

# **Explaining with Mental Content**

## **Prospects for Content Expressivism**

Allan Gibbard

*University of Michigan*

*Ann Arbor, Michigan 48109*

*U. S. A.*

I'll begin with some truisms, and then strive to accommodate them. We explain things that people do in terms of what they think. But people now appear to be complex physical systems, upshots of natural selection and interactions with the environment. The movements that comprise, say, punching the buttons on a money machine, proposing marriage, or setting off on a vacation, these are resultants of complex physical goings-on in the brain and elsewhere. To the extent that they are explainable, we might then think, they should be explainable in physical terms. Where, the worry is, does that leave ordinary explanations in terms of mental content? She signed the credit card slip to pay for the dinner and tip the waiter. Does this amount to a physical explanation? Or, would it be made superfluous by the true explanation, which is physical? Or what?

This is a central puzzle we encounter in trying to understand ourselves. I hope that the way I have put it is naïve, and that there are more sophisticated and helpful ways to pose our questions—but we philosophers are finding it extremely difficult to find which sophisticated strategies of analysis might yield greater understanding. In this paper, I'll be exploring broad directions, and mostly not keeping careful track of how they tie in with extensive, wonderful philosophical treatments we already have.

Explanation may not be the only thing that meaning is good for. We want to speak sense, most of the time, and can get upset when people try to pass off nonsense for sense, empty cant for wisdom. Thoughts of meaning, then, can play a critical as well as an explanatory role in our lives. Critique is a part of philosophy, and one form a critique can take is to ask about meanings: in the first part of the past century, analytic philosophers and logical positivists placed this critical use at the core of philosophy. It is still a major tool of analytic philosophy—and of science, social criticism, and indeed almost any discipline that aims at understanding. Attributions of meanings and the thoughts behind them are crucial in tracking a conversation, in treating others people and not as machines.

These loose musings suggest a broad strategy for elucidating the concepts of meaning and of mental content. Consider how we use these concepts in such activities as tracking a conversation or a piece of reasoning, and then devise accounts of the concepts that will account for these uses. I'll consider one particular class of such strategies, the class I label "expressivistic", which stem especially from the study of ethical concepts. Give up on any straight analysis of the concept in question, the strategy is, and consider instead what it is to have the concept and use it. Ethical emotivists, for instance, started from the dictum that to think something is good is to approve of it. They aimed to characterize the mental state of approval, and then use the dictum to offer an indirect account of the concept GOOD. I'll be exploring two attempts to explain, along such lines, the concept of what a person is thinking. One starts with stances of agreement and disagreement. The other starts with claims that "meaning is normative", that the concept of meaning is a normative concept. My aim won't be to work up such proposals in detail, but to sketch aspects of them and then explore some worries.

What, then, of explanations in terms of mental content, explanations of things that happen, like explosions, population movements, and global warming? In the background of what I say will be an assumption: that even if we identified a concept of mental content that worked well in such explanations, this concept might still not meet the other demands we place on a concept of mental content. Natural happenings should be explainable naturalistically, and it may well be that some naturalistic successor to our everyday concept of mental content will do a fine job in such naturalistic explanations. Still, we might find, this naturalistic concept doesn't serve adequately for such purposes as tracking a conversation; perhaps no naturalistic concept will serve these other purposes. So I explore alternatives to naturalistic concepts of mental content—though I won't try to motivate the assumption that naturalistic concepts are inadequate for these purposes. Of course if we did find an alternative to naturalistic concepts of mental content, we would have to say whether and how such concepts can figure in explanations of natural events such as global warming. I'll touch on this issue briefly.

### **1. Naïve Attributions of Content**

Begin with ordinary, relaxed conversation. Jack stands in the hallway looking lost, and Jill asks if he needs help. He tells her he's looking for the philosophy department. Jill gestures and tell him it's down the hall to the left. Implicitly, we can say, each is attributing meaning to the other's utterances, and attributing content to states of mind that the other's words express. Jill, after all, isn't surprised when Jack heads off down the hall to the left, though she would be surprised if he went the other way. If we asked her, she might offer an obvious explanation of Jack's action—though she'd be baffled why

we asked at all. It would be, in philosophers' jargon, an explanation in terms of mental content. As the scenario stands, however, the two haven't been overtly thinking about meanings and states of mind in this interchange; Jack has been thinking about where to find the department. We observers, in the business of theorizing, may ourselves need to attribute sophisticated mental processes to these two in order to explain the interchange. When Jack says, "I'm looking for the philosophy department," Jill implicitly knows it's he and not she who's lost. She doesn't sympathize and ask him how long he's been looking; she tells him how to get there. Certain mental defects, we now know, would leave Jill incompetent in this situation, defects, we may try saying, in her capacities for "theory of mind". But she's not explicitly theorizing as she offers directions and watches without surprise as he goes off.

My aim in belaboring the obvious here is theoretical. I aim to ask what we, as theorists, might mean when we attribute mental content to Jack and Jill. Now Jill, of course, doesn't have to turn high powered theorist to have thoughts about meaning. If Jack says, "I'm seeking infrastructure for loving wisdom," she'll gape and may explicitly ask him what he means. "I'm looking for the philosophy department" could be the answer. Accepting this, she has now made an explicit attribution of meaning: that when he said that strange thing, he meant that he was looking for the philosophy department. We observers can note what has happened: now she can respond to his earlier words more or less as if he had says, "I'm looking for philosophy department." An explicit meaning claim redirects her responses, those responses that constitute, as we say, attributing meanings implicitly.

Taking the relaxed conversational stance, she hears Jack's words directly as expressions of thoughts—maybe we should even say, as the thoughts themselves. She simply *hears* him as saying that he's lost, say, and takes this perception at face value. Call this stance *direct comprehension*. At times this stance comes to stumble: Jill can fail to process what Jack is saying, and so wonder what he means. She can hear him as thinking he's lost, and then find herself wondering whether that is really what he could have meant. (Other things he does, for instance, might not fit what she took him to be saying.) Statements can be ambiguous, or in a dialect that isn't hers. Or he may be speaking a foreign language. When the stance of direct thought fails or its deliverances come into question, she need to interpret. The situation now calls for an *interpretive stance*: she now asks herself what he meant, and looks for an answer.

How should we conceive of this interpretive stance? To ask this question is to take a further stance, a stance that calls ordinary notions of meaning and interpretation into question. Call this the *meta-interpretive* stance. With this stance, we reach philosophy.

As a start at pondering matters from this metastance, picture the interpretive stance as a substitute: it substitutes for the mode of direct comprehension. Akua speaks a tongue that's foreign to me, and Kwasi interprets: "She says that no bus will come on Sunday." I accept his interpretation at face value, unreflectively. I now see matters as if I had myself understood Akua as saying this. I code matters much the way I would if I could take the stance of direct comprehension with her, and just hear her as saying that no bus would come on Sunday.

I don't mean this as a full philosophical account of the interpretive stance. We'd need to say more about the ways I do and the ways I don't think just as if I'd directly heard her say that no bus would come. We'd have to say, moreover, what would vindicate Kwasi's gloss, and what it would be for the gloss to be mistaken. That, indeed, is the topic of this present study. For now I offer just a vague suggestion: that we won't understand interpretation unless we somehow find it a surrogate for direct, unquestioned perception of what a person is saying—hearing the person directly *as* speaking certain thoughts.

The interpretive stance can apply to one's own thoughts. Ordinarily, you just think your thoughts and think nothing more about them. Call this the mode of *direct thinking*. You may also know what you've been thinking, from a stance of direct comprehension: no question of interpretation arises in your recall. But also you can be drawn up short in your thinking, and wonder whether what you've been thinking makes any sense. You resolve this with an interpretation of your own thoughts: not thoughts that you are whispering in your head at the very moment of interpreting them, but thoughts in recent memory that now bother you. Interpretation, though, as the saying goes, comes to an end: the thoughts that constitute the interpretation are thought directly, and only in a separate operation could be themselves brought up for scrutiny and interpretation. The self-interpretation, like an interpretation of another, acts as a surrogate: in this case, as a surrogate for just recalling yourself as thinking that such-and-such.

In this study, I work to develop a view of thoughts from a philosophical, meta-interpretive stance. I fret over interpretation: what do interpretive statements mean, and what justifies or validates them? In asking these questions, however, I strive to take a relaxed conversational stance with my readers. I hope that my readers can simply read me as saying what I say, can take a stance of direct comprehension toward what I write. Of course in this I must expect often to fail: Readers are sure to be brought up short, despite my efforts, and so to turn to an interpretive stance. And I myself must take an interpretive stance toward my writing from time to time, partly to check on how things will look to a reader, and partly to check on my own thoughts and their cogency.

My goal in the end, though, would be to have an account of what we're doing when we interpret my words or someone else's: what our interpretive claims mean, and what would

validate them as correct. This might even bring with it an account of what constitutes an error in direct comprehension.

## 2. Expressivist Renderings

You and I agree that Henri is a physical system, let us suppose, and we agree on what's physically happening with him. Imagine indeed that we are omniscient on these matters. We each have a way of matching physical states of his to ways of thinking that we ourselves can share. My scheme of interpretation  $\mathcal{M}$ , imagine, satisfies this constraint: that when I apply it to myself as a physical system, it gives the very thoughts I am having. Your scheme  $\mathcal{M}^*$  satisfies a like constraint. This constraint, though, doesn't force our schemes to agree.

Henri, as I interpret him, is a believer in a non-physical *elan vital*. When he sees a rabbit, he's convinced that it is not a pure physical system, but, as I'll put it, a *vitabunny*. He sees a rabbit and says, "Tiens, un lapin!" I interpret him as saying something with which I disagree: "Lo, a vitabunny!" You think this interpretation far-fetched; indeed, you don't interpret him as having any views at all on vitalism. You interpret him as saying, "Lo, a rabbit!" with which you agree. We are both physicalists; we think that rabbits are physical and that there's no such thing as a vitabunny. We both agree that Henri is presented with a rabbit, and we both agree, in physical terms, how he reacts to it. We disagree, though, on what his words mean, and on the content of the thinking that his words evince. What does our disagreement consist in?

I have framed the question in a way reminiscent of debates in metaethics in the first half of the twentieth century. (I'll render the issues and doctrines freely, changing some aspects to suit my own purposes.) I am an egoistic hedonist, imagine: I think that pleasure is the good, and that reason commands us to seek it. You are a perfectionist; you deny these tenets. We agree on all the physical and psychological facts of the world; we agree, for instance, that an opium dream is, in itself, pleasant. I, though, think that this makes the state of mind good in itself—though the later consequences of opium use may be bad. I think that the pleasantness rationally favors the use of opium (though other consequences may favor abstinence). You, as a perfectionist, reject these claims about goodness and rational considerations in acting. We agree on all the psychological facts but not on goodness and rationality in action.

G.E. Moore concluded that goodness is a non-natural property, though a property that supervenes on natural properties. Emotivists like Ayer and Stevenson proposed an alternative diagnosis: that we disagree in attitude. I favor pleasure and pleasure alone, and tell myself to seek my own pleasure. You favor perfection, and favor pleasure only as it contributes to perfection. Our disagreement, then, concerns what to go for in life.

Attributions of goodness and rationality are expressions of favoring, settling on a goal, and the like. I call this kind of analysis *expressivistic*.

Might a like, expressivistic strategy explain what's at issue in questions of interpretation? You and I disagree on what Henri is thinking and saying. Our disagreement doesn't concern the physical facts of how his brain works and the like. What, then, does it consist in? Perhaps, in some attitude or stance that I take toward him and his words. My musing at the start of this study might suggest some possibilities. In regarding Henri as meaning things, I'm taking his thoughts into my own stream of thinking, treating them subject to agreement or disagreement. If I had found myself thinking "Lo! a vitabunny," I'd immediately have second thoughts. I'd reject the thought; I'd disagree with myself of the previous moment. Perhaps it's a stance like this that I'm taking to Henri when I interpret him as meaning, "Lo! a vitabunny." What's at issue between us is whether or not to agree when Henri says "Tiens! un lapin." A scheme of interpretation, we might say, is a policy for agreeing or disagreeing with people. When I say that Henri means VITABUNNY by 'rabbit', I'm setting myself to agree with him or not according as there's a vitabunny.

Before exploring further such an expressivistic treatment of mental content, let me touch on explanations of actions and the like in terms of mental content. Jack asks where the Philosophy Department is, and on hearing Jill's answer, walks down the hall to the left. Schematically, we explain his response in terms of a belief and a goal: his belief that the Philosophy Department is to the left, along with his goal of reaching the department. It might be thought that an expressivist can't offer such explanations. The physical story of what goes on is the full causal story. Attributions of content, according to the expressivist, aren't physical descriptions; they aren't to be explained as descriptions at all, but as, say, policies for agreeing or disagreeing.

Like worries arise with expressivism in ethics. People use opium, a hedonist might claim, because the mental state it produces is good. If we explain the concept of good expressivistically, though, so that thinking something good consists in favoring it, are such explanations so much as intelligible?

Expressivists, though, can understand such explanations. The story is a long one, but briefly, we can say this: If I'm an ethical hedonist, favoring pleasure alone, and I think that people smoke opium because its effects are pleasant, then I accept that they smoke it because its effects are good. You, as a perfectionist who agrees with me on the psychological facts, will reject this explanation; my accepting the explanation consists in a combination of accepting a psychological explanation and favoring pleasure—and you don't favor pleasure. And what if someone is unsettled on what to favor? She might still think that people smoke opium do so because the results are good. This amounts

to restricting what combinations of favorings and facts she can come to accept without changing her mind. She can come to agree with me without change of mind, but not with you who are a perfectionist and don't count the pleasures of opium as perfections. We explain explanations in terms of goodness expressivistically, by elucidating the state of mind that constitutes accepting the explanation.

Suppose, then, that a scheme of interpretation consists in this: a set of policies for taking a stance of agreement or disagreement in certain physical circumstances. Then in thinking that Jack's belief helps explain his walking to the left down the hall, I'm doing this: I'm restricting myself to (i) policies for agreement and disagreement and (ii) beliefs about the physical layout of the world, which pair jointly have a certain property. The property is roughly this: that the believed physical layout includes a state such that the policy is, if that state obtains, to agree with Jack just in case the Philosophy Department is to his left, and the state helps explain his walking to his left. Or here is the strategy put in another way: a *maximally decided* state would consist both a maximally opinionated belief as to the physical layout of the world and its causal relations, and a maximally detailed policy for agreeing or disagreeing with a person given the physical state of the world. We first explain how, in a maximally decided state, one counts as agreeing or disagreeing with an explanation couched in terms of mental content. We then explain what it is to accept such an explanation: it is to rule out being in any maximally decided state that doesn't count as agreeing with the explanation.

I have sketched an expressivistic theory of content ascription, but the sketch is extremely rough, and it raises many issues. Let me raise just one such issue, a qualm I think quite serious. The theory helps itself to the notion of agreeing or disagreeing with a person. When Henri says "Tiens! un lapin" and we disagree on how to interpret him, what's at issue is whether to agree with him. A scheme of interpretation amounts to a policy for agreeing or disagreeing with people on the basis of their naturalistically specified characteristics and dispositions.

Now a philosopher, it is often said, can speak with the vulgar in their own terms, while secretly disagreeing with them in his serious philosophical doctrines. Maureen, imagine, is an atheist who thinks in English. The language she hears around her sounds like an English pervaded with theistic claims. When a rabbit appears, everyone says, "The gods have sent a rabbit." Maureen interprets them as speaking English, and regards them as making mostly false claims. She could therefore object at every turn, and she'd be sincere—but that would make life terribly inconvenient. She therefore plays along: she treats them as if they were just saying, "Lo! a rabbit." and instead of saying, "Yes, here's a rabbit, but no, it wasn't sent by the gods," she replies, "Yea verily, the gods have sent a rabbit." For convenience, she treats his fellows and herself as if they were making

non-theistic claims—though she’s convinced that actually, their claims are theistic and systematically false. She’s an error theorist on this talk she’s surrounded by.

Compare Maureen to Richard, also an atheist who thinks in English. Richard takes the theistic sound of the seeming English he hears as formulaic. He takes their words “The gods have sent a rabbit” just to mean “Lo! A rabbit.” Richard and Maureen disagree systematically, then, in their attributions of meanings to utterances and thoughts to people. What, we now ask, is at issue between them? In what does their disagreement consist?

The expressivist looks for some difference in attitude or stance that constitutes their disagreement. In the respects I have focused on, though, Maureen and Richard react alike to the utterance “The gods have sent a rabbit.” They think to themselves that here’s a rabbit, with greater or less confidence as they think their fellows sincere and good at telling rabbits. They both respond as if with a stance of direct comprehension to “Here’s a rabbit” in their own language English. Both fit my initial description of treating their fellows as meaning, by “The gods have sent a rabbit,” just that here’s a rabbit.

What should we say, then, of expressivism for meaning and mental content? The qualm I have raised is only for expressivism in one particular kind of version. I started with the stance of direct comprehension. When this stance isn’t available, I proposed, meaning ascriptions substitute for it. They consist in responding as if one were directly comprehending different words from the ones that are spoken. This proposal, I have complained, fails to distinguish serious ascriptions from convenient playing along with the vulgar. I haven’t seen how to distinguish playing along with the vulgar as if they meant sensible things from a stance of direct comprehension that takes them really to mean sensible things.

### 3. Oughts of Belief

Meaning and mental content tie in with norms. If I mean PLUS by ‘+’, it doesn’t follow that I *will* accept ‘ $58 + 67 = 125$ ’, but it does follow that I should. It seems to be somehow built into the concepts involved that I should accept that  $58 + 67 = 125$ —or at least shouldn’t reject it. It would be *incorrect* to reject it, we can say. This suggests the slogan that “meaning and content are normative,” that the concepts of meaning and of mental content are normative concepts. This position indeed has moved some way toward being conventional wisdom among some philosophers.

Other philosophers argue, however, that these claims collapse on examination. After all, any concept whatsoever can figure in a normative statement. If it’s raining, you ought to take an umbrella, perhaps, but this won’t make RAIN a normative concept. Perhaps the tie of meaning PLUS to answering 125 is like the tie of rain to taking an umbrella. *Oughts*,

then, aren't built into the concept of meaning PLUS, any more than they are built into the concept of rain. It tends to be a good thing to have an umbrella when it's raining, but that's because an umbrella can keep you dry and, usually, you ought to keep dry. It's also a good thing not to say that  $58 + 67 = 5$ , but that too is a matter of consequences. It is because if you do things like answering 5, banks dishonor your checks, buildings you design fall down, and all sorts of other bad things tend to happen. Or perhaps it's a matter of intrinsic values: to accept an answer 5 would be to believe falsely, and it is desirable to believe the truth.

I myself don't think it is always desirable to believe the truth or shun believing falsehoods. There is sense of the term 'ought', though, I want to claim, in which you *ought* not to believe a falsehood. Perhaps you ought to *want* to believe it, but still, if it's false, you ought not to believe it. I'll try to elucidate this sense of 'ought' and make some distinctions.

Start with reasons, with reason to believe or disbelieve. We can ask whether there's reason to believe the theory of evolution by natural selection. We're wondering, then, about evidence for and against the theory. True enough, perhaps, disbelieving it would make you more effective in winning elections, but that is irrelevant to the question. We are asking about reason to believe or disbelieve the theory, and that's different from the advantage or desirability of believing it. Is the theory to be believed? Is it credible? Is belief in the theory warranted? That's a matter of whether there's sufficient reason all told to believe it. Now in one sense of the term, whether you "ought" to believe the theory is this question: whether it's to be believed. This sense ties in with reasons to believe it or not, as opposed to reasons to *want* to believe it. Perhaps, in this sense, you ought to believe the theory, given the evidence, but in light of the career advantages of disbelieving it, you ought to *want* to disbelieve it.

One diagnosis of these phenomena is this: there are purely epistemic senses of the terms 'reason' and 'ought'. For these senses, we consider the evidence alone, and take as the aims in play only believing the truth and not falsehoods. I think, however, that there is a better account to be had. One notion of reason to do something, I propose, is primitive. This notion is tied to how reasons combine and weigh together and against each other, and what we *ought* to do, in the sense I have in mind, is the resultant. These basic normative concepts, though, apply to a variety of kinds of doings broadly construed: to action, to belief, to desires or preferences, and to feelings or attitudes. Distinguish, then, what I have reason to *believe* and what I have reason to *want* to believe. In one sense of the term 'ought', what I have reason on balance to believe is what I ought to believe, given my evidence. It's what it "makes sense" for me to believe, what belief is "warranted" for me. Let's use the term 'ought' in this primitive, vanilla sense; we can now define other

normative notions in terms of it. What's *desirable*, we can say, is what one ought to desire. So a belief is desirable if one ought to desire to have it, to want to have it.

To illustrate, take the stock example of a man with indications that his wife is unfaithful. "Evidence" that she is unfaithful just *means*, we might say, reason to believe that she is, and what he ought to believe is a matter of how the reasons to believe stack up. So in this sense, that he ought to believe in accordance with the evidence is just analytic. That doesn't at all settle the question, though, of whether believing her unfaithful is desirable, of whether he ought to *want* to believe her unfaithful. That's a matter of how reasons to want to believe or not stack up. Perhaps, in this sense of the term 'ought', he ought, given his evidence, to think her unfaithful—but in light of the dire effects this belief will have on his life and hers, he ought, none the less, to *want* to think her faithful.

Ought I always to believe the truth? Believing the truth isn't always desirable—or if it is, that's an extreme claim about the importance of believing the truth, not something that falls out of the very concepts we are employing. That's what the stock fidelity example illustrates. But we are now considering the basic sense of 'ought', the sense tied to warrant and to reasons to believe or to do. So in this sense, ought I always to believe the truth? Two worries about this claim still need investigating. First, I have been using senses of 'ought' and 'reason to' that are tied to evidence, and evidence doesn't always support the truth. I have been using terms 'ought' and 'reason', we can say, in a *subjective* sense. Evidence can be misleading or absent. There's a truth about the number of partridge eggs in Ann Arbor this morning, but I have no reason to believe it. This truth is not something I ought, in any subjective sense, to believe. At most, I ought to believe in accordance with my evidence. Second, life is short, and most questions, it seems, I ought not to bother with. The number of partridge eggs in Ann Arbor is surely one of these.

Still, I think we can drive notions of ought, belief, and meaning or content closer together. Think of epistemic *oughts* as deliverances of a kind of planning for belief. I think what to believe, and I plan for various epistemic contingencies, thinking what to believe in circumstances that might arise, as news comes in. Now I can separate at least two kinds of consideration: questions of the costs of inquiry, and questions of epistemic policy ignoring cost. Why do this? Because plans for what to believe costs aside may be far more straightforward than questions of how to employ one's limited human resources for planning. Answers to planning questions that ignore costs may bear, in complex but important ways, on how to form beliefs when we factor in questions of costs of inquiry.

Clearly with costs factored in, we ought not to bother believing unimportant truths. But isn't that a matter of the costs of cluttering one's mind? Costs aside, perhaps, we ought to believe all truths and disbelieve all falsehoods. That is perfection in belief.

I have considered two objections to the claim that special oughts apply to belief, in a special way, that one ought, as such, to believe truths and disbelieve falsehoods. One is that we can understand the *oughts* that attach to meanings as instrumental, on the model of the *oughts* that attach to rain. I'm arguing that once we distinguish a primitive ought from an ought to want, reason to believe from reason to want to believe, we see that these primitive normative notions tie in with belief in a more intimate way than any normative notions tie in with rain. The other objection is that we ought not always to believe in accordance with logic and evidence, because it's burdensome and impracticable. The primitive ought, I say, is the one that ignores costs of thinking.

I have left another big question dangling, though. Distinguish "subjective" and "objective" oughts: What I ought to do in light of my evidence, and what I ought to do in light of all the facts, whether or not I have any way of knowing them. According to egoistic hedonists, I ought objectively to maximize my net pleasure. But I have no way of knowing how to do this, and so subjectively, what I ought to do is to maximize my prospects for pleasure, the subjectively expected value of pleasure. What is it to take my own net pleasure as my goal, when I can't know for sure what will promote it? Presumably, it is to respond to evidence in certain ways, somehow maximizing my prospects for pleasure. Roughly, in terms of subjective oughts, if I ought to believe, given my evidence, that an act will maximize my net pleasure, then that's what I ought to do.

What is it, then, to take true belief (and avoiding false belief) as my objective goal. From the pattern for seeking a goal, we get this: If I ought to believe that believing that snow is white would give me true belief, then I ought to believe that snow is white. But that amounts to the empty claim, if I ought to believe that snow is white, then that's what I ought to believe.

The injunction to seek the truth, I think we can argue, does have some bite. It does require formal epistemic coherence, that one's subjective probabilities satisfy the axioms of probability theory. And that will be enough to rule out such things as believing one's spouse faithful because the belief is comforting. But all this is a long story. The upshot, if I am right, is this: There are norms that govern belief. But these norms cash out as subjective norms of coherence and reason. Correctness is the standard of belief, in that true beliefs are the ones we ought to have, in the objective and cost-ignoring sense of ought that I have been sketching. But that amounts to a set of subjective oughts: that our credences ought to be formally coherent and reasonable.

#### 4. Subjective Oughts and Logic

One ought not to believe both that snow is white and that nothing is white. This seems clearly correct for an ought that is subjective and cost-ignoring. It might, in wild

circumstances, be desirable to violate this ought, but the ought in question is not one of desirability. The ought primitively applies to belief, not to wanting to believe.

Here, then, is how a “normativity of content” claim might now go: Henry believes that nothing is white, suppose. What does this mean? It is built into the very meaning of this claim, we can say, that Henry is in a state that figures in certain normative constraints. One of these constraints applies jointly to this state and another: that one ought not both to believe that nothing is white and believe that snow is white. Logical relations of inconsistency and the like characterize what each item of content is. These, the proposal is, cash out as normative constraints on beliefs.

Such a proposal is analogous to a functional role semantics as a theory of mental content. But the ties between items of belief aren’t causal ties, the claim is; they aren’t tendencies to go from one belief to another. They are normative ties. To believe that nothing is white just *is* to be in a state with certain *ought* relations to other states you could be in, such as believing that snow is white.

Such a program raises many questions, and I won’t be able to get far into any of them. One question concerns explanations that invoke mental content: how do a general’s plans explain the movements of tanks? If claims about mental content are fraught with ought and we explain ought-thoughts as deliverances of planning, how can oughts explain movements? How can what to believe and intend explain the movements of machines and buildings turned to rubble. I gestured earlier at a complex kind of account I would offer of such explanations—but instead of elaborating, I pass on to worries I find less tractable.

If I’m right in what I have sketched, there is indeed a normative tie between believing that snow is white and believing that nothing is white, a tie of incompatibility. It isn’t a tie that is exhausted by considerations of desirability of the kinds that tie umbrellas to rain. The hypothesis we’re examining is that these normative ties are built into the very meaning of content claims. It’s part of the meanings of the two claims, “Jack believes that snow is white” and “Jack believes that nothing is white” that those are states he shouldn’t jointly be in.

How do we settle, though, that oughts are part of the very *meaning* of claims of mental content? Partly, of course, by eliminating contrary explanations of the oughts involved. It isn’t, I’ve argued, like rain and umbrellas, and it isn’t just the intrinsic desirability of believing truths. There is, however, another kind of explanation to consider. Justification comes to an end. An ought can be grounded in other oughts, a reason in other reasons, but eventually reasons come to an end. For actions, reasons of suffering seem to stand on their own, with a force that is perhaps not grounded in further reasons. An alternative to the claim that certain prohibitions hold just in virtue of the meanings of belief claims is this: Perhaps these prohibitions are substantive and basic.

Why mustn't I accept both that nothing is white and that snow is white? "This sort of thing is just what it *means* to have NOTHING beliefs"—that's the answer we're experimenting with. The alternative answer is: "This is the point at which oughts and reasons come to an end. Clearly one mustn't do such a thing, and there's no more to be said." Here we reach the problem that Quine made central: How shall we distinguish truths of meaning from substantive truisms?

To be sure, anyone competent in the use of English term 'nothing' implicitly accepts requirements like this. She tends away from states that violate this requirement, and tends to registers shock or bafflement when this requirement is violated. It does seem, then, that we won't credit you as meaning NOTHING by the term 'nothing' unless you implicitly recognize requirements like these. How, though, do we go from this to the claim that oughts are built into the meaning of 'believes that nothing is white'?

In some cases, *oughts* that are credible on their own clearly aren't *oughts* that define a concept, and their violation doesn't impugn one's conceptual competence. Consider the *oughts* of action, of what it makes sense to do. If an egoistic hedonist—call her Hedda—is right, one ought to pursue one's own pleasure, for no further reason. This is an *ought* of independent credibility, she says, at which justification comes to an end. A perfectionist Percy, though, can deny this and still have full command of the concepts involved. Hedda can't respond, "I just don't know what you mean." She knows what he means; their disagreement is genuine. It is, I would claim, disagreement on how to live, on what to pursue in life.

My worry is whether the norms that govern the concept NOTHING have the same status. On the proposal I've been trying out, the claim that

By 'rien' Jacques means NOTHING (1)

is a normative claim; MEANING such-and-such is a normative concept. And indeed we have succeeded in finding a norm that applies if one means NOTHING, that one ought not to think that snow is white and that nothing is white. But is this norm tied to the concept of MEANING such-and-such in a way that is purely conceptual? Or if two observers disagreed whether this norm applied to Jacques, would their disagreement be genuine and substantive—a disagreement on how to think, on how to reason with the concept NOTHING?

It's a little hard to invent a plausible dispute involving the concept NOTHING, so as to ask what's at issue, and so let me shift to the concept OR. Classical logic endorses the pattern,  $P \vee Q, \neg Q, \therefore P$ . This is tied to various normative patterns, such as this pattern of permissions: If permissibly you accept both  $P \vee Q$  and  $\neg Q$ , then it is permissible to accept  $P$ . Does such a normative pattern help define the concept OR? The pattern is

controversial; intuitionists reject it. We might interpret a rejection, though, in either of two ways: One is that this normative pattern indeed helps define the classical concept, and the critic is maintaining that we should drop the classical concept. This is the “meaning is normative” position. The alternative is that classical logicians and the critic share the concept OR, and disagree how to reason using it. Their disagreement is substantive; label this interpretation of the dispute “substantialist”.

If we could help ourselves to talk of agreeing and disagreeing, we could distinguish these alternatives clearly enough. Claudia accepts classical logic, suppose, and Ian rejects it. Claudia thinks  $P \vee Q, \neg Q, \therefore P$ . On the second, substantialist alternative, Ian the critic can agree that  $P \vee Q$  and that  $\neg Q$ , and think it permissible to accept these things. He, though, unlike Claudia, refuses to accept that  $P$ , and thinks it impermissible for Claudia to accept it. He shares a concept with Claudia but disagrees how to deploy it. On the first, “meaning is normative” alternative, Ian neither agrees nor disagrees with Claudia’s claim ‘ $P \vee Q$ .’ He doesn’t share the concept that she expresses with the symbol ‘ $\vee$ ’, and regards it as defective; thus he can’t straightforwardly agree or disagree with her when she deploys the concept. They don’t disagree about how to think with Claudia’s concept; they disagree about *whether* to think with Claudia’s concept; Ian endorses a different concept.

Agreement and disagreement seem to tie in with attributions of content, inescapably and perhaps primitively. If I accept that snow is white and attribute to you the thought that snow is white, then I agree with you; otherwise I’m incoherent. If we could identify stances of agreeing or disagreeing, I proposed earlier, we could use that to construct an expressivistic account of the meaning of mental content ascriptions. My problem was that I couldn’t independently characterize such a response.

With norms, the tie to content is more elusive. If two people mean OR by ‘or’, then the same logical norms apply to both. Claudia and Ian agree on this dictum. The norms Claudia attributes to her own use of ‘or’ differ from the norms Ian attributes to his use of ‘or’. Does it follow that what she thinks she means by ‘or’ differs from what he thinks he means by ‘or’? If what she means by ‘means OR’ has the norms she accepts for ‘or’ built into it, and Ian rejects these norms, then he is committed to denying her claim that she “means OR by ‘or’”. But is he committed to denying this?

I leave this as a problem for the reader and for myself.