Representational Content and the Keys to Success.

Justin C. Fisher

This shortened version of Chapter 5 of my dissertation has been accepted for presentation at the upcoming Pacific APA.

Abstract.

I consider the question of whether success-linked theories of content – theories like those of Ramsey (1927), Millikan (1984) and Blackburn (2005) which take there to be a definitional link between representational content and behavioral success – are consistent with the plausible claim that we can use content-attributions to explain behavioral success. Peter Godfrey-Smith (1996) argues that success-linked theories of content are too closely linked to success to be able to explain it. Against this, I present a plausible account of how content-attributions make available good explanations of behavioral success, and argue that if we want our content-attributions to be able to do this explanatory work, then we actually need to *embrace* a success-linked theory of content.

1. Introduction.

My primary opponent in this paper is Peter Godfrey-Smith (1996). Godfrey-Smith and I agree that attributions of representational content are pragmatically useful because they help us to explain behavioral success and behavioral failure. For example, Godfrey-Smith writes,

> [It is] possible to give a certain type of explanation for the success of behavior: Jake arrived safely *because* he knew the lie of the land. He thought the quicksand was to the right, and that is where it was. (Godfrey-Smith, 1996, pg. 172).

Godfrey-Smith and I disagree about whether or not this plausible view is compatible with *success-linked theories of content*, which Godfrey-Smith characterizes as follows:

> The relevant common feature of [success-linked theories of content] is their making some link between an inner [representational] state and behavioral success constitutive, or partially constitutive, of the state's having a particular truth condition (Godfrey-Smith, 1996, pg 178).

Success-linked theories of content are defended by Frank Ramsey (1927), Ruth Millikan (1984), Anthony Appiah (1986), David Papineau (1987), Fred Dretske (1988), Jamie Whyte (1990), and Simon Blackburn (2005). Given this list of advocates, it is an interesting question whether such theories are compatible with the plausible view that content attributions help us to explain success.[1] Godfrey-Smith thinks that if we want

[1] There are independent reasons to be interested in this question. For one thing, this question is closely related to long-standing questions about mental causation. For another, there is a general methodology that I elsewhere (in prep-a) call "Pragmatic Conceptual Analysis." This methodology seeks to explicate

content attributions to play this explanatory role, then we must *reject* all success-linked theories of content. I will argue that, on the contrary, if we want content-attributions to play this role, we actually must *accept* a success-linked theory of content.

Success-linked theories take there to be a constitutive link between content and success. This link ensures that content at least involves something relevant to success. This might seem to bode well for our using it to explain success. But, on the other hand, one might worry that this link actually disqualifies success-linked notions of content from doing this sort of explanatory work. Just as some *political* tasks require a political outsider, many *explanatory* tasks seem to require that the proffered explanation be conceptually independent of the thing being explained. Since success-linked notions of content are conceptual cronies of success, they seem poorly suited to explain it.[2]

Godfrey-Smith's objection runs roughly along these lines. I will turn to his objection momentarily (in section 2), and attempt to divert a portion of its force. I then offer a plausible account (in sections 3 and 4) of how our attributions of content explain success. I then argue (in section 5) that success-linked theories of content are not just *compatible* with giving these sorts of explanations; instead, they are required.

2. Godfrey-Smith's objection.

Godfrey-Smith (1996) presses the worry that success-linked notions of content are too closely related to success to explain it. He writes:

> The difficulty is that [success-linked theories of content] are not so much incompatible with the idea that truth leads to practical success – they are in a sense *too* compatible with it. These theories make the connection between truth and success an *ingredient* in what it is to represent the world. […] The problem is that explanations of success in terms of truth threaten to become vacuous, or at least much reduced in content. (Godfrey-Smith, pg 173-4)

Godfrey-Smith has in mind a very particular idea of what would count as a fully-contentful, not-at-all-vacuous explanation – namely *causal explanations* that cite only active causes of the thing being explained. He captures this idea metaphorically by saying that the goal of what he calls "orthodox naturalistic" theories of content is to offer explanations that depict truth as a sort of "resource" or "fuel", which plays an active causal role in bringing about success (pg 171-2). His primary objection against various

---

concepts by considering what explications we would need to adopt in order for these concepts to continue to deliver the sorts of pragmatic benefits that we have been using them to deliver. The usefulness of this methodology gives us reason to be interested in the question of which formal notions of representational content would be capable of doing the sorts of explanatory work we commonly call upon our folk concept of representational content to do.

[2] I'm not entirely convinced that we should always disqualify any proposed explanation that employs a crony of its explanandum. I'm inclined to think that, if the crony is embedded in a framework that is rich and informative enough, then the framework might still constitute a good explanation. However, we needn't delve into these issues here because I think that content-attributions may point us towards explanations that are quite free of cronies, and it is that sort of response that I defend below.

success-linked theories is that, since their notion of content bears a definitional rather than a causal relation to success, they are unable "to keep the idea of truth as a fuel or resource that generates success" (pg 182), and hence that they can't deliver the sorts of explanatory mileage that we "orthodox naturalists" want.

I find these presumptions very peculiar in two ways. First, I think Godfrey-Smith is employing a peculiarly narrow notion of *explanation*.[3] I have defended a broader notion of explanation elsewhere (in prep-b) and I draw upon it in my positive proposal below.

Second, Godfrey-Smith is attributing to the "orthodox naturalist" a very peculiar understanding of the role that truth plays in the world. I agree with Godfrey-Smith's earlier (pg 168) depiction of "naturalists" as believing that, whatever truth is, it must be some sort of *correspondence relation* that obtains between representational tokens and things in the world. But, so far as I can tell, this view *by itself* precludes the thought Godfrey-Smith attributes to the "orthodox naturalist": that truth plays a fuel-like causal role. Two things that correspond to one another may each play a causally-active fuel-like role, but correspondence relations, *themselves*, do not play such roles. They aren't the right sort of thing to play such roles – they aren't even things at all! Godfrey-Smith's "orthodox naturalist" is committed to a straight-forward category mistake, and I doubt that many truly-orthodox naturalists have fallen victim to this.

This is not to say that truth and other correspondence relations aren't *explanatorily* relevant, nor even that they aren't *causally relevant*. For example, I think the correspondence between the protrusions on my key and the tumblers in my lock is both causally and explanatorily relevant to my door's opening. I just don't think that any reflective naturalist should commit herself to thinking that *the way* in which various correspondence relations are explanatorily and causally relevant is the same as the way in which *fuel* is explanatorily and causally relevant.

So here, then, is a better metaphor. *Representational tokens are keys to success*. Like keys, representational tokens are causally active things. And just as a key normally won't be much good unless it corresponds, in the relevant way, with the lock you put it in, our representational tokens normally aren't much good unless they correspond, in the relevant ways, with the circumstances we use them in. Just as a key that doesn't fit the relevant lock may easily lead you to do something that turns out to be fruitless or worse, a representation that doesn't fit your circumstances may easily lead you to behave in ways that are fruitless or worse.

It is good to have improved our metaphor, but we still face the question: *How exactly <u>can</u> accuracy of representation be explanatorily relevant, especially if we're employing a notion of representational content that has definitional links to the successes and failures*

[3] At some points Godfrey-Smith seems willing to grant that success-linked theories might successfully give other sorts of explanations besides the peculiar sort of causal explanations he has in mind. However, he clearly thinks of these other sorts of explanations as being of an inferior class, as is evidenced by the words he uses to describe them: "explanatory, in a certain sense" (pg 172), "vacuous, or much reduced in content" (pg 174), "…do not say as much as they seem to" (pg 174), "a weak explanatory role" (pg 182), "an explanation in the sense of a subsumption under a historical generalization, but no more"(pg 183).

we'd like to explain? I will focus on a simple example. A good account for it may serve as a good model for understanding more complicated cases.

3. A simple example.

Suppose S is an astronomy buff, who each evening forms beliefs about whether the weather conditions are the sort C* of conditions where one can see stars, or the sort of conditions C where no stars are visible. These beliefs determine whether she will engage in a behavior B* that requires visible stars for its success like going to the observatory and gazing upwards, or instead engage in some behavior B that doesn't involve stars, like reading a book.

We might naturally explain S's successfully avoiding a fruitless trip to the observatory by saying she believed, correctly, that the conditions C were such that no stars would be visible. And, we naturally explain cases of failure – e.g., S's missing an interesting celestial alignment – by saying she incorrectly believed (C) no stars would be visible, and hence decided (B) to stay home.

Such explanations presuppose that S might have one or the other of two (syntactic[4]) types of representational tokens. S's belief that she is in C is a token of type {C}; had S instead believed that she was in C*, that belief would have been a token of type {C*}. I have given each of these types a name that includes a helpful reminder of the content that was ascribed to its instances. It is worth stressing, however, that each instance of type {C} is just a particular token – constituted, perhaps, by an active assemblage of neurons – which almost certainly does not, in fact, "wear its content on its sleeve."

A good explanation depicts its explanandum as an instance of a robust pattern, one that specifies how the explanandum and various contrasting alternatives depend upon which of various antecedent-properties is instantiated.[5] The above content-attributions naturally suggest that the following 'Master Pattern' is displayed by S.[6]

---

[4] There are difficult questions involving how various particular tokens should be grouped into types. My own inclination is to try to spell out the notion of 'syntactic types' in terms of teleo-functional role. Within an agent, a type consists of the various different tokens that are all supposed to be treated as meaning the same thing by the given cognitive system. Sorting tokens across agents would be more challenging, but might be done by looking for certain sorts of correspondence in teleo-functional role.

[5] Throughout this section and the next, I will draw upon the account of explanation given in my (###), though this particular point may also be supported, e.g., by Wesley Salmon who writes:

> "[The world is endowed] with patterns that can be discovered by scientific investigation… To explain an event … is to fit [it] into a discernible pattern" (Salmon 1984, pg 17).

[6] The claim that this pattern is robustly displayed should be limited to a certain sort of circumstances – e.g. ones where background beliefs, desires, and circumstances are relevantly like those S has. This pattern might also be claimed to hold for other agents who are relevantly like S. There will be significant challenges in spelling out the notions of 'relevant likeness', but those are challenges for another paper.

	{C}	{C*}
C	**B, SUCCESS**	**(B*, failure)**
C*	**(B, failure)**	**B*, SUCCESS**

Figure 1. The Master Pattern. This pattern describes how an agent's behavior and success (listed in the cells) both depend upon both what circumstance (row) she is in, and upon what sort of representational token (column) she has, and it says that cells on the major diagonal are more likely to occur than the other cells.

This Master Pattern doesn't *explicitly* say anything about content – instead it just describes ways in which behavior and success depend upon circumstances and (syntactic) types of representational tokens. This allows us to evade Godfrey-Smith's worries about political cronyism. Admittedly, we were led to consider the Master Pattern by a content-attribution that may (for all we've said so far) have had very close conceptual links to success. But even if it was a political crony that recommended it to us, the Master Pattern's own political record is squeaky clean – it includes no references to *content*, nor does it contain any illicit reference to *success*.[7] Hence it is immune to any worries about cronyism.

However, since we were prompted to consider this pattern by content-attributions, it is quite natural to characterize what is going on in the Master Pattern in terms of content. This yields three generalizations:

> **Content determines behavior:** An agent's behavior (B or B*) will match the content (C or C*) that U_{S+} attributed to her internal representational state.
>
> **Accuracy determines success.** One's success or failure depends upon whether one's representational contents (and hence one's behavior) match one's circumstances (C or C*) – i.e., upon the *accuracy* of one's representations.
>
> **Inaccuracy is unlikely.** Agents usually get things right, and only rarely misrepresent their circumstances and suffer failure.

These generalizations are as close to truisms of folk psychology as one could reasonably expect to get – we all tacitly know that, at least in ordinary cases, generalizations like these have many more instances than counter-instances. This suggests that content-attributions really do suggest (albeit tacitly) to competent 'folk psychologists' that patterns like the Master Pattern obtain.

---

[7] The Master Pattern does of course mention success *within its cells*, but these cells correspond to potential *explananda*, and hence these references are perfectly acceptable. More worrisome would be if the row-headings mentioned success or a success-linked notion of content. For then one could reasonably question whether success and being-in-a-certain-row are independent enough for one to explain the other.

4. Are such explanations *good* explanations?

We now have a proposal regarding what sort of pattern lies at the core of simple explanations evoked by content attributions. An explanation of S's success (or failure) depicts this as an instance of the top-left (or bottom-left) cell of the Master Pattern. Explanations like this are immune to Godfrey-Smith's cronyism worry. *But are they good explanations*?

They do at least satisfy many desiderata for good explanations.[8] Patterns like the Master Pattern evidently do reflect real connections between representational tokens, behaviors, and the conditions in which those behaviors will succeed. Such patterns also seem simple enough, natural enough, and general enough; or at least no worse in these respects than many other explanations we usually count as good.

But a good explanation can't just highlight an appropriate pattern; it must also provide links to relevant practical[9] and theoretical information.

Patterns like the Master Pattern naturally suggest practical recipes for affecting which cell a subject well end up in by affecting which row or column she is in. For example, the victim of a classic 'shell game' can succeed only if she finds a particular pea. If you can get this victim to have a different belief about where the pea is, even while the pea remains unmoved – i.e. if you can change her column while leaving her row unchanged – this is a good way to alter her behavior and her success. And if you can secretly move the pea without altering her beliefs about it – i.e., if you change her row without changing her column – this too is a good way to alter her success.

What remains is to show that explanations like the one proposed above might offer links to good theoretical accounts for *why* their central patterns are robustly displayed. Without such links, these proffered explanations will seem suspiciously like *mere instruction manuals* that tell us what to expect and how to bring about various consequences but don't provide the *understanding* we expect good explanations to deliver.[10]

---

[8] These plausible desiderata play a central role in the view of good explanations defended in my (in prep-b).

[9] This demand is articulated by Woodward (2004) as well as in my (in prep-b). This demand captures what is correct about the simple causal theory of explanation that Godfrey-Smith presumes (see section 2 above). However, the demand for practical informativeness allows for a broader notion of explanation than does the simple causal theory. E.g., we may explain a statue's current color by noting its current chemical composition. This explanation is practically informative because practical recipes for modifying chemical composition are great ways to modify color. However, since causation is a diachronic relation, current chemical composition cannot be a cause of current color. Further counter-examples are the dynamical explanations discussed below.

[10] It might not be that uncomfortable to hold that this is indeed all that folk content-attributions provide – instruction manuals can be well worth having, and it might not be all that important to the folk *why* these instruction manuals work. However, it still is reasonable for scientists and philosophers to ask *why* the instruction manuals work; and it certainly would be preferable (ceteris paribus) to adopt a theory of content that allows good answers to this question, rather than leaving these regularities mysterious.

What we need, then, is an account for why the generalizations embodied in the Master Pattern should hold. Why would agents like S be such that their contents determine their behavior? And why would their representations tend to be accurate rather than inaccurate? [11] A good answer to these questions can be put quite simply: *because this works – i.e., because this enables people to be successful.*

This answer is a *teleological* explanation. To understand how such explanations work, let us consider a simple analogy. A few minutes ago, I gently tossed a marble into an empty salad bowl. Why does the marble now rest at the bottom of the bowl? A good explanation notes that friction guarantees that the marble must come to rest at some point, while the pull of gravity ensures that it won't stay at rest at a point that is not a local minimum. As a consequence of this 'bottom-seeking dynamics', the marble was bound to end up resting at the bottom of the bowl. [12, 13]

Simple teleological explanations refer to 'success-seeking dynamics' rather than 'bottom-seeking dynamics.' They depict their explananda as being *states that tend to succeed* in a dynamical system that won't rest long in states that tend comparatively poorly to succeed. Such explanations rely upon two dynamical features of the systems in question: (1) whenever such systems are in states that tend comparatively poorly to succeed they will tend not to remain in those states, and (2) such systems would eventually arrive at a state that does work comparatively well.[14]

---

[11] The robustness of the third generalization mentioned above ('accuracy determines success') will be quite straight-forward to account for, given that 'content determines behavior'. There should be no special problems in explaining, regarding particular cases, why one's success would hinge on there being a match between one's *behavior* and one's circumstances. E.g., there is no particular problem explaining why stargazing behavior B* succeeds only when C* stars are visible, and the less exciting (to the astronomy buff) alternative behavior B really succeeds only when C there are no stars to be seen. From this and 'content determines behavior' it follows that one's success will hinge upon one's *content* matching one's circumstances.

[12] This explanation is practically informative: recipes for changing the local minimum in the bowl – e.g., tilting the bowl – are good recipes for bringing about changes in where the marble (soon enough) rests.

[13] This dynamical explanation of the marble's resting place is not some sort of second-class explanation, nor is it just a feeble gesture towards an explanation in terms of the actual sequence of causes that led the marble to rest where it did. It would *not* improve this explanation if we added to it (in any prominent way) a further specification of the path the marble actually took or of the particular forces that acted upon it along the way. These additions would make the explanation *worse*, by adulterating it with details that didn't make a difference to what happened – it doesn't matter that the marble experienced precisely the causes that it did, for any other 'nearby' causes would have served just as well. Hence, this dynamical explanation is a first-class explanation in its own right, even if it fits poorly with a causal theory of explanation.

[14] This general structure of teleological explanation is neutral across many particular ways in which success-seeking dynamics might be instantiated. These dynamics might be instantiated by natural selection upon genetic diversity, in which case 'success' would involve producing a large number of offspring. They might also be instantiated by cultural selection upon various artifactual designs, by trial-and-error learning mechanisms acting upon various attempted kinds of behavior, or by the tendency of foresightful reasoners to reject plans that obviously won't work. Regardless of how these success-seeking dynamics happen to be instantiated, they will underwrite the robustness of a pattern of frequent successes, where 'success' in each case must be interpreted as involving the sorts of states that that particular sort of dynamical system would be willing to rest with. Since these various sorts of success often coincide, at least roughly, we often do

The Master Pattern is stably maintained by a particular sort of feedback loop in a success-seeking dynamical system. Again, an analogy will help. Market economies exhibit a sort of success-seeking dynamics.[15] In such economies, consumers tend not to engage long in optional undertakings that require goods whose producers are quite unreliable. Similarly, producers tend not to continue producing goods that don't fit the needs of consumers' current undertakings. The result is a feedback loop – consumers tend to undertake projects that depend upon the goods provided reliably by producers, and producers tend to produce the sorts of goods that consumers' current undertakings require.

A very similar feedback loop arises in success-seeking dynamical systems that produce and consume *representational tokens*. Given the ways representational tokens are currently consumed (namely, as a guide to behavior), success-seeking dynamics won't rest long with ways of producing these tokens that needlessly fail to make such patterns of consumption successful. And given the ways these tokens are produced, success-seeking dynamics won't rest long with ways of consuming them that fail to capitalize upon their actual informational value. Patterns of token production and token consumption are constrained by success-seeking dynamics to dance closely together for as long as these tokens are used. (See Figure 2.) And so long as they dance closely together, patterns like the Master Pattern will be robustly displayed.

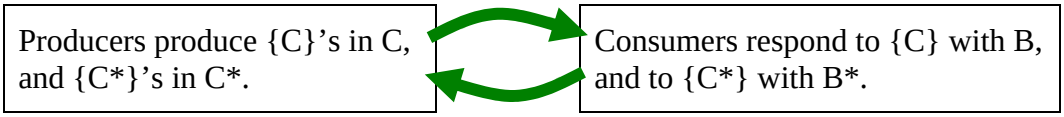

Figure 2. The feedback loop maintaining the robustness of the Master Pattern.

5. Why content must be success-linked.

I have sketched what I hope is a plausible account of how content attributions make available good explanations of success and failure. In laying this story out, I did not *explicitly* draw upon any particular theory of content. However, I will now argue that this plausible story actually requires a success-linked theory of content.

Our choice of a theory of content commits us to constraints upon which contents we may (correctly) attribute in various cases, and hence upon which particular choices of C may occur in the Master Pattern that is naturally evoked by such attributions. In turn, this will affect (1) whether the Master Pattern will robustly obtain in the world; (2) whether the Master Pattern will sustain the sorts of intervention required for it to be a good explanation; and (3) what sorts of supporting theoretical accounts we might give of the Master Pattern's robustness.

Because of (1), (2) and (3), the plausible expectation that the Master Pattern be a good explanation constrains our choice of C in two significant ways:

---

fairly well not worrying about exactly which notion of 'success' is at play, and not worrying about what exactly is implementing success-seeking dynamics in the systems we consider.

[15] They tend not to rest long in 'Pareto-suboptimal' states that involve needless unhappiness.

(a) C must be a sort of circumstance whose presence or absence is correlated with the presence of {C}-tokens, at least when we restrict our attention to a certain range of 'normal' agents and circumstances; and

(b) C must be a constitutive or causal prerequisite for the success of behavior B in the background context in question.

[For brevity, I omit the derivation and discussion of (a).]

Now, let us establish (b). If the Master Pattern is to constitute a good explanation, then (2) it would need to admit recipes for altering success *by way of* altering the presence of C. (It should admit, for example, that *secretly moving the pea* is a way of changing who will win the shell game.) For this to work out, C can't be a *mere* correlate of success in circumstances where B is performed – instead C needs either to be a *constituent* or a *cause* of the success of B.

We may independently support (b) by considering what is required for (3) the feedback loop in the teleological explanation for the Master Pattern's robustness. If C is not a pre-requisite for the success of the behavior B that is reliably generated in response to {C}-tokens, then we lose our teleological explanation of the accuracy of these tokens (the bottom arrow in figure 2). Why should the producer of {C}-tokens reliably track the presence of C, if the presence or absence of C makes no difference to the success of the consumer of these tokens?[16]

Conclusions (a) and (b) together place a strong constraint that any theory of content must meet if it is to allow that our content attributions make available the sorts of good explanations described above. This constraint involves a fairly essential reference in (b) to what is required for certain behaviors to be successful. I can see no promising way to spell out a particular notion of content that will be consistent with this constraint, but not itself make reference to some sort of success. Hence, I conclude that we will need to embrace a *success-linked* theory of content if we want to allow that our content attributions make available the sorts of good explanations described above.

[16] Of course there may be alternative explanations available in particular cases, but there wouldn't be the sort of teleological explanation available that we can usually presume is there.

References.

Appiah, A. 1986. "Truth Conditions: A Causal Theory." In *Language, Mind and Logic*, Thyssen Seminar Volume, Jeremy Butterfield (ed.) Cambridge: Cambridge UP, pp. 25-45.

Blackburn, S. 2005. "Success Semantics". In *Ramsey's Legacy*, H. Lillehammer and D. H. Mellor, eds. Oxford: Oxford UP.

Dretske, F. 1988. *Explaining Behavior*. Cambridge, MA: MIT Press.

Fisher, J. (in prep-a) *Pragmatic Conceptual Analysis*. University of Arizona dissertation.

_____. (in prep-b) "Why Ask Why? What Good Explanations Deliver and How They Deliver It." (Chapter 4 of *Pragmatic Conceptual Analysis*.)

Godfrey-Smith, P. 1996. *Complexity and the Function of Mind in Nature*. Cambridge: Cambridge UP.

Millikan, R. 1984. *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.

Papineau, D. 1987. *Reality and Representation.* Oxford: Blackwell.

Ramsey, F. 1927. "Facts and Propositions." In *The Foundations of Mathematics, and Other Logical Essays*. R.B. Braithwaite, ed. London: Routledge and Kegan Paul, 1931, pp. 138-55.

Salmon, W. 1984. *Scientific Explanation and the Causal Structure of the World*, Princeton: Princeton UP.

Whyte, J. 1990. "Success semantics." *Analysis* 50: 149–57.

Woodward, J. 2004. *Making Things Happen: A Theory of Causal Explanation*, Oxford: Oxford UP.