

Raleigh Miller
Georgia State University
2008

Two Dimensions of Moral Responsibility

Abstract

Shaun Nichols and Adina Roskies (2008) show that intuitive evaluations of moral responsibility in a deterministic world vary depending on whether a possible world is considered as actual or as counterfactual. I develop and extend the two-dimensionalism of Nichols and Roskies's own two-dimensional account of the concept of MORAL RESPONSIBILITY by providing a Strawsonian interpretation of MORAL RESPONSIBILITY's primary intension, inspired by Amy Glaser's two-dimensional moral semantics. I show that it is rationally defensible to employ a concept of MORAL RESPONSIBILITY according to which determinism is compatible with moral responsibility in *actual* deterministic worlds, *and* determinism is incompatible with moral responsibility in *counterfactual* deterministic worlds. I show that these seemingly inconsistent intuitions are rationally consistent, and that this concomitance does not indicate that the folk conceive of moral responsibility in an irrational way.

§I. Introduction

In this paper I argue, building upon a position outlined by Nichols and Roskies (2008), that if we understand MORAL RESPONSIBILITYⁱ two-dimensionally, we can provide a rational justification for simultaneously holding that determinism and moral responsibility are compatible in deterministic worlds considered as actual *and* that determinism and moral responsibility are incompatible in deterministic worlds considered as counterfactual.

Recent work in moral psychology has explored whether (and in what circumstances) people think that moral responsibility is compatible with determinism. Determinism is the theory that all phenomena are completely caused by prior phenomena in accordance with invariant laws of nature. If determinism is true, and if we had perfectly complete knowledge of all physical facts and all laws of nature, we could perfectly predict all future phenomena, including human behavior.ⁱⁱ The possibility that determinism is true has led to some anxiety over whether we have *free will*, and whether we can be *morally responsible* for our actions.ⁱⁱⁱ We can consider the following 'incompatibility conditional':

IC: If determinism is true, then people are not morally responsible.

Compatibilism holds that free will and moral responsibility are compatible with determinism. A compatibilist will reject IC. *Incompatibilism* holds that free will and moral responsibility are incompatible with determinism. An incompatibilist will endorse IC. Nichols and Roskies have produced data that suggests that *the folk* might

endorse IC as an *analytical conditional*. Analytical conditionals ($P \gg Q$) take the form:

$(P \gg Q)$: *If P is false, then $P \rightarrow Q$ is true and Q is false. If P is true, then $P \rightarrow Q$ is false and Q is still false.*

If Nichols and Roskies are correct, and the folk endorse IC as an analytical conditional, then the folk endorse the following set of claims, which I will henceforth refer to as AC:

AC: If the actual world is indeterministic, then moral responsibility is incompatible with determinism; else in the actual world compatibilism is true of moral responsibility and is true in other worlds. (2008; p.384)

Nichols and Roskies' hold that people endorse AC because they have the intuition that the consequent of IC (that people are not morally responsible) is false in the actual world. Nichols and Roskies suggest that "the intuition that we are, in fact, morally responsible, is a non-negotiable intuition" (384). That is, the folk simply will not (perhaps cannot) broadly assent to the consequent of IC when they are evaluating the actual world. Nichols and Roskies recommend a *two-dimensional* interpretation of MORAL RESPONSIBILITY in order to make sense of AC. Two-dimensionalism holds that the referent of some concepts in a given world may vary, depending on whether the world in which the concept is tokened is considered as actual or as counterfactual. The central idea is that "which world one considers actual affects at least some philosophical judgments" (ibid., 371). Nichols and Roskies suggest that correctly specifying the *conditions* of successfully tokening MORAL RESPONSIBILITY in a world may depend on whether the world is considered as actual or as counterfactual.

Nichols and Roskies aim to account for folk intuitions concerning moral responsibility without alleging that the folk employ their MORAL RESPONSIBILITY concept irrationally.^{iv,v} The objective is to account for folk intuitions concerning moral responsibility, if possible, without "imply[ing] that subjects are mistaken in their judgments", or adopting a "presumption of irrationality" (Nichols and Roskies, 2008; 382). However, if the story ends with the analytical conditional, it is quite difficult to see how the folk's rationality has been vindicated. If the folk fix the extension of IC as an analytical conditional, they now stand accused of 'believing' a conditional statement to be true, if and only if the antecedent is false. Such capriciousness in one's willingness to endorse conditional statements is hardly becoming of the folk's rationality. If I allegedly commit to *If P then Q*, while denying *Q* in the case that *P*, I seem to have violated a basic principle of rational thought. If I persist in

believing that “if I drank coffee, I would fall asleep” even while acknowledging that I cannot sleep after I drink coffee, I seem to have made a mistake. Could such concomitants (“if p then q” and “not q, even though p”) be rationally tenable? *Maybe*, but we are right to be skeptical. Before two-dimensionalism can demonstrate that AC is plausibly rational, we must have a clearer picture of the conceptual components of MORAL RESPONSIBILITY. What we require is an account of the *primary intension* of MORAL RESPONSIBILITY that could justify endorsing \sim IC in the case that determinism is true, and IC in the case that determinism is false. I will provide such an account.

My conclusion will be (1) that moral reactive attitudes, as elucidated by P.F. Strawson, play an important role in establishing the content of MORAL RESPONSIBILITY. Particularly, I will argue that the function which fixes the *primary intension* of MORAL RESPONSIBILITY is “*that which obtains in agents towards whom we comport ourselves with moral reactive attitudes.*” Further, I will show that construing the primary intension of MORAL RESPONSIBILITY in this way will suggest an account MORAL RESPONSIBILITY that has the potential to capture relevant folk intuitions, and allow for a *rational justification* of AC.

In §II, I will discuss recent empirical research that is relevant to the debate over in/compatibilism. This research illuminates the way people employ their MORAL RESPONSIBILITY concept. I will introduce Nichols and Roskies’ data, and the role that AC plays in their interpretation. I will conclude that Nichols and Roskies’s account is inviting, but impoverished in the absence of an account of the primary intension of MORAL RESPONSIBILITY. To provide such an account, I will (§III) outline the basic claims and motivations of two-dimensionalism. David Chalmers’s model of two-dimensionalism is the most fully developed, so I will borrow my basic conceptual tools from him.^{vi} I will introduce Amy Glaser’s model (2005, 2008) of two-dimensional *moral* semantics. My own analysis of MORAL RESPONSIBILITY will borrow from Glaser. In §IV, I will argue for a Strawsonian interpretation of the primary intension of MORAL RESPONSIBILITY, and will show that AC, accordingly, is not an irrational position.

§II. Moral Psychology and Folk Intuitions; Nichols and Roskies

Experimental philosophers (e.g., Nahmias and Nadelhoffer (2007); Nichols and Knobe (2007)) approach IC as a theory of how MORAL RESPONSIBILITY refers. Subjecting various “intuition premises” to experimental scrutiny, these philosophers investigate the referential tendencies of normal language users, affectionately referred to as “the folk”.^{vii} The goal of experimental philosophy is to discover how people actually employ their concepts, and to provide an empirical basis for adjudicating philosophical arguments that rely on ordinary-language conceptual analysis or intuition premises.

Nichols and Knobe (2007) have shown that when an abstractly described world is characterized as deterministic, the majority of participants say of agents in that world that they are not responsible for their actions. Nahmias *et.al.* (2006) have gathered data that more strongly suggests that participants respond that free will and moral responsibility are compatible with determinism across a variety of cases. Notably, however, most of these studies ask subjects questions about *other* worlds, *counterfactual* worlds (usually in the form of fictional universes).^{viii} It is worth considering that a world’s status as *counterfactual* may influence the way that our concepts refer in that world, and that our concepts might behave differently in an *actual* world. This has led Nichols and Roskies (2008) to compare folk intuitions about worlds considered as actual with worlds considered as counterfactual. They have collected surprising data that suggests that people locate the extension of MORAL RESPONSIBILITY in an actual determinist world differently they do in a counterfactual deterministic world. Subjects were told that a world was causally deterministic, and asked to rate their agreement with the following claims:

1. It is impossible for a person to be fully morally responsible for their actions
2. People should still be morally blamed for committing crimes
3. It is impossible for people to make truly free choices.

Some subjects were asked to rate their agreement with (1), (2), and (3) when considering an *alternate* deterministic world. Another group of subjects were asked to rate their agreement with (1), (2), and (3) when considering an *actual* deterministic world (described as their own world). Subjects were strongly inclined to agree with (1) in the alternate world, and moderately inclined to disagree with (1) in the actual world. Subjects

were strongly inclined to disagree with (2) in the alternate world, and moderately inclined to agree with (2) in the actual world. Subjects were strongly inclined to agree with (3) in the alternate world and very slightly inclined to agree with (3) in the actual world (only slightly higher than “neutral”).^{ix} If this data is representative, then it appears that subjects are more inclined to express intuitions that support IC when the world they are discussing is an alternative, counterfactual world. They are less inclined to express intuitions that support IC when they are discussing the actual world.^x

To accommodate this data, Nichols and Roskies propose several interpretations. They suggest that the *actual* scenario may increase *affective* involvement, which may increase compatibilist intuitions (see also Nichols and Knobe (2007)). They also suggest that the actual scenario may have a particular *motivational* salience, since the subject is motivated to “hang on” to her own free will. The final interpretation that Nichols and Roskies proffer is that the folk endorse AC, rather than IC or \sim IC. The proposed advantages of AC are that it (1) adequately captures (or “respects”) apparently inconsistent intuitions, and (2) it suggests that the truth of compatibilism will turn, in part, upon what turns out to be true of the actual world (384-5). It is unclear what Nichols and Roskies mean by “respect”. If they only mean that a theory of folk intuitions ought to be consistent with what the folk actually say, then AC appears to be successful. However, Nichols and Roskies also refer to their interpretation as not “imply[ing] that subjects are mistaken in their judgments” (382). It seems that they also (or should also) intend to “respect” folk intuitions by accounting for them in a way that avoids, when possible, the implication that the folk are irrationally applying their concepts.

AC takes the belief that people are morally responsible in the actual world as a nonnegotiable intuition, one that simply will not (or cannot) be given up. But Nichols and Roskies do not provide a *justificatory* reason for this intuition’s *being* nonnegotiable. For all that is said, people may regard this intuition as nonnegotiable out of *stubbornness*. This would be an unsatisfactory answer to the irrationality charge. If the folk are applying their concepts inconsistently out of stubbornness, then they are, plausibly, being irrational. What is required, and what has not yet been provided, is a theory of MORAL RESPONSIBILITY’s conceptual composition which, if correct, allows that endorsing *IC if and only if determinism is false of the actual world* is rationally defensible.

Nichols and Roskies have advanced AC as a *two-dimensional* interpretation of MORAL RESPONSIBILITY, as it acknowledges that our intuitions about the extension of moral responsibility may be different when applied to actual and counterfactual worlds. Nichols and Roskies do not, however, give reason to believe that the intuition *that people are morally responsible in the actual world is justifiably non-negotiable*. I believe that such reason can be given by extending and developing the two-dimensionalism of Nichols and Roskies's account, and by providing a plausible account of MORAL RESPONSIBILITY's primary intension. I move on in the next section to introduce and develop the two-dimensionalist model of semantic content, which I will later use to render AC rationally defensible.

§III. Two-Dimensionalism

Two-dimensionalism has come to prominence as an attempt to synthesize competing theories of semantic content, which we may call *externalism* and *internalism*. The *semantic content*, or meaning, of a concept may be viewed as an *intension*. The intension of a concept or sentence is a function from a world, *W*, to the extension of the concept or sentence in *W*.^{xi} A function from *A* to *B* is a way to determine *B*, given *A*. The intension of WATER is a way to determine the extension of WATER (what is and is not water) in a world. Internalists have argued that the extension of a concept is fixed by states internal to the speaker tokening the concept. A prominent candidate for a reference-fixer on an internalist model is *cognitive significance*, or the functional role that a particular concept plays in the subjects cognitively organizing the world. On this model, the extension of a concept is fixed by a particular internal state, and any instance of that state is a tokening of the concept. Externalists have argued that the intension of some concepts (e.g., proper names (Kripke 1980) and natural kinds (Putnam 1973, 1975)) *does not* fix an extension via *any* internal states of a speaker, but instead fixes rigidly upon a certain entity or state of affairs in the world, external to the speaker tokening the concept. A standard argument for content externalism is Putnam's *Twin Earth* argument. On Twin Earth, all of the watery stuff is XYZ, rather than H₂O. Because XYZ is *not* water, in spite of playing WATER's cognitive functions, WATER picks out an external state of affairs, H₂O, rather than an internal cognitive significance.

Two-dimensionalism accepts externalism up to a point. The two-dimensionalist is willing to

acknowledge that the *correctness conditions* for some concepts are not “in the head”. However, “many would still like to hold that [internalism]^{xiii} was right about *something*” (Chalmers 2004). Like many externalists, the two-dimensionalist holds that we discover what water *is* by discovering *what* plays the cognitively significant roles that water plays. Unlike the externalist, however, the two-dimensionalist takes this cognitive significance to be partly constitutive of the concept, by assigning it to the *primary intension* of the concept. The externalist content of a concept is the concept's *secondary intension*. So, while the *primary intension* of WATER is watery stuff, the *secondary intension* of WATER (relative to the actual world, Earth) is H₂O.

The *primary intension* is a function from worlds considered as actual to an extension, and is fixed via the internal states that accompany tokening episodes. The primary intension of WATER is the set of cognitively significant roles that our WATER concept plays, and picks out the “clear, odorless, drinkable liquid that fills the rivers and lakes”, or, in short, the watery stuff. When an extension is fixed by the primary intension of a concept, different worlds, considered as actual, will yield different extensions. On earth, the extension picked out by the primary intension of WATER is H₂O. On Twin Earth, the extension picked out by the primary intension of WATER is XYZ. The *secondary intension* is a function from worlds considered as counterfactual to extensions. If we fix earth as actual, we determine that the primary intension of WATER always picks out H₂O. We can then consider the extension of WATER in counterfactual worlds. On all counterfactual worlds, WATER refers to H₂O, regardless of what other watery stuff might be present, and regardless of whether H₂O is watery in other worlds. Thus (holding fixed earth as actual) the secondary intension of WATER in all possible worlds (respectively considered as counterfactual) is H₂O. This generates our intuition that there is no water on Twin Earth.

Like internalism, two-dimensionalism acknowledges that the internal states that accompany a concept-tokening episode are important to understanding how the concept fixes its referent. By relativizing any secondary extension to a particular actual world, two-dimensionalism can account for the intuition that if the actual world had been different, water could have been XYZ. Like externalism, however, two-dimensionalism acknowledges that the correctness conditions of a particular concept tokening (once an actual world has been fixed) need not

rely upon any states internal to the speaker.

§IIIa. Moral Concepts on Two Dimensions

My claim, that MORAL RESPONSIBILITY can be usefully interpreted two-dimensionally, will be supported by showing that moral concepts *in general* are amenable to the two-dimensional model. Amy Glaser (2005, 2008) has developed a provisional model for understanding moral concepts two-dimensionally. Glaser is responding to a dispute in metaethics. On one side of this dispute is the claim that moral concepts like GOOD or RIGHT fix their referents upon external states of affairs (like pleasure maximization or divine command). Alternatively, some have claimed that moral concepts must always fix their extension by a function that is internally determined (like emotive approval).^{xiii} Glaser has suggested that the *primary intension* of our moral concepts is their inherent *prescriptive, motivational* character. That is, prior to fixing any external features in the world to which our GOOD concept refers, we know that GOOD picks out objects which possess the feature of *to-be-doneness*. Suppose Fred and Ginger disagree diametrically about what things are good. Fred thinks that all and only the good things are A (e.g. pleasure satisfying), and Ginger thinks that all and only the good things are \sim A (e.g. ascetic self-denial). Plausibly, Fred and Ginger disagree over the feature shared by objects being picked out by a bare notion of *what is to be done*. Ginger does not think that Fred is wrong because she denies that pleasure satisfying actions are pleasant. Rather, Ginger thinks that pleasure satisfying actions are not *to-be-done*. Our GOOD concept *primarily* refers to whatever is *to-be-done*, or *to-be-striven-for*. The *primary intension* of GOOD fixes upon that which satisfies a cognitively significant role (that which is *to-be-done*).

Glaser has suggested that our moral concepts *also* behave as if they are picking out objects by implicit reference to a non-moral feature that is shared by all and only the good things. That is, speakers take themselves, in calling an action good, to mean the action has a particular non-moral property (God's approval, or pleasure maximization). Our moral terms will often behave as if it is *not* mere *to-be-doneness* that is being picked out, but an underlying non-moral feature that tokens of *good* characteristically have in common. This position ought to be familiar. Fred and Ginger both presuppose such an externalism by even bothering to argue over what feature (pleasure, self-denial) picks out all the good things. The traditional ethicist assumes an

externalism of this kind as she wades through intuitive moral beliefs in search of such a feature.^{xiv}

These intensions cooperate in constituting moral concepts in a manner analogous to Chalmers's analysis of WATER. Our GOOD concept is, in the first place, determined internalistically. Even if all the good things have some external feature in common, we could only discover such a feature by observing its reliable co-instantiation with our bare notion of *to-be-doneness*. If it were the case that *that which maximizes utility* picked out the same extension as *that which is to be done* (as fixed by motivational states), we could accurately say that the secondary intension of GOOD is utility maximization. If the secondary intension of GOOD is utility maximization, all of the utility maximizers on counterfactually-regarded possible worlds are the good things, *regardless* of the motivational states of Twin Earthlings. At the same time, we would be able to recognize that the extension of GOOD at Twin Earth *considered as actual* could include those actions and states of affairs which are, for instance, divinely commanded. In the event that some action, A, is within the set of divinely commanded actions, but not within the set of utility maximizing actions, we would be able to recognize, without contradiction, that on Twin Earth considered as actual, A is good, but considered as counterfactual, A is not good.

My analysis of MORAL RESPONSIBILITY will borrow from Glaser's two-dimensional interpretation of moral concepts. I will suggest that the primary intension of MORAL RESPONSIBILITY is a motivational stance one takes in interpersonal relations, along the lines of P.F. Strawson's reactive moral attitudes. Fixing the primary intension accordingly will show that AC is a rational position to adopt in adjudicating the extension of MORAL RESPONSIBILITY in deterministic worlds.

§IV. Strawsonian Reactive Attitudes and the Primary Intension of MORAL RESPONSIBILITY

Before taking the final step, I will briefly recapitulate. I have (§II) introduced Nichols and Roskies's data, which suggests that subjects' adjudication of the compatibility of determinism and moral responsibility in an actual deterministic world is more compatibilist than their adjudication of the compatibility of determinism and moral responsibility in a counterfactual deterministic world. I have introduced Nichols and Roskies's AC, which attempts to rescue folk intuitions from charges of irrationality by positing the existence of moral

responsibility on the actual world as a non-negotiable intuition. I have suggested that, absent an account of the *primary intension* of MORAL RESPONSIBILITY, it is quite unclear that AC gives us any reason to doubt that the folk are irrational. In pursuit of such an account, I have introduced (§III) the basic conceptual tools of two-dimensionalism. In particular, I have shown that two-dimensionalism can provide a plausible account of moral concepts, by understanding the primary intension of such concepts motivationally. Turning now to moral responsibility, I argue that MORAL RESPONSIBILITY is partially constituted by an internalist, motivational content. This will lead to a plausible justification of AC, and the conclusion that, should AC in fact be the position people endorse, such a state of affairs would not warrant the conclusion that the folk are irrational.

Nichols and Roskies recommend a two-dimensional interpretation of MORAL RESPONSIBILITY.

However, in attempting to characterize the primary intension of moral responsibility, they say very little:

“one might maintain that the [primary] intension for the proposition that people are morally responsible is never rendered false by the status of determinism when evaluated in a world taken as actual; that is, regardless of whether the world is determinist or indeterminist, the [primary] intension yields compatibilist judgments.” (2008, 385)^{xv,xvi}

This, however, merely restates the claim that the existence of moral responsibility is a nonnegotiable intuition. It does not provide a story as to why such an intuition would be nonnegotiable, or a satisfactory account of MORAL RESPONSIBILITY’s primary intension. By providing such an account, I will show that the non-negotiability of the intuition *that we are morally responsible in the actual world* is rationally justifiable.

Peter Strawson (1962) has famously argued that *interpersonal moral reactive attitudes* uncontroversially exist, and that this sufficiently demonstrates that moral responsibility obtains in the actual world. Strawson characterizes moral reactive attitudes as

...the non-detached attitudes and reactions of people directly involved in transactions with each other; of the attitudes and reactions of offended parties and beneficiaries; of such things as gratitude, resentment, forgiveness, love, and hurt feelings. (1962, 4)

In the absence of moral responsibility, such attitudes would be unintelligible. Strawson suggests that FREE WILL and MORAL RESPONSIBILITY, are not, first and foremost, metaphysical concepts. They are primarily *characterizations*, and quite accurate characterizations, of the ways in which we interact with one another. We do not comport ourselves towards all objects, even all sentient beings, by adopting moral reactive attitudes. We do not comport ourselves this way towards trees, animals, infants or the severely mentally handicapped, for

instance, because we do not regard them as fully moral. We often do not comport ourselves this way towards intoxicated people, or overly emotional people, because we regard them as not in full possession of themselves. However, we cannot suspend such attitudes entirely. We must retain the practice of adopting moral reactive attitudes towards those whom we take to be mature agents who are reasonably under their own control. Moral reactive attitudes are necessary components of any sensible model of human interaction. Strawson elaborates,

...in the absence of *any* forms of these attitudes it is doubtful whether *we* should have anything that we could find intelligible as a system of human relationships, as human society...[it] *is* wrong...to forget that these practices, and their reception, the reactions to them, really *are* expressions of our moral attitudes and not merely devices we calculatingly employ for regulative purposes. Our practices do not merely exploit our natures, they express them. (1962, 24-5)

The practices which fundamentally constitute MORAL RESPONSIBILITY *may* have some interesting physical basis, such as indeterminism, or agent causation. However, the basic reference fixer for MORAL RESPONSIBILITY, by which we determine that some agent is or is not morally responsible, is not a physical configuration. It is, rather, a feature of interpersonal exchange. It is a way of acting towards others and a way of regarding ourselves. I take Strawson's compatibilism to suggest a worthy candidate for the *primary intension* of MORAL RESPONSIBILITY. We recall that Glaser recommended we understand the primary intension of GOOD as *to be done*, or *to be striven for*. Likewise, we can understand the primary intension of MORALLY RESPONSIBLE as *to be comported towards with moral reactive attitudes* and the primary intension of MORAL RESPONSIBILITY as *that which obtains in agents towards whom we comport ourselves with moral reactive attitudes*. For brevity, I will refer to this latter intension as *MRA*.

We can then understand the in/compatibilist debate as concerning the *secondary intension* of moral responsibility. However, if, as I have claimed, moral reactive attitudes are required in order to have “anything that we could find intelligible as a system of human relationships,” (Strawson, quoted above) the secondary intension of MORAL RESPONSIBILITY *must yield a nonempty extension in the actual world*. Any candidate secondary intension which declares moral responsibility to be an illusion is barred from consideration. This contention may seem preposterous; to argue that the extension of moral responsibility could not be empty is to beg the question against the free will skeptic. To alleviate such concerns, we might compare MORAL

RESPONSIBILITY with WATER. If the incompatibilist is correct, and indeterminism is required in order for anyone to be genuinely moral responsible, then upon learning that the world is deterministic we would be rationally compelled to deny that people have moral responsibility. But imagining an analogous case concerning WATER will show that we would not be inclined, or rationally obliged, to accommodate such a discovery accordingly. Suppose all the chemists have been mistaken, and it turns out that there is no hydrogen in the world, and consequently, no H₂O. If we learned this, we would *not* conclude that there was no water on earth. We would conclude that water must be something else. The belief *that there is water in the actual world* is non-negotiable. This is compatible with saying that there is no water on Twin Earth, *so long as* earth is fixed as actual and Twin Earth is considered as counterfactual. H₂O is *only* an acceptable secondary intension of WATER *so long as* we know that there is H₂O in the actual world, and consequently know that by taking WATER to be H₂O we are providing an account of what all the watery stuff is. But in the event that the *secondary intension* of our WATER concept implies that WATER's primary intension picks out nothing at all, something has gone quite wrong, and we ought to revise the secondary intension. I am suggesting that MORAL RESPONSIBILITY behaves the same way. Any function that we are considering as a candidate secondary intension for MORAL RESPONSIBILITY is only permissible as a candidate in the event that it successfully picks out all or most of the tokens of *MRA* in the actual world. If, having agreed upon a secondary intension of MORAL RESPONSIBILITY, we come to realize that this secondary intension has the consequence that the primary intension of MORAL RESPONSIBILITY picks out nothing at all, we would reasonably decide that something has gone quite wrong, and that we ought to reevaluate the secondary intension.

Though my own compatibilist leanings may be apparent, I want to emphasize that I am *not* assuming that compatibilism is true. Neither are the folk. Not, that is, *unless and until* it is discovered that determinism is true. Suppose that there is some set of physical and psychological facts, Δ , which causally underlies *MRA* in the same way that H₂O causally underlies the watery stuff.^{xvii} So long as determinism is false (or its truth is unknown), the proposition *that determinism is false* may be in Δ . If there turns out to be no H₂O on the actual world, then water is not H₂O, but something else. We can accept such a conditional, without asserting that water

is not H₂O. If determinism turns out to be true in the actual world, then the proposition *that determinism is false* is not in Δ. We can accept such a conditional, but still allow that the proposition *that determinism is false* could be in Δ if determinism *is* false.

There is no convergence concerning *what* physical or psychological facts might belong in Δ. The folk do *not* have a robust theory as to what makes all the morally responsible things morally responsible. The implicit assumptions of philosophers, scientists, and writers at the *New York Times*^{xviii} coax the folk into believing that the proposition *that determinism is false* must be in Δ. Lacking any robust theory of their own, the folk are cooperative. The suggestion is sensible, after all. If we couldn't calculate what Ginger will do in five minutes (because her decision making is indeterministic) then perhaps this is *why* Ginger is free, and what *makes* her responsible for what she does. Given the initial plausibility of indeterminism's membership in Δ, it is not difficult to say of other beings, in another world in which *determinism* is true, that those beings, lacking what allegedly brings about our own moral responsibility, are not themselves morally responsible. In the event that determinism is *actually* true, then Δ cannot contain the proposition *that determinism is false*, because this would result in *MRA* yielding an empty set in the actual world, and the above analysis has shown that this is an unacceptable consequence. For this reason, if the folk learn that determinism is true in the actual world, they will deny that determinism is incompatible with moral responsibility.

Whatever the *secondary* intension of MORAL RESPONSIBILITY is, it will pick out the same things on the actual world as the *primary* intension of MORAL RESPONSIBILITY picks out.^{xix} I have suggested that the primary intension of MORAL RESPONSIBILITY is *MRA*. Because *MRA* must pick out *something*, the secondary intension of MORAL RESPONSIBILITY, whatever it may be, must pick out something. If, on the assumption that the world is indeterministic, we believe that indeterminism is a component of Δ, we would say of counterfactual persons whose physical constitution is deterministically lawful that they are not actually morally responsible. However, if our own world turns out to be deterministic, then indeterminism cannot be in Δ. If determinism is false of this world, then it may be incompatible with moral responsibility. However, if determinism is true of this world, then it must be compatible with moral responsibility. This, we recall, is the

analytical conditional.^{xx} This position no longer has the *prima facie* implausibility that was assigned to it above. Fixing the primary intension of MORAL RESPONSIBILITY motivationally, incorporating Strawsonian reactive attitudes, I hope to have shown that AC is a reasonable position, and that the folk's endorsement of AC does not illustrate that they are irrational.

§V. Conclusion

Incompatibilism requires that the secondary intension of MORAL RESPONSIBILITY must be a set of physical and psychological facts that, minimally, includes the proposition *that determinism is false*. However, there is reason to think that the folk are flexible on this demand when considering deterministic worlds as actual. This is not a manifestation of irrational concept tokening. The folk are justified in believing that there is *some* set of facts that accompany tokenings of *MRA*, *some* Δ . Insofar as that set may include indeterminism, the folk are willing to assert of *counterfactual* determined persons that they are not morally responsible. However, insofar as indeterminism cannot serve such a role (for instance, if it is false of this world) then moral responsibility is no worse off. Deterministic *actual* worlds are worlds in which we assert that the secondary intension of MORAL RESPONSIBILITY must be constituted by something else, and MORAL RESPONSIBILITY must be compatible with determinism.^{xxi}

References

- Chalmers, D. (2002) "On Sense and Intension". (J. Tomberlin, ed.) *Philosophical Perspectives 16: Language and Mind*. Blackwell. pp. 135-82.
- Chalmers, D. (2004) "The Foundations of Two Dimensional Semantics".
<http://consc.net/papers/foundations.html>
- Kripke, S. (1980): *Naming and Necessity*. Cambridge, MA: Harvard University Press.
- Glaser, A. (2005) "A Two-Dimensional Analysis of Ethical Language".
<http://www.unc.edu/~adglaser/mathesis.pdf>
- Glaser, A. (2008) "Facts that Tell Us What to Do: The Semantics and Ontology of Moral Discourse". (Draft)
http://www.unc.edu/%7Eadglaser/draft11_2_08.pdf
- Horgan, T. & Timmons, M. (Forthcoming) "Analytical Moral Functionalism Meets Moral Twin Earth"
Forthcoming in a collection of essays on the philosophy of Frank Jackson.
- Fischer, J.F. and Ravizza, M. (1999) *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press.
- Jackson, F. (1998) *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. Oxford University Press.
- Knobe, J. & Doris, J. (Forthcoming) "Strawsonian Variations: Folk Morality and the Search for a Unified Theory". *The Handbook of Moral Psychology*. (Ed. Doris) Oxford: Oxford University Press
- Knobe, J. & Nichols, S. (2008) "An Experimental Philosophy Manifesto". *Experimental Philosophy*. Oxford University Press.
- Mackie, J.L. (1977). *Ethics: Inventing Right and Wrong*. Harmondsworth: Penguin.
- Nadelhoffer, T. & Nahmias, E. (2007) "The Past and Future of Experimental Philosophy". *Philosophical Explorations*, Vol. 10, No. 2. June 2007
- Nahmias, E. & Nadelhoffer, T. & Morris, S. & Turner, J. (2006) "Is Incompatibilism Intuitive?" *Philosophy and Phenomenological Research* 73(1): 28-53.
- Nichols, S. and Knobe, J. (2007). "Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions". *Nous*, 41, 663-685.
- Nichols, S. and Roskies, A. (2008) "Bringing moral responsibility down to earth" *Journal of Philosophy*. CV:7. pp.371-88.
- Overbye, D. (2007) "Free Will: Now You Have It, Now You Don't". *New York Times*. 1/2/2007.
- Putnam, H. (1973), "Meaning and Reference" *Journal of Philosophy* 70: 699-711.
- Putnam, H. (1975) "The Meaning of 'Meaning'," in Keith Gunderson, ed., *Language, Mind and Knowledge*, Minnesota Studies in the Philosophy of Science, VII, Minneapolis: University of Minnesota Press.
- Sayre-McCord, G. (1997) "'Good' On Twin Earth". *Philosophical Issues*, Vol. 8, Truth (1997), pp. 267-292.
- Soames, S. (2005) *Reference and Description*. Princeton University Press.
- Stalnaker, R (1999). *Context and Content: Essays on Intentionality in Speech and Thought*. Oxford University Press
- Strawson, P. (1962) "Freedom and Resentment" *Proceedings of the British Academy*. 48: pp. 1-25.

Notes

ⁱ Words written in all capital letters refer to the concept to which the word refers. Italics will often be used to indicate either that the word is a constituent of a proposition under consideration, but will sometimes be used merely for emphasis.

Context should successfully disambiguate these two uses.

ⁱⁱ Much ink has been spilt trying to accurately formulate the thesis of determinism. The term does seem to be the subject of some abuse. Nichols and Knobe (2007) characterize determinism in terms of complete causation. Nahmias et.al. (2007) have characterized determinism in terms of theoretically possible perfect prediction. I have no desire to adjudicate between these characterizations, so I use both.

ⁱⁱⁱ Many notable theorists have considered it appropriate to separate the claim that moral responsibility is compatible with determinism from the claim that free will is compatible with determinism. The most notable example of this is Fischer's

semi-compatibilism (See Fischer and Ravizza's 1999) according to which determinism is compatible with moral responsibility but incompatible with free will. For the sake of my own argument, it is sufficient to consider FREE WILL as the *sort* of free will that is required for moral responsibility, while allowing that other possible interpretations of FREE WILL may be orthogonal to my analysis entirely.

^{iv} For an example of the sort of account to which Nichols and Roskies are providing an alternative we might consider Knobe and Nichols (2007). Knobe and Nichols show that people appear to endorse IC unless they are asked to evaluate the moral responsibility of a person who has committed a heinous crime in a deterministic world. Though Knobe and Nichols do not explicitly commit to this claim, they proffer the idea that people are only compatibilist when affect (or the desire to punish) produces conceptual error.

^v I regret the ambiguity inherent in a word like "rationality." I'm attempting to use it in a very straightforward way. I consider an instance of concept use to be irrational iff elucidating all of the beliefs to which one commits oneself in so using the concept would yield some belief of the form "*p* and *not p*".

^{vi} Nothing in my account, so far as I can see, is tied to Chalmers's particular account of two dimensionalism. Notably, Nichols and Roskies invoke Jackson's two-dimensionalism. Nothing in their account, so far as I can see, is tied to Jackson's particular account of two-dimensionalism. In both cases, the two-dimensional tools being employed are very general, and amenable to many of a number of candidate two-dimensionalisms.

^{vii} I take for granted that the experimentalist approach to the free will debate has some merit. This is not the appropriate paper to broadly defend the experimentalist methodology. For such a defense, see for instance Knobe and Nichols (2008), Nahmias *et.al.* (2006) and Nadelhoffer and Nahmias (2007).

^{viii} The notable exception is Nahmias *et.al.* (2006), whom I address below in note (ix).

^{ix} Some may be concerned that an apparent lack of convergence dampens the appropriateness of the two-dimensional interpretation that Nichols and Roskies suggest, and which I extend. Provisionally, this worthy concern simply calls for further empirical investigation, as Nichols and Roskies admit (385n.36). As it relates to my own thesis, I address this concern further below in note (xv).

^x In further support of the thesis that the folk endorse AC, Nahmias *et.al.*(2006) found that when subjects were asked to say whether psychologically determined subjects (for whom all actions are completely caused by their intentional states, which were in turn caused entirely by prior intentional states) were morally responsible, they were drastically less likely to do so when the agents were described as living on a fictional planet, Earta, than when subjects were told that the agents were human beings in the actual world.

^{xi} I am following Chalmers in using the word 'intension' to refer to the function by which a concept's referent is picked out (c.f. Chalmers (2002)). This is distinct from Fregean intensions, which stand in for cognitive significance. On the Chalmersian model, the *Fregean* intension is the *primary* intension.

^{xii} Chalmers says that "many would like to hold that Frege was right about something."

^{xiii} To see this debate in action, one could consult Timmons & Horgan (forthcoming) as an example of the "internalist" ethical theorist, and Sayre-McCord (1997) as an example of the "externalist" ethical theorist.

^{xiv} Agnosticism or skepticism about the existence of such a feature is quite compatible with this claim. This can be construed as accepting Mackie's (1977) claim that some moral sentences do purport to make factual claims about the world, while being open to the possibility that, as Mackie also claims, there are no corresponding facts. Eddy Nahmias has suggested (in correspondence) that this feature of two-dimensional moral semantics may pose a problem for me, insofar as we can understand moral concepts to refer as if there are factual propositions that constitute their secondary intension *while being open* to the possibility that nothing is picked out by this secondary intension. In allowing this, perhaps I am required to allow that nothing is picked out by the secondary intension of MORAL RESPONSIBILITY either. This would make my later claim, that the extension of MORAL RESPONSIBILITY must yield a non-empty set on the actual world, indefensible. This is a worthwhile point; a full answer to this charge will have to be deferred to a future work. However, notice that in claiming that there are no moral facts, Mackie has at the very least provided an account of the structure that such facts would have to assume, and is then, arguably, in a better position to suggest that there are no such facts. In the case of MORAL RESPONSIBILITY, we lack any such account. An analysis like my own, which explicitly provides the primary intension of MORAL RESPONSIBILITY and some account of the entailed secondary intension would be a prerequisite for any position that claims that MORAL RESPONSIBILITY has, properly speaking, no referent. In short, future work may show that some of my later claims about the extension of MORAL RESPONSIBILITY could be wrong, but such an argument would (a) need to work within a two-dimensional model like the one I have suggested, and would (b) need to provide either (i) an alternative account of the primary intension of MORAL RESPONSIBILITY or (ii) a refutation of Strawson's own claims concerning the indispensability of moral reactive attitudes.

^{xv} It is worth mentioning that Nichols and Roskies may be *wrong*, and the existence of moral responsibility in the actual

world *may be* negotiable. In this case, AC would be ruled out as a candidate for a correct account of folk intuitions concerning moral responsibility and its compatibility with determinism. We might take Nahmias *et.al.* (2006) to illustrate that this is the case. When presented with an account of *mechanism* or full-blown *ontological reductionism* with respect to mental states, and told that this account was true of the actual world, people were inclined to say that agents were not morally responsible in the *actual world*. If this data is representative, then the belief that people are morally responsible in the actual world may be negotiable. To this potential concern, I would offer two responses. First, I have not argued that the folk endorse AC. I have merely argued that in the case that the folk endorse AC, they could plausibly be rational in doing so. If it turns out that folk intuitions are not tracked by AC, then my conclusion becomes less interesting, but it does not become false. Second, I have my doubts that the subjects of Nahmias *et.al.*'s survey took the reductionist-actual-world *as actual*. I think people are largely skeptical of reductionism (for reasons such as phenomenology, or religion) no matter what the scientists tell them. It seems to me that even if told in a survey that reductionism is true of the actual world, subjects may consider that world counterfactually, and take the same license in conceiving of agents on that world as they would with a citizen of Twin Earth. This matter could be empirically explored by replicating the study and following up with the question "Do you believe that the world fits the [mechanistic] description given above?" Preliminary data has suggested that people answer this question in the negative, even when the prompt unequivocally asserts the description to accurately portray the actual world.

^{xvi} I actually think that Nichols and Roskies misspeak here. They write: "regardless of whether the world is determinist or indeterminist, the [primary] intension yields compatibilist judgments". But this is simply not what AC says. AC says that if the world is indeterministic, the primary intension yields an incompatibilist judgment. However, overlooking this misstep, I still take this quote to best represent the thoughts that Nichols and Roskies have offered on the primary intension of MORAL RESPONSIBILITY.

^{xvii} I have introduced delta (Δ) in order to avoid the incautious claim that indeterminism *is* the secondary intension of moral responsibility. This is implausible, since we could acknowledge that electrons may behave in an indeterministic way without our being tempted to say that electrons are morally responsible. Thanks to Amy Glaser and Dylan Murray on this point.

^{xviii} This reference is only intended to nod at the general tendency of journalists to publish articles with titles like "Science Shows People Not Free!" However, the choice of the *New York Times* is not entirely arbitrary. For instance, see Overbye (2007).

^{xix} I have remained agnostic on the secondary intension of MORAL RESPONSIBILITY, the actual constituents of Δ . Δ may be different in different cases; Δ may not qualify as a single kind (see Knobe and Doris, forthcoming). Δ may even be theoretically undiscoverable. In the event that the secondary intension may pick out more than one kind, it can still be the case that MORAL RESPONSIBILITY has a secondary intension. The secondary intension would simply pick out its referent disjunctively. Similarly, JADE does not appear to pick out a single natural kind. Jadeite and Nephrite, two substances composed of distinct molecular configurations, are both picked out by what we have, historically, called 'jade'. Nonetheless, JADE has a secondary intension, which could be different on different worlds considered as actual. Even though JADE picks out [Jadeite or Nephrite] on this world, it could still be the case that TWIN JADE on Twin Earth picks out some third substance, which we would then say is not jade. Neither should the potentially undiscoverable nature of the secondary intension of MORAL RESPONSIBILITY be worrisome. It may be the case, for instance, that CONSCIOUSNESS has a secondary intension, and that its secondary intension is undiscoverable.

^{xx} Or, rather, this is the analytical conditional with some slight modal adjustments. I think that it must be acknowledged that indeterminism may both be true and *not* belong in Δ . However, because the folk have been led to believe that determinism and moral responsibility exist in an antagonistic relationship with one another, they are likely to consider indeterminism an essential constituent of Δ , so long as it is true, thus yielding the stronger modal claim (indeterminism *does* belong in Δ) that appears in AC.

^{xxi} I am indebted to Dan Bernstein, Justin Bernstein, Amy Glaser, Dylan Murray, Eddy Nahmias, Paul Pfeilschifter, and Ben Sheredos for their invaluable assistance in refining the ideas of this paper and commenting on a previous draft.