[In *The Philosophy & Theory of Artificial Intelligence* (Ed.) Vincent C. Müller, New York: Springer, 2012, pp. 312-333.]

NOTE: Special Draft of Penultimate Version with images not included in published version.

Machine Intentionality, the Moral Status of Machines, and the Composition Problem

David Leech Anderson (Illinois State University)

ABSTRACT: According to the most popular theories of intentionality, a family of theories we will refer to as "functional intentionality," a machine can have genuine intentional states so long as it has functionally characterizable mental states that are causally hooked up to the world in the right way. This paper considers a detailed description of a robot that seems to meet the conditions of functional intentionality, but which falls victim to what I call "the composition problem." One obvious way to escape the problem (arguably, the only way) is if the robot can be shown to be a moral patient – to deserve a particular moral status. If so, it isn't clear how functional intentionality could remain plausible (something like "phenomenal intentionality" would be required). Finally, while it would have seemed that a reasonable strategy for establishing the moral status of intelligent machines would be to demonstrate that the machine possessed genuine intentionality, the composition argument suggests that the order of precedence is reversed: The machine must first be shown to possess a particular moral status before it is a candidate for having genuine intentionality.

In this paper, an answer will be sought to the question:

> "What properties must a machine have if it is to have genuine beliefs about objects in the world?"

It would seem to be an altogether different question – about the foundations of moral personhood rather than about intentionality – to ask:

> "What properties must a machine have if it is to be an object of moral concern (if I am to have any obligations with respect to it)?"

Most believe that an answer to the first question does not engage issues of morality or moral agency. The common assumption is that we must first determine what intentionality is, and whether non-biological machines are capable of possessing it, before we ask what, if any, our moral obligations might be with respect to such systems. I will argue that the order of conceptual priority is reversed. The second question must be answered *before* the

first. An argument in defense of this claim will support the theory that the very nature of intentionality must be grounded in the moral status of the cognitive agent.

1.

Very few philosophers are in the market for a fundamentally new theory of intentionality. There are some areas within the philosophy of mind where there is no overwhelming consensus and it seems that every week there is a new theory gaining a hearing. *Phenomenal consciousness* is one such area. It is a difficult nut to crack and many philosophers and scientists not only doubt that any current theory has solved the "hard problem" of consciousness, many are less than sanguine about the prospects of ever resolving it satisfactorily. The same cannot be said for intentionality. Intentionality is one of the cognitive domains where there is a dominant theory which many consider unassailable. I will refer to the dominant theory as "functional intentionality" (borrowing from Jaegwon Kim). This is a broad view that has room for a variety of different versions. There continues to be a vigorous debate about precisely what form the final, ideal version of this theory should take, but a majority of philosophers and an overwhelming majority of cognitive scientists are confident that intentionality can ultimately be *functionalized.* Kim expresses this confidence about the functionalizing of intentionality in an article where he also admits his skepticism about functionalizing phenomenal consciousness. Although willing to flirt with epiphenomenalism – which would require abandoning a physicalist reduction of phenomenal consciousness – he has in no way lost confidence in a physicalist reduction of intentionality. He says:

> In short, if a group of creatures are behaviorally indistinguishable from us, we cannot withhold from them the capacity for intentional states . . Intentional states, therefore, supervene on behavior. . .
>
> . . intentional states are functional states specified by their job descriptions. (Kim, J. 2007, p. 415).

I offer the following not as a definition, but as one popular example of how the theory might be characterized.

> *Functional Intentionality*: A complex system will be in the intentional state of "believing that p" so long as some state of the system serves as a representation "that p" and is playing the belief-role in the overall economy of the system. The

> content will be fixed in information-theoretic terms by the fact that the state "carries the information that p."

There is no question that functional intentionality dominates the field. Yet there is a minority voice within the intentionality debate that has only gained traction within the past decade. This is the view I will call *phenomenal intentionality*.[1] Very simply, a theory will qualify as a version of phenomenal intentionality if it makes phenomenal consciousness, either actual or potential, a necessary condition for genuine, underived intentionality.

> *Phenomenal Intentionality*: A complex system will be in the intentional state of "believing that p" if and only if that system is (actually or possibly) phenomenally conscious "that p."

I find the disagreement between the functional intentionality camp and the phenomenal intentionality camp to be one of the most engaging within the study of cognition because it is a watershed issue where it may finally be decided whether phenomenal consciousness is merely an inconsequential byproduct of contingent forces on this planet and ultimately inessential to the appearance of genuine cognition or whether it is instead an essential characteristic of intentionality, a prerequisite for mental states bearing genuine cognitive content. The arguments of this paper do not directly confront this question. However, if successful, these arguments will introduce a new dimension to this dispute which could determine which side ultimately prevails.

2.

If the promise of functional intentionality is to be fulfilled in a non-revisionist way (which rules out eliminativism), it must, at the very least, be able to specify, solely in physical terms (including structural and functional terms) what we know to be the proper content of the complete range of contentful mental states. On the one hand, popular versions of this theory make the strongest case in their favor by offering physically-kosher content-fixing properties that seem to include within their extension, those objects which we in fact judge to be within the wide content of the relevant mental states. On the other hand, most defenders and detractors alike would admit that the real challenge for these theories is not

[1] Two excellent surveys of arguments advanced in defense of phenomenal intentionality are Horgan and Kriegel (forthcoming) and Siewert (2009).

inclusion of the right objects or properties, but exclusion of the wrong ones. The vulnerability of these theories is that there are simply too many physical objects (or properties) that are candidates for doing the reductive work. This is usually characterized as an "indeterminism" problem of one kind or another. The worry is that the reductive properties lack the resources to draw a (non-arbitrary) line between the semantically irrelevant objects or properties and the ones we know (pre-theoretically) to be the correct ones. The two most important domains where the threat of indeterminism arises are the fixing of cognitive content & identification of cognitive agents.

The most discussed type of indeterminism is *content indeterminism.* This was originally raised not by detractors but supporters of the behaviorist branch of the functional intentionality family – the first branch to really flourish. The charge was raised in 1960 by W.V.O. Quine (1960) one of the most famous and earliest defenders of functional intentionality. Quine argued that linguistic behavior alone left it ambiguous whether the meaning of a particular natural language term (*gavagai*) was justifiably translated by the concept "rabbit" as opposed to the extensionally equivalent concepts of "undetached rabbit part" or "manifestation of rabbithood" (Quine 1960, 152-154). This poses the threat of *content* indeterminism.

> *Content Indeterminism:* When a theory of intentionality specifies multiple and incompatible assignments of content to one and the same state of a complex system.

Quine did not flinch in the face of what many consider an intolerable kind of indeterminism. He used it as grounds for eliminating "meaning," traditionally conceived, from his ontology, replacing it with the more ontologically parsimonious "disposition to *verbal behavior.*" He often reveled in the fact that his was an eliminativist position more than a reductionist one. Conclusions he drew from this view include the "indeterminacy of translation" and the "inscrutability of reference" and, it even played a contributing role in his defense of "ontological relativity."

A second, closely related, kind of indeterminism I will call *agent indeterminism* and is defined at the level of cognitive agents rather than token mental states. Again, this issue is most powerfully raised *not* by an opponent of the view but by one of its most famous

contemporary defenders: Daniel Dennett. Exhibiting his dry humor, Dennett asks us how we can tell the "true believers" from those that lack genuine intentionality. Dennett says you can't – not at a metaphysical level. Of course, he acknowledges that there are pragmatic grounds – relative to particular interests – to justify our taking the "intentional stance" with respect to some system; he simply denies that there are any non-relative metaphysical grounds for such a distinction. Or, what comes to the same thing, he is an instrumentalist when it comes to the interpretation of sentences that seem to express an ontological commitment to genuine intentional agents. He made the eliminativist implications of this theory explicit when he denied that there exists such a thing as intrinsic (non-derivied) intentionality. It isn't merely the thermostat's contentful states that are derivative, ours are equally so (Dennett 1987, Chapter 8).

We might call the Dennett version, *global agent indeterminism*, since the indeterminacy afflicts the entire range of potential candidates that might qualify for the label, "intentional agent." There is a related form of the same phenomenon, less familiar, that afflicts a single complex system. Consider any complex system, C, that is a potential cognitive agent and that consists of multiple subsystems, $S_1 – S_N$. Then, consider another mechanism, M, that bears causal connections to C. A theory of intentionality must have the resources to determine whether M is indeed an object external to C, or is instead another constitutive element of C, properly classified as, S_{N+1}. Agent indeterminism arises when a theory of intentionality lacks the resources to specify whether M is distinct from C, or is partly constitutive of C.

> *Agent Identity Indeterminism:* When a theory of intentionality lacks the resources to specify which causally connected elements are constitutive of the agent and which are external to it.

Most adherents of functional intentionality have considerably less tolerance than do Quine and Dennett for a level of indeterminacy that undermines the very ideas of intentional content and intentional agency. Defenders of functional intentionality, then, dedicate considerable energy to an explanation of how the resources of functional intentionality can, indeed, deliver determinate cognitive content and identify determinate conditions for cognitive agency.

This paper will raise a pair of test-cases that: (1) put functional intentionality's indeterminism problems in the severest light (with a single case that manifests both *content indeterminism* and *agent identity indeterminism*); and, (2) justify the claim that judgments about intentionality, if they are to escape debilitating indeterminism, must appeal to the moral status of the cognitive agent.

3.

As an alternative to functional intentionality, I offer the following theory:

> *Moral Status Intentionality*: The content of token mental states and the identity conditions of intentional agents are grounded in those properties by virtue of which a complex system is an object of moral concern (i.e., at least has the status of a *moral patient*).

By "moral status" I mean to invoke a concept familiar in moral philosophy. While there are undoubtedly many subtle gradations of moral status that carry ethical import, only the two most commonly used classifications are necessary to understand the theory.

> *Moral Patient*: An entity that deserves to be an object of moral concern
>
> *Moral Agent*: An entity that (1) deserves to be an object of moral concern, (2) has the capacity to make moral judgments, and therefore (3) has the right of self-determination

A dog is a moral patient. It has a privileged status which requires me to take its interests into account as well as my own before I act. I have even more obligations with respect to human beings who qualify as moral agents. I am morally permitted to lock my dog in the house, but I am not permitted to lock *you* in my house. Dogs lack the cognitive abilities that would otherwise earn them the right of self-determination.

According to moral status intentionality, if a complex system is to have genuine, intrinsic intentionality (unlike a thermostat that has only derived intentionality) it needs to achieve the status of being a moral patient. A thermostat "carries the information that *p*" (for *p*: "The room temperature is 70 degrees Fahrenheit.") but it does not "believe that *p*." I will argue that only a moral patient can have beliefs. When a dog walks to its supper dish to eat, it is possible that it genuinely *believes* that *p,* for some *p* roughly like, "Bowl has food" (but substituting more primitive, doggy-concepts for our concepts of 'bowl' and 'food'). On this theory, dogs are perfectly reasonable candidates for genuine intentionality.

Obviously enough, the appeal to *moral status* needs to be grounded in a moral theory. Any theory will do, but I will begin by assuming *utilitarianism* both because it is the theory most widely held by defenders of functional intentionality and because it is the most straightforward. According to utilitarianism, acts are morally evaluated according to the overall balance of happiness and suffering they cause in objects of moral concern. Assuming utilitarianism, moral status will be grounded in the morally evaluable conscious states of pain, pleasure, happiness, sadness, etc. (Moral status can also be grounded in a Kantian, deontological theory which takes respect for rational agents as a fundamental principle. Space does not permit a detailed examination of how *moral status intentionality* would be conceived on a deontological model.)

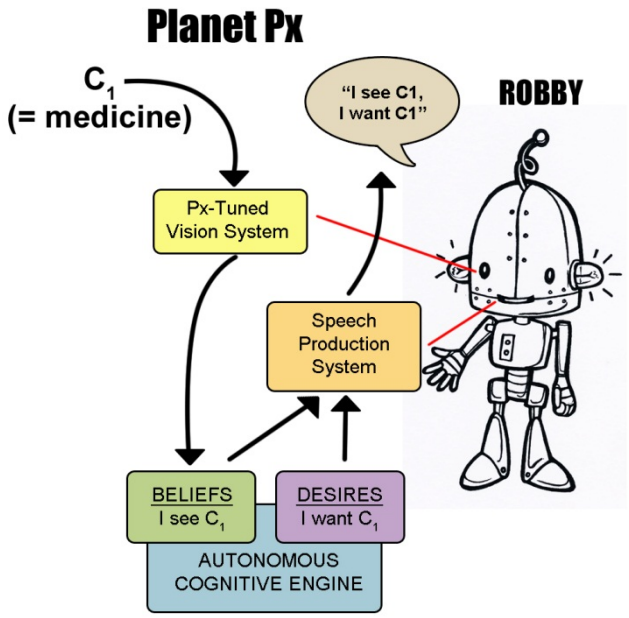
4.

With the important background concepts now in place, it is time to introduce a story about a robot in the year 2050. The only assumptions that I will make are that the robot described (below) meets the conditions required by functional intentionality and that the robot lacks phenomenal consciousness. If you believe the robot's behavior (as described below) is insufficiently complex to meet the first condition, then simply re-imagine the robot with additional capacities. Nothing in the argument hinges on details of that sort. If you believe there is no possible world in which a robot has behavior sufficiently complex to meet the conditions of functional intentionality and yet lacks phenomenal consciousness, then you are in that minority group for whom the following argument will be a non-starter. But you *are* in the minority. Many defenders of functional intentionality think it is a strength of the theory that it has room for intentional systems without phenomenal consciousness. It is this view against which the argument is directed.

CASE #1: Robby the Robot: It is the year 2050. John is the sole support of an extended family of 25 people. He has worked his entire adult life and has exhausted his family fortune to build a complex robot ("Robby") that will travel in space to visit a planet, Px, in a distant galaxy. The robot is remarkably autonomous and has a built-in operational parameter (a "desire") to find and gather a scarce chemical compound, C_1, that has

remarkable medicinal properties and that is known to exist nowhere else in the universe. When Robby locates a cache of C_1 Robby says things like:

> "There (pointing to the ground) are approximately 4 kilograms of C_1 between .5 and .75 meters below the surface of the ground. Should I extract it?"

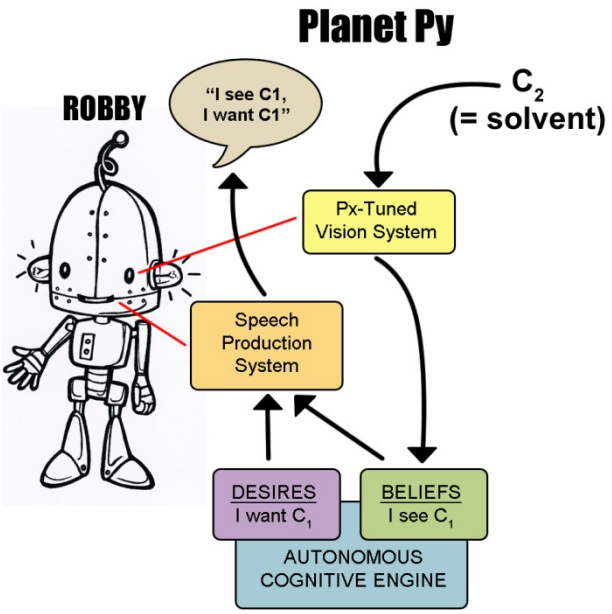


Building Robby to successfully perform this task is complicated by the conditions on the planet's surface. The environment on Px is wildly different than on earth because it has a dense, inhospitable atmosphere filled with all manner of gases, unrelenting electrical storms, and rhythmic gravitational fluctuations. These conditions are so extreme that they materially affect the operations of standard sensory devices. Devices that would be effective for locating and extracting C_1 on earth, would be wholly ineffective on Px. Robby has been equipped with special sensory systems that compensate for the distorting effects of the wild forces at work on the surface of Px and that give Robby the capacity to reliably identify and extract C_1 in Px's strange environment – which he did for several years.

In time, though, the bottom dropped out of the C_1 market and Robby's work on Px was no longer profitable. John's financial situation was dire. He needed a new source of income or his family of 25 was threatened with homelessness. Happily, it turned out that there was another planet, Py, filled with a powerful and equally valuable industrial solvent, C_2. Py also had an epistemically inhospitable environment. Earth-calibrated sensors seriously malfunctioned on Py, but in very different ways than they did on Px. So Robby's sensory apparatus were as useless on Py as on earth. Until, that is, John discovered that, by an accident of cosmic good fortune, Robby's C_1-detector-in-Px could, with some accommodation, function as a C_2-detector-in-Py. The end result is that when Robby is on Py he will speak the sentence

"There (pointing to the ground) are approximately 4 kilograms of C_1 between .5 and .75 meters below the surface of the ground. Should I extract it?"

only when there are 40 pounds of C_2 (the valuable solvent) between 50 and 75 feet below the surface of the ground. Robby is functioning, on Py, exactly the way John wants him to function except for the one inconvenience that many of the words that Robby utters must be re-interpreted if they are to have meanings which result in Robby's utterances actually being *true.*



John discovers an easy way to solve Robby's false-utterance problem. Radio Shack has a $3 language translator module that John installs so as to intercept signals sent from Robby's linguistic processing center to the voice-box (where it is articulated). The translator module performs a translation function on all sentences with words like 'C_1,' 'meters,' and 'kilograms.' It also does a transformation on numerals, using one function when the numerals are syntactically tied to weights and another function for when they are syntactically tied to distances.

A further benefit of this simple change is that John need not tinker with Robby's so-called, "beliefs" and "desires." Robby's central cognitive system still outputs the sentence

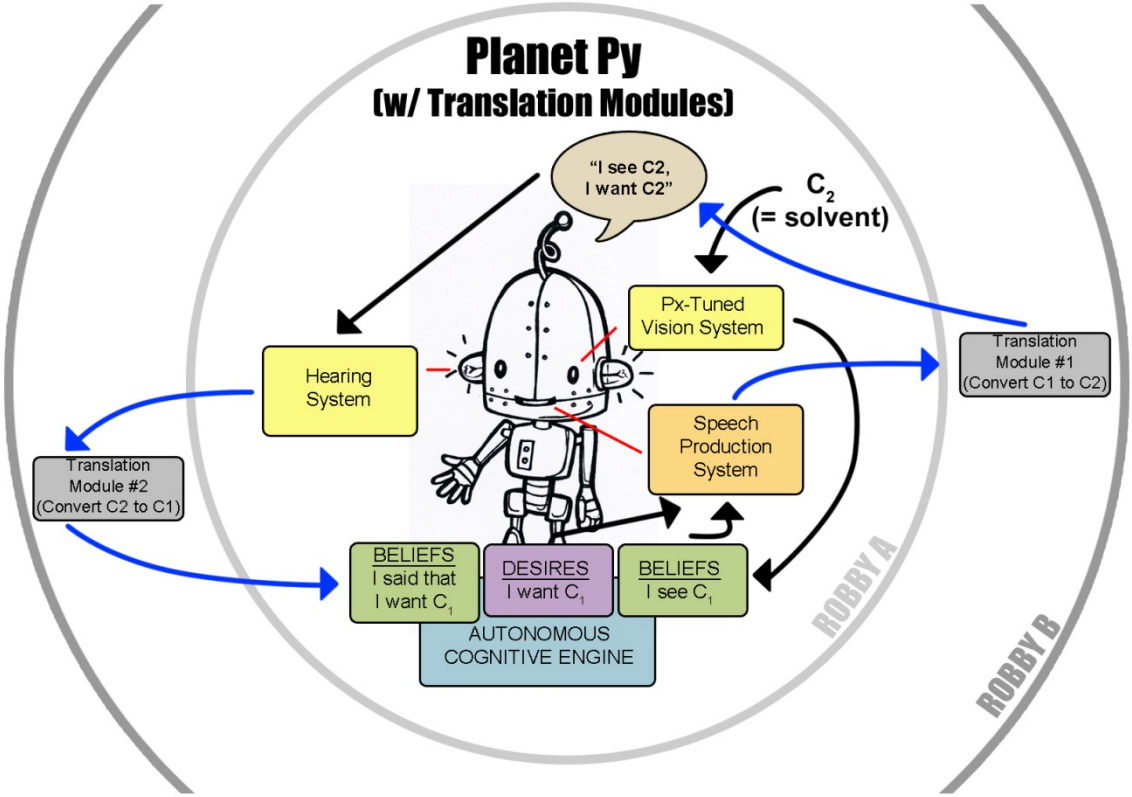
S_1: "I want to extract C_1 with the goal of providing the world with an effective medicine."

But that is not what anyone hears. What Robby is heard to say is:

S_2: "I want to extract C_2 with the goal of providing the world with an effective solvent."

To prevent a "cognitive" dissonance that might diminish Robby's performance, John buys a second translation device (they're cheap) and inserts it between Robby's auditory sensory devices (i.e., microphones) and the linguistic processing center so that all references to C_2 are re-converted back to references to C_1. The result is that what Robby now actually

"hears" (or should we say what Robby "thinks" he hears) is what he "thinks" he said, which is S_1.



The result is that Robby's autonomous cognitive engine (which includes his "beliefs" and "desires") outputs sentences like "I love mining C_1." What we all hear Robby say is "I love mining C_2" but what Robby hears himself say is "I love mining C_1."

As viewed from the outside, Robby is now effectively receiving information about Py and acting intelligently upon that information by extracting C_2. His verbal behavior now coheres with his practice. John describes what he did when he added the two translation modules as *repairing a defect in Robby's word-world coupling module.* And Kim would be bound to agree since, "intentional states . . supervene on behavior" and given Robby's current behavior, Robby has C_2 not C_1 intentions.

5.

CASE #2: Sally my neighbor. Consider now a second case, in many respects similar to the first. I have a neighbor named "Sally." I kidnap Sally and whisk her off to an alien planet where her sensory and cognitive faculties completely malfunction, but due to a

bizarre coincidence, they malfunction in a way that is isomorphic to proper cognitive functioning. When placed on this alien planet, her *sensory experience* makes her (falsely) believe that she is still in an environment very similar to her old home on planet earth and to (falsely) believe that she is *doing* the thing she most likes to do – raising flowers. What she is actually doing is mining a valuable compound, C_2, which will make me a great deal of money. All I need to do to exploit this situation over the long run is to secretly implant two translation devices in her brain and . . . . voila. Sally now has genuine C_2-desires and I have reason to claim that I have *successfully repaired a defect in her word-world coupling module*.

Notice that in both the Robby and the Sally cases, functional intentionality has implications not only for the interpretation of the intentional content, but for the very identity conditions for "being Robby" and "being Sally." In the Robby Case, we get what I will call "the composition problem." There seem to be no, non-arbitrary means of determining whether Robby's two implanted translation modules are constitutive sub-systems of Robby himself, or external devices distinct from Robby. If that question has no determinate answer, then questions about intentional content will also remain hopelessly indeterminate.

Raising "the composition problem" in the Sally Case proves morally offensive. If the implantation of the translation modules were, indeed, a way of repairing a defect in Sally's word-world coupling module, then those translation modules were *not* external devices of manipulation that kept her imprisoned within a constrictive, epistemic cocoon. Instead, they were an integral part of her identity, quite literally constitutive of what it *means* to "be" Sally.

The results that functional intentionality give in the Sally case are quite outrageous. If this is indeed, what the theory prescribes, then the Sally case is a refutation of the theory. While there are undoubtedly strategies that the defender of functional intentionality will try, the fact that they have to give a plausible outcome in both the Robby and the Sally cases, limits the kind of rebuttals that will be possible. In contrast to the challenge that these cases pose to functional intentionality, moral status intentionality, produces a very

plausible result. It says that intentional content is grounded in those states of Sally that are morally evaluable – in particular, her phenomenally conscious states.

It is absurd (and morally reprehensible) to say that she no longer has a desire to raise plants but now has an authentic desire to mine a chemical compound. According to moral status intentionality, Sally could be on the alien planet for 50 years and at the point at which she discovers her situation and screams: "You have wronged me. I thought I was raising plants all these years and I was *not*!" she will not be uttering a trivially false statement about C_2 (remember 'plant' means 'C_2' according to functional intentionality) she will be uttering a true and morally damning judgment of the suffering I caused her by thwarting her authentic intentions.[2] It *matters to her* that we properly interpret her utterances and her mental states. And it is *only* because they matter to her (since she is a moral patient) and she matters to us (since we have obligations to her), that we are able to identify the proper and determinate content of her utterances and that we have grounds for saying that the dual translation modules are, indeed, external devices causing moral harm.

6.

A defender of functional intentionality might attempt to resist the foregoing argument by appeal to an externalist theory of reference-fixing. One of the perceived strengths of the theory is its ability to account for reference-shifting when intentional agents move from one environment to another. Might not an externalist theory of reference eliminate the indeterminism problems just raised? I will seek to show that the Robby argument is not vulnerable to any commonly accepted features of externalist reference-fixing. Furthermore, the externalist account of reference-fixing itself succumbs to *precisely* the same type of counterexamples that we have seen in the Robby and Sally cases. The situation requires explanation.

---

[2] As I have claimed here, Sally's moral utterance condemning my vicious betrayal – even if she has lived on the alien planet for 50 years – demands that we interpret her term 'plant' as referring to biological plants back on earth and *not* C_2 on the alien planet. This is compatible, however, with the view that there may well be other utterances of Sally's, also using the linguistic token, 'plant,' that might indeed refer to C_2. This does not lead to contradiction, but in fact solves many semantic dilemmas, if one embraces the view that I have labeled *semantic dualism.* That theory is explained and defended in (Anderson 1995 & 2002).

Consider Putnam's famous case about Oscar on Earth and Twin Oscar on Twin earth (Putnam 1975). Twin Earth differs from earth only in this regard: Every place you find water (H_20) on Earth you find twin-water (XYZ) on Twin Earth, but the two liquids are qualitatively indistinguishable. Assume that Oscar is transported to Twin Earth without his knowledge (analogous to Robby's move from Px to Py). Since Twin-water is indistinguishable from water, Oscar will continue to use the English term, 'water,' to refer to it. But what does the word, 'water,' *mean* when he points to a lake filled with XYZ and says: "That is beautiful water"? The standard position is that on the first day that Oscar arrives on Twin Earth 'water' still refers to H_20, and so his utterance is false. But if he stays on Twin Earth long enough, the reference will eventually shift (just as the meaning of 'Madagascar' shifted over time in our own language). Words refer to whatever it is that causally regulates their use, and if that causal connection shifts (for long enough) so too will the extension of the term.

Let's consider how this account of reference-fixing might alter the Robby case. When Robby first arrives on Py, C_1-utterances will continue to refer back to C_1 on Px, until some length of time, *t,* after which Robby's tenure on Py will have been long enough to shift the reference from C_1 on Px to C_2 on Py. Of course, after *t*, it is not simply the meaning of Robby's words that have shifted, it is also the content of his propositional attitudes. All of his hopes, dreams, and desires that used to be about C_1 are now about C_2. And, of course, all the same things apply to the Sally Case.

At this point the reader might be wondering why the author went to so much trouble to generate the overly complicated story of Robby and the dual translation modules. Doesn't the direct theory of reference generate the same phenomenon by itself? If I wanted to "get away with murder," as it were, by deceiving Sally so that she would become my "unwitting slave," all I needed to do would be to get Sally to the alien planet long enough so that the extension of all of her terms shifted. At that point, her word, 'plant,' would – through no work of my own – come to refer to C_2 and "voila" she will have instantly gained a new, genuine desire to mine C_2 replacing her desire to raise flowers. And if she has a genuine desire to mine C_2 then I have instantly been redeemed from being a

heartless monster, and am now an innocent bystander, receiving her "gifts" of C_2 freely given, no longer coerced without her consent. (The reader should also note that the fact that Sally is no longer being deceived even though she has no first-person awareness that she is sending a chemical compound rather than flowers is *precisely* the threat to "self-knowledge" that has plagued externalist theories of reference and that has generated literally hundreds of articles and books on the topic.)

What should now be clear is that what plagues functional theories of intentionality also plagues externalist theories of reference. If all that was needed was a counterexample to functional intentionality, it would have been sufficient to offer straight, non-translation-module versions of both the Sally and the Robby cases. Functional intentionality gets them both wrong, even *without* translation modules. Functional intentionality says that after being on the alien planet for some period of time, *t* (a week, a month, 6 months?), Sally's mental representations of and linguistic references to 'plants' shift and their intentional content becomes C_2 (on the alien planet), rather than biological plants (on earth). Robby's content shifts similarly from C_1 to C_2.

Moral status intentionality gives a different result. In Sally's case, intentional content is determined by the phenomenally conscious states that ground Sally's status as a moral patient/agent. If at some future point, her epistemic situation was no longer defective, and she became (phenomenally) aware of the truth about the environment she inhabited, it would be obvious to everyone (regardless of what theory of intentionality you officially profess), that the sentences of outrage that she would utter at that point using the word, 'plant,' would *not* refer to C_2.

Likewise, there is a problem with functional intentionality's interpretation of Robby. There is also no time, *t*, at which his intentional content shifts ineluctably from C_1 to C_2. As argued above, there exists no no-arbitrary grounds upon which one could assign *any* duration to *t.* It is perfectly reasonable for John to interpret Robby's C_1-utterances as meaning, C_2, the moment Robby steps on planet Py. The only kind of content that Robby's utterances carry is *derived* content based on the assignments that John has an interest in making. Externalist theories of reference-fixing (absent the contributions of moral status

and/or phenomenal intentionality) lack the resources to even articulate a theory of reference-shifting, because there are no resources available to assign any *determinate* value to *t*. Whatever intuitions we have about a reasonable length for *t,* they clearly derive from our judgments about human semantic interests (in cases like, 'Madagascar') that already smuggle in the utilitarian harm and benefit that come from different assignments of content. The indeterminism about the length of *t* is just one more species of content indeterminism.

If the important work could have been done without them, why does the primary argument of this paper rely on the seemingly gratuitous features of Robby (and Sally) with two implanted translation modules? Very simple. Discussions of *content* indeterminism are legendary and ubiquitous. Discussions of *agent* indeterminism are considerably less so. I believe that is primarily a consequence of the fact that Dennett is one of the few who raises the issue of agent indeterminism and does it in the form of *global* agent indeterminism which leads to the outright rejection of non-derived intentionality. Most philosophers respond by saying that (1) intentionality obviously exists (so we must reject Dennett's arguments), and (2) it is just one more "line-drawing" problem and everyone has line-drawing problems so there is no reason to think it has to be fatal.

My goal in introducing the "Robby-plus-dual-translation-modules version" of the argument is to confront defenders of non-eliminativist functional intentionality with "the composition problem" – a problem of *agent identity* indeterminism that is not so easily dismissed.

7.

And now for two concluding remarks. First, if you want to build a robot with genuine intentionality, then build one that has a moral status – create one that is an object of moral concern. I have grounded moral concern in phenomenal consciousness, because I cannot conceive of anything else that might be the source of intrinsic moral value. If there is nothing else, then moral status intentionality has the same extension as phenomenal intentionality (at least in the actual world). But I could be wrong. There might be some other property that has intrinsic value (and/or disvalue) – call it "v-ness." If there is, then

phenomenal consciousness is not the only property that could determine that a thought had C_1 rather than C_2 content and phenomenal consciousness will not be a necessary condition for intentionality. (Maybe it is even possible that there be a conscious creature whose phenomenal states have no intrinsic moral status because they have no valence: being wholly neutral with no positive or negative value. In that case, phenomenal consciousness would not even be *sufficient* for disambiguating the C_1 vs. C_2 indeterminism. Maybe this is possible – I'm undecided.)

And finally, I would like to suggest that there is one more way that moral status might have the last word even if it loses the theoretical battle. I believe that there is no escaping the influence of moral status even if it turns out that functional intentionality is true and moral status intentionality is false. In an attempt to make the issue vivid, let me concede everything in the theoretical realm to the functionalist. Your theory is true; mine is false. I was wrong in thinking that there are only two fundamental categories of being: (A) things lacking both intentionality and phenomenal conscious, and (B) things possessing both phenomenal consciousness and intentionality. In concede the possibility of a third category, (C) things (like Robby) that lack phenomenal consciousness yet possess genuine intentionality.

Now let us return to John's moral dilemma. What will change in his judgments about the situation if he concedes that Robby is an example of category (C)? John now believes that Robby lacks moral status (because he lacks phenomenal consciousness) yet possesses genuine intentionality. What should John do? If he doesn't attach the translation devices, 25 people in his family suffer and may die. If he attaches the translation devices, he will not cause any phenomenal suffering but he will prevent Robby from achieving the objects of Robby's intentional states (at least prior to the passing of time, *t*). Is he wrong to do this? This is a *moral* question, not a metaphysical one. In John's moral reasoning, what moral weight should be given to "thwarting the genuine intentions of a system lacking phenomenal consciousness"? A little? A lot? None? And to whatever answer is given, *Why*? John must weigh the harm that he does to his family (by making all 25 of them homeless)

against the harm he does to Robby (by implanting the translation devices). How would you make that judgment?

I suggest that it doesn't matter if we attribute intentionality to Robby. If the *kind* of intentionality we confer on Robby is a morally irrelevant (or, at the most, a morally negligible) non-phenomenal kind of intentionality, then *as a matter of practical fact* and *for very compelling moral reasons* we will treat systems like Robby (which have intentionality but no phenomenal consciousness) more like thermostats than like humans. Robby will have gained an intentional status above thermostats, but if that is not accompanied by a similar gain in moral status, then it will be a hollow victory. It will have attained the "status" of an intentional agent, but no one will much care because then the salient property will be "moral status + intentionality," even if we assume that intentionality can be functionalized.[3]

REFERENCES

Anderson, D (1995). A Dogma of Metaphysical Realism. *American Philosophical Quarterly, 32*, 1-11.

Anderson, D (2002). Why God is not a Semantic Realist. In Alston, W (Ed.) *Realism & Antirealism*, Ithica, NY: Cornell University Press.

Dennett, D (1987). *The Intentional Stance*. Cambridge, MA: The MIT Press.

Dretske, F (1981). *Knowledge and the Flow of Information*. Cambridge, MA: The MIT Press.

Horgan, T and Tienson, J (2002). The Intentionality of Phenomenology and the Phenomenology of Intentionality. In Chalmers, D (Ed.), *Philosophy of Mind: Classical and Contemporary Readings*. Oxford: Oxford University Press.

Kim, J (2007). The Causal Efficacy of Consciousness. In Velman, M and Schneider, S (Eds.), *The Blackwell Companion to Consciousness*, Oxford: Wiley-Blackwell.

Kriegel, U and Horgan, T (Forthcoming) The Phenomenal Intentionality Research Program. In Kriegel, U. (Ed.) *Phenomenal Intentionality: New Essays*. Oxford: Oxford University Press.

Kripke, S (1972). Naming and Necessity. In Harman, G and Davidson, D (Eds.), *Semantics of Natural Languages*. Dordrecht: Reidel.

Loar, B (2002). Phenomenal Intentionality as the Basis of Mental Content." In Chalmers, D (Ed.), *Philosophy of Mind: Classical and Contemporary Readings*. Oxford: Oxford University Press.

[3] Thanks for helpful discussions with Pat Francken, Mark Siderits, Peter Boltuc, Thomas Polger, Daniel Breyer, Chris Horvath, Todd Stewart, Jim Swindler and Helen Yetter Chappell.

Millikan, RG (1984). *Language, Thought, and Other Biological Categories*. Cambridge, MA: The MIT Press.

Putnam, H (1975). The Meaning of 'Meaning'. *Minnesota Studies in the Philosophy of Science 7*, 131-193

Quine, WVO (1960). *Word and Object*. Cambridge, MA: The MIT Press.

Shannon, CE and Weaver, W (1949). *The Mathematical Theory of Communication*. Champaign, IL: University of Illinois Press.

Siewert, CP (2009). Consciousness and Intentionality. *Stanford Encyclopedia of Philosophy,* http://plato.stanford.edu/entries/consciousness-intentionality.

Siewert, CP (1998). *The Significance of Consciousness*. Princeton, NJ: Princeton University Press.