THE COMPATIBILIST FALLACY

Jaap Hage
Universities of Maastricht and Hasselt
jaap.hage@maastrichtuniversity.nl

1. Introduction

There is an issue with free will and responsibility. Some believe that humans lack a free will and that free will is a necessary condition for responsibility. The conclusion they validly draw from these two premises is that humans cannot be responsible for their doings. Others believe that humans can be, and normally are, responsible for what they do and in support of this belief they either assume that humans do have the free will that is necessary for responsibility, or that free will is not necessary for responsibility. These others are called 'compatibilists', because they assume that responsibility is compatible with a lack of free will.

The main conclusion of this article will be that compatibilists are right and wrong at the same time. They are right in their claim that responsibility is compatible with the absence of free will, but they are wrong to assume that compatibility can be founded on our social practice. This assumption would involve the naturalistic fallacy, and the compatibilist fallacy would be the nth instantiation of this naturalistic fallacy.

The argument that leads to this conclusion that compatibilists are both right and wrong is based on the starting point that there are two fundamentally different ways of looking at humans as agents and at their acts. One way starts from the way in which people subjectively experience their acts, including their own role as agents who perform these acts. I will call this the 'phenomenological view'. The second way starts from the facts as they can be established by the sciences, facts which are assumed to be independent of our knowledge of them or the way we experience them. I will call this the 'realist view'. The main message of this article is that these two views are hard to combine into a single approach to responsibility, but that a separation, as compatibilists propose, is not well possible.

The argument of this article is structured as follows. In the sections 2 and 3, the realist and the phenomenological view of acts and agency are described. The problems that result if one attempts to mix the two views are illustrated in section 4 by means of the paradox that both the assumption and the denial of determinism lead to the conclusion that there cannot be responsibility based on free will. Compatibilism, the attempt to safeguard the phenomenological view by keeping it separated from the realist view is discussed in the sections 5-7. The article is concluded in section 8.

2. The realist view

As human beings, we have experiences. Some of our experiences, such as anger, free floating anxiety, or nausea, are pure experiences, which means that they are not experiences about something else. However, some other experiences are about something else. A persons sees a chair, fears an

exam, is indignant about the way he[1] was treated, or doubts whether he will catch the train. This 'aboutness', which philosophers call intentionality (Jacob 2014), is reflected in the experience itself, as when a person sees a chair, or hears a song by Placebo, and not merely has a 'chair-experience' or a 'music of Placebo-experience'. We will call these experiences intentional experiences.

Many of our intentional experiences are sensory experiences, which means that we experience them as being brought to us via our senses. We hear, see, feel, smell or taste something. Perhaps it is the intentionality of our sensory experiences that has made us postulate the existence of an external world which we experience by means of our senses. This external world causes – at least that is what we assume – our sensory experiences and through those experiences our beliefs about what is real. Building upon such beliefs about the external world we erect comprehensive theories about what this external world must be, including beliefs about the laws that connect events in the external world.

Realism is a position in ontology according to which things exist independent of, amongst others, our knowledge or beliefs about them (Miller 2014). People tend to be realists about some parts of their knowledge, and non-realists about other parts. For example, people tend to be realists about cars, chairs, other people, mountains, and seas, and about many of their characteristics. Many people are non-realists with regard to the taste of food, moral rightness of acts, the beauty of works of art, the quality of football-matches, and experiences such as pain, joy and sense experiences.

The realist view is characterized by its emphasis on reality, where reality consists of those objects in the world and those characteristics of these objects about which people tend to be realists. This reality can then be opposed to the world, which is then taken to be a more comprehensive collection. The world, as it is defined here, consists of everything that is described by true descriptive sentences. For instance, the world would contain organizations, leaders, money, torture, crimes, sounds, colors, causes and effects, cruel acts, and beautiful paintings, while none of these would make it to reality because they cannot exist independent of human recognition or experience.

Because reality is abstracted from the experiences that gave rise to it, reality is assumed to be the same for everybody. Obviously, people may disagree about what is really the case, but that would be a sign that at least one of them is wrong. If something can be different for different persons, this is a sure sign that it is not part of reality.

The basis of reality maybe found in those things that we experience through our senses, but that is not the whole story. The beliefs that are based on sensory experience are the foundation for the elaborate mental constructions that go under the name 'theories'. For instance, we have theories about the life of dinosaurs which are amongst others based on physical objects which we consider to be remnants of these animals. There is a huge distance between our sensory experiences of what we believe to be dinosaur bones and our theories about how dinosaurs lived. Still we would consider the objects of these latter theories as parts of reality. The same holds for the distance between our theories about sub-atomic particles and the sensory experiences on which they are based, and also for our theories about the functioning of (clusters of) neurons and our sensory experiences of different kinds of brain scans.

Two characteristics of our practice of theorizing are important to mention in the present connection. The one is the attempt to find regular connections between elements of theories. 'Regular connec-

[1] In this article I follow the convention that references to persons whose gender is not relevant should reflect the gender of the author.

tions' is an expression that stands for what are usually called physical or causal laws. The expression is introduced to avoid the connotation of one thing bringing about another thing and of the idea of manipulation which attaches to the latter way of description. This bringing about or manipulation cannot be perceived in reality as Hume as made clear to us. When this element is stripped away, regular connections between elements of theories remain, and the existence of these connections is a reason for adopting a theory. Lack of regular connections to other things –or, to state it more traditionally, absence from the chain of causes and effects - may be a reason to deny things a place in reality.

The other characteristic is reductionism. High level theories can sometimes be derived from lower level theories. An example would be the derivation of Kepler's laws of planetary motion from the Newtonian theory of gravitation. The possibility of such derivations lends credibility to a realism about the elements of the low level theories, such as gravitational mass, and even to the belief that the lowest level elements of theories are the most 'real' ones, and that higher level elements merely supervene upon the lower level ones.

These two characteristics, regular connections and reductionism, are important for our present purposes, because they seem to take away the room for entities and relations which figure in our experiences, such as selves, causation, acts, agency and responsibility, but which do not fit into the picture of reality sketched by scientific theories.

3. The phenomenological approach

Our experiences themselves are tinged with feelings and emotions, but our theories about the external world distinguish between those aspects of our experiences that are caused by 'real' things and events and the aspects that do not stem from the external world but which somehow have been added by our minds. Classic examples from the history of philosophy of what has been attributed to our minds are secondary properties such as sound and colors (Locke), obligatoriness and valuation (Hume), causality itself (again Hume, and Kant), and space and time (Kant). We distinguish between what is 'objective' or 'real' in the sense of belonging to reality, and 'subjective', in the sense of being added by our minds. If the things we allegedly add by our minds are nevertheless ascribed to the outside world, we say that these phenomena are 'projected' onto the world (Joyce 2009). The phenomenological approach to knowledge, including self-knowledge, is characterized by its focus on the world as experienced, not on the reality which we take to underlie many of our experiences (Smith 2013).

People tend to experience themselves in, amongst others, the experience of doing something. The paradigm of this phenomenon in the philosophical literature is Descartes' argument in which he derived his own existence from *his* thinking: '*Je pense, donc je suis*' (Descartes 1973, first meditation). Characteristically, Descartes does not experience his self *tout court*, but he experiences himself as thinking. From this experience he derives that there must be a thinking subject, a self, although he does not call it that way (but rather a *res cogitans*). In a similar way, people experience themselves as performing different kinds of acts, such as reasoning, listening, walking, whistling, or closing a door. In all these experiences an acting self plays a role, and the existence of this self can be derived, in the vein of Descartes, from its role in action.

In particular when the self is experienced in thinking or doing, the experience includes a sense of control. It is not a thought occurring to a self as a pain would; it is the person himself doing the think-

ing.[2] Similarly, it is the person or self that listens, walks, whistles, or closes the door. Moreover, as the example of closing the door illustrates, the self is also the originator of causal chains. It is a self who closes the door, makes an end to the draught, and avoids catching a cold. This involvement of a self distinguishes typical acts from events that occur to somebody. If somebody falls from the stairs, it is literally some body that falls from the stairs, but if somebody runs down the stairs it is a person, a self, who does the running, not his body, even though the running consists of bodily movements.[3] Acts and agency occur when they are ascribed to events. An event counts as an act if this act-status is ascribed to it, and this ascription goes hand in hand with the identification of an agent who performed the act: no act without agent. Moreover, the starting point for the ascription of agency is the experience of oneself acting. This experience can be extrapolated to other entities to which agency is ascribed: first and foremost other human beings, but also (higher) animals, organizations, and even computer programs as the word-processor which formats my text as I type it.[4]

Many will assume that ascription of agency to a computer program is merely metaphorical. That may well be the case, but it raises the urgent question what distinguishes agency that is merely metaphorically ascribed from agency that is 'really' ascribed. If agency were 'real', rather than ascribed, there would be a simple test to distinguish between metaphorical ascription and non-metaphorical description. However, if agency is ascription all over, the difference between metaphorical and non-metaphorical ascription is in need of further substantiation.

The step from the experiences of a single person to the collective ascription of acts and agency is crucial for the phenomenological approach. What happens is that the starting point, subjective experiences such as the experience of person of himself doing something, is transformed in ascription of characteristics that are admittedly not to be found in reality (in the technical sense used here) and which are somehow bound to personal experiences, but which are nevertheless not anymore experiences themselves. When people ascribe acts and agency, they do not describe personal experiences, but the ascribed acts and agency are to be found in a world that is basically a world as experienced, a meaningful world.[5]

Causation, to the extent that it is considered to be more than mere regularity, also finds its basis in our experience of a self that manipulates its environment. As Hume pointed out, our sensory experience cannot provide us with more than mere regular succession of kinds of events.[6] In particular it cannot give us a necessary connection between cause and effect. And yet, in obvious cases – colliding billiard balls might be such a case – we experience the one event as bringing about the other event. A crowing rooster does not bring about the sunrise, but my pushing the vase does bring about its tumbling. In the former case there may be regular succession, but there is no causal connection,

---

[2] For a Buddhist inspired different view, cf the songtext by The Beatles (Across the Universe): 'Pools of sorrows, waves of joy, are drifting through my open mind, possessing and caressing me'. Even there we find a 'me' that is possessed and caressed, but this 'me' is not in active control.

[3] The idea that a body, rather than a person, acts is rightly denounced by Bennett and Hacker as the 'mereological fallacy' (Bennett and Hacker 2003; see also Pardo and Patterson 2013).

[4] Many will assume that ascription of agency to a computer program is merely metaphorical. That may well be the case, but it raises the question what distinguishes agency that is merely metaphorically ascribed from agency that is 'really' ascribed.

[5] The idea that the world consists of meaningful facts, rather than of facts on which meaning is projected, takes a central place in my doctoral dissertation (Hage 1987) and, - more accessibly – in Hage 2016.

[6] Actually, Hume (1978, I, III, II)) adds contiguity, but that is of no concern here.

while in the latter case there is causation, even though there is no regular connection. (The author does not make a habit out of pushing vases.) The experience of bringing about, just as the experience of agency, finds its origin in the experience of a self manipulating its environment and thereby causing events. This experience can then be extrapolated to other agents, including lifeless 'agents' such as earthquakes, which cause buildings to tumble.

4. The paradox of determinism

The issues about free will and responsibility find their cause in attempts to mix the realist and the phenomenological approach to agency. According to the phenomenological approach agents determine what they will do and acts are the result of decisions taken by the agent. There may be exceptions, as will be discussed briefly in section 6 as the absence of capacity control, but these exceptions are exceptional. On the realist approach there is no room for agents who decide what to do and acts that flow from these decisions. First it is difficult to make room for decision making agents in a realist theory, because it is not at all how clear how agents relate to firing neurons. Second, it is not clear how acts can play a role in the realist approach. What is an act, apart from the bodily movements that constitute it? And third, even if decisions to act and acts are given a place in a realist theory, there are reasons to doubt that the causal connection between a decision to act and the act itself really exists (Libet 2011).

Perhaps even more convincing than these theoretical considerations is that a discussion of free will and responsibility in terms of the question whether human behavior is determined by the facts of the past leads to the conclusion that free will and responsibility cannot exist, whether human behavior is determined or not. Apparently, the very fact that free will and responsibility are discussed from a realist perspective makes them disappear, independent of the findings of the realist discussion. We will discuss this in some detail, because of the light it sheds on the difference between the two approaches to agency.

4.1 Determinism

Many people argue that determinism makes responsibility impossible. Their argument goes as follows. A person can only be held responsible for acts that were the result of his free will. It must have been 'up to this person' exercising his free will whether he performed the act. If determinism holds for mental facts and events, a person's will is something that merely happens to him, not something over which he has control. Consequently, what a person does is not subject to his control either, and therefore determinism precludes responsibility. Is this correct?

Simply formulated, determinism holds that all facts and events[7] are necessitated by facts and events from the past, on the basis of regular connections. For instance, given the facts that this bar is made of iron, that it was 20 centimeters long, that it was heated during 5 minutes at a temperature of 500 degrees Centigrade, and that the air pressure was 1050 mBar (and possibly some other relevant facts), it could not have been otherwise than that the bar is now, say, 21 cm long. Given the facts as they presently are and given the regular connections that govern physical nature, there can be only one set of facts in the near future. Since the facts of the near future similarly necessitate the facts of

[7] Strictly speaking it is necessary to distinguish between facts and events and since determinism applies to both facts an events, I should properly speaking all the time write about 'facts and events'. To make the text more readable, I write instead about 'facts' or about 'events', depending on what is more suitable at that moment.

the somewhat later future, these latter facts are also determined by the present facts. Moreover, the present facts were necessitated by the facts that immediately preceded them. According to determinism, the history of the physical world is one long chain of facts that necessitate their successors in time, in accordance with physical laws.

Some people may believe that science has proven, or at least made highly plausible, that determinism is true. That is not the case. As a matter of fact, determinism is not something that can be proven because it is a theory about what is necessary, while evidence can only relate to what is actually the case. Probably it is better to consider determinism to be a paradigm, a kind of preliminary assumption of the physical sciences. We do physical science research on the assumption that all facts can be explained from other facts on the basis of physical laws, and research largely aims at finding those laws. Suppose for instance that there is some domain in which events occur that could not be predicted on the basis of what went before and that appear to be completely at random. We cannot find a law (regular connection), but still we do not believe that there is no law, but only that we did not discover it yet. The unwillingness to interpret the failure to find a law as evidence that there is no law signals that we presuppose that all events have a cause, whether we already discovered it or not. Determinism is a research strategy: interpret the impossibility to find regular connections between facts and events as a sign that we still lack relevant information. Whether this strategy is a useful one is something that needs to be established in research and it may turn out that it is a good strategy for some domains, but not for other domains.

4.2 Determinism and the mind

At first sight, determinism only applies to physical nature and not obviously to mental phenomena, such as decisions and intentions. If determinism is to be applied to mental processes too, there must be a way in which the mind is 'determined' by the brain. There are at least two ways to account for this determination. One is to *identify* mental phenomena with brain states. A mental phenomenon such as the will to push a button is on this view nothing else than the flip side of a particular brain state. The same thing can both be described in physical terms, as a brain state, and in mental terms as the will to push a button. If the brain state as a physical state is determined by earlier physical facts and events, so is the mental state, since this mental state is, on this identity theory, identical to the brain state.[8]

The other way to make mental states subject to determinism is to adopt epiphenomenalism. Epiphenomenalism is the view that mental states such as pain, anger, doubt, knowledge and the will to do something are merely side-effects of brain states.[9] A person with a certain brain state will also have this mental state, but the mental state does not affect the brain state. The relation between a brain state and the corresponding mental state is one-direction and comparable to that between the light reflecting characteristics of an item and its color. Whether some item is red or green is completely determined by the light that this item reflects. The other way round, the color of an item has no influence whatsoever on the light that the item reflects. This color is merely an 'epiphenomenon', an added characteristic, to the reflective properties. As epiphenomena of brain states, mental states would be determined completely by their underlying brain states.

---

[8] Different variants of the identity theory are discussed more elaborately in Rosenthal 1994.

[9] Epiphenomenalism is discussed more elaborately in McLaughlin 1994, and in Walter 2009.

If brain states are completely determined by earlier physical facts and regular connections, so are the mental states. Therefore, so goes the argument, determinism also applies to mental states. Mental phenomena are, according to this view, completely determined by the facts of the past and, given the past, could not have been different from what they actually are. This means that it is not up to an agent to determine what his mental phenomena will be. A person's will is determined by the past, not by the agent. therefore there is no free will and, to the extent that responsibility is based on free will, no responsibility.

4.3 If determinism is irrelevant

The argument above that determinism excludes the existence of a free will presupposes that determinism applies to mental phenomena. That presupposition may be doubted, but suppose that determinism does not apply to mental processes or states, what would that mean for the possibility of a free will? It would mean that there are brain events that are not the result of the past. Suddenly one or more neurons 'fire' without any cause, and that leads to contraction of muscles and an event that is classified as an act. Would the random nature of the neuron firing be a reason to ascribe a free will to the agent? Randomly firing neurons do not necessarily lead to a conscious phenomenon such as a will to act to begin with. But suppose that the random firing of neurons does lead to a will. Such a will would probably be experienced as a will that merely happened to the agent. He would, for instance, suddenly feel a strong urge to buy an ice cream, completely out of the blue. If he then acts on that urge, would that be a typical exercise of free will? The contrary seems true; the agent seems to be the victim of a will that merely happens to him and which he is certainly not free to adopt or reject. An uncaused will is not a free will.

4.4 The dilemma

Apparently we are stuck with a dilemma. Either our will is determined by the brain state underlying it and its causes, or it is not. In the former case, there is no free will because there is no room for an agent to choose what he will want. In the latter case there is not free will either, because his allegedly free will is something that merely happens to the agent. So it seems that purely on logical grounds there cannot be a free will.

If an argument based on one premise leads to the same conclusion as an argument based on the contradictory premise there must be something wrong.[10] A possible explanation is that the determinism and the indeterminism argument share a presupposition that is incorrect. Compare it with the public prosecutor who asks the defendant whether he did or did not spend the money he stole on a necklace for his girlfriend. It does not matter whether the defendant admits that he spent the money that way, or denies it, the defendant seems to admit that he did steal the money, which was of course the intention of the prosecutor. An incorrect presupposition can make seemingly contradictory claims both false.

Let us hypothesize that both the argument from determinism and the argument that some events are not determined share a wrong presupposition. What might this presupposition be? Possibly that both the determinist and the indeterminist story belong to the realist approach to mental phenomena. The implicit assumption is that real facts and events are tied to each other by regular connections. Where this is the case, determinism applies, while events are purely random where this determinism does not apply. This story has no room for a person who, exercising free will, intervenes in the regular connections between real facts and events. If one nevertheless tries to make room for

[10] For logicians: the possibilities that a premise is redundant or self-contradictory are ignored.

such an intervening agent, the regular connections as they would be without this agent are interrupted and instead of a free will randomness appears.

The problem at issue seems to be the mix between the phenomenological and the realist approach to acts and agency. On a realist approach, the insertion of entities from the phenomenological approach such as an intervening agent or free will can cause logical paradoxes, while on a phenomenological approach realist assumptions distort what we seem to know from experience, for instance that we are persons who most of the times freely decide what we will do. The simple solution for the dilemma that seems to be posed by determinism is to keep the phenomenological and the realist approach to agency apart, and that is exactly what so-called compatibilists do.

5. Compatibilism

Compatibilists keep the realist and the phenomenological approach to agency separated by assuming that freedom of the will is not something that exists objectively in reality, to be discovered by science, but a status assigned by human culture to exercises of the will. The assignment of the status 'free' to the will goes hand in hand with two other assignments, namely the assignment of the status 'act' to an event, and the status 'agent' to a person involved in this event.

If people attribute responsibility to an agent they hold that the agent whom they assign responsibility for an act is the one who should take the blame, or – more seldom – the praise for this act. Usually the reason is that they also attribute the act to this agent: he did it and therefore he is responsible for the act and often also for its consequences. The following quotation gives an impression (Morse 2000):

> "In brief, the law's concept of the person is a creature who acts for reasons and is potentially able to be guided by reason. […]
>
> The law's conception of the person as a practical reasoner is inevitable if one considers the nature of law. At base, law is a system of rules and standards expressed in language that are meant to guide human behavior. The law therefore presupposes that people are capable of using rules and standards as premises in the practical syllogisms that guide action. […]
>
> The law's concept of responsibility follows from its view of the person and the nature of law itself. Unless human beings are rational creatures who can understand the applicable rules and standards, and can conform to those legal requirements through intentional action, the law would be powerless to affect human behavior. Legally responsible agents are therefore people who have the general capacity to grasp and be guided by good reason in particular legal contexts. They must be capable of rational practical reasoning. The law presumes that adults are so capable and that the same rules may be applied to all people with this capacity. The law does not presume that all people act for good reason all the time. It is sufficient for responsibility that the agent has the general capacity for rationality, even if the capacity is not exercised on a particular occasion. Indeed, it is my claim that lack of the general capacity for rationality explains precisely those cases, such as infancy or certain instances of severe mental disorder or dementia, in which the law now excuses agents or finds them not competent to perform some task.
>
> The general capacity for rationality in a particular context is thus the primary criterion of responsibility and its absence is the primary excusing condition."

Morse wrote this about responsibility, but his argument can easily be expanded into an argument about free will: if people attribute agency to some person, they typically assume that this person had

a free will, because in the absence of a free will agents cannot conform to legal requirements through intentional action.

The people who attribute responsibility and free will also determine on what grounds they will do so. Responsibility and free will are not to be found in a mind-independent reality, but are the outflow of people experiencing themselves as a person doing things, and as free to determine what to do. The standards for determining whether somebody is responsible are developed in a social group from such experiences. They are part of what might be called the 'practice of agency'. This practice consists in the use of standards that determine which evens count as acts, which persons (or other entities, such as organizations) count as agents, who is responsible for which acts, which acts count as causes of which facts (including facts involving damage), and which agents are liable for which damage caused by their acts.

Because standards are not to be found in an objective reality, they can theoretically have any content. It is possible to hold an agent responsible for his own doings; it is possible to hold parents responsible for what their children did. It is possible to hold teachers responsible for what their pupils did, and to hold dog owners responsible for what their dogs did. It is also possible to hold dog breeders responsible for what dogs from their kennels did, and to hold dog breeders as a collective responsible for what any dog in the country did. And it is even possible to hold paranoid persons responsible for what they did during a psychotic attack. In short, given the 'right' standard, it is possible to hold anybody responsible for anything. All that is needed is the adoption, preferably collectively, of a standard that makes the relevant persons responsible for the relevant acts.

Logically speaking, there is nothing that prevents the adoption of a standard that makes people responsible for acts that they could not influence at all, or acts that they could not help performing because they were determined to perform them. In short, given that responsibility is the result of attribution, it is compatible with determinism. *Compatibilism is obviously true, but it is also trivially true.*

6. Dworkin's argument

If we reason from our own experiences as agents who determine what they will do, we know that we have a free will. The social practice in which we attribute a free will to agents who are not ill, drugged, or otherwise influenced in an extraordinary way is based on this experience. Implicitly this practice is based on the assumption that our judgment on the freedom of the will should take its starting point in our personal experiences. But should we make this assumption? One argument that we should make it was provided by Ronald Dworkin (Dworkin 2011, 219-252), who was one of the more influential defenders of compatibilism. It is worthwhile to take a closer look at his argument, because it provides us with a nice illustration of the compatibilist fallacy.

6.1 Causal control and capacity control

Dworkin starts his argument with the assumption that we only have responsibility when we are in control of our behavior, and for the things that we could help. This assumption seems to lead immediately to the conclusion that there cannot be responsibility if determinism is correct, because determinism seems to exclude control. To avoid that conclusion, Dworkin distinguishes between two kinds of control. Causal control only exists when a person's decisions are not determined by external forces in the way that determinism holds all behavior is. In other words, determinism makes causal

control impossible. This means that if causal control is necessary for responsibility, determinism makes responsibility impossible.

The other kind of control is capacity control. An agent has capacity control over his act if he is conscious of facing and making a decision, when no one else is making that decision through and for him, and when he has the capacities to form true beliefs about the world and to match his decisions to his normative personality - his settled desires, ambitions and convictions. The capacity control that Dworkin defines comes close to our actual practice of holding people responsible under normal circumstances and not holding them responsible if certain exceptional circumstances apply. What counts as normal and exceptional in this connection is answered by our social practice of holding people responsible.

Dworkin emphasizes, rightly, that it is not a matter of hard fact which kind of control is required for responsibility. It is in his opinion an ethical issue; the question at stake is what is the best social practice for holding people responsible. Should we require causal control, or should we require capacity control? If we require causal control and if determinism is applies to the mind, we should never hold anybody responsible anymore. Our practice of holding people responsible would not make sense then. However, if we merely require capacity control, we can continue our actual practice, perhaps with some fine-tuning to take away minor inconsistencies. So we have to choose between a practice based on causal control and a practice based on capacity control. How should we make this choice?

6.2 Interpretation

Dworkin is very much aware of the fact that the way in which this choice is made determines which kind of control is adopted as essential for responsibility. The way we choose therefore also determines whether our present practice of holding people responsible under certain circumstances makes sense. It is therefore somewhat surprising that Dworkin writes that we should make this choice by finding the *best possible interpretation of our actual practice*. According to Dworkin we should start from our present practice, try to find its underlying ideas, including its underlying image of man, even though Dworkin does not mention that explicitly. From that starting point we should try to determine which kind of control best fits our actual practice. It should not come as a surprise that capacity control best fits with our actual practice, because capacity control was *defined* as the kind of control that is required by our actual practice of assigning responsibility.

6.3 The naturalist fallacy

From a logical perspective, the argument presented by Dworkin is an instance of the fallacious derivation that something ought to be the case from the fact that it is actually the case. When all the elaborations are stripped away, Dworkin's argument boils down to it that we should choose for capacity control for our practice of assigning responsibility, because that choice fits best with our actual practice. We do it this way and therefore we should do it this way. That Dworkin's argument consists of a naturalistic fallacy does not mean that his conclusion is false; it only means that the argument that Dworkin offered for the continuation of our actual practice of assigning responsibility does not support its conclusion. It only convinces those who were already convinced to begin with.

The weakness of Dworkin's argument becomes more clear when we take a look at a similar argument about a practice which most of us would not support: drawing cards to predict the future. Suppose there exists a community in which the practice has arisen to predict the quality of an upcoming marriage by drawing playing cards. The prospective groom drinks a 'predictive potion', a magic formula is pronounced, and then the groom draws, one by one, at most five playing cards from a shuffled deck.

The rules are that if from the five cards three or more are red, the marriage will be happy, and otherwise not. However, if the very first card happens to be the Ace of Spades, the marriage will be happy anyhow, and the drawing of cards is not continued.

Suppose that this practice has existed for some time, when unexpectedly a 'hard case' arises. The first card drawn by the groom is the Ace of Hearts, and the second card is the Ace of Spades. On one interpretation of the rules, the groom should continue the drawing until he has five cards. Some, however, favor a different interpretation. The Ace of Hearts is the most important red card, and as such has clearly predictive power, they say, for a happy marriage. And then the second card is the Ace of Spades, which should have predicted a happy marriage when it would have been the first card! Such a combination surely indicates that the marriage will be happy, and continuation of the card drawing procedure would be useless.

Which side is right? If this practice of card drawing is comparable to law as Dworkin sees it, we should try to understand the practice from within. Why do people believe that red cards predict a happy marriage (ask them!) and why do they assign a special role to a single black card, the Ace of Spades, when it is drawn as the first card? We should try to find the best possible interpretation of the actual practice and then use this interpretation to determine which side is right in the dispute about the hard case. What we should NOT do according to Dworkin is to step outside the practice and to ask whether the very practice of card drawing to predict the quality of the marriage makes sense. We work within a practice and we should interpret the practice to determine what is the best way to deal with a hard case arising from it.

Not many would agree that in the case of this example we should take the practice as a whole for granted and only argue from within the practice to find the best solution for the hard case. Most would say that drawing cards to predict the quality of marriages does not make any sense and that arguing from the presumption that it does, is misguided. The proper way to deal with the hard case is to use it as an opportunity to stop doing what was nonsensical all the time! In a similar way we should ask whether the very practice of holding people responsible makes sense, and we should not answer that question by merely looking at the practice as it actually is and by giving the practice its best possible interpretation. The practice of holding people responsible should be evaluated in the light of *all* available knowledge. If that knowledge includes the applicability of determinism to mental phenomena then determinism should play a role in judging our actual practice of holding people responsible. Then we might use the notion of causal control to determine whether a person is responsible for what he did, and the outcome might be that nobody is ever responsible for any of his doings, and that the very practice of holding people responsible makes no sense. A hard case, for example a case about somebody who only seems accountable to a diminished degree, should not be seen as an opportunity to interpret our present practice, but as an opportunity to raise the question whether our existing practice as a whole makes sense.

7. The capacities approach

Dworkin's argument for the capacity approach to responsibility may be fallacious, but that does not mean that the capacities approach is wrong. We should therefore independently investigate what its virtues are. The underlying assumption of the capacities approach was formulated well by Morse (2000): Legally responsible agents have the general capacity to grasp and be guided by good reason in particular legal contexts. They must have the capacity to use rules to guide their action. This capac-

ity is general, shared by most adult humans, and therefore human beings can generally be held responsible for their doings. However, sometimes there are special circumstances which make that an agent lacks this capacity to have his conduct guided by legal rules. When such circumstances are present, this may be a reason not to hold a human agent responsible for his acts.

The test for responsibility if a legal rule was violated is, allegedly, therefore whether in the concrete case there were special circumstances that took away the agent's general capacity to be guided by the relevant rule. The same point was made more concrete by Dworkin when he assumed that an agent has capacity control over his act if he is conscious of facing and making a decision, when no one else is making that decision through and for him, and when he has the capacities to form true beliefs about the world and to match his decisions to his settled desires, ambitions and convictions.

The capacities approach is used to defend compatibilism, the view that our practice of assigning responsibilities is compatible with determinism. At first sight the compatibility of the capacity approach and determinism is obvious. According to determinism a human agent who violated a rule could not have had the capacity the obey the rule. All behavior was necessitated by regular connections and the preceding facts, and therefore the rule violation was also necessitated. There could not have been a capacity not to violate the rule. Since the human agent apparently lacked the capacity to comply with the rule, he should not be held responsible. Therefore the capacities approach and determinism lead to the same conclusion: nobody should ever be held responsible for his doings.

Clearly this is not what adherents of the capacities approach have in mind. They assume that our present practice of holding most human agents responsible for most of their acts, is right. To do so consistently, they must also assume that most human beings who violated rules in particular circumstances had the capacity to comply with these rules *under those circumstances.* Such an assumption seems incompatible with determinism, and therefore the question needs to be addressed how compatibilists can assume that the actual practice of assigning responsibilities can go together with determinism. To that purpose we must delve a little deeper into the nature of capacities and possibilities.

7.1 What is a capacity?

An agent has the capacity to do something if he can do it. But what does that mean? If Katarzyna actually undersigned her exam because the rules required that, it is obvious that Katarzyna could undersign her exam. More in general, if an agent performed some act, he had the capacity to do so. However, we are more interested in capacities in cases where an agent did not do what he had the capacity to do. If Katarzyna violated the exam rules and did not undersign her exam, how can it be established whether she had the capacity to undersign?

Capacities, and more in general, possibilities are the most interesting in case they were not realized. It is notoriously difficult, however, to establish the existence of possibilities, including capacities, in case they were not realized. To deal with this problem, a thinking device was constructed: possible worlds theory.[11] The basic idea underlying possible worlds theory is that something is necessary when it is the case whatever else may be the case. For instance, whatever the other facts may be, in any case every colored object has a surface and whatever the other facts may be, the number 5 is bigger than the number 3. Therefore, necessarily every colored object has a surface and necessarily 5

[11] The idea of possible worlds theory can be traced back at least to the German Philosopher Leibniz (1646-1716), who in his Theodicee defined necessity as that what is the case in all possible worlds.

is bigger than 3. A different way of expressing that something is the case whatever else may be the case is to say that it is the case *in all possible worlds*. In all possible worlds every colored object has a surface and in all possible worlds the number 5 is bigger than the number 3.

The real world consists of all the facts as they actually are, while a different possible world contains a set of all facts as they might have been under different circumstances. Actually - in the real world – Bartosz has brown hair, but under different circumstances, in some other possible world, Bartosz is red-headed. Because there is some alternative possible world in which Bartosz is red-headed, it is possible that Bartosz is red-headed. In fact, he is not, but he might have been. Something is possible if it is the case in some possible world. That may be the actual world, but that is not necessary. In the actual world, Katarzyna undersigned her exam, but in some other possible world she did not. Therefore, actually Katarzyna undersigned, but it would have been possible that she did not undersign. This captures the notion of a capacity quite well. *We might say that an agent has the capacity to do something if there is a possible world in which the agent does it.* That would mean that Katarzyna has the capacity to undersign her exam if there is some possible world in which she undersigned her exam.

7.2 Possible worlds and constraints

We now have a definition of what it means that a person has a certain capacity, but it may seem that this definition has replaced one problem – the nature of capacity – with another problem, the nature of a possible world. What makes a set of facts a possible world? Here the notion of a constraint plays a role.[12] Not all sets of facts can go together. To give an obvious example: the fact that it is raining (here and now) cannot go together with the fact that it is not raining. Incompatible facts cannot be part of one and the same possible world. That is a constraint on possible worlds. A logical constraint in this case, because it is a matter of logic that a fact and its denial cannot go together. Next to logical constraints, there can also be physical constraints. The laws of physics can be interpreted as constraints on worlds that are physically possible. It is, for instance, physically possible that a metal bar is red, but it is physically impossible that a metal bar is heated but does not expand. There is no physically possible world, no world that satisfies all the physical constraints, in which a metal bar is heated but does not expand. And neither is there a physically possible world in which something travels faster than light in vacuum.[13]

We are now in a position to define possible worlds more precisely. A possible world is a world that satisfies a set of constraints. A logically possible world satisfies the laws of logic; a physically possible world satisfies the laws of physics. A world that is both logically and physically possible needs to satisfy boh sets of constraints. A particular world counts as possible if it satisfies one or more sets of constraints. Only relative to constraints does it make sense to ask whether something is possible or necessary. Necessity or possibility *tout court*, without being made relative, does not make sense. Every time when somebody claims that something is possible, it is legitimate to ask relative to which

---

[12] The notion of a constraint as used here is closely related to that of a regular connection. This is not the place to explore similarities and differences between the two, however.

[13] Obviously these examples of physical possibility work with generally available knowledge of physical laws. This knowledge may turn out to be false, and then our ideas about what is physically necessary or possible turn out to be false too. This goes to show that necessity and certainty are not the same things. Something may be uncertain, but if it is true, also necessarily true. See Kripke 1972.

set of constraints it is possible. If the set of constraints cannot be specified, the claim about possibility is too obscure to make sense.

Both logically and physically it is possible that Bartosz is red-headed, but is it still possible if we take into consideration that Bartosz just finished dying his hair brown? That is apparently not the case, and it is worthwhile to consider more closely why that is not the case. Both with logical and with physical necessity (and possibility) the necessity is the result of constraints that consist of laws (regular connections), the laws of logic and of physics respectively. A law expresses a necessary general connection between types of facts, for instance the type of fact that something is a metal bar that is being heated and the type of fact that this something expands. When we speak of possible worlds, such laws are the most obvious constraints to take into account. However, it is not necessary to take only laws into account as constraints. There is no fundamental reason why particular facts should not be considered as constraints too. One such a fact might be that Bartosz just finished dying his hair brown. Given that fact, it is necessarily the case that Bartosz's hair is brown, and impossible that his hair is red. And given the fact that the train Dobrochna was on departed five minutes ago, it is impossible that she was seen at the railway station one minute ago. In particular in connection with the claims of determinism it is important not to take only laws into account as constraints on possible worlds, but also facts. If it is claimed that Katarzyna could not help submitting the exam without having undersigned it, this claim will probably not only be based on the laws of nature (purely physical necessity), but also on facts concerning Katarzyna's personal history.

7.3 The relativity of capacity

An agent has the capacity to do something if there is a possible world in which the agent does it. Now we know that this specification of capacity is still too vague: we also need to specify relative to which set of constraints the capacity exists. The crucial question is: which set of constraints should be taken into account in determining whether a particular agent had the capacity to perform some act, or to refrain from performing it. Here I will not attempt to answer this question in abstract, but focus merely on the characteristics of individual agents.

It is clear that in determining the capacities of a particular agent, we should take some personal characteristics of this agent into account. Going only by the laws of physics which are the same for everybody, every agent would have the same capacities. That would be an unattractive finding, and to avoid it, we must take personal characteristics into account in determining which capacities some agent has. But which personal characteristics should be taken into account? If the agent cannot write, we should most likely take that into account. So if Katarzyna could not write, she did not have the capacity to undersign her exam and most likely she should not be held responsible for not undersigning it.[14] Should we also take into account that the agent was strongly motivated to violate a norm? Suppose that a kidnapper held Katarzyna's baby and required from Katarzyna that she would not undersign the exam. Almost paralyzed by fear that something would happen to her baby, Katarzyna does not undersign. Did she have the capacity to sign? Would that have been different if Katarzyna was a drug addict who could only score if she did not undersign the exam? If we want to distinguish between the latter two cases, would that be a distinction based on a moral judgment regarding what *ought to* motivate Katarzyna?

[14] That might be different if it was Katarzyna's fault that she cannot write, but for now I will ignore the possibility of responsibility without capacity.

Stepping back from this casuistry, the general issue raised by determinism is the following: if all facts regarding an agent are taken into account, as well as all physical laws, the only thing that an agent could do is what he actually did. The distinction between what an agent did and what he had the capacity to do makes only sense if not all facts are taken into account as constraints on what is possible. Then the question arises which facts should be taken into account, and which facts should not. Capacity becomes a normative issue, the issue which facts *should* be left out of consideration to determine what else the agent could have done next to what he actually did. Perhaps this seems an acceptable approach; after all it is what lawyers are actually doing when they ask whether a criminal suspect could have acted differently than he actually did. We should realize, however, that if we make capacity a normative notion, we cannot anymore adduce the capacity of an agent as a reason for holding the agent responsible. What we actually do is to give one single normative judgment concerning both the capacities and the responsibility of the agent. Either we judge the agent to have the relevant capacities and to be responsible, or we judge him to lack the capacities and not to be responsible. This judgment cannot be founded in the capacities of the agent, because these capacities are themselves part of the judgment.

The last observation that there may be a single judgement, covering both the presence of a capacity to have acted differently and the assignment of responsibility for the act that was actually performed, touches the core of the compatibilist approach. The argument from determinism to the conclusion that there can be no responsibility in a sense works from the 'bottom' upward: everything is determined, therefore an agent does not have a free will and therefore the agent could not have acted differently and therefore the agent cannot be held responsible. The 'bottom' of this argument is taken to express a hard fact about the world we live in and the rest follows from this hard fact.

Compatibilists work in a fundamentally different way. Responsibility is something we attribute to agents, and in doing so we attribute the status of an act to an event that took place, the status of agent to the person who was causally involved in bringing about this event, we attribute free will to the agent and with free will also the capacity to have acted differently. These are in the compatibilist view not different argument steps that build upon each other, but one act of assigning meaning to ourselves and to the world that surrounds us. This meaning encompasses acts and agency, free will and capacities, responsibility and liability.

7.4 Concerning the capacities approach

We saw that Dworkin's argument for the capacities approach to responsibility rested on a naturalistic fallacy: this is how it is done and therefore this is how it should be done. In this section we took a closer look at the capacities approach to see whether it is attractive apart from Dworkin's argument. Central to the capacities approach is the assumption that human beings are normally capable to comply with the rules of law which constitute for them reasons for action, they are 'reason-responsive', and that therefore they should normally be responsible for norm violations. However, there may be special circumstances in which the agent lacks the capacity to comply with the applicable rules, and that would be a reason not to hold the agent responsible for eventual violations.

The central question is what it means that an agent lacked the capacity to comply with a norm. If the capacities approach is to lead to different results than the approach, inspired by determinism, not to hold anybody responsible under any circumstances, it must assume that sometimes an agent violated a norm even though he had the capacity to comply. We saw that this assumption would only cut ice if, in determining the capacities of an agent, not all facts about the agent are treated as con-

straints on what counts as possible. Some facts should be left out of consideration to allow the agent the 'freedom' to choose between norm compliance and norm violation.

The problem here is that there are no obvious criteria to determine which facts should, and which facts should not be treated as constraints on what the agent could do. If the choice which facts are treated as constraints is the outcome of normative decision making, it is not possible anymore to adduce the capacities of the agent as a reason for holding the agent responsible. Doing this anyway would amount to a circular argument along the following line: we want to hold the agent responsible for what he did and therefore we do not treat the facts that caused him to violate the norm as constraints that define the agent's capacities. For now we may conclude that the capacities approach to holding agents responsible is the outcome of normative decision making without foundation in an independent notion of capacity. The argument based on determinism that our practice of holding people responsible does not make sense therefore applies equally to the capacities approach as to the traditional image of man as rational decision maker. This should not come as a surprise, since the capacities approach is based on this traditional image of man.

8. Conclusion

There are two ways to look at the world or reality, the phenomenological and the realist way. And if we try to kook at acts and agency in both ways at the same time, the results may be paradoxical. This is illustrated by the example of free will and determinism. Working from the realist point of view determinism either applies or does not apply to agency. However, in neither case there is room for a free will or the responsibility based upon it. Acts are either determined through the determination of the will, and then there is no free will. Or the will underlying acts originates in an arbitrary way, and then there is no free will either, or – if one wants to call an arbitrary will free – that kind of free will cannot be a basis for personal responsibility.

There seem to be two alternatives for such a mixed, and paradox-generating, view of agency. One is to adopt a strict realist perspective and only allow those entities in one's views of reality of which one can seriously believe that they are mind-independent. That would imply that acts, agency, free will and responsibility disappear from one's picture of reality. The problem has been solved by defining away the entities that make the problem possible.

The second alternative is to assign an independent realm to phenomenological entities. The relations between these entities are defined in terms of how we experience them (an agent feels free to decide what to do) and in terms of attribution by means of social standards (an act is only an act if we count it as an act). This is a compatibilist approach, and a prominent version of it is the capabilities approach that was adopted by, amongst others, Morse and Dworkin. The capabilities approach to acts, agency, free will and responsibility makes these phenomena by definition compatible with the facts of a mind-independent reality. Compatibilism is true, but trivially so. The question remains unanswered whether the social practices which define what we count as acts, as agents, as free will and as responsibility make sense. The justification of the practice by the invocation of the practice itself would be a variant on the naturalistic fallacy, a variant that we might call the compatibilist fallacy. The question whether the agency practice makes sense will not be answered in this article, at least not in more details than the general observation that one's view of reality and one's view of mind-dependent entities need to be coherent. Separating the two domains without giving other reasons than that is what we actually do, does not lead to such a coherent theory.

References

Bennett and Hacker 2003

M.R. Bennett and P.M.S. Hacker, *Philosophical Foundations of Neuroscience*, Oxford: Blackwell 2003.

Davidson and Harman 1972

Donald Davidson and Gilbert Harman (eds.), *Semantics of Natural Language*, 2nd ed., Dordrecht: D. Reidel Publishing Company.

Descartes 1973

René Descartes, M*editations Metaphysiques*, (1st edition 1641), Paris: Larousse

Dworkin 2011

Ronald Dworkin, *Justice for hedgehogs*, Cambridge: harvartd University Press 2011.

Guttenplan 1994

Samuel Guttenplan (ed.), *A Companion to the Philosophy of Mind*, Oxford: Blackwell 1994.

Hage 2016

Jaap Hage, 'Facts and Meaning. How a rich ontology facilitates the understanding of normativity', in Stelmach, Brozek and Kurek 2016,13-44.

Jacob 2014

Pierre Jacob, 'Intentionality', The Stanford Encyclopedia of Philosophy (Winter 2014 Edition), Edward N. Zalta (ed.), http://plato.stanford.edu/archives/win2014/entries/intentionality/.

Joyce 2009

Richard Joyce, 'Moral Anti-Realism', *The Stanford Encyclopedia of Philosophy* (Summer 2009 Edition), Edward N. Zalta (ed.), http://plato.stanford.edu/archives/sum2009/entries/moral-anti-realism/.

Kripke 1972

Saul A. Kripke, 'Naming and Necessity', in Davison and Harman 1972, 253-355.

Libet 2011

Bejnamin Libet, 'Do We Have Free Will?', in Sinnot Armstong and Nadel 2011, 1-10

McLaughlin 1994

Brian P. McLaughlin, 'Epiphenomenalism', in Guttenplan 1994, 277-288.

McLaughlin, Beckermann and Walter 2009

Brain P. McLaughlin, Ansgar Beckermann, and Sven Walter (eds.), *The Oxford Handbook of Philosophy of Mind*, Oxford: Clarendon Press 2009.

Miller 2014

Alexander Miller, 'Realism', *The Stanford Encyclopedia of Philosophy* (Winter 2014 Edition), Edward N. Zalta (ed.), <http://plato.stanford.edu/archives/win2014/entries/realism/.

Morse 200

Stephen J. Morse, 'Rationality and responsibility', *Southern California Law Review*, 2000, 251-268.

Pardo and Patterson 2013

Michael S. Pardo and Dennis Patterson, *Minds, Brains and Law*, Oxford: Oxford University Press 2013.

Rosenthal 1994

David M. Rosenthal, 'Identity Theories', in Guttenplan 1994, 348-355.

Sinnott-Armstrong and Nadel 2011

Walter Sinnott-Armstrong and Lynn Nadel (eds.), *Conscious Will and Responsibility*, Oxford: Oxford University Press 2011.

Smith 2013

David Woodruff Smith, 'Phenomenology', in *The Stanford Encyclopedia of Philosophy* (Winter 2013 Edition), Edward N. Zalta (ed.), http://plato.stanford.edu/archives/win2013/entries/phenomenology.

Stelmach, Brożek and Kurek 2016

Jerzy Stelmach, Bartosz Brożek and Lukasz Kurek (eds.), *The Emergence of Normative Orders*, Kraków: Copernicus Press 2016.

Walter 2009

Sven Walter, 'Epiphenomenalism', in McLaughlin, Beckerman and Walter 2009, 85-94.