factors like those mentioned above that constrain interpretation, such a reading is "simply interpretively inappropriate" in puzzle cases (2015, p. 15).

A similar question can be raised with respect to exhaustivity: why, in a puzzle case like *Death*, a deviant claim is exhaustively grounded? In other words, what makes the difference between *Death* and *Denise* in terms of exhaustivity? Again, Stear's answer seems to be that it is ultimately a matter of interpretation.

Recall his proposal that the fictionality puzzle arises when "*on the best interpretation(s)*, a work attempts to prescribe a claim that, fictionally, is *inadequately* and *exhaustively* grounded" (2015, p. 12, emphasis added). According to this proposal, the fictionality puzzle eventually boils down to interpretation.

I believe Stear is right that, in a puzzle case, the deviant claim is inadequately and exhaustively grounded. I also agree that whether a claim is adequately or exhaustively grounded in a fiction is a matter of interpretation, which is affected by a variety of contextual factors.[65] In explicitly foregrounding interpretation and its dependence on context, Stear's solution marks an improvement over Weatherson's.

Stear's account, however, still falls short of a complete solution to the fictionality puzzle. His solution does not tell the whole story unless we can answer the question: *why* is the best interpretation(s) of a puzzle case such that the deviant claim is inadequately and exhaustively grounded, while deviant claims are interpreted differently and thus manage to be fictional in non-puzzle cases such as science fiction and fantasy? Stear remains silent on this question. No doubt, the fictionality puzzle eventually boils down to interpretation. But to simply accept

[65] The term interpretation may mean different things. The kind of interpretation relevant here is interpretation with respect to fictional truth, not literary meaning. Although literary meaning and fictional truth may overlap, some instances of literary meaning go beyond fictional truth, as when we say that the novel *American Psycho* is a critique of the shallow capitalist society. Similarly, some instances of fictional truth do not seem to be part of literary meaning. For instance, the proposition that gravity exists is true in many realistic fictions, but few of them would have the proposition as part of their literary meaning. For discussion on the distinction and relationship between fictional truth and literary meaning, see Folde (2015).

this fact without exploring why puzzle cases are interpreted the ways they are interpreted is to simply outsource the explanatory task required to fully solve the fictionality puzzle.

In fairness to Stear, he acknowledges that his solution is not meant to provide "complete theories of fictional grounding and import" (2015, p. 12, fn.27). He is also skeptical that such theories could ever be formulated. If what he means by "complete theories" here are complete algorithms that determine matters of fictional grounding and importation in any fiction, he is probably right that they are unlikely to be found. However, this does not mean that no principled answer whatsoever can be given as to why deviant claims in puzzle cases, unlike deviant claims in non-puzzle cases, are interpreted to be exhaustively and inadequately grounded. In the remaining parts of this essay, I attempt to offer a principled answer to this question.

Note that this question has two parts, regarding exhaustivity (or fictional importation) and adequacy (or fictional grounding). Thus, there are two questions to answer. First, why is the deviant claim in a puzzle case interpreted to be exhaustively grounded? To put it in another way, why is it not permissible in a puzzle case to import an unwritten lower-level fact that would ground the deviant higher-level claim? Second, why is the deviant claim in a puzzle case interpreted to be inadequately grounded? In other words, why, in a puzzle case, does the relevant interlevel relation remain the same as its ordinary counterpart in reality? Let us call the first question the *exhaustivity question* and the second the *adequacy question*. Note that the adequacy question is essentially the same

question as the one I raised regarding Weatherson—how to explain possible exceptions to the default assumption that interlevel relations remain the same across fiction and reality. Both questions call for a principled explanation of how interlevel relations in fiction are to be interpreted.

The exhaustivity and adequacy questions are left unanswered by the interlevel relation solution in its current form. Since these questions need to be answered to fully solve the fictionality puzzle, the interlevel relation solution is, in its current form, incomplete. In the following sections, I answer these two questions, thereby providing a complete solution to the fictionality puzzle. To do so, I first propose in Section 3 a general account of fictional truth, temporarily setting aside the fictionality puzzle. Since the exhaustivity and adequacy questions essentially concern how fictional truth is generated, we need to discuss the topic of fictional truth in general to answer them. In Section 4, I return to the fictionality puzzle and answer the exhaustivity and adequacy questions using the proposed account of fictional truth.

3. Fictional Truth and Inference to the Best Explanation

In this section, I propose my own account of fictional truth. Before doing so, let me briefly discuss why we need a new account of fictional truth to solve the fictionality puzzle. To my knowledge, virtually no extant accounts of fictional truth have the resources to fully solve the fictionality puzzle.[66] Some extant accounts lack the resources in that they do not offer a mechanism for generating fictional truth that is

[66] Miyazono (2017) makes a similar claim, though not on exactly the same grounds.

specific enough to answer the exhaustivity and adequacy questions.[67] Two examples that belong to this category are Walton's account, according to which a proposition is fictional if and only if it is to be imagined (1990, §1.5), and Gregory Currie's, according to which a proposition is fictional if and only if "it is reasonable for the informed reader to infer that the fictional author of [the fiction] believes [the proposition]" (1990, p. 80).[68] These accounts can tell us that propositions like "Craig did the right thing" in puzzle cases are not to be imagined, or that it is not reasonable for the informed reader to infer that the fictional author believes such propositions. But these accounts, on their own, do not explain exactly why this is the case, which is what we need to know to answer the exhaustivity and adequacy questions.

On the other hand, some other extant accounts of fictional truth offer more specific mechanisms for generating fictional truth, but these mechanisms fail to accurately capture what is going on in puzzle cases. David Lewis's account belongs to this category. He offers two analyses of fictional truth without adjudicating between the two, one of which is as follows:

> A sentence of the form "In the fiction f, ϕ" is non-vacuously true iff some world where f is told as known fact and ϕ is true differs less from our actual world, on balance, than does any world where f is told as known fact and ϕ

[67] This does not necessarily mean that these theories are defective for this reason. Those theories do not aim to offer such a specific generation mechanism.
[68] Walton later admits that to be imagined is only a necessary, not sufficient, condition for a proposition to be fictional (2015).

is not true. It is vacuously true iff there are no possible worlds where f is told as known fact. (Lewis, 1983, p. 270)[69]

The problem with this theory in relation to the fictionality puzzle is that it does not allow what is explicitly written in a fiction to be false in the fiction (note that this problem persists whether or not there is a possible world where the fiction is told as known fact).[70] But that is exactly what happens in puzzle cases like *Death*.

Lastly, some extant accounts of fictional truth imply that the fictionality puzzle is not a real puzzle, so they are not helpful in *solving* the puzzle.[71] Kathleen Stock's Extreme Intentionalism belongs to this category (2017). She argues, very roughly speaking, that all that is needed to make a proposition fictional is the author's having a certain kind of intention. So, if the author of *Death* had the right intention, it is fictional that Craig did the right thing. This means that the fictionality puzzle does not arise, at least in the way it is commonly characterized in the literature.

To solve the fictionality puzzle, an account of fictional truth must offer a specific mechanism for generating fictional truth, and the mechanism must provide a principled explanation of the ways whereby what is explicitly written in a fiction determines (or fails to determine, as in puzzle cases) what is fictional. My account offers such a mechanism.

[69] I will not discuss the other analysis since the difference between the two is not important for the present purpose—the other analysis also has the problem I point out.
[70] Miyazono (2017) makes the same point.
[71] I am assuming in this essay that the fictionality puzzle is a real puzzle.

My proposal is that the generation mechanism of fictional truth is based on inference to the best explanation. What is explicitly written or shown in a fiction is *evidence* or *data* to be explained, and the set of propositions that together best explain or make sense of the data as a whole is the set of propositions that are true in the fiction.[72] More specifically, my account goes as follows: a proposition is true in a fiction if and only if it is part of the best explanation(s) of the entire text, by which I mean the sentences explicitly written (in the case of literary fiction) or the images and sounds explicitly presented (in the case of films and paintings) in the entire fiction.[73]

As for what makes an explanation good, I follow the criteria widely accepted in the literature: a good explanation not only fits with the data but also has explanatory virtues such as coherence with background assumptions, unification, and simplicity, among others. An explanation coheres better with background assumptions than another explanation if it is more plausible in relation to those background assumptions, is more unifying if it explains more

[72] The basic idea that what fits or explains the text is fictional is not new. See, e.g., Alward (2010) and Dworkin (1985). What is new about my proposal is that it develops this idea in detail in terms of inference to the best explanation, with the purpose of solving the fictionality puzzle.

[73] A few clarifications are in order: (1) a proposition is false in a fiction if and only if its negation is part of the best explanation(s) of the text. Thus, my account allows the possibility of a proposition being neither true nor false in a fiction—this happens just when neither the proposition nor its negation is part of the best explanation(s); (2) There may be more than one best explanation, in which case the fiction is undetermined regarding the matters about which the best explanations disagree. Also, it is possible that no explanation is good enough to count as the best explanation. In such a case, there is no fact of the matter about what is true in the fiction; (3) Note that the best explanation is of the *entire* text as a whole. Thus, fictional truth is determined holistically, not just based on some sentences or passages in the fiction; (4) the kind of explanation I have in mind concerns what is the case in the fiction, not why the author wrote this and that in the text or, to put it another way, what was the author's intention. I believe that fictional truth cannot be fully analyzed in terms of authorial intention, whether actual or hypothetical, although I cannot further discuss this issue in this essay. I do not deny that *literary meaning* can be analyzed in terms of (actual or hypothetical) authorial intention.

diverse data without making an ad hoc assumption, and is simpler if it requires fewer resources to explain the data.[74] The best explanation, then, is the explanation that exhibits these explanatory virtues to the highest degree among all competing explanations of the data.

Inference to the best explanation (hereinafter, IBE) is the form of reasoning that infers that something is true on the grounds that it best explains the given data.[75] We use this form of inference all the time, not only in the context of scientific inquiry but also in commonsense reasoning. Suppose, for instance, that you leave your office for a few hours, and upon returning you discover that the door lock is broken and your laptop is missing. You would judge that your laptop was stolen. This judgement is based on IBE—you make the judgment because it best explains the given data. Other explanations (e.g., the laptop disappeared by itself, you actually left your laptop at home, etc.) are less explanatorily virtuous. They do not cohere well with background assumptions (e.g., things do not disappear by themselves, your memory is reliable, etc.) and are less unifying in that they need some additional assumption to explain why the door lock is broken. Of course, you may not consciously consider all these potential explanations. But this might be because we are so used to employing IBE that we do not need to do so with conscious effort. At the very least, IBE would be the kind of reasoning

[74] There is no complete agreement on the criteria for a good explanation as well as what terms to use to refer to them. It is also controversial exactly how to spell out these explanatory virtues. I am here following the terms used by Mackonis (2013). For more on the criteria for a good explanation, see, e.g., Lipton (2004); McMullin (1996); Quine & Ullian (1978); Thagard (1978). Also note that unification and simplicity are closely related to one another. Depending on how to spell them out, they may mean the same thing (Mackonis, 2013). I will ignore these complexities regarding explanatory virtues, since they are unimportant for the purposes of this essay.

[75] For more on IBE, see, e.g., Lipton (2004).

you would invoke if asked to justify your judgment. My suggestion is that fictional truth is generated through the same form of inference.

Let me illustrate the suggestion with an example. Consider the scene in the film *Birdman* (2014) in which, as far as what is explicitly shown goes, the main character Riggan jumps off of a building and flies around New York City. Intuitively, it is not fictional that he is flying. But why? This is because, I suggest, Riggan's flying, which would mean he has a super power, is not part of the best explanation of the film's text. Although it is *an* explanation, there is a better one: Riggan is delusional, and the scene simply exhibits the content of his delusion.

Why is this delusion explanation better than the super power explanation? Although both manage to explain the film's text, the delusion explanation is more explanatorily virtuous. Consider what is shown after the flight scene.[76] After Riggan lands on the ground, a taxi driver accuses him of not paying his taxi fare. The delusion explanation can explain this datum nicely—Riggan did not fly but rather took a taxi and, lost in his delusion, did not pay the fare. The super power explanation, by contrast, requires an ad hoc assumption to account for this datum—Riggan flew, but it just so happens that a random taxi driver wrongly accuses him of not paying the taxi fare. This is compatible with the datum, but it is clearly a worse explanation. It is less unifying, and it coheres less with background assumptions about film (a film does not usually show us meaningless coincidences). Therefore, the super power explanation is not the best explanation,

[76] Although there are several other scenes that suggest that Riggan is delusional, I cannot discuss all of them here.

and that, I suggest, is the reason why it is not fictional that Riggan flew, even though it is explicitly shown.

According to my account, what is explicitly written or shown is fictional only when it is part of the best explanation of the entire text. Of course, what is explicitly written or shown is usually part of the best explanation, but not always; the explanation which regards what is explicitly written or shown as true may not be the most explanatorily virtuous one. Sometimes, as in *Birdman*, an explanation that regards what is explicitly written or shown as false is more unifying, is simpler, and/or coheres better with background assumptions.[77] (This point is important for solving the fictionality puzzle, as I will illustrate in the next section.)

What are the reasons to accept my account of fictional truth? A full-fledged defense of it against other competing accounts would require a separate essay.[78] My aim here is merely to motivate my account, showing that it at least merits a serious consideration so that the account can be used to solve the fictionality puzzle in the next section.

---

[77] Similar cases exist in other applications of IBE. Suppose, for instance, historians newly discover a journal written by Napoleon during the battle of Waterloo, and they are trying to infer new historical facts based on the text, using IBE. They would probably infer that many things written in the journal are true, but it is possible that they judge some of the things explicitly written are false, perhaps because those things do not cohere well with what they already know about the battle. Thus, the best explanation of what is written in the journal may not contain the truth of everything written in it. Even in science, scientists sometimes ignore part of the data when the best explanation does not accommodate all the data.

[78] It is worth nothing that my account is compatible with several extant accounts of fictional truth, including Walton's and Currie's. In fact, my account nicely complements theirs. My account can explain exactly when a proposition is to be imagined, or it is reasonable for the informed reader to infer that the fictional author believes a proposition—it is when the proposition is part of the best explanation(s) of the text.

One important motivation for my account comes from the fact that we indeed use IBE when figuring out what is fictional.[79] I believe that for most viewers of *Birdman*, figuring out what is actually going on in the flight scene would involve an inference process just like the one described above, although the inference may not be done explicitly. At the very least, viewers would invoke IBE if asked to justify their judgment that Riggan is delusional.

That we use IBE to infer fictional truth becomes more apparent if we think about the similarities between inferring fictional truth and real-life uses of IBE. Consider, for instance, the kind of inference a detective would make when investigating a crime. She would first gather all the available evidence, consider different potential hypotheses about who committed the crime, and then choose among those hypotheses by looking to see which one best explains the evidence. Compare this case to how we interpret films like *The Usual Suspects* (1995). This film begins with a crime scene without revealing the identity of the criminal, inviting the viewers to guess who the criminal is. While watching this film, the viewers would consider potential hypotheses and choose one based on which best explains what is explicitly shown in the film. Both cases involve the same form of inference IBE.[80]

[79] Of course, this does not necessarily mean that fictional truth is *determined* by IBE, which is a metaphysical, not epistemic, claim about fictional truth. I discuss this point below.
[80] Of course, IBE in these two cases would involve different background assumptions. The real detective would use background assumptions regarding real crimes (e.g., the reliability of finger print evidence), while the viewers of the film would use background assumptions regarding the fiction genre (e.g., there is usually a twist). But this does not mean that different kinds of inference are used in these cases.

In fact, it is unclear whether there are any other forms of inference that we can plausibly use when we infer what is true in a fiction based on the text. To begin with, such an inference is certainly not deductive. What is explicitly shown in *Birdman* does not *entail* that Riggan is delusional. What about using a conditional in a broadly Lewisian way—that is, may we infer that a proposition is fictional just when, if the text were told as known fact, the proposition would be true? The problem with this idea, as I mentioned earlier, is that it cannot account for cases like *Birdman* wherein what is explicitly written or shown in some part of the text is false in the fiction.[81] There might be other possible forms of inference that can be used to interpret a fictional text, but I cannot consider all of them here. Regardless of what other possibilities remain, it should be clear by now, I hope, that IBE is essential to inferring fictional truth from a fictional text.

Of course, that we use IBE when interpreting fictional truth does not necessarily mean that fictional truth is *determined* by IBE. The former is an *epistemic* claim about how we figure out fictional truth, while the latter is a *metaphysical* claim about fictional truth itself. My account of fictional truth makes the latter, metaphysical claim. However, these epistemic and metaphysical claims about fictional truth are more closely connected than they might at first appear. This is due to the interpretive nature of fictional truth. Many philosophers agree that fictional truth does not transcend, or exist independently of, how we

[81] A Lewisian may reply that the text being told as known fact does not mean that everything that is explicitly written in the text is told as known fact. But then the question remains as to how we figure out what is told as known fact (figuring that out would probably involve IBE).

understand and interpret fiction.[82] In other words, fictional truth is not discovered in the way scientific truth is discovered (assuming some objective view of scientific truth). [83] As Stear nicely put it, fictional truth is "determined interpretatively, not investigatively" (2015, p. 14). This interpretive view of fictional truth provides a motivation for my account, assuming that the epistemic claim I put forward above is true. If fictional truth is interpretive by nature, its epistemic dimension (i.e., how we interpret fictional truth) should be closely linked to its metaphysical dimension (i.e., how fictional truth is determined) in some way. My account naturally provides such a link. If, as I argued above, readers use IBE to interpret fictional truth, and authors create fiction with the knowledge that their readers interpret fictional truth this way, we do not need to look further to find such a link. Fictional truth is determined through IBE. In other words, in the case of fictional truth, IBE is not merely truth-conducive but truth-determining.

I have so far motivated my account of fictional truth. In the next section, I will illustrate how this account helps us to solve the fictionality puzzle.

4. Solving the fictionality puzzle

In Section 2, I argued that to completely solve the fictionality puzzle, we need to answer what I called the exhaustivity and adequacy questions. In this section, I

[82] See, e.g., Alward (2010); Lamarque (1990); Stear (2015); Todd (2009).
[83] I do not mean that fictional truth is subjective in the sense that there is no fact of the matter about what is fictional. What I mean is just that fictional truth does not exist independently of how we interpret fiction. The right interpretation(s) of a fiction—which, I argue, is determined by the best explanation(s) of the text—fixes what is fictional. Thus, we can be wrong about what is fictional because what we think is fictional or the best explanation of the text may not in fact be the best explanation.

offer my own answer to these questions, based on the account of fictional truth proposed above. I will first discuss why deviant claims in puzzle cases are interpreted to be exhaustively and inadequately grounded. I will then explain why deviant claims in non-puzzle cases are interpreted differently.

4.1. Interpreting puzzle cases

Consider *Death* as an example puzzle case. I have argued that what is true in a fiction is determined by what best explains the text. What, then, best explains the text of *Death*, especially the sentence that Craig did the right thing? There are several potential explanations: (1) Craig did the right thing because there is some unwritten lower-level fact that morally justifies his action (e.g., Jack and Jill were about to kill everyone in the scene); (2) Craig did the right thing because killing people for holding up traffic is morally right; (3) Craig did not do the right thing. Note that explanation (1) preserves the ordinary interlevel relation by importing an unwritten lower-level fact. Explanation (2), on the other hand, violates the ordinary interlevel relation. I will show below that explanation (3) is the best explanation. This accounts for why it is not fictional that Craig did the right thing. More specifically, I will elaborate on why explanation (1) is not the best explanation, which will answer the exhaustivity question, and why explanation (2) is not the best explanation, which will answer the adequacy question.

4.1.1. Answering the Exhaustivity Question

Explanation (1) is not as good as explanation (3) because explanation (1) involves an ad hoc assumption (and it is also not more explanatorily virtuous than

explanation (3) in other aspects). Whether the imported lower-level fact is non-specific (i.e., there was *some* unwritten morally justified reason for Craig to kill Jack and Jill) or specific (e.g., Jack and Jill were about to kill everyone in the scene), it is imported into explanation (1) only for the sake of accommodating the sentence that Craig did the right thing. Given the data (i.e., what is explicitly written in the entire text), there is no independent reason to think that such a lower-level fact holds in the fiction.[84] In fact, there is data that goes against importing such a lower-level fact. Recall the last sentence of *Death*: "So Craig did the right thing, because Jack and Jill should have taken their argument somewhere else where they wouldn't get in anyone's way." This sentence strongly suggests that the reason Craig's action is right is that Jack and Jill held up traffic. This textual evidence conflicts with importing an unwritten lower-level fact that morally justifies Craig's action.[85]

Contrast this case with *Denise*, in which an unwritten lower-level fact, i.e., that Denise died before the burial, is imported to ground the higher-level claim that burying Denise was the proper thing to do. In that case, the imported assumption is not ad hoc because we have independent reasons in support of it. What reasons? First, according to the text, Denise had "a strong headache," indicating she was ill. Furthermore, Denise was buried "in the grounds of the old

[84] Importing a generic proposition that is true in the actual world (e.g., gravity exists) may not be ad hoc even when there is no direct textual evidence if it is part of background assumptions (I say more about background assumptions in Section 4.1.2.). However, the kind of lower-level proposition that, if imported, would morally justify Craig's action is not a generic proposition that is true in the actual world.

[85] The statement "Craig did the right thing" on its own cannot provide a reason for importing such an unwritten lower-level fact because the statement equally supports the interpretation that the relevant interlevel relation in the fiction differs from that in reality.

church," which suggests that the burial was her funeral. If Denise's friends and family had buried her alive, they would probably had buried her somewhere else. Given these data points, it is not ad hoc to assume that Denise died before the burial. In fact, this assumption unifies the data, such as her strong headache, her being buried in the ground of the church, and the burial being the proper thing to do. Thus, the assumption is part of the best explanation of the text, given that no other potential explanation is equally explanatorily virtuous. This makes it fictional that Denise died before the burial.

Let me now summarize my answer to the exhaustivity question. A claim is interpreted to be non-exhaustively grounded (i.e., some unwritten lower-level fact grounds it) just when the best explanation of the text includes such a lower-level fact (and thus the lower-level fact is fictional). Such is the case in *Denise*. There, the best explanation of the text, given the textual evidence, includes the unwritten lower-level fact that Denise died before the burial. This makes the claim that burying Denise was the proper thing to do non-exhaustively grounded. In the case of *Death*, an explanation that includes such an unwritten lower-level fact is ad hoc and thus fails to be the best explanation, given the lack of textual evidence in support of such a fact. Thus, the claim that Craig did the right thing is exhaustively grounded. This point is generalizable to other puzzle cases: the reason the deviant claim in a puzzle case is interpreted to be exhaustively grounded is that the best interpretation of the text does not include an unwritten lower-level fact that grounds the claim due to the lack of textual evidence in support of it.

4.1.2. Answering the Adequacy Question

Let me now turn to explanation (2), which violates the ordinary interlevel relation. Why is this explanation less good than explanation (3), according to which Craig did not do the right thing? The problem with explanation (2), I argue, is that it does not cohere well with background assumptions. As mentioned earlier, coherence with background assumptions is one of the widely accepted explanatory virtues.[86] Thus, what counts as the best explanation is partly a function of background assumptions, which range from already established theories or facts to assumptions about explanatory standards themselves (i.e., assumptions about what counts as a good explanation).[87] To illustrate how background assumptions affect IBE, recall the missing laptop example. The explanation that the laptop is stolen is better than other explanations (e.g., the laptop disappeared by itself, you actually left the laptop at home, etc.) only on certain background assumptions (e.g., things do not just disappear by themselves and your memory is reliable). If the background assumptions were different (e.g., you suffer from dementia), then the best explanation could be different.

Similarly, what counts as the best explanation of a fictional text is partly a function of background assumptions. Consider, again, the flight scene in *Birdman*. The delusion explanation is better than the super power one only given certain background assumptions. I earlier mentioned the assumption that a film does not usually show us meaningless coincidences, but that is just one of the many

[86] Some philosophers believe that other explanatory virtues like unification and simplicity are also closely connected to background assumptions. See, e.g., Day & Kincaid (1994); Sober (1988).
[87] See, e.g.,, Day & Kincaid (1994); Lipton (2004).

background assumptions involved in this case. Other important background assumptions include assumptions specific to the film's genre. *Birdman* is a realistic film, and this inheres certain background assumptions that hold in that genre (e.g., basic facts about physics, human psychology, and social practices are not different from those in the actual world—hence, gravity exists, people pay for a taxi ride, etc.). If *Birdman* belonged to a different genre, such as science fiction or fantasy, different background assumptions would become relevant, and, accordingly, some other explanation could turn out better than the delusion explanation.

The case of *Death* can be explained in a similar way. Explanation (2), which violates the ordinary interlevel relation, is not as good as explanation (3), according to which Craig did not do the right thing, because explanation (2) does not cohere with background assumptions as well as explanation (3) does.[88] Assuming *Death* belongs to an ordinary, realistic fictional genre (I believe this is implicitly assumed in the literature on the fictionality puzzle[89]), background assumptions specific to that genre become relevant to what counts as the best explanation of the text. It is plausible that one such background assumption is that interlevel relations, including the relations between moral and non-moral facts, in the fiction are the same as those in reality. Realistic fiction may diverge from

[88] This idea is in line with Liao's (2016) solution to the fictionality puzzle, according to which the deviant proposition in a puzzle case fails to be fictional because it is genre convention-discordant. My solution develops his idea further by explaining *why* a proposition's genre convention-discordance affects its fictionality—it is because genre conventions are part of background assumptions that affect what counts as the best explanation of the fictional text, which determines what is fictional.

[89] But, to be more precise, it is difficult to say which genre(s) *Death* belongs to, since it is a toy example made up by a philosopher.

reality with respect to particular events or people, but one of the things that characterize realistic fiction is that interlevel relations, which constitute the backbone of reality, are preserved within the fiction. In fact, there are few fictions, whether realistic or not, in which the relations between moral and non-moral facts differ from those in reality.[90] Explanation (2) thus conflicts with background assumptions.

Explanation (3), on the other hand, coheres well with background assumptions. It is not a background assumption about (realistic) fiction that what is explicitly asserted is always fictional. There are many extant (realistic) fictions in which what is explicitly asserted is not fictional—for instance, when the narrator is being ironic or a textual assertion describes what is commonly believed in the fictional world rather than what is true in it.[91] Thus, explanation (3) coheres better with background assumptions than does explanation (2). Since these two explanations are equally good according to other criteria—neither of them is particularly simpler or more unifying than the other— explanation (3)'s better coherence with background assumptions makes it the better explanation.

One might wonder why the sentence "Craig did the right thing" does not override the background assumption that the relations between moral and non-

[90] There are, of course, many cases in which a fiction invites us to take an immoral attitude (e.g., to admire an evil character) or the people in a fiction hold wrong moral beliefs. There are also cases in which the world of a fiction looks morally exotic because of exotic *non-moral* facts (e.g., in some cartoons, violence is not morally bad because violence does not cause suffering). But these are not necessarily cases in which the relations between moral and non-moral facts differ from those in reality.

[91] It seems undetermined, in the case of *Death*, which of these possibilities is the case. *Death* is too short to provide enough data to adjudicate on this matter.

moral facts in fiction do not differ from those in reality. After all, the sentence is part of the data, and background assumptions are just one of the many factors that determine what is the best explanation. When we look at the history of science, background assumptions have often been overridden by innovative explanations. For instance, the assumption that causes cannot work at a distance was overridden by Newtonian physics, and quantum physics is altering the deterministic world view that scientists had previously assumed. One might wonder, then, why the sentence "Craig did the right thing" cannot override the assumption that the relations between moral and non-moral facts do not differ across fiction and reality.

The reason why those innovative explanations can override background assumptions is that they are more explanatorily virtuous than the competing explanations that conform to existing background assumptions: they are simpler, more unifying, and/or fit better with available evidence. Explanation (2) of *Death*, however, is not more explanatorily virtuous than its competing explanations. There is no textual evidence that is consistent only with explanation (2), and explanation (2) is not more unifying or simpler than explanation (3). Therefore, explanation (2) cannot override the background assumption that conflicts with it.

If the text of *Death* were different, of course, explanation (2) might be the best explanation despite its conflict with background assumptions. Imagine, for instance, a modified version of *Death* (call it *Death**) with the following sentence added to the beginning: "This story is about a world in which what you believe is morally right is in fact morally wrong and what you believe is morally wrong is in

fact morally right."[92] Now, explanation (2) is better than explanation (3). Explanation (3), according to which Craig did not do the right thing, becomes less good than before because the proposition that Craig did not do the right thing conflicts with the newly added sentence. So, in order for explanation (3) to make sense, either (a) there should be some extra explanation for the conflict or (b) the added sentence must be false. Either way, explanation (3) becomes less good because, in the case of (a), it would be less simple or unifying, and in the case of (b), it would not cohere well with background assumptions about fiction—a sentence like the newly added one, which is a meta-statement about the story and appears at the beginning of the fiction, is almost always true. On the other hand, explanation (2) becomes much better than before. It accommodates the modified text in a simple and unifying way. Explanation (2) still conflicts with the background assumption that the relations between moral and non-moral facts in a fiction do not differ from those in reality, but it now has other strong explanatory virtues that make it better overall than explanation (3). Thus, explanation (2) is the best explanation of *Death**, and, accordingly, it is fictional in *Death** that Craig did the right thing because killing people for holding up traffic is morally right. The case of *Death** shows that an explanation of a fictional text that violates ordinary interlevel relations *can* be the best explanation if it has strong

[92] Stear (2015) also observes that an added sentence can cancel ordinary interlevel relations (p.14). He, however, does not explain exactly why such a cancelation happens. That is what I am trying to explain here.

explanatory virtues that outweigh its conflict with background assumptions.[93] *Death*, however, is not such a case.

Let me now summarize my answer to the adequacy question. The reason the deviant higher-level claim in a puzzle case is interpreted to be inadequately grounded is that, in puzzle cases, an explanation of the text which violates the relevant ordinary interlevel relation is not the best explanation because it conflicts with background assumptions (and also because its other explanatory virtues are not strong enough to compensate the conflict). This point applies to non-moral puzzle cases as well. Whether the higher-level claim concerns morality, humor, or the property of being a television, an explanation of the text that violates the relevant ordinary interlevel relations does not cohere well with background assumptions because such interlevel relations in (realistic) fiction rarely differ from those in reality.

I have so far argued that both explanation (1) and explanation (2) with respect to *Death* are defective compared to explanation (3). This makes explanation (3) the best explanation, given that there are no other potential explanations. This is the reason it is not true in *Death* that Craig did the right thing. This point can be generalized to other puzzle cases: puzzle cases arise when an explanation according to which the explicitly asserted deviant claim is false is the best explanation of the text.[94]

---

[93] This answers the question I raised in regard to Weatherson—when and how interlevel relations can differ across fiction and reality.

[94] It is in principle possible that, in some puzzle cases, the reason such an explanation is the best explanation has nothing to do with interlevel relations. For instance, there might a puzzle case in

4.2. Interpreting non-puzzle cases

What accounts for the interpretive difference, then, between puzzle cases, in which deviant claims are interpreted to be exhaustively and inadequately grounded and thus fail to be fictional, and non-puzzle cases, such as fantasy and science fiction, in which deviant claims are fictional? That the deviant claim in a non-puzzle case is fictional means, according to my account of fictional truth, that an explanation that makes the claim true is the best explanation of the text. But which explanation? Is it the explanation that imports an unwritten lower-level fact to ground the claim while preserving the ordinary interlevel relation (i.e., an explanation that corresponds to explanation (1) in the case of *Death*)? Or the explanation according to which the relevant interlevel relation in the fiction differs from the ordinary one in reality (i.e., an explanation that corresponds to explanation (2) in the case of *Death*)? Intuitively, for many non-puzzle cases, neither interpretation seems to be correct, although there are some non-puzzle cases in which one of the two interpretations is correct (e.g., in the case of *Death**, the second interpretation is correct). In other words, deviant claims in many non-puzzle cases seem to be fictional without thereby making any grounding details fictional.

Take, for instance, the film *Last Action Hero* (1993), in which a young boy Danny is transported into a movie, meets his favorite action hero Jack Slater

---

which the deviant claim concerns a primitive fact. If there is indeed such a case, then my solution to the fictionality puzzle can handle it, while the interlevel relation solution cannot. (My solution is compatible with and complements the interlevel relation solution, but it is not a version of interlevel relation solution itself).

there, and then returns to the real world with Slater, where they chase some bad guys who also escaped from the movie into the real world.[95] Several deviant propositions are true in this fiction: Danny as well as fictional characters like Slater can travel between a fictional world and the actual world, an actual person and a fictional character can interact with each other, etc. These propositions are certainly fictional—denying this would be interpreting the fiction incorrectly. How, then, do they manage to be fictional, considering that the propositions are not about primitive facts? Are they fictional because there are some unwritten lower-level facts that ground them? No: there seems to be no lower-level fact whatsoever that, if imported, would ground those deviant claims while preserving the relevant ordinary interlevel relations (as well as the ordinary meaning of "fictional" and "actual").[96] We cannot even import a *non-specific* lower-level fact (i.e., *some* lower-level fact that could ground those deviant propositions) since there is simply no such fact. (So, this case is different from the case of *Death*, in which there *are* some lower-level facts that, if imported, would ground the claim that Craig did the right thing, although such importation is prohibited.) Does this mean, then, that the relevant interlevel relations in *Last Action Hero* are different from their ordinary counterparts in reality? If so, exactly how are they supposed to be different in order to make the deviant claims fictional? The intuitive response to this question, I believe, is that there is no fact of the matter as to what the interlevel relations fictionally are in *Last Action Hero*. The film is simply

[95] This example is discussed in Proudfoot (2006), though not in relation to the fictionality puzzle.
[96] In the film, Danny uses a "magic ticket" when traveling between fiction and reality. But this is not a lower-level fact that can ground the deviant claim that he can travel between fiction and reality. Rather, it is another deviant higher-level claim that itself requires lower-level grounds.

undetermined on this issue, and there is no fictional truth about it. I do not believe I am the only one who has this intuition. Weatherson, for instance, seems to make a similar point when he states, "sometimes we simply ignore, either in fiction or imagination, what goes on at some levels of detail," for example when a cow jumps over the moon in fiction (2004, p. 17).

Assuming that this intuition is correct, we have another question we must answer to fully solve the fictionality puzzle: how, in many non-puzzle cases but in no puzzle cases, do deviant higher-level claims manage to be fictional without thereby making further details about relevant lower-level facts or interlevel relations fictional? My answer, simply put, is that, in such non-puzzle cases, the best explanation of the text does not need to concern such details. This is because what counts as a good explanation is partly a function of what we are interested in or care about, or, to put it another way, what we seek to understand.[97] In many non-puzzle cases, the details about how a deviant claim is true (e.g., how Danny and Slater can travel between reality and fiction) are not what we are interested in. Rather, our interest lies in what will happen in the fiction *if* the deviant claim is fictional. Therefore, the claim is simply assumed to be true through idealization. Let me elaborate.

[97] See, e.g., Achinstein (1983); Potochnik (2017). The claim that whether an explanation is good is relative to interest may be controversial in science. In particular, one might worry that the claim blurs the distinction between explanations in the ontological sense (i.e., explanations as mind-independent facts in the world) and explanations as communicative acts. However, this is not a problem in the case of fictional truth as I characterize it, since there is no sense in which fictional truths as explanations are mind-independent facts in the world.

Scientists often set aside certain complicating factors so they can focus on the factors that are central to their research interests. For instance, physicists often assume that planes are frictionless, and economists often assume that human beings are perfectly rational agents. I suggest that a similar kind of idealization occurs in many non-puzzle cases, such as in science fiction and fantasy. In such genres, deviant claims are simply assumed to be true in the best explanation of the fictional text because the details about what grounds those claims are not of interest to us. Our interest lies instead in what the characters do and what happens to them and the significance of their actions, assuming that the deviant claims are true. The truth of such deviant claims serves as a backdrop against which other important events and actions take place.

For instance, in the case of *Last Action Hero*, what is of interest to us is what happens to Danny and Slater and the actions that they perform, not the question of how exactly they are able to travel between fiction and reality. Thus, an explanation of the film's text may simply suppose that it is possible to travel between fiction and reality, and no further explanation needs to be included in the best explanation of the text, just as a scientific explanation of an object's movement on a plane does not need to include any details about how the plane is frictionless. (It would be pointless to ask whether there is some special lower-level fact that grounds frictionless planes or whether the relevant interlevel relations are different from the ordinary ones.) A scientific explanation that assumes planes to be frictionless simply ignores the complications that arise from the assumption. Similarly, an explanation of *Last Action Hero*'s text can simply

ignore the complications that arise from the deviant claims. Ignoring such complications does not make the explanation less good, since they are not what is of interest to us. In fact, the best explanation of the text should ignore such details to better reflect what is of interest to us, which is what the characters do and what happens to them, assuming that the deviant claims are true.

It is worth clarifying what exactly I mean by interests, concerns, or what we seek to understand. I do not mean the interests or concerns of an individual. If that were the case, what counts as the best explanation of a fiction's text, and thus what is fictional, would be partly dependent on an individual's subjective interests and concerns, which is unacceptable. Thus, my suggestion is not undermined by the fact that some viewers may indeed be interested in the details about how Danny and Slater can travel between fiction and reality. This only shows that they do not understand how fictions like *Last Action Her*o work. What I mean by interests or concerns are the collective, shared interests or concerns of people engaging with or creating the particular genre of fiction in question. Such interests, again, do not lie in how Danny and Slater are able to travel between fiction and reality. In fact, one might even say that the lack of interest in such matters and the idealizations of deviant claims partly define science fiction and fantasy.

An important question, then, is why such an idealization does not occur in puzzle cases. For instance, why is it not simply assumed to be true in the best explanation of *Death*'s text that Craig did the right thing? It is not assumed because deviant claims in puzzle cases concern things that are of interest to us

with respect to the fiction in which they appear. The significance of human action, including moral significance, is very often precisely what we seek to understand through fiction (or at least realistic fiction, to which puzzle cases are supposed to belong). The primary themes of many fictions explore moral issues, and, even when a fiction does not explicitly address a moral issue, morality is still among the things that are of interest to us as long as the fiction deals in some way with human actions. Therefore, the claim that Craig did the right thing cannot simply be assumed to be true. Whether and how his action is right is exactly what we care about or seek to understand when engaging with *Death*.[98]

Non-moral puzzle cases can be explained in a similar way. Whether a nonjoke is extremely funny and whether something indistinguishable from a knife is a television are among the things that are of interest to us when we engage with fiction (though probably to a lesser extent than moral matters). Thus, they cannot simply be assumed to be true. Of course, I am here assuming that these examples belong to a genre where such matters are of interest to us. If this is not the case (that is, if we assume that those deviant claims are found in a fantasy like *Alice in Wonderland*), then our intuition changes—these examples no longer seem like puzzle cases. In a fantasy like *Alice in Wonderland*, it may well be fictional that a nonjoke is extremely funny or something indistinguishable from a knife is a television. This change in our intuition is consistent with my proposal that

[98] I do not deny that morally deviant propositions can, in principle, be idealized in some other cases in which moral matters are not of interest to us and our interest lies in exploring something else that requires the fictional truth of morally deviant propositions (e.g., what the social system would be like if moral facts were different).

different genres are related to different interests and concerns, which shape what aspects of the text need to be explained and what can be simply assumed to be true in the best explanation of the text.

4.3. Summing up

Let me now summarize my solution to the fictionality puzzle. I suggest that the asymmetry between puzzle and non-puzzle cases eventually results from the difference in our interest or concern regarding (the relevant genre of) fiction. In many non-puzzle cases, the details about how the deviant claim in question is fictionally true are not of interest to us. Rather, what is of interest to us in such cases is often what will happen in the fiction if the deviant claim is fictionally true. So, deviant claims in many non-puzzle cases can simply be assumed to be true in the best explanation of the text through idealization. Thus, those claims need not be adequately or non-exhaustively grounded to be fictional; questions about adequacy and exhaustivity simply do not arise.

By contrast, in puzzle cases, deviant claims concern matters that are of interest to us. Therefore, they cannot simply be assumed to be true in the best explanation of the text, and adequacy and exhaustivity begin to matter. Once adequacy and exhaustivity begin to matter, a deviant claim in a puzzle case fails to be fictional because it is interpreted to be inadequately and exhaustively grounded. This is because the best explanation of the text, in a puzzle case, is neither that there exists some unwritten lower-level fact that grounds the claim (hence the claim is exhaustively grounded) nor that the relevant interlevel relation differs from that in

reality (hence the claim is inadequately grounded). Rather, the best explanation is that the deviant claim is false, which makes the claim fail to be fictional.

5. Conclusion

I have argued that fictional truth is generated through IBE, and this way of understanding fictional truth provides a complete solution to the fictionality puzzle. According to the solution I have proposed, the key to solving the puzzle essentially lies in the contextuality of IBE: what counts as a good explanation depends on background assumptions and interests, which vary with context. In this respect, my solution to the fictionality puzzle is a development of Stear's solution that emphasizes contextuality. One challenge to such a contextual approach, which acknowledges the influences of multiple factors, is to find a principled way to explain such complex influences. I have shown in this essay that this challenge can be met by construing fictional truth in terms of IBE.

Chapter 4. Aptness of Fiction-Directed Emotions

1. Introduction

An emotion can be subject to different norms. It can be moral or immoral, prudentially good or bad, and epistemically appropriate or inappropriate.[99] An emotion can also be appropriate all things considered, in the sense that it is the emotion one ought to feel, considering all these different norms. This essay concerns epistemic appropriateness of emotions—I will use 'aptness' as a technical term for this concept.[100] More specifically, the question I shall pursue is how to understand the aptness of emotions directed towards fictional entities, such as characters and events in fiction (hereinafter, fiction-directed emotions). Intuitively, some fiction-directed emotions (e.g., pity for Anna Karenina) are apt while others (e.g., fear of Anna Karenina) are not. My goal in this essay is to elucidate what criteria govern the aptness of fiction-directed emotions and, in particular, whether and how they differ from the aptness criteria of emotions directed towards real entities (hereinafter, reality-directed emotions).

I will defend the so-called continuity thesis regarding this issue. This thesis claims that the criteria for aptness of fictional-directed emotions are *analogous*—in a sense to be specified later—to the criteria for aptness of reality-directed emotions. This thesis is contrasted with the discontinuity thesis, the denial of the continuity thesis.[101]

---

[99] On this distinction, see D'Arms & Jacobson (2000).
[100] In using this term, I am following Gilmore (2011).
[101] This issue as well as the terms were introduced by Gilmore (2011). To my knowledge, this issue has not received much attention yet.

This essay will proceed as follows. In section 2, I clarify what exactly the continuity thesis amounts to. I do so by first analysing the criteria for aptness of reality-directed emotions and then articulating in what sense those criteria are analogous to the criteria for aptness of fiction-directed emotions. In Sections 3 and 4, I defend the continuity thesis by considering and rejecting two arguments against it. In Section 3, I consider the argument that contrary emotions are apt towards the same object, depending on whether the object is real or fictional. In Section 4, I consider the argument that different kinds of considerations count as justificatory reasons for reality-directed and fiction-directed emotions. Section 5 concludes the essay.

2. Aptness of emotions

2.1. Aptness of reality-directed emotions

In this section, I analyse the notion of aptness in the domain of reality-directed emotions, to provide a basis for understanding the aptness of fiction-directed emotions. This analysis of the aptness of reality-directed emotions is meant to capture our general understanding of the notion, while remaining neutral between continuity and discontinuity theorists.

As I mentioned earlier, aptness is epistemic appropriateness, and there are roughly three dimensions in which a reality-directed emotion can be epistemically appropriate: fittingness, justification, and being salience-tracking.[102]

[102] There is another dimension that I will ignore for simplicity, regarding an emotion's duration and intensity being proportionate to its object.

(1) *Fittingness*: a reality-directed emotion is fitting if and only if its evaluative content fits the world, that is, when its object has the evaluative quality the emotion presents it as having.[103] Each emotion type is associated with a certain evaluative quality, often called the emotion type's formal object or criterial quality (e.g., danger in the case of fear, loss in the case of sadness). An emotion presents its object as having the criterial quality of the emotion type (e.g., fear presents its object as dangerous).[104] An emotion, then, is fitting if and only if its object has the emotion type's criterial quality. For example, fear is fitting if and only if its object is dangerous. In this respect, an emotion's fittingness is analogous to a belief's truth.

(2) *Justification*: a reality-directed emotion is justified if and only if it is in the right justificatory relation to the object's having the emotion's criterial quality. The kind of justification at issue here is *epistemic* justification or, to use Gilmore's term, justification for emotions 'in their representational dimension' (2011, p. 471). Emotions can be justified in other senses, for instance, prudentially or morally, but these kinds of justifications are not what is at issue in this essay. For example, although the fact that a joke is told by your boss gives you a reason to find it amusing, this is just a prudential reason. It is irrelevant to whether the joke is amusing or not, so it does not justify amusement in its representational dimension (Gilmore, 2011, p. 471).

---

[103] This definition is based on D'Arms & Jacobson (2000).

[104] To be more precise, when an emotion presents an object as having a certain evaluative quality, it is not the object itself but the object in a specific context that is seen as having the quality. For instance, when you fear a bear you encounter in the woods, what you fear is, strictly speaking, not the bear itself but the bear in that specific context. If the same bear is behind the bars of a zoo, the bear-in-the-zoo may not be dangerous, and fear may not be fitting.

I will not take a stand on what exactly it means for a reality-directed emotion to be in the right justificatory relation. The expression is used as a placeholder for whatever relation between the emotion and the object's having the criterial quality that counts as necessary and sufficient for the emotion to be justified on a particular theory of justification. Whichever particular theory of justification one endorses, some instances of reality-directed emotions are intuitively justified or unjustified. For example, if I fear a bear based on my perception of it roaring right in front of me, my fear is justified. If I fear a dog despite perceiving that it is small, old, and has no teeth left, my fear is unjustified. Any theory of justification should be able to accommodate such cases, and this intuitive understanding of justification is sufficient for the purposes of this essay.

As should be apparent, an emotion's justification is distinct from its fittingness, just as a belief's justification is distinct from its truth. An emotion can be justified and unfitting, as in the case of my fear of an insect that is in fact harmless but looks very similar to another kind of insect that I know to be dangerous—assuming that my fear is based on the insect's appearance—, or unjustified and fitting, as in the case of my fear of an insect that is in fact dangerous but looks very similar to another kind of insect that I know to be harmless.

(3) *Being salience-tracking*: a reality-directed emotion is salience-tracking if and only if it is a response to the salient object(s) or feature(s) thereof in the context the emotion is experienced. By a salient object or feature, I do not mean whatever object or feature the agent happens to find salient. Rather, it means what *ought to* be salient to the agent in the context. In other words, a salient object or feature is

what is important to the agent in the context, in some objective sense of importance that is partly a function of the agent's values, desires, and interests without being reducible to her own evaluation of what is important in the context.

Unlike fittingness or justification, this dimension of aptness has received little attention from philosophers. But being salience-tracking is an important epistemic virtue an emotion can have, independently of whether it is fitting or justified. For instance, if I am amused at a bear based on my perception of it roaring in front of me, my amusement is unfitting and unjustified, but my amusement is at least salience-tracking in that it is a response to what is important in the context, namely, the bear roaring in front of me. Conversely, an emotion that is fitting and justified may not be salience-tracking. Suppose, for instance, that my partner announces that he has been seeing someone else and wants to break up with me, and I, instead of feeling sadness or anger, feel amusement at his wearing unmatched socks. My amusement is fitting (unmatched socks are indeed amusing) and justified (my amusement is based on my perception of the socks' amusement-making features), but there is still a sense in which it is epistemically inappropriate. My amusement gets something wrong: it fails to track what is important in the context. This problem is not of a moral or prudential nature. There is nothing morally wrong with finding the unmatched socks amusing in the context, and it could be prudentially better for me to feel amusement in such a distressing situation. The problem is epistemic.

I shall use the term aptness to encompass all these three dimensions of epistemic appropriateness. That is, I stipulate that a reality-directed emotion is apt if and only if it is fitting, justified, and salience-tracking. Note that this analysis of

aptness is not meant to be a particular theory of aptness that competes with other theories. My purpose here is merely to provide a basis for our later discussion on the differences between the aptness criteria of reality- and fiction-directed emotions, and I believe that my rough analysis is compatible with any reasonable theory of the aptness of reality-directed emotions. Also note that both continuity and discontinuity theorists would have no problem accepting this analysis of the aptness of *reality-directed* emotions. The debate between them concerns the aptness of *fiction-directed* emotions.

2.2. Aptness of fiction-directed emotions

In this section, I articulate what exactly the continuity thesis amounts to, in light of the analysis of the aptness of reality-directed emotions I just provided. But, first, a clarification is in order. One might argue that all fiction-directed emotions are inapt because their objects do not exist—at least in the way ordinary real objects do—and, thus, the continuity thesis cannot even get off the ground. For instance, it might be claimed that pity for Anna Karenina is inapt because, while pity is apt for an object having the property of suffering undeservedly, Anna Karenina is not the kind of entity that can have that property. Since Anna Karenina is not an actual sentient being, she cannot suffer, not to mention suffer undeservedly.

In response, let me point out that there are two different senses in which a fictional entity has a property. In one of them, Anna Karenina has properties such as being a fictional character created by Tolstoy, and thus does not have the property of suffering undeservedly. But there is a different sense in which Anna

Karenina does have the property of suffering undeservedly. This is in the sense of what is *fictional* or *true in fiction*—it is fictional or true in the relevant fiction that she suffers undeservedly.[105] It is in light of these properties of hers that are within the scope of the fictionality operator that pity for her is apt.

This assumption—that the aptness of fiction-directed emotions should be assessed in terms of what is fictional rather than what is in fact true—is shared by both continuity and discontinuity theorists. Thus, the debate between them is not about the aptness of fiction-directed emotions *as a class*—discontinuity theorists grant that it is not the case that fiction-directed emotions *as a class* are inapt (Gilmore, 2011, p. 471). Rather, the debate is about the criteria for aptness that make some *particular* fiction-directed emotions apt and others inapt, and whether those criteria are analogous to the criteria for aptness that make some particular reality-directed emotions apt and others inapt.[106]

Let me now articulate how I understand the continuity thesis. To do so, recall the earlier analysis of the aptness of reality-directed emotions, according to which a reality-directed emotion is apt if and only if it is fitting, justified, and salience-tracking in the following sense:

> A reality-directed emotion is fitting iff its object has the emotion's criterial quality.
>
> A reality-directed emotion is justified iff it is in the right justificatory relation to the object's having the emotion's criterial quality.

[105] On fictional truth, see, e.g., Currie (1990b), Lewis (1978), and Walton (1990).
[106] Thus, this issue is distinct from the paradox of fiction.

> A reality-directed emotion is salience-tracking iff it is a response to the salient object(s) or feature(s) thereof in the context the emotion is experienced.

The continuity thesis claims that a fiction-directed emotion is apt if and only if it is fitting, justified, and salience-tracking in the following sense:

> A fiction-directed emotion is fitting iff its object *fictionally* has the emotion's criterial quality.

> A fiction-directed emotion is justified iff it is in the right justificatory relation to the object's *fictionally* having the emotion's criterial quality.

> A fiction-directed emotion is salience-tracking iff it is a response to the salient object(s) or feature(s) thereof in the context the emotion is experienced (i.e., in the context of *engaging with the relevant fiction*).[107]

Note that, although the aptness criteria of reality- and fiction-directed emotions are not *identical*, they are *analogous*—that is, they are *structurally* identical: in both cases, aptness is characterized in terms of fittingness, justification, and being salience-tracking, and each of these notions is understood in the same way across reality- and fiction-directed emotions, except that, in the case of fiction-directed emotions, fictional truth replaces truth, and salience in the context of engaging with the fiction replaces salience in the real context. Thus, the continuity thesis allows *some* differences in aptness criteria between reality- and fiction-directed emotions, in terms of the difference between truth and fictional truth, as well as the difference between salience in real and fictional contexts. For instance,

[107] What is salient in this context, which is a function of what is important to the *audience* while engaging with the fiction, may differ from what is salient *in* the fiction. What is relevant to the aptness of a fiction-directed emotion is the former.

fictional truth might have different kinds of grounds (e.g., genre conventions) from truth, and salience in the context of fiction might be determined by different factors (e.g., whether an object in a film is shown at the center or on the periphery of the screen) from salience in a real context. Such differences are compatible with the continuity thesis.

Thus, the disagreement between continuity and discontinuity theorists is not about whether there are any differences between the aptness criteria of reality- and fiction-directed emotions. Both parties agree that there are some differences. Rather, the disagreement is about exactly where the differences lie. Continuity theorists believe that the differences lie only in the differences between truth and fictional truth, and between salience in the real and fictional contexts. Discontinuity theories, on the other hand, claim that there are other differences that make the aptness criteria in the fictional and real domains structurally dissimilar. For instance, if there is an additional criterion for aptness of fiction-directed emotions besides fittingness, justification, and being salience-tracking, or if different kinds of relations count as the right justificatory relation depending on whether the emotion is reality- or fiction-directed, the continuity thesis would be refuted.

This is, of course, just one way to characterize the continuity thesis, and others might do it differently. This disagreement is not problematic, however, because the terms continuity and discontinuity are introduced only as a convenient way into the real subject, namely, exactly where to locate the differences between the aptness criteria of reality- and fiction-directed emotions. My real thesis, thus, is that the differences lie only in the differences between

truth and fictional truth as well as the differences between salience in real and fictional contexts. Whether this thesis should be called the continuity or discontinuity thesis is not of any substantial importance. In the remainder of the essay, I defend my thesis by considering and rejecting two arguments that the aptness criteria of fiction- and reality-directed emotions are disanalogous.

3. Contrary but apt emotions

One reason to think that the aptness criteria of fiction- and reality-directed emotions are disanalogous is that contrary emotions seem apt towards the same object, depending on whether the object is real or fictional. For instance, amusement towards Hynkel, a Hitler-like dictator in Charlie Chaplin's comedy film *The Great Dictator*, seems apt, although fear or indignation rather than amusement would be apt towards a real dictator like Hitler.[108] Examples like this are ubiquitous. Consider, for instance, dark comedies such as the Coen brothers' *Fargo* and Stanley Kubrick's *Dr. Strangelove*. In one scene in *Fargo*, a murderer kills another murderer and feeds the body into a wood chipper, and the final scene of *Dr. Strangelove* depicts a nuclear holocaust. Many viewers find these scenes amusing, and this emotional reaction seems apt, even though if those events were real, horror rather than amusement would be apt. Similar cases can be found in cartoons. *Tom and Jerry*, for instance, depicts serious violence as amusing. If such violence were real, horror would be apt. But amusement rather than horror seems apt towards the fictional violence.

[108] The example is from Gilmore (2011, p. 479).

One way to articulate what these examples suggest is as follows: the aptness of a fiction-directed emotion is a function of how the object is *represented*. Fictions are mere representations, not alternative worlds, so there is nothing beyond representations that determines the aptness of a fiction-directed emotion.[109] Thus, a fiction-directed emotion is apt as long as it accords with how the object is represented. Therefore, the discontinuity theorist might argue, the aptness criteria of fiction- and reality-directed emotions are not analogous.

I believe the discontinuity theorist is too quick in drawing this conclusion. Even if the aptness of a fiction-directed emotion is a function of how the object is represented, this does not necessarily conflict with the continuity thesis. There are different mechanisms through which representation determines aptness, and when we examine those mechanisms more carefully, it will become apparent that the above examples are in fact compatible with the continuity thesis.

In the case of *The Great Dictator*, representation affects aptness by influencing *salience*. Amusement rather than fear is apt towards Hynkel because, although Hynkel is both amusing and fearsome and we see both sides of him throughout the film, the amusing side is made salient by the way he is represented. The film highlights his amusing side by spending much more time depicting his antics than his atrocities. This salience of Hynkel's amusing side is what makes amusement an apt emotional response towards him. Fear and indignation towards him are fitting and justified but not salience-tracking. It is amusement that satisfies all three requirements of aptness.

[109] More on this point, see Currie (2014).

Note, however, that salience plays the same role in the case of reality-directed emotions, and thus salience-based examples like *The Great Dictator* fail to undermine the continuity thesis.[110] Even in the case of a real dictator, amusement can be apt if the dictator's amusing side is salient in the context (assuming that amusement is also fitting and justified), as when we are watching a comedian making a joke about Hitler's appearance. Although fear and indignation towards Hitler are fitting and justified, they are not salience-tracking in such a context. Conversely, amusement towards a fictional dictator may not be apt if their atrocities, rather than antics, are the focus of the fiction. Therefore, there is no real asymmetry here between fiction- and reality-directed emotions.

In the case of *Tom and Jerry*, representation affects aptness by influencing what is true in fiction rather than what is salient. Since Tom and Jerry are represented as recovering from injuries immediately, without any serious suffering, it is not true in *Tom and Jerry* that violence causes serious suffering.[111] Unlike *The Great Dictator*, in which Hynkel is indeed fearsome but his fearsomeness is simply not salient, Tom and Jerry do not suffer in the fiction. For this reason, horror is unfitting towards the fictional violence—because it is not horrifying in the fiction—, and accordingly inapt, although horror would be fitting towards real violence that causes serious suffering. Thus, *Tom and Jerry* does not undermine the continuity thesis. The amusement directed towards the fictional violence does not share an analogous object with horror directed towards real violence.[112] The objects of the two emotions have different properties: while real

[110] I owe this point to Gilmore (2011, pp. 477-480).
[111] I owe this point to Smuts (2013).
[112] Livingston and Mele (1997) make a similar point.

violence causes suffering, it is not true in the fiction that the same physical act causes suffering.

The most compelling cases for the discontinuity thesis are perhaps examples like the wood chipper scene in *Fargo* and the final scene in *Dr. Strangelove.* Let me first clarify how these cases differ from *The Great Dictator* and from *Tom and Jerry*. I will then explain why even these cases fail to support the discontinuity thesis.

Firstly, unlike *The Great Dictator,* in which the dictator's atrocities are not salient, violence is salient in the wood chipper scene in *Fargo*. Indeed, the very focus of the scene is the graphic violence of feeding a body into a wood chipper. The salience of violence is not as clear in the final scene of *Dr. Strangelove*, which only indirectly depicts the nuclear holocaust through images of mushroom clouds. But the clouds still signal numerous deaths and extreme suffering, though less graphically.

Fargo and *Dr. Strangelove* are also different from *Tom and Jerry*. While the asymmetry between fiction- and reality-directed emotions in the case of *Tom and Jerry* results from an exotic fictional truth—it is not fictional that Tom and Jerry die or suffer from violence—, people do die and suffer in *Fargo* and *Dr. Strangelove*.

One might argue that *Fargo* and *Dr. Strangelove* involve an exotic fictional truth regarding what counts as amusing: the events depicted in the wood chipper scene and the nuclear holocaust scene are amusing in the fictions, while their real counterparts are not amusing. This suggestion is mistaken. As often

noted in the literature on imaginative resistance, it is very difficult to make a fictional world exotic regarding an evaluative matter, such as what counts as amusing, as far as its non-evaluative basis remains the same across the fiction and reality (note that the exotic fictional truth in the case of *Tom and Jerry*—violence does not cause suffering or death—concerns a non-evaluative matter).[113] It might be possible to make such a fiction, but *Fargo* and *Dr. Strangelove* are not such a case, being realistic fictions. Thus, it is fictional in *Fargo* and *Dr. Strangelove* that those events are amusing if and only if their real counterparts are amusing.

How, then, should the continuity theorist explain the intuition that contrary emotions are apt towards those fictional events depicted in the wood chipper and nuclear holocaust scenes and their real counterparts? I suggest that amusement in those cases is directed not towards the fictional events depicted *in* the scenes but towards the *scenes* (i.e., the parts of the fictions or works). The amusement, then, is not a fiction-directed emotion because scenes or fictions are real, not fictional, entities.[114] A fiction-directed emotion apt in these cases is horror, not amusement; amusement is apt only towards the scenes. If so, *Fargo* and *Dr. Strangelove* are not counterexamples to the continuity thesis. Let me elaborate.

Filmmakers sometimes generate amusement through certain stylistic choices rather than by presenting a funny fictional event. Edgar Wright, for instance, often depicts a mundane event such as a person traveling from one city to anther through a series of quick cuts, thereby triggering laughter. In one scene

[113] See, e.g., Stear (2015).
[114] The term 'fictional-directed emotion' is ambiguous between an emotion directed towards the *content* of a fiction and an emotion directed towards the *fiction itself*. This is unfortunate, but I cannot think of a better term. I use the term to mean the former.

in Lars von Trier's *Nymphomaniac*, a female character compares three voices in Bach's music to her three lovers, and the filmmaker composed this scene by juxtaposing images of her sexual intercourses with her lovers with images of different parts of a pipe organ, with Bach's music in the background. The resulting combination is quite amusing. In these cases, our amusement is directed not towards what is going on *in* the scenes but towards the scenes. One consideration in support of this claim is that what we find amusing in these cases is what the filmmakers did in reality rather than what the characters do in the fictions. Edgar Wright chose to represent a mundane even in an unusual way, and Lars von Trier chose to combine incongruent images and sounds. Since these choices are what we find amusing, our amusement is directed towards the scenes, which manifest those choices, rather than their fictional contents. This is, of course, not to say that the fictional contents are irrelevant to the amusement. Rather, the amusement arises from a tension between the fictional contents and how they are represented. This, however, is yet another reason to think that the amusement is about the scenes, which include both the contents and the vehicles of representation.

Similarly, what we find amusing in the wood chipper scene in *Fargo* and the nuclear holocaust scene in *Dr. Strangelove* is not the fictional events but the scenes. In the case of *Fargo*, what triggers laughter is the Coen brothers' choice to use certain props—a wood chipper, as a means to dispose of a dead body, and a foot with a white sock on sticking out of it. Our amusement comes from recognizing this choice to use those unexpected and incongruous props to present a horrendous act. In other words, our amusement is about what the Coen brothers

did in reality rather than what the murderer does in the fiction. Similarly, my amusement at the final scene of *Dr. Strangelove* is not about what happens in the film but about what Stanley Kubrick did: he chose to end the film with an annihilation of the human race—which does not happen very often in fiction—and made a joke out of it—for instance, by adding cheerful background music. The fact that *Dr. Strangelove* is a political satire commenting on the real world also contributes to the amusement. This adds yet another reason to think that our amusement is not directed towards the fictional event depicted *in* the scene. A fictional event cannot comment on the real world; a scene can. Therefore, the amusement in these cases is not a fiction-directed emotion and thus does not conflict with the continuity thesis.

One might doubt my claim that our amusement in these cases is only about the scenes; perhaps it is also about the fictional events. In *Dr. Strangelove*, for instance, the nuclear holocaust results from the absurd behaviour of politicians and soldiers. And some people might find that behaviour, as well as it leading to an annihilation of humanity, amusing. Similarly, in the case of *Fargo*, some might find what is fictionally going on *in* the wood chipper scene amusing, such as the stupidity of using a wood chipper to get rid of a dead body—it actually makes more mess—or the incongruity between the horrendous act and a mundane white sock. Our amusement might partly be about these features, even when they are not considered as the filmmakers' choices as I suggested.

In response, let me first grant, for the sake of argument, that our amusement in these cases is partly, or even completely, about the fictional events, and that they are amusing. This, however, does not undermine the continuity

thesis because if the fictional events are amusing, their real counterparts would be amusing as well. Suppose that, as a result of some foolish actions of politicians, a nuclear war occurs in reality (this should not be too difficult to imagine given the recent political situation involving the US and North Korea). Amusement would be fitting and justified towards such an event, since it does have amusing features, although horror would also be fitting and justified. There is no reason to think that ambivalent emotions towards the same object cannot both be fitting and justified.[115] Suppose, for instance, you see a person slip on a banana peel and die. Both amusement and pity would be fitting and justified towards the death, although amusement might be *morally* inappropriate. Since the death has an amusing feature and we are in the right justificatory relation to it (which, in this case, involves perception), our amusement is fitting and justified. Similarly, both amusement and horror can be fitting and justified towards a real nuclear war that has amusing features (as well as towards its fictional counterpart in *Dr. Strangelove*).

The same point applies to the wood chipper scene. If using a wood chipper to dispose of a dead body is amusing in virtue of its stupidity, it should be amusing whether the act is fictional or real. So, while reading a newspaper article that a murderer killed another murderer and dumped the body into a wood chipper, you might well be amused as well as horrified. Both emotions are fitting and justified in this case, whether the object is real or fictional, since the object

[115] In fact, this possibility is considered as a feature that distinguishes emotion from judgment. See, e.g., Greenspan (1980).

has both amusing and horrifying features (assuming that you are in the right justificatory relation to them).

Of course, when it comes to *salience*, not just fittingness and justification, there is some asymmetry between the fictional and actual cases. When you encounter a nuclear holocaust or a murderer feeding a dead body into a wood chipper in reality, horror rather than amusement would be salience-tracking, assuming the events have a serious impact on what you care about, such as your own well-being (recall that salience is partly a function of the agent's desires and interests). Then horror rather than amusement would be, overall, apt towards those real events. By contrast, the fictional counterparts of those events do not impact what we care about in the same way. Furthermore, the amusing sides of the fictional events are made salient by the ways they are represented, for instance, by the cheerful background music in the nuclear holocaust scene in *Dr. Strangelove*. Thus, both amusement and horror (or possibly just amusement alone in the case of the nuclear holocaust scene, which is less graphic than the wood chipper scene) are salience-tracking and, thus, apt towards the fictional events.

This asymmetry, however, does not undermine the continuity thesis. The thesis as I characterized it earlier allows salience to be determined in different ways in fictional and real domains. While salience is a function of, among other things, our desires and interests, we have different desires and interests towards fictional and real objects (and this is a difference in desires and interests, not the aptness criteria of emotion). Additionally, salience in the fictional domain is affected by factors absent in reality, such as background music. Thus, it is unsurprising that different features of the same object are salient depending on

whether it is real or fictional. What the continuity thesis claims, with respect to salience, is merely that an apt emotion must be salience-tracking both in the fictional and real domains. It does not require the same object or feature to be salient in both domains.

Therefore, even *Fargo* and *Dr. Strangelove* fail to support the discontinuity thesis. Amusement in those cases is not a fiction-directed emotion if, as I have suggested, it is directed towards the scenes rather than their fictional contents. Even if the amusement is directed towards the fictional contents, and thus contrary emotions are apt towards the analogous fictional vs. real objects, this asymmetry stems from the different ways in which salience is determined in fictional and real contexts, which is compatible with the continuity thesis as I characterize it.

4. Different Kinds of Justificatory Reasons

This section concerns another argument for the discontinuity thesis, offered by Gilmore (2011). He argues that different kinds of considerations can count as reasons in justifying fiction- versus reality-directed emotions: only in the case of fiction-directed emotions, 'the fact that we are merely caused to feel a given way may be a reason to feel that way', provided that the work is designed to solicit the emotion and the solicitation is successful (p. 485). That is, 'if a work of art or exercise of the imagination is designed or intended to evoke an emotion toward some object it represents, and succeeds in doing so, that is a justification for the emotion in its representational dimension even if it is not supported by any other reasons that speak to the object's possession of the emotion's criterial qualities'

(p.481). Gilmore offers several examples in support of this claim, including a menacing soundtrack causing fear towards a fictional object and the beauty of an elegy evoking sadness towards the object the poem describes.

There are two different ways to unpack Gilmore's argument. First, his idea might be that the fact that we are merely caused to feel a certain fiction-directed emotion may provide a justification for the emotion because *this fact may speak to the object's having the emotion's criterial quality*. Or, he might mean that the fact provides a justification for the emotion even when it does *not* speak to the object's having the criterial quality, if the work is designed to solicit the emotion and the solicitation is successful. It is not completely clear which claim Gilmore in fact has in mind, but I will show below that neither undermines the continuity thesis.

1. Suppose Gilmore's claim is that in the case of fiction-directed, but not reality-directed, emotions the fact that we are merely caused to feel a certain emotion may provide a justification for the emotion *insofar as the fact speaks to object's having the criterial quality*. For instance, the fact that a menacing soundtrack causes fear towards a fictional object may be a reason to feel fear towards it because the fact speaks to the object being fearsome. Now, there are different ways in which such a fact may speak to the object's having the criterial quality.

1.1. First, such a fact may speak to the object's having the criterial quality because the emotion's cause may be *reliably linked* to the object's having the quality. For instance, at least in certain film genres, a menacing soundtrack is

usually accompanied by something fearsome, such as a monster's attack. Then, it might be the case that such a reliable link is sufficient to justify fear, even in the absence of any other justificatory reasons.

This, however, does not support the discontinuity thesis because reality-directed emotions, too, can be justified in the same way.[116] For instance, if there is a reliable link between walking alone in a dark alley and something fearsome happening, my fear towards an object I encounter while walking alone in a dark alley may be justified even when it is merely caused by the environmental trigger. Thus, if Gilmore's argument is understood in terms of a reliable link between causes of fiction-directed emotions and their objects' possession of relevant criterial qualities, it fails to support the discontinuity thesis.

1.2. Another way in which the fact that we are merely caused to feel a certain fiction-directed emotion may speak to its object's having the emotion's criterial quality is by *affecting fictional truth*. That is, it is not just that a menacing soundtrack is *reliably linked* to a fictional object's fearsomeness; the soundtrack or the fact that it causes fear might be part of *what makes the object fearsome* in the fiction (that is, the supervenience or grounding base of the fearsomeness). And that is why, Gilmore might argue, fear is justified by merely being caused by the soundtrack. There are places in which he seems to endorse this idea. For instance, he says, 'something depicted in a fiction can be fearsome (it can be part of the content of the fiction that it is fearsome) just because we are *caused* to feel fear

[116] See, e.g., Greenspan (1988).

toward it, if that is in accordance with the ends with which the work was designed' (2011, p.483, emphasis in original).

I agree with Gilmore that fictional truths can be determined this way. But I disagree that this undermines the continuity thesis. If fear towards a fictional object is justified in virtue of being caused by the supervenience or grounding base of the object's fearsomeness, then fear towards an actual object (e.g., a dog) is also justified in virtue of being caused by the supervenience or grounding base of the object's fearsomeness (e.g., the big teeth of the dog).[117] There are, of course, differences between the fictional and actual domains in what kinds of things can be the base of an object's fearsomeness. There is no non-diegetic soundtrack in reality, so such a soundtrack or the fact that fear is caused by it cannot be the base of a real object's fearsomeness. This asymmetry, however, lies merely in a difference between truth and fictional truth, not in a difference in the aptness criteria of emotion. No doubt, truth and fictional truth are constituted in different ways—for instance, a fiction's stylistic features partly determine what is true in it, whereas such features are absent in reality. But this is not the kind of asymmetry at issue in the debate between the continuity and discontinuity theses. The continuity thesis as I characterize it allows differences between truth and fictional truth.

Having said that, it is possible that Gilmore and I merely characterize the continuity thesis differently. Perhaps Gilmore takes the thesis to be incompatible with truth and fictional truth having different kinds of grounds. If so, there is no

[117] In fact, Gilmore seems to admit that reality-directed emotions can be justified this way (2011, pp. 474-475).

substantial disagreement between us.[118] The discussion so far is still meaningful, however, because it shows that both he and I agree about where to locate the difference between the aptness criteria of fiction- and reality-directed emotions: it is located in the different ways truth and fictional truth are determined, rather than different kinds of justificatory reasons for emotions. In both fictional and actual domains, justificatory reasons for an emotion speak to the object's (fictionally or actually) having the emotion's criterial quality. The difference is just that the object's criterial quality may have different grounds depending on whether it is fictional or actual.

2. Let us now turn to the second way to unpack Gilmore's argument. His claim might be that, in the case of fiction-directed, but not reality-directed, emotions, the fact that we are caused to feel a certain emotion provides a justification for the emotion *even when it does not speak to the object's having the emotion's criterial quality*, as long as the work is designed to solicit the emotion and succeeds in doing so. This interpretation is consistent with the following passage:

> Many fiction- and imagination-directed emotions, however, are generated in activities that are defined by ends – such as pleasure, entertainment, and absorption – that don't require that the emotions always be rationalized by the facts of the objects to which they respond. One may feel warmly disposed toward a figure in a painting only because the beauty of the work is designed to cause one to feel that way. Even if that feeling is not

[118] Let me add, though, that the victory of the discontinuity thesis would be quite uninteresting if the thesis held simply in virtue of the differences between truth and fictional truth, which are, unsurprisingly, different.

> justified by any facts that are represented, it may be still be justified if one of the internal purposes of engaging with that work of art is to respond with the emotions the work is designed to solicit (assuming that solicitation is successful) (Gilmore, 2011, p. 484-485).

This second way of understanding Gilmore's argument, too, fails to undermine the continuity thesis. To begin with, the fact that a work is designed to solicit an emotion towards an object and succeeds in doing so—if it does not speak to the object's having the criterial quality—seems to be irrelevant to the emotion's justification *in its representational dimension*, which Gilmore and I agree is what is at issue. As he puts it, justification for an emotion, say amusement, in its representational dimension is justification 'in the merited sense that would make intelligible *what it is* for an object to be amusing' (p. 471, emphasis in original). Given this, the fact that a work is designed to solicit an emotion and succeeds in doing so is irrelevant to the emotion's justification in the representational dimension, if the fact does not speak to its object's having the relevant criterial quality. This fact might be relevant if the emotion's object were the *work* rather than a fictional object in it. When a work is designed to solicit an emotion and succeeds in doing so, this speaks to the *work* having the relevant criterial quality. However, an emotion directed at a work is a reality-directed, not fiction-directed, emotion.[119]

Of course, there is a sense in which the fact that the work is designed to solicit the emotion may justify and rationalize feeling that emotion, even when it is directed at a fictional object: feeling the emotion is conducive to something else

[119] On the appropriateness of emotions directed at artworks, see, e.g., Choi (2003).

that the agent has a reason to do, namely, engaging with the work for pleasure, entertainment, or absorption. But this instrumental sense of justification is not the kind of justification at issue.

There is a further problem with Gilmore's argument. Suppose, for the sake of argument, that the fact that a work is designed to solicit an emotion and succeeds in doing so does provide a justification for the emotion in its representational dimension. The problem, now, is that reality-directed emotions can be justified in the same way. To begin with, some artworks solicit emotions towards real objects. For instance, one might feel a certain emotion towards a real person in a non-fictional photograph only because the work is designed to cause that emotion. In such cases, the solicited emotion may be justified merely by the fact that the photograph is designed to solicit the emotion and succeeds in doing so. There are also non-art examples. Suppose that you are engaging in a religious ritual designed to solicit admiration towards a religious leader (who is a real person), and it succeeds in doing so. This mere fact, even when you have no other reason that speaks to the person being admirable, may justify your admiration in the same way a fiction-directed emotion is justified by the fact that the work is designed to solicit the emotion and succeeds in doing so. There are other similar cases in which reality-directed emotions can be justified in terms of the purpose or design of the activity one engages in, such as sadness towards someone's death merely caused by the funeral one is attending, and fear one experiences while riding a (perfectly safe) roller-coaster. Thus, there is no discontinuity. *If* fiction-directed emotions can be justified by reasons that do not speak to their objects'

having relevant criterial qualities, reality-directed emotions, too, can be justified by such reasons.

Therefore, Gilmore's argument, whichever way it is understood, fails to undermine the continuity thesis. Whether his claim is that the fact that we are merely caused to feel a certain fiction-directed emotion may provide a justification for the emotion only when this fact speaks to its object's having the emotion's criterial quality or even when it does not, it fails to show that there is asymmetry between fiction- and reality-directed emotions that is incompatible with the continuity thesis.

5. Conclusion

I have so far argued that fiction- and reality-directed emotions have analogous criteria for aptness. In both cases, an emotion is apt if and only if it is fitting, justified, and salience-tracking, and these notions are understood in analogous ways. The seeming discontinuities between the fictional and actual domains result from the different ways truth and fictional truth, as well as salience in fictional and actual contexts, are determined. These differences, however, are not differences in the aptness criteria for emotions *per se*.

Bibliography

Achinstein, P. (1983). *The Nature of Explanation*. Oxford: Oxford University Press.

Alward, P. (2010). That's the fictional truth, Ruth. *Acta Analytica*, 25(3), 347–363.

Anderson, J., & Jeffrey, D. (1998).Moderate autonomism. *British Journal of Aesthetics 38*, 150–166.

Audi, P. (2012). Grounding: Toward a theory of the in-virtue-of relation. *Journal of Philosophy, 109* (12), 685-711.

Beardsley, M. (1974). The descriptive account of aesthetic attributions. *Revue Internationale de Philosophie*, *28*(109), 336–352.

Bender, J. (1987). Supervenience and the justification of aesthetic judgments. *The Journal of Aesthetica and Art Criticism*, *46*(1), 31–40.

Bender, J. (1997). On Shiner 's "Hume and the Causal Theory of Taste ". *The Journal of Aesthetics and Art Criticism*, *55*(3), 317–320.

Benovsky, J. (2012). Aesthetic supervenience versus aesthetic grounding. *Estetika:The Central European Journal of Aesthetics*, (2), 166–179.

Carroll, N. (1996). Moderate moralism. *British Journal of Aesthetic*s, 36, 223–238.

Carroll, N. (1998). Moderate moralism versus moderate autonomism. *British*

Journal of Aesthetics, 38, 419–424.

Carroll, N. (2000). Art and ethical criticism: An overview of recent directions of research. *Ethics, 110*, 350–387.

Carroll, N. (2006). Ethics and aesthetics: Replies to Dickie, Stecker, and Livingston. *British Journal of Aesthetics, 46*, 82–95.

Carruthers, P. (2015). Perceiving mental states. *Consciousness and Cognition, 36*, 498–507.

Choi, J. (2003). All the right responses: Fiction films and warranted emotions. *British Journal of Aesthetics, 43*, 308-321.

Currie, G. (1990a). Supervenience, essentialism, and aesthetic properties. *Philosophical Studies*, *58*, 243–257.

Currie, G. (1990b). *The Nature of Fiction*. Cambridge: Cambridge University Press.

Currie, G. (2014). Emotions fit for fiction. In S. Roeser & C. Todd (Eds.), *Emotion and Value* (pp. 146–166). Oxford: Oxford University Press.

Dancy, J. (2004). *Ethics without Principles*. Oxford University Press.

D'Arms, J. & Jacobson, D. (2000). The moralistic fallacy: On the 'appropriateness' of emotions. *Philosophy and Phenomenological Research, 61*, 65-90.

Day, T., & Kincaid, H. (1994). Putting inference to the best explanation in its

place. *Synthese, 98*(2), 271–295.

Dworkin, R. (1985). Law as interpretation. In *A Matter of Principle* (pp. 119–180). Cambridge, Massachusetts: Harvard University Press.

Eaton, A. (2012). Robust immoralism. *Journal of Aesthetics and Art Criticism, 70*, 281–292.

Eaton, A. (2013). Reply to Carroll: The artistic value of a particular kind of moral flaw. *Journal of Aesthetics and Art Criticism, 71*, 376–380.

Folde, C. (2015). Grounding interpretation. *British Journal of Aesthetics*, *55*(3), 361–374.

Franklin-Hall, L. (2015). Explaining causal selection with explanatory causal economy: Biology and beyond. In P. Braillard & C. Malaterre (Eds.), *Explanation in biology - an inquiry into the diversity of explanatory patterns in the life sciences* (pp. 413–438). New York: Springer.

Friedman, M. (1974). Explanation and scientific understanding. *Journal of Philosophy, 71*, 5–19.

Gaut, B. (1998). The ethical criticism of art. In J. Levinson (Ed.), *Aesthetics and Ethics: Essays at the Intersection* (pp. 182–203). Cambridge University Press.

Gaut, B. (2007). *Art, Emotion and Ethics*. Oxford: Oxford University Press.

Gendler, T. S. (2000). The puzzle of imaginative resistance. *Journal of*

Philosophy, *97*(2), 55–81.

Gendler, T. S. (2006). Imaginative resistance revisited. In S. Nichols (Ed.), *The Architecture of the Imagination* (pp. 149–174). Oxford: Oxford University Press.

Giovannelli, A. (2007). The ethical criticism of art: A new mapping of the territory. *Philosophia*, *35*,117–127.

Giovannelli, A. (2013). Ethical criticism in perspective: A defense of radical moralism. *Journal of Aesthetics and Art Criticism, 71*, 335–348.

Gilmore, J. (2011). Aptness of emotions for fictions and imaginings. *Pacific Philosophical Quarterly, 92*, 468-489.

Goldman, A. (2006). There are no aesthetic principles. In M. Kieran (Ed.), *Contemporary Debates in Aesthetics and the Philosophy of Art* (pp. 299–312). Oxford: Blackwell Publishing.

Greenspan, P. (1980). A case of mixed feelings: Ambivalence and the logic of emotion. In A. Rorty (Ed.), *Explaining Emotions* (pp. 223– 250). Los Angeles: University of California Press.

Greenspan, P. (1988). *Emotions and Reasons: An Inquiry into Emotional Justification*. New York: Routledge.

Harold, J. (2008). Immoralism and the valence constraint. *British Journal of Aesthetics, 48*, 45–64.

Harold, J. (2011). Autunomism reconsidered. *British Journal of Aesthetics, 51*,137–147.

Hart, H., & Honoré, T. (1959). *Causation in the Law*. Oxford: Oxford University Press.

Hempel, C. & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science 15* (2), 135-175.

Hick, D. H. (2012). Aesthetic supervenience revisited. *British Journal of Aesthetics*, *52*(3), 301–316.

Jacobson, D. (1997). In praise of immoral art. *Philosophical Topics, 25*, 155–199.

Jacobson, D. (2005). Ethical criticism and the vice of moderation. In M. Kieran (Ed.), *Contemporary Debates in Aesthetics and the Philosophy of Art* (pp. 342–355). Blackwell.

Kieran, M. (2003). Forbidden knowledge: The challenge of immoralism. In J. Bermúdez & S. Gardner (Eds.), *Art and Morality* (pp. 56–73). Routledge.

Kieran, M. (2005). *Revealing Art.* Routledge.

Kivy, P. (1975). What makes "aesthetic" terms aesthetic? *Philosophy and Phenomenological Research*, *36*(2), 197–211.

Lamarque, P. (1990). Reasoning to what is true in fiction. *Argumentation*, *4*(3), 333–346.

Levinson, J. (2011). Aesthetic supervenience. In *Music, Art, Metaphysics* (pp.

134–158). Oxford: Oxford University Press.

Lewis, D. (1978). Truth in fiction. *American Philosophical Quarterly, 15*, 37–46.

Lewis, D. (1983). Truth in fiction. In *Philosophical Papers Volume I* (pp. 261–275). Oxford: Oxford University Press.

Lewis, D. (1986a). Causal explanation. In *Philosophical Papers II* (pp. 214–240). Oxford: Oxford University Press.

Lewis, D. (1986b). Causation. In *Philosophical Papers II* (pp. 159-213). Oxford: Oxford University Press.

Lewis, D. (2004). Causation as influence. In J. Collins, N. Hall & L. A. Paul (Eds.), Causation and Counterfactuals (pp. 75-106). Cambridge, MA: MIT Press.

Liao, S. (2016). Imaginative resistance, narrative engagement, genre. *Res Philosophica*, *93*(2), 461–482.

Lipton, P. (2004). *Inference to the Best Explanation* (2nd ed.). New York: Routledge.

Livingston, P., & Alfred M. (1997). Evaluating emotional responses to fiction'. In M. Hjort & S. Laver (Eds), *Emotion and the Arts* (pp. 157–176). Oxford: Oxford University Press.

Mackonis, A. (2013). Inference to the best explanation, coherence and other explanatory virtues. *Synthese*, *190*(6), 975–995.

Levinson, J. (2005). What are aesthetic properties? *Proceedings of the Aristotelian Society, Supplementary Volumes*, *79*, 211–227.

McKeever, S., & Ridge, M. (2011). Aesthetics and particularism. In Michael Brady (Ed.), N*ew Waves in Metaethics* (pp. 264-285). Palgrave-Macmillan.

McMullin, E. (1996). Epistemic virtue and theory appraisal. In I. Douven & L. Horsten (Eds.), *Realism in the Sciences: Proceedings of the Ernan McMullin Symposium, Leuven 1995* (pp. 13–34). Leuven: Leuven University Press.

Mill, J. S. (1974). *System of Logic Ratiocinative and Inductive Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation BOOKS I-III*. Toronto: University of Toronto Press.

Miyazono, K. (2017). Truth in fiction and imaginative resistance. Retrieved May 2, 2018, from https://junkyardofthemind.com/blog/2017/9/18/truth-in-fiction-and-imaginative-resistance

Nannicelli, T. (2014). Moderate comic immoralism and the genetic approach to the ethical criticism of art. *Journal of Aesthetics and Art Criticism, 72*, 169–179.

Nisbett, R., & Wilson T. (1997) Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 231–295.

Potochnik, A. (2017). *Idealization and the Aims of Science*. Chicago: University of Chicago Press.

Proudfoot, D. (2006). Possible worlds semantics and fiction. *Journal of*

Philosophical Logic, *35*(1), 9–40.

Quine, W. V., & Ullian, J. S. (1978). *The Web of Belief* (2nd ed.). New York: McGraw-Hill, Inc.

Schaffer, J. (2012). Causal contextualisms: Contrast, default, and model. In M. Blaauw (ed.), *Contrastivism in Philosophy* (pp. 35–63). Routledge.

Shiner, R. A. (1996). Hume and the causal theory of taste. *The Journal of Aesthetics and Art Criticism*, *54*(3), 237–249.

Sibley, F. (1965). Aesthetic and nonaesthetic. *The Philosophical Review*, *74*(2), 135–159.

Sibley, F. (1974). The inaugural address : Particularity, art and evaluation. *Proceedings of the Aristotelian Society, Supplementary Volumes*, *48*, 1–21.

Simoniti, V. (2017). Aesthetic properties as powers. *European Journal of Philosophy*, 25(4), 1434–1453.

Smuts, A. (2013). The ethics of singing along: The case of "mind of a lunatic"'. *Journal of Aesthetics and Art Criticism, 71*, 121-129.

Sober, E. (1988). *Reconstructing the Past: Parsimony, Evolution, and Inference*. Cambridge: MIT Press.

Song, M. (2018). The nature of the interaction between moral and artistic value. *Journal of Aesthetics and Art Criticism, 76*(3), 285-295.

Stear, N.-H. (2015). Imaginative and fictionality failure: A normative approach.

Philosophers' Imprint, *15*(34), 1-18.

Stecker, R. (2005). The interaction of ethical and aesthetic value. *British Journal of Aesthetics, 45*, 138–150.

Stecker, R. (2008). Immoralism and the anti-theoretical view. *British Journal of Aesthetics, 48*, 145–161.

Stock, K. (2017). *Only Imagine*. Oxford: Oxford University Press.

Stokes, D. (2018). Rich perceptual content and aesthetic properites. In A. Bergqvist & R. Cowan (Eds.), *Evaluative Perception* (pp.19–41). Oxford: Oxford University Press.

Thagard, P. (1978). The best explanation : Criteria for theory choice. *The Journal of Philosophy*, *75*(2), 76–92.

Todd, C. S. (2009). Imaginability, morality, and fictional truth: Dissolving the puzzle of "imaginative resistance." *Philosophical Studies*, *143*(2), 187–211.

Walton, K. (1990). *Mimesis and Make-Believe*. Cambridge, MA: Harvard University Press.

Walton, K. (1994). Morals in fiction and fictional morality. *Proceedings of the Aristotelian Society, Supplementary Volumes*, *68*, 27–50.

Walton, K. (2006). On the (so-called) puzzle of imaginative resistance. In S. Nichols (Ed.), *The Architecture of the Imagination* (pp. 137–148). Oxford: Oxford University Press.

Walton, K. (2015). Fictionality and imagination—mind the gap. In *In Other Shoes: Music, Metaphor, Empathy, Existence* (pp. 17–35). Oxford: Oxford University Press.

Weatherson, B. (2004). Morality, fiction, and possibility. *Philosophers' Imprint*, *4*(3), 1–27.

Winkielman et al. (2005). Unconscious affective reactions to masked happy versus angry faces influence consumption behavior and judgments of value. *Personality and Social Psychology Bulletin, 31*, 121-135.

Zangwill, N. (1995). Moral supervenience. *Midwest Studies in Philosophy, 20*, 240–262.