This is a preprint of an article published in *Acta Analtica* (2021 36: 323-339). The final authenticated version is available online at: https://doi.org/10.1007/s12136-021-00461-5

The Prejudice of Freedom: An Application of Kripke's Notion of a Prejudice to our Understanding of Free Will

James Cain

This essay reframes salient issues in discussions of free will using conceptual apparatus developed in the works of Saul Kripke, with particular attention paid to his little-discussed technical notion of a prejudice. I begin by focusing on how various forms of modality (metaphysical, epistemic and conceptual) underlie alternate forms of compatibilism and discuss why it is important to avoid conflating these forms of compatibilism. The concept of a prejudice is then introduced. We consider the semantic role of prejudices, in particular conditions in which prejudices turn out to express metaphysically necessary truths. With that as background I discuss a set of prejudices involving the notion of choice. We consider the role these prejudices might play, should they turn out to be true, in determining the answer to various compatibility questions concerning the nature of moral responsibility and choice.

Compatibilism

Though the literature on free will tends to focus on the question of whether free will is compatible with determinism, a broader collection of related compatibilist questions is really at stake. It helps to frame these as being of the form "Is X compatible with Y?", where X and Y stand in for statements (or perhaps formulas with free variables) and it is being asked whether it is possible for X and Y to be true together (or jointly satisfied). Thus, a compatibilist question takes the form of asking whether $\Diamond$(X & Y) is true (or satisfiable) for a given reading of X and Y. To fully specify a compatibility question, we need to identify X and Y and identify the modality of the possibility operator. Typical pairs, X/Y, include: Free will exists/Determinism holds; Moral responsibility exists/Determinism holds; Agent A is responsible for making choice C at time t/Prior to t it was already determined that A would make choice C at t.

There are a number of modalities in which interesting compatibility questions can be raised. I will say that X and Y are *metaphysically compatible* (or $\Diamond_M$ (X & Y)) iff there is a possible world (a way the world could have been) such that (X & Y). X and Y are *conceptually compatible* (or $\Diamond_C$ (X & Y))—iff it is conceptually

possible that (X & Y) (that is, "~(X & Y)" is not an analytic truth; or equivalently, it is not the case that "~(X & Y)" is true in all possible worlds in virtue of its meaning[1]).[2] X and Y are *epistemically compatible* for person P at time t (or $\Diamond_{E,P,t}$ (X & Y)) iff, at t, P's total epistemic body of evidence does not rule out its being the case that (X & Y).[3] For any of the modalities, to say that X is *incompatible* with Y is to say that it is not the case that X is compatible with Y.

Let us consider the entailment relationships between various modes of compatibilism and incompatibilism. Metaphysical compatibility entails conceptual compatibility: if "(X & Y)" is true in some possible world, then it is conceptually possible that (X & Y). By contraposition, it follows that conceptual incompatibility entails metaphysical incompatibility. Conceptual compatibility does not entail metaphysical compatibility: If we let X = "This contains water" and Y = "This contains no H_2O," then X is conceptually compatible with Y (it is not the case that "~(X & Y)" is true in every possible world *in virtue of its meaning*) but X is not metaphysically compatible with Y. It follows that metaphysical incompatibility does not entail conceptual incompatibility.

Neither metaphysical nor conceptual compatibility entails epistemic compatibility (and thus epistemic incompatibility does not entail metaphysical or conceptual incompatibility). If we let X = "The inventor of bifocals never lived in Philadelphia" and Y = "Benjamin Franklin was the inventor of bifocals," then X is metaphysically and conceptually compatible with Y, but they are not epistemically compatible for those who know that Franklin took up residence in Philadelphia: their epistemic body of evidence rules out its being the case that (X & Y).

Epistemic compatibility does not entail metaphysical compatibility (and thus metaphysical incompatibility does not entail epistemic incompatibility). Let X = "This is a sea anemone" and Y = "This is a

---

[1] Kripke (1980: 39) stipulates that "an analytic statement is, in some sense, true by virtue of its meaning and true in all possible worlds by virtue of its meaning."

[2] If X or Y contains free variables the analysis of their conceptual compatibility will have to be rephrased in terms of satisfaction: For any assignment to its variables, "$\Diamond_C$ (X & Y)" is satisfied iff it is not the case that "~(X & Y)" is satisfied in all possible worlds in virtue of its meaning.

[3] Usually when I speak of epistemic possibility, I will leave the values for "P" and "t" unstated. Sometimes I may index epistemic possibility to a group rather than an individual (as in "for all scientists know it may be the case that life first evolved on Mars").

plant." X and Y may be epistemically compatible for an agent at a given time, and yet they are not metaphysically compatible—sea anemones are by nature animals, not plants.

One might think that conceptual incompatibility entails epistemic incompatibility. X and Y are conceptually incompatible iff "~(X & Y)" is an analytic truth. If in general analytically true statements can be known *a priori*, then conceptual incompatibility entails epistemic incompatibility. Kripke in fact held that analytic statements are *a priori*. He wrote:

> … let's make it a matter of stipulation that an analytic statement is, in some sense, true by virtue of its meaning and true in all possible worlds by virtue of its meaning. Then something which is analytically true will be both necessary and *a priori*. (That's sort of stipulative.) (1980: 39)

Kripke is thus committed to holding that conceptual incompatibility entails epistemic incompatibility—and, by contraposition, that epistemic compatibility entails conceptual compatibility. This treatment of the relation between analyticity and *a prioricity* depends on the unargued assumption that, in general, statements that are true in all possible worlds in virtue of their meaning can be known *a priori*. I argue elsewhere (reference omitted for blind review) that, given the apparatus developed in *Naming and Necessity*, there may be counterexamples to this claim. But if we confine our attention to cases in which, if "~(X & Y)" is analytically true, then it can be known *a priori*, then if X and Y are conceptually incompatible they will also be epistemically incompatible.

Relevance of varieties of compatibilism

Though I believe that the above remarks should be fairly obvious, it is easy to overlook their importance. I will illustrate this with a couple of examples. Sometimes people worry that future findings of science might show that there is no moral responsibility or that other deep-seated views about responsibility are false (e.g., it is false that punishment is ever morally deserved). If it were shown that, say, the existence of responsibility is conceptually incompatible with determinism; then it would follow that moral responsibility is epistemically and

metaphysically incompatible with determinism[4], and thus that if it were discovered that determinism holds that would require us to abandon the view that we are morally responsible.

On the other hand, suppose we were to come to the conclusion that the existence of moral responsibility is epistemically compatible with a future discovery that determinism holds. Though epistemic compatibilism does not generally entail metaphysical compatibilism, might we have special grounds for holding that the epistemic compatibility of determinism and moral responsibility gives us a reason to expect that determinism is metaphysically compatible with moral responsibility? We find a recurrent argument in the writings of John Martin Fischer (1999: 129-30; 2012: 118; 2016; Fischer and Ravizza 1998: 253-54) that might seem to suggest this. Fischer writes as follows in support of his treatment of "semicompatibilism," the view that moral responsibility is compatible with determinism:

> I believe that our fundamental status as agents…should not depend on certain subtle ruminations of theoretical physicists. That is, I do not think that our status as genuine *agents* should hang on a thread—that it should depend on whether natural laws have associated with them (say) probabilities of .99 or 1.0. In my view, *that sort* of empirical difference should not *make a difference* as to our moral responsibility. … Now this is simply one consideration, and it specifies a desideratum of an adequate theory of moral responsibility. In my view, it counts in favor of a theory of moral responsibility—it is a reason to accept such a theory—if the theory does not conceptualize moral responsibility as hanging on a thread (in the indicated way). (Fischer 2012: 118)

One might take these considerations to heart and hold that it counts in favor of an account of responsibility that it allows for the metaphysical possibility of responsibility existing in a deterministic world.

I believe it would be a mistake to take a commitment to the epistemic compatibility of determinism and responsibility to count as an independent reason to favor theories that hold responsibility to be metaphysically compatible with determinism. An example illustrates the problem. In the early days of chemistry, a proponent of the view that water is H_2O (and thus water's existence is metaphysically incompatible with the nonexistence of

[4] In saying this I am not assuming that conceptual incompatibility by itself entails epistemic incompatibility. Rather I am saying if X and Y are conceptually incompatible and have been *shown* to be conceptually incompatible, then (for those to whom this has been shown) X and Y are epistemically incompatible.

H_2O) might worry that the science would deny the existence of H_2O. But he would not feel that the world's water supply was "hanging on a thread;" at most the correctness of his theory would hang on a threat. We do not count the epistemic possibility that science will disconfirm the existence of H_2O as a strike against the view that water is by nature H_2O—and thus also a strike against the view that water's existence is metaphysically incompatible with the nonexistence of H_2O.

One who theorizes that moral responsibility ultimately depends on our having a kind of agency that is metaphysically incompatible with determinism need not hold that a discovery of determinism would show that we lack moral responsibility; she might instead allow for the possibility that such a discovery would cast doubt on the correctness of her account of responsibility, and claim that the fact that her theory could ultimately be shown to be incorrect by new evidence does not give us grounds on our current body of evidence to think that her theory is wrong. Science would be hobbled if we felt a need to try to develop our hypotheses concerning the essential nature of things so that what are now open epistemic possibilities (even if their likelihood is low) come to be counted as metaphysical possibilities. We should be worried that our investigation of the nature of responsibility would be similarly hobbled if so constrained (van Inwagen 1983: 223; Kane 2007: 181; Steward 2012: 124). We will return to this idea below.

Another example in which we need to carefully distinguish the modality of compatibilist claims is found in discussions of the problem of evil. For instance, in the "logical problem" of evil an attempt is made to disprove the existence of God by appealing to the existence of evil. The free will defense can be seen as providing a partial account of how evil could be present in a world created by and under the providence of an all good, almighty God.[5] The free will defense maintains that it may be impossible for God to allow for the exercise of free will in certain circumstances and guarantee that no wrong choices will be made. It is sometimes held that compatibilism undermines the free will defense. For example, David Lewis (1993) argues as follows:

> Compatibilism says that our choices are free insofar as they manifest our characters (our beliefs, desires, etc.) and are not determined via causal chains that bypass our characters. If so, freedom is

[5] Free will defenses vary and it goes beyond the scope of this paper to discuss the details of free will defenses. My purpose here is merely to highlight the importance of being sensitive to the modality of compatibility claims in discussions of the problem of evil. For a more detailed treatment of the free will defense as it bears on these issues see [omitted for blind review].

> compatible with predetermination of our choices via our characters. *The best argument for compatibilism is that we know better that we are sometimes free than that we ever escape predetermination; wherefore it may be for all we know that we are free but predetermined* [emphasis added]. (1993: 155)
>
> It seems that free-will theodicy must presuppose incompatibilism. God could determine our choices via our characters, thereby preventing evil-doing while leaving our compatibilist freedom intact. Thus He could create utopia, a world where free creatures never do evil. (1993: 156)[6]

Note that Lewis's "best argument" for compatibilism appears to be a straightforward argument for an epistemic form of compatibilism. Yet presumably the free will defense only needs to take as a premise the claim that free choices of the sort considered in the free will defense are metaphysically incompatible with predetermination; it does not need to appeal to epistemic or conceptual incompatibility. The reason for saying this is that presumably God cannot bring about that which is metaphysically impossible (e.g., the existence of water that is not H_2O) even if it is epistemically possible (for all some people know there might be water that is not H_2O) or it is conceptually possible (it is not analytically false that there is water that is not H_2O).

It may help to illustrate the connection between epistemic and metaphysical compatibility more fully here. Imagine that Jane is undecided with respect to whether God exists, whether determinism holds, and whether exercises of free will are even possible under conditions in which they are determined to take place by causes lying outside the agent (even if the causal chains do not bypass the agent's character). She is however well aware of the fact that exercises of free will (of the sort singled out in the free will defense) take place. For Jane each of the following claims may be an open epistemic possibility even though they cannot both be true:

(i) Exercises of free will (of the sort singled out in the free will defense) take place and are determined to take place by a cause or causes lying outside the agent.

[6] In fairness to Lewis, I should note that his paper also presents arguments that attempt to raise difficulties for the free will defense even if we assume that its libertarian views of free will are correct.

(ii) It is metaphysically impossible for exercises of free will (of the sort singled out in the free will defense) to take place and to be determined to take place by a cause or causes lying outside the agent.

Given Jane's current epistemic situation she is unable to rule out either hypothesis. Since she cannot rule out (i), it follows that for Jane the exercise of free will (of the sort singled out in the free will defense) is epistemically compatible with its determination by causes lying outside the agent. Yet, since she cannot rule out (ii), she has no basis to assert the corresponding metaphysical compatibility claim. Thus, merely on the basis of the compatibility claim she can accept, she is in no position to assert that, if God exists, then it is within God's power to determine our exercises of free will (of the sort singled out in the free will defense) in a way that precludes their misuse.

The above examples illustrate the need to bring considerations of metaphysical compatibility into our discussions of freedom and responsibility. I believe that Kripke's account of prejudices, our next topic, may prove helpful in this endeavor.

Kripke's notion of a prejudice

Mario Gómez-Torrente (2011: 301) introduces Kripke's concept of a prejudice as follows:

> Beliefs such as the belief that most paradigmatic cases of gold belong to a single substance, or that most paradigmatically red things share a certain nondispositional property, are examples of what Kripke calls *prejudices*. Although not perforce a priori, a prejudice is a belief that we hold onto pretty firmly, and that we try to retain with as little modification as possible in the face of pressures from empirical data. For Kripke, our language is replete with working prejudices. Many of them have a semantic role, though not the role of an analytic definition. One semantic role of the belief that most paradigmatic cases of gold belong to a single substance is that of setting a condition for the assignment of an extension to "gold," in case the prejudice is true. If we find out that it is not true, the semantics

> we give to "gold" will depend to a great extent on our new decisions in view of the empirical data, and one semantic role of the relevant remaining prejudices will be to guide these decisions.[7]

Kripke (Gómez-Torrente, 2011: 301) illustrates the idea of a prejudice by reminding us of a disagreement between Arthur Eddington and Susan Stebbing. Science has found that things we normally think of as solid are really collections of particles separated by gaps.[8] Call such objects *gappy*. Eddington (1933: 342) wrote as if science had shown us that the things we normally think of as solids are not really solid after all. It may appear that he is appealing to an *a priori* analytic truth that solids are not mostly empty space (in the way that ordinary things are mostly empty space given Rutherford's model of the atom). Stebbing, on the other hand, held that science has merely shown us that the solids we normally encounter are gappy. So, she would certainly not regard the claim that solids are not gappy as analytically true. Kripke treats the dispute between Eddington and Stebbing as involving a clash of prejudices. Eddington gives more weight to the prejudice that solids are not gappy whereas Stebbing gives more weight to the prejudice that most things we take to be paradigms of solidity really are solid.

While I suspect that most people, myself included, would side with Stebbing, it is important to see that the prejudice Eddington appeals to is not without weight. Suppose that it had it turned out that the vast majority of objects we ordinarily think of as being solid were made of continuous matter without any gaps. In that case,

---

[7] In his essay, Gómez-Torrente summarizes and discusses Kripke's unpublished lectures on color. Though Kripke makes extensive use of the notion of a prejudice in these lectures, Gómez-Torrente's discussion is the only published account of this topic of which I am aware. A video of Kripke's lecture "No Fool's Red? Some Considerations on the Primary/Secondary Quality Distinction," in which he develops his notion of a prejudice (and at which Gómez-Torrente is the commenter), is available at the Saul Kripke Center website: https://saulkripkecenter.org/index.php/videos/. In the video Kripke covers much of the material on prejudices that Gómez-Torrente summarizes in his essay. [This part of the note is omitted for blind review.]

[8] Eddington (1933: 1-5) was particularly impressed by the separation of the subatomic components of atoms. According to Eddington the space occupied by those components is miniscule compared to the size of the whole atom. "The atom is as porous as the solar system. If we eliminated all the unfilled space in a man's body and collected his protons and electrons into one mass, the man would be reduced to a speck just visible with a magnifying glass." (Eddington, 1933: 1-2)

Kripke holds (Gómez-Torrente, 2011: 309), if we were to discover abnormal cases in which there were gappy objects that looked like ordinary solids, we would probably not count them as falling under the extension of our word "solid." On the analogy with "fool's gold," we might speak of these objects as "fool's solids."[9]

Though the concept of a prejudice was not explicitly developed in *Naming and Necessity*, an example discussed by Kripke in that work nicely illustrates the tendency to confuse a prejudice with an *a priori* truth or an analytic truth. At the time Kripke gave his *Naming and Necessity* lectures in 1970, it was commonly held that statements like "cats are animals" are analytic and thus *a priori*. Kripke (1980: 122, 125-26) maintained that "cats are animals" is neither analytic nor *a priori*, but rather is a necessary truth that must be discovered empirically. Though, epistemically speaking, it might have turned out to be the case that cats were something other than animals, as things actually turned out we found that cats are in fact animals and that necessarily cats are animals. Kripke says:

> 'Cats are animals' *has* turned out to be a necessary truth. Indeed of many such statements, especially those subsuming one species under another, we know *a priori* that if they are true at all, they are necessarily true. (1980: 138)

We might try to put this in terms of the notion of a prejudice by saying that we have an initial prejudice that cats are animals. The prejudice is not an analytic truth but may have a role in determining the semantics of the term "cat." One semantic role of the belief that cats are animals may be that of setting a condition for the

[9] The fact that Kripke is willing to treat the claim that solids are non-gappy as a prejudice that carries some initial weight (even if we do not ultimately accept it) shows that a prejudice is not to be understood as a stereotype (in Putnam's (1970) sense), for the linguistic competence of a speaker with respect to the term "solid" does not require that the speaker have a stereotype that has implications with respect to the micro-structure of solids (e.g., that there are not gaps between subatomic particles in the way Eddington thought). Nor should prejudices be understood as simply conveying features by which we pick out paradigm cases of the application of an expression. While it might contribute to our picking out solids that they are not *observably* gappy, we don't pick them by their being fully non-gappy—after all, the paradigms of solidity we pick out are all gappy.

assignment of an extension to “cat” (in both the actual world and in other possible worlds), in case the prejudice is true with respect to the cats we actually encounter.[10]

Another prejudice that we have with respect to cats is that most paradigmatic instances that we are inclined to call “cats” really are cats. Had it turned out that most of the things we called “cats” were really demons (had that epistemic possibility turned out to be the case) then, depending on the details, this prejudice might have led us to abandon the prejudice that cats are animals. Perhaps—again depending on the details—under those circumstances the extension of the term “cat” would have applied to a certain category of demons.

Prejudices relevant to the will

There appears to be a broad range of prejudices that pertain to the will. Many of these may be platitudes of folk psychology. I will give some examples that I expect are common prejudices. Keep in mind that, as we saw in the Eddington-Stebbing dispute, prejudices need not be true; they may be defeasible; they may come into conflict with each other, and the weight of a prejudice may vary from person to person—what is given significant weight by some may be given little or no weight by others. Also keep in mind that prejudices are not usually formulated explicitly and attempts at their formulation may to an extent go wrong by making them appear overly precise. My formulations should rather be taken as giving approximations of prejudices.

I am particularly interested in two sets of related prejudices. The first set, which I call the *paradigms prejudice with respect to choice*, includes the following:

PP(i) Most paradigmatic instances of the application of the term "choice" really are cases in which there is a choice.

We may come to recognize that some of what we take to be paradigmatic applications of the term "choice" will not in fact be choices. There may turn out to be “fool’s choices” (analogous to fool’s gold)—apparent choices that are not genuine. Perhaps an example of a “fool’s choice” would be an apparent choice made as a result of hypnotism.

[10] Cf. the previous quote from Gómez-Torrente on the semantic role of prejudices.

I suspect, with the qualification given below, that there is tendency to hold a related paradigms prejudice, namely:

PP(ii) Most paradigmatic cases of choosing belong to a natural kind of psychological phenomena that occurs in the formation of intentions.

The needed qualification concerns what I have in mind when I speak of a natural kind of psychological phenomena. As I use the expression “natural kind,” it will suffice for a kind to count as a natural kind that it can be successfully picked out by a term characterized by the features set out below that Kripke identifies in his treatment of natural kind terms. I will argue that when we speak of "making a choice" we tend to use the term "choice" in conformity with these features and that there is a prejudice that in doing so we succeed in demarcating a kind of psychological phenomena.

Kripke (1980: 118-19) notes some common characteristics of natural kind terms. Our initial use of a natural kind term typically involves an attempt to single out a kind (a kind of substance, species, quality, phenomena, etc.) instantiated by (most) members of a paradigmatic set of items picked out through certain identifying marks thought to be possessed by the objects in that set. Though the kind is initially picked out in this way, the identifying marks or properties need not be essential features of the kind. It may be the case that some items that have the identifying marks fail to be of the kind, and items of the kind may fail to have some, or even any, of the identifying marks. But if the kind term is successfully introduced it will pick out a kind whose essential features are fixed by the actual features of paradigmatic instances of the kind. The essential properties of the kind need not be readily obvious; some may only be discoverable *a posteriori*—perhaps simply by becoming familiar with the kind, or perhaps though detailed scientific investigation.

How might the features of natural kind terms just outlined apply to the term "choice"? The identifying marks by which we single out paradigmatic instances that we take to be choices include the following:

- There is a typical phenomenological “feel” that we find when we observe ourselves making a choice.
- When one makes a choice one’s intentions with regard to some matter become fixed, at least for a while.
- The agent who makes a choice has a subjective awareness of the fixing of the intention and its content.

- The occurrence of a choice typically falls between a period in which the agent deliberates about what it might be like were an action to take place or not take place and a period in which the agent has an intention to do or not do the action.

However, we recognize that choices can be found that lack one or more of these identifying marks: We may choose without having any such special phenomenological feel; we may choose and immediately change our minds; perhaps we may choose without an awareness of our true intentions; a choice may be made without deliberation.

Furthermore, we should recognize the epistemic possibility that something may have the identifying marks of a choice without really being a choice. For example, suppose that (a) we learn to stimulate a subject's brain so that she "feels" like she makes a choice, though the process by which this happens is independent of any normal processes that lead to a choice. Maybe, by stimulating the brain a certain way, *a**, a person has the inner experience of saying to herself, "That's what I'll do!" along with the kind of feelings one might be expected to have on making a choice. If stimulation *a** were to take place in a context in which no action was up for consideration, the agent might, after the experience, simply wonder, "What was that about?" and have no idea what the demonstrative in "That's what I'll do" could possibly be referring to. Surely that would not count as a choice. Next suppose that (b) by stimulating the brain in a certain way it causes the subject to intend to perform a given action and does this independently of the usual ways in which a person comes to form an intention. For example, by stimulating the brain in manner *b** the subject immediately has an intention to climb onto the roof. If the subject were, say, watching television with no thoughts of climbing onto the roof and stimulation *b** occurred the subject would instantly have the intention to climb onto the roof. Here it would hardly be correct to say that a choice had been made. Suppose further that (c) by stimulating the brain in certain ways we can bring it about that the person has specific beliefs where the coming about of these beliefs is independent of any justifying evidence. For example, by stimulating the subject's brain in a certain way, *c**, it comes about that the subject believes that she had chosen to climb onto the roof. Now imagine a context in which a subject (whose brain was secretly implanted with three devices, each capable of bringing about one of *a**, *b**, or *c**) is asked, "Would you like to climb up to the roof?" and then the subject's brain is given stimulations *a**, *b** and *c**, which work independently to produce the effects noted above. We have here a set of identifying marks that would normally lead one to say a choice had been made. Yet we should not automatically count this as a case of choosing. Since the process by which the intention was formed differs significantly from the kind of psychological process found in paradigmatic choices, it is not clear that what took place should be

counted as a choice. Our use of the term "choice" here seems to fit nicely with Kripke's account of how natural kind terms are intended to function. Of course, that does not settle the question of whether in using the term we successfully pick out a psychological kind whose essential features are fixed by the actual features of paradigmatic instances of the kind. The expectation that we do pick out such a kind is reflected in PP(ii).

Another prejudice, or cluster of prejudices, may be termed the *prejudice of freedom*: It includes the following:

PF(i) Often when we make a choice, we feel that we do not have to make that choice. When this happens it really is open to us whether we make the choice. It is consonant with the laws of nature that we do not make that choice under those circumstances.

PF(ii) There is a class of paradigmatic choices of the sort we ultimately appeal to when we make moral assessments of people's actions. A person making a choice belonging to this class could, under the very same circumstances, have not made that choice.[11]

Note that in stating this prejudice, I am not saying that we have a prejudice that morality depends on choices that are free in some sense. Perhaps we do, but that is not at issue here. Regardless of whether we are justified in our moral judgments, we often do make them and in making them there is a class of choices we often appeal to as fundamentally justifying our judgments. PF(ii) says of the choices in that class that they are open in the specified sense.

PF(iii) If we make a choice between X and Y and feel that we can choose either way then we really are able to choose either way unless prior to choosing we are prevented from exercising our choice. (We might, for example, be prevented by a heart attack, or a compelling distraction.) Under those circumstances it is consonant with the laws of nature that we do not choose X and we either choose Y or something prevents our choosing Y, and it is consonant with the laws of nature that we do not choose Y and we either choose X or something prevents our choosing X.

[11] The qualification "we ultimately appeal to" has been added to exclude from PF(ii) consideration of choices for which we hold the agent merely derivatively responsible.

The next prejudice employs a concept of control that derives from Fischer's writings. Fischer (2006: 6) identifies *regulative control* as "the sort of control that involves genuine metaphysical access to alternative possibilities." Fischer (2006: 8) is inclined to accept a version of van Inwagen's Consequence Argument from which it follows that determinism precludes regulative control, but he does not think that this conclusion is "indisputably true." I will employ the term "strong regulative control" to refer to a kind of control that determinism definitely precludes.

> An agent has *strong regulative control, at time t, with respect to action X* provided that, at t, she has 'the sort of control that involves genuine metaphysical access to alternative possibilities'; in particular she has access (consonant with the laws of nature and the state of the world at t) to an alternative in which she does X, and she has access (consonant with the laws of nature and the state of the world at t) to an alternative in which she does not do X, and it is up to the agent which way that control is exercised.

Using this notion, we may formulate the following prejudice of freedom:

PF(iv) Normal humans who have reached the age of reason typically have strong regulative control over a wide range of their actions and choices; in particular they have strong regulative control with respect to the class of paradigmatic choices of the sort we ultimately appeal to when we make moral assessments of people's actions.

Do we also have a prejudice that whenever we have made a choice, we could have not made that choice? Some may have a prejudice that, even under extreme torture, if we are capable of making a given choice, then we are also capable of not making that choice; and, if we cannot help what we are doing, then we are not really choosing. Others may disagree and allow that we may make choices and be unable to do otherwise. The first response will be labelled the *strong prejudice of freedom:*

SPF Whenever we make a choice, we do not have to make that choice. It is consonant with the laws of nature that we not make that choice under those circumstances.[12]

Prejudices and metaphysical necessity

We saw earlier that our prejudices, taken in conjunction with empirical facts, may play a role in determining the semantics of our terms: Had the objects we call "solids" turned out to be non-gappy, then the extension of our term "solid," even with respect to possible worlds containing gappy objects, would not have included gappy objects. We also saw that the confirmation of a prejudice may be associated with the recognition of a metaphysical necessity: What led us to confirm that cats are animals also led us to confirm that cats are necessarily animals. I will be particularly interested in considering two questions: (i) How might our prejudices concerning choice play into the semantics of the term "choice" (and related terms) if it turns out that our choices are in fact undetermined?[13] (ii) If it turns out that the choices for which we hold people ultimately responsible are undetermined, what implications might that have with respect to the metaphysical compatibility of responsibility and determinism? I will begin, however, by briefly considering the role our prejudices might play should it turn out that determinism is true.

If determinism is found to be true there will be a clash of prejudices. To the extent that the paradigms prejudices predominate we would tend to keep intact our commitment to the existence of choice and abandon the prejudice of freedom.[14] To the extent that we give prominence to the prejudice of freedom we might reject or

[12] I am skeptical about the truth of SPF. In addition to the above worry about torture, examples given by Eleonore Stump (1999: 323) cast serious doubt on SPF.

[13] I do not want to give the impression that our prejudices together with the discovered empirical facts need always fully determine an answer to this question. See the quotation from Gomez-Torrente with which I introduced the notion of prejudice. It speaks of prejudices as guiding our semantic decisions.

[14] For examples of approaches to discussions of free will and free action that give a central place to an appeal to paradigms, see Flew's (1955) classic statement, and Heller's (1996). Heller (1996: 336) holds that if there is a single kind of which the paradigms of free actions are all instances, then if determinism holds compatibilism is true; if there is "no single kind of which the paradigms are all instances", then it is not actually the case that free acts exist; and what counts as essential to an action's being free is determined by the kind (if it exists) of which

downplay the existence of choice. For one who held onto SPF, accepting the truth of determinism would call for denying the existence of choice. Even though the other prejudices of freedom (PF i-iv) do not entail that all choice is undetermined, one strongly committed to them might be inclined to say that it turns out—given determinism—that we do not make "genuine" choices after all. The thought here could be that prior to discovering that determinism holds we believe that the reference of the kind term "choice" is fixed by core cases of undetermined choice; cases of "determined choices" might then be countenanced given that they bear sufficient similarity to the core cases. But once we can no longer recognize any core paradigms of undetermined choice, we must (according to this line of thought) give up the idea that genuine choices exist at all.[15]

Suppose, on the other hand, that determinism was found to be false. Such a discovery would not clash with the prejudices listed above.[16] The question arises whether evidence confirming any of the prejudices of

---

the paradigms are instances (and thus it could turn out that "what is in fact essential to an action's being free is that there be some undetermined event in that action's causal history"). Heller correctly warns us not to confuse epistemic possibility and metaphysical possibility: "Assuming that we do not know determinism to be false and even assuming that we do know that 'free act' refers, what follows is that there is an epistemic possibility of a deterministic world in which our actions are free. It does not follow that there really is any such possible world."

[15] Perhaps an analogy may help here. Suppose that "bacteria" was initially introduced as a natural kind term intended to pick out a kind of organism that had been viewed under microscopes. At that time, it may have been an open question whether there exist bacteria with a much different superficial appearance (say, that of being doughnut shaped) than the appearance of the paradigms used to fix the reference. Being doughnut shaped would have been epistemically compatible with being a bacterium, and—given that bacteria form a natural kind—it may also have been metaphysically compatible. But had it turned out that what were initially taken to be paradigms of bacteria were not microscopic organisms, but were instead scratches on microscope slides, then the term "bacteria" would have been withdrawn and not recognized as a successful natural kind term. It would lack reference in every possible world. Even if doughnut-shaped microorganisms were discovered later, they would not count as falling under the extension of the term "bacteria" as it was originally introduced.

[16] An anonymous reader has suggested that there might be some 'compatibilistic' prejudices (e.g., uncaused = random = unfree) pertaining to the determination of actions or events that would clash with libertarian accounts of freedom. If so, then if it were discovered that what are taken to be paradigm cases of freely choosing are not

freedom would warrant our acceptance of significant forms of metaphysical incompatibilism with regard to determinism and either choice or responsibility. I will first present considerations that might seem to indicate that this would be the case. We will then consider a difficulty that would need to be faced before such a conclusion could be drawn.

Let us begin with SPF: Suppose it becomes confirmed that whenever we actually make a choice it is consonant with the laws of nature that, under those very circumstances, we do not make that choice. Then it would be tempting to think that any possible phenomena superficially resembling a choice but determined to take place would simply be too different from the kind of phenomena that we call "choosing" to qualify as a choice, and thus it would turn out to be metaphysically necessary that choices are undetermined. On this line of thought determinism would be metaphysically incompatible with free choice and with choice-dependent responsibility.

Next, suppose we were to discover that PF(ii), PF(iii) or PF(iv) holds. This might raise serious doubts concerning the metaphysical compatibility of determinism and moral responsibility. For the question would arise whether any choice-like phenomena that took place in a deterministic world would be similar enough to the actual-world undetermined paradigmatic choices that in fact ground our moral responsibility to count as responsibility grounding. If they were not adequately similar then it is hard to see how responsibility could be grounded in a deterministic world.

One might try to push the argument further and claim that if PF(ii-iv) were shown to hold that would not merely cast doubt on the metaphysical compatibility of responsibility and determinism; it would allow us to conclude that responsibility is metaphysically incompatible with determinism. But, as just noted, there is a difficulty that needs to be confronted by anyone who wishes to endorse this claim, and I am unsure how to handle it. To that we now turn.

To introduce the problem, it will help to reconsider Kripke's claim that we know the following *a priori:* if it is true that cats are animals then it is necessarily true that cats are animals.[17] A related example

---

caused by anything prior, there would be a clash between the compatibilistic prejudices and the prejudice that what are taken to be paradigms of free choice really are cases of free choice.

[17] I am unconcerned with whether the claims I consider regarding determinism, choice and responsibility are *a priori*. But it will help to illustrate the problem I have in mind to see how a similar problem arises with Kripke's treatment of *a prioricity*.

illustrates a difficulty facing this claim. Imagine that we are living in an earlier time in which we have a less extensive knowledge of biology. We have become aware of sea anemones as a kind of sea life and use the term "sea anemone" to pick out this kind, but we do not know much about the nature of sea anemones; in particular, we do not know whether they are plants or animals.[18] While they appear to be animals, we realize that it is an empirical question whether they really are. Perhaps, we think, they may turn out to be plants. Despite our ignorance we might still have a reasonable expectation that if sea anemones are in fact animals then it is metaphysically necessary that sea anemones are animals and being an animal is essential to being a sea anemone. Nevertheless, we would not know *a priori* that:

(1) If sea anemones are animals then it is metaphysically necessary that sea anemones are animals.

I say this because under the circumstances we could not rule out *a priori* certain hypotheses which hold that though all sea anemones are in fact animals it could have been the case that some sea anemones were not animals. Here is an example of such a hypothesis:

> Sea anemones are formed when a certain kind of fish begins to digest its prey and then spits it out, oddly mutilated, as a sea anemone which then lives on the sea floor rather than as it used to. One sea anemone might be a misshapen eel, another a misshapen turtle, etc. The fish responsible for forming sea anemones does not like to eat plants; so all sea anemones are in fact animals. But were the fish to spit out a partially digested plant, it too would come to live on the sea floor and have the same appearance and behavior that sea anemones are usually observed to have; it would in fact be a sea anemone that is a plant.

If this hypothesis held, it would be a contingent truth that all sea anemones are animals. Since we cannot rule out this hypothesis *a priori*, we do not know *a priori* that if sea anemones are in fact animals then necessarily sea anemones are animals.

[18] Aristotle (1984: 922; *History of Animals*, Bk.VIII, 588b20) took them to be intermediate between plants and animals.

Being a plant is conceptually compatible with being a sea anemone. Furthermore, prior to the discovery of appropriate details about the nature of sea anemones, being a plant was epistemically compatible with being a sea anemone. Given our *a posteriori* knowledge of how the biological world tends to work, it would have been reasonable to expect (a) that bizarre hypotheses like the one just considered are false, (b) that if sea anemones turn out to be animals then this is because they form something along the lines of a biological subspecies, a species, a family, or an order of animals, and this being so, (c) it is metaphysically necessary that sea anemones are animals.

What I have said for sea anemones holds for cats as well. We do not know *a priori* that if all cats are animals then it is metaphysically necessary that all cats are animals. But we had good *a posteriori* reason to believe that if cats are animals then necessarily cats are animals. Furthermore, though it might be hard to precisely formulate the knowledge, it is probably correct to say that that we have some sort of *a priori* knowledge along the lines suggested by Kripke. It would seem, for example, that we know *a priori* something along the following lines: if cats as a group evolved as suggested by current evolutionary accounts forming reproductively viable populations of animals, then it is metaphysically necessary that to be a cat a thing must be an animal.[19] Similar remarks hold for sea anemones: We seem to know *a priori* that if sea anemones as a group evolved as suggested by current evolutionary accounts forming reproductively viable populations of animals, then it is metaphysically necessary that to be a sea anemone a thing must be an animal.

Let us return to questions involving determinism, choice and responsibility. If SPF were verified, would we be warranted, as suggested above, in holding that any possible phenomena superficially resembling a choice but determined to take place would be too different from the kind of phenomena that we call "choosing" to qualify as a choice? More specifically, should we affirm the following (the antecedent of which is implied by SPF)?

[19] To be precise I would have to spell out, as part of the antecedent, the details of the evolutionary theory that I have in mind, otherwise I could not know *a priori* that the conditional was true—my knowledge of the conditional would depend upon *a posteriori* knowledge of what the current evolutionary accounts are. If we let *E* abbreviate a statement of current evolutionary theory, then I would have *a priori* knowledge that if cats evolved as suggested by E forming reproductively viable populations of animals, then it is metaphysically necessary that to be a cat a thing must be an animal. Perhaps further qualifications would be needed, but, for present purposes, my approximate statement should be adequate.

(2) If all choices are undetermined, then it is metaphysically necessary that choices are undetermined.[20]

Here a problem arises analogous to the one found in the sea anemone example. In order to be justified in holding (2) one needs to rule out its being a mere coincidence that actual choices happen to be undetermined. But consider the following hypothesis:

> All cases of choosing involve brain processes. Sometimes indeterministic "static" from other brain processes "interferes" with the normal processes involved in choosing. Because of this the outcome of the choosing process is always somewhat indeterministic. Were the part of the brain involved in choice to be insulated from the parts involved in the interfering processes, choosing could still take place, but it would happen in a deterministic fashion.

Unless we can rule out hypotheses of this sort, we do not have grounds to accept (2). Furthermore, there is a difference between this case in which the question arises whether indeterminacy is essential to choice and the earlier case in which the question arose whether animality is essential to being a sea anemone. In the latter case it was fairly easy to imagine circumstances in which it would be clear that being an animal is essential to being a sea anemone. But in the case of a brain process, or a mental process tied to a brain process, it is difficult to see ahead of time just what determines whether the presence of indeterminacy is an essential feature of the process. That, of course, does not mean that science will never be able to show us whether choice is an essentially indeterministic phenomena.

A similar difficulty would seem to arise were one to claim that, if either PF(ii) or PF(iii) were shown to be true, that would justify us in holding that determinism is metaphysically incompatible with moral responsibility, or at least that it would justify us in holding that determinism is metaphysically incompatible with instances of responsibility grounded in choice in the way that our responsibility is so grounded. For the fact that PF(ii) or PF(iii) held would not by itself settle the question whether an essentially indeterministic kind of choice underlies (our) responsibility.

[20] If it were an analytic truth that choice is undetermined then (2) would be analytic. I am assuming that neither SPF nor the claim that choices are undetermined is analytic.

As we saw previously, sometimes the verification of a prejudice (sea anemones are animals) along with the addition of extra facts (e.g., sea anemones as a group evolved as suggested by current evolutionary accounts forming reproductively viable populations) may warrant us in asserting a metaphysical necessity (being a sea anemone is metaphysically incompatible with being a plant). Perhaps something along those lines may apply here. Suppose PF(iv) was found to be true, and furthermore it was discovered that whenever people actually exercise regulative control in a choice or action, they exercise agent-causation in doing so, where agent causation is understood as it is set out by, say, O'Connor (2000) or Steward (2012).[21] If that were so, then I should think we would be warranted in saying that determinism is metaphysically incompatible with moral responsibility, or at least that determinism is metaphysically incompatible with instances of responsibility grounded in choice in the way that our responsibility is so grounded.

Conclusion

Debate over compatibilism in its various forms is central to discussions that come under the heading of free will. The concept of compatibility is a modal notion and is subject to the sorts of distinctions and clarifications that those working in the philosophy of language have developed with regard to the notions of possibility and necessity. I have tried to show how Kripke's insights concerning modality have application to free-will problems. In particular I consider how clarity can be brought into those discussions by distinguishing conceptual, epistemic and metaphysical notions of compatibility, and how an understanding of prejudices concerning choice closely relates to important questions concerning the metaphysical compatibility of moral responsibility and determinism.

[21] As I will use the term, to say that an exercise of strong regulative control is an instance of agent causation requires that in that exercise the agent, as a substance, acts as a cause, where the agent's role as a cause is not ontologically reducible to causation by some event, property, or state involving the agent. Furthermore, in agent causation the agent initiates a new causal chain that is not simply the working out of past causal influences. See Clarke (2019) for a discussion of the notion of irreducible substance causation, as well as a review of several versions of agent causation.

As we saw, a question of the form "Is X compatible with Y?" asks whether it is possible for X and Y to be true together, and thus to fully specify the question one must indicate the kind of possibility that is at stake. Unfortunately, the modal issues involved in discussions of compatibility are difficult to untangle.

If we are aware of a conceptual incompatibility, then we automatically have epistemic incompatibility and we have grounds to assert metaphysical incompatibility. However, one must be careful not to confuse what is merely a strong prejudice with a conceptual truth, as that could lead one to mistakenly claim that there is conceptual incompatibility.

On the other hand, if we have conceptual or epistemic compatibility, that does not by itself give us metaphysical compatibility. Yet, as we saw, there can be a temptation to use epistemic or conceptual compatibilism as a basis for drawing conclusions that really need the support of metaphysical compatibilism. In just this way it can be tempting—though I believe misguided—to accept Fischer's view that it counts in favor of a theory of moral responsibility that it does not imply that determinism precludes responsibility. Similarly, it can be tempting to follow Lewis in holding that if God existed, then God could causally determine how creatures exercise free will. It both cases, I have argued, the temptation should be resisted.

We saw that sometimes the verification of a prejudice can lead us to knowledge of metaphysical necessities. This suggests that an exploration of prejudices, especially what I have called the prejudices of freedom, may help to indicate a pathway to finding metaphysical incompatibilities. For example, if we were to verify that in the actual choices underlying our moral responsibility we exercise agent causation in a way that involves strong regulative control, that might be an indication that responsibility is metaphysically incompatible with causal determination. But, as we saw, even if a prejudice of freedom were verified there may be special difficulties that arise here in an attempt to move from the verification of a prejudice to the recognition of a metaphysical necessity.

There is much work to be done, but I hope that the ideas we have been considering may go some distance toward untangling the modal issues involved in compatibilism.

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Aristotle. 1984. History of Animals. In *The Complete Works of Aristotle: The Revised Oxford Translation*, ed. J. Barnes. Princeton: Princeton University Press.

Clarke, R. 2019. Free Will, Agent Causation, and "Disappearing Agents". *Noûs* 53(1): 76-96.

Eddington, A. 1933. *The Nature of the Physical World.* New York: The MacMillan Company.

Fischer, J. M. and M. Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

Fischer, J. M. 1999. Recent Work on Moral Responsibility. *Ethics* 110(1): 93-139.

Fischer, J. M. 2006. *My Way: Essays on Moral Responsibility*. New York: Oxford University Press.

Fischer, J. M. 2012. Semicompatibilism and its Rivals. *Journal of Ethics* 16(2): 117-43.

Fischer, J. M. 2016. Libertarianism and the Problem of Flip-flopping. In *Free Will and Theism: Connections, Contingencies, and Concerns*, ed. K. Timpe and D. Speak, 48-61. Oxford: Oxford University Press.

Flew, A. 1955. Divine Omnipotence and Human Freedom. In *New Essays in Philosophical Theology*, ed. A. Flew and A. McIntyre, 144-69. New York: Macmillan.

Gómez-Torrente, M. 2011. Kripke on Color Words and the Primary/Secondary Quality Distinction. In *Saul Kripke* ed. A. Berger, 290-323. Cambridge: Cambridge University Press.

Heller, M. 1996. The mad Scientist Meets the Robot Cats: Compatibilism, Kinds, and Counterexamples. *Philosophy and Phenomenological Research* 56(2): 333-337.

Kane. R. 2007. Response to Fischer, Pereboom, and Vargas. In *Four views on free will*, ed. J. M. Fischer, R. Kane, D. Pereboom, and M. Vargas, 166-183. Oxford: Blackwell Publishing.

Kripke, S. 1980. *Naming and Necessity*. Cambridge: Harvard University Press.

Lewis, D. 1993. Evil for Freedom's Sake? *Philosophical Papers* 22(3): 149-172.

O'Connor, T. 2000. *Persons and Causes*. New York: Oxford University Press.

Putnam, H. 1970. Is Semantics Possible? *Metaphilosophy* 1(3): 187-201.

Stebbing, L.S. 1937. *Philosophy and the Physicists.* London: Methuen & Co. Ltd.

Steward, H. 2012. *A Metaphysics for Freedom*. Oxford: Oxford University Press.

Stump, E. 1999. Alternative Possibilities and Moral Responsibility: The Flicker of Freedom. *The Journal of Ethics* 3(4): 299-324.

van Inwagen, P. 1983. *An Essay on Free Will*. Oxford: Oxford University Press.