THE COMPATIBILIST FALLACY

Jaap Hage

Professor of Law at the Faculty of Law, University of Maastricht

E-mail: jaap.hage@maastrichtuniversity.nl

1 Introduction

There is an issue with free will and responsibility. Some believe that humans lack 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 necessary for responsibility, or that free will is not necessary for responsibility. The latter are called 'compatibilists', because they assume that responsibility is compatible with a lack of free will.

The main conclusion of this article is 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 practices. Their assumption involves the naturalistic fallacy, and the compatibilist fallacy is then the nth instantiation of the naturalistic fallacy.

The argument that leads to the 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 call this the 'phenomenological view'. The other way starts from facts as established by the sciences, facts which are assumed to be independent of our knowledge of them or the way we experience them. I call this the 'realist view'. The main message of this article is that it is difficult to combine these two views into a single approach to responsibility, and that it is not possible to separate the two as has been proposed by compatibilists.

The argument of this article is structured as follows. In sections 2 and 3, the realist and the phenomenological views of acts and agency are described. The problems that arise 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, an attempt to safeguard the phenomenological view by keeping it separate from the realist view, is discussed in sections 5 to 7. The article's conclusion is brought 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 person sees a chair, fears an exam, is indignant about the way he[1] has been treated, or doubts that he will catch the train. This 'aboutness', which philosophers call intentionality (Jacob 2014), is reflected in the experiences themselves, as when a person sees a chair or hears a song by Placebo; a person does not merely have a 'chair-experience' or a 'song-by-Placebo-experience'. We 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

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

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 this 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 like, including beliefs about the laws that connect events in the external world.

Realism is a position in ontology, according to which things exist independently of, amongst other things, our knowledge or beliefs about them (Miller 2014). People tend to be realists about some parts of their knowledge and non-realists about some 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, the 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 characterised 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 here defined, consists of everything that is described by true descriptive sentences. For instance, the world contains organisations, leaders, money, torture, crimes, sounds, colours, causes and effects, cruel acts, and beautiful paintings, but none of these make it to reality because they cannot exist independently of human recognition or experience.

Because reality is abstracted from the experiences that give 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 to different persons, this is a sure sign that it is not part of reality.

The basis of reality may be found in the things that we experience through our senses, but that is not the whole story. Beliefs based on sensory experiences are the foundation for elaborate mental constructions that fall under the name ‘theories’. For instance, we have theories about the life of dinosaurs, which are, amongst other things, based on physical objects which we consider to be remnants of these animals. There is a huge gap between our sensory experiences of what we believe to be dinosaur bones and our theories about how dinosaurs lived. Still, we consider the objects of these theories to be part of reality. The same holds true for the gap between our theories about subatomic particles and the sensory experiences which they are based on, as well as for our theories about the functioning of (clusters of) neurons and our sensory experiences of different kinds of brain scans.

With regard to the above, it is important to mention two characteristics of our practice of theorising. One is the attempt to find regular connections between elements of theories. ‘Regular connections’ is an expression that stands for what are usually called physical or causal laws. The expression is introduced to avoid the connotation that one thing brings about another, and that 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 made clear. When this element is stripped away, regular connections between elements of theories remain, and the existence of these connections is a reason to adopt a theory. A lack of regular connections to other things – or, to put it more traditionally, an absence of a 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 is 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 concerning the elements of lower-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 lower-level ones.

These two characteristics, that is, regular connections and reductionism, are important for our present purposes, because they seem to do away with entities and relations which figure in our experiences, such as the self, 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 the aspects of our experiences which are caused by 'real' things and events, and the aspects which do not stem from the external world, but which have somehow 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 sounds and colours (Locke), obligatoriness and valuation (Hume), causality itself (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 characterised by its focus on the world as experienced, and not on reality which we take to underlie many of our experiences (Smith 2013).

People tend to experience themselves in, amongst other things, the experience of doing something. The paradigm of this phenomenon in philosophical literature is Descartes's argument in which he derives his own existence from *his* thinking: "*Je pense, donc je suis*" (Descartes 1973, first meditation). Characteristically, Descartes does not experience his self *tout court*; rather, 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 so (but, rather, a *res cogitans*). Similarly, 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 his role in action.

More specifically, when the self is experienced in thinking or doing, the experience includes a sense of control. What occurs to a self is not a thought in the way that pain would; what occurs is the person himself doing the thinking.[2] Similarly, it is the person or the self that listens, walks, whistles, or closes a door. Moreover, as the example of the closing of a door illustrates, the self is also the originator of causal chains. It is the self who closes the door that puts an end to a draught, and thus avoids catching a cold. It is this involvement of the self that distinguishes typical acts from events that occur to somebody. If somebody falls down a flight of stairs, it is literally some body that falls down a flight of stairs. But if, on the other hand, somebody runs down the stairs, it is a person, a self, who does the running, and not his body, even though this running does consist of bodily movements.[3] Acts and agency occur when they are ascribed to events. An event counts as an act if an act-status is ascribed to it, and this ascription goes hand in hand with identifying the agent who performed the act: there is no act without an agent. Moreover, the starting point for the ascription of agency is the experience of oneself acting. This experience can be extrapolated to other entities that

[2] For a different, Buddhist-inspired view, cf. the lyrics of *Across the Universe* by The Beatles: "Pools of sorrow, waves of joy, are drifting through my opened mind, possessing and caressing me". Even here, there is a 'me' that is possessed and caressed, but this 'me' is not in active control.

[3] The idea that acts are performed by bodies and not by persons is rightly denounced by Bennett and Hacker as the 'mereological fallacy' (Bennett & Hacker 2003; see, also, Pardo & Patterson 2013).

agency is ascribed to: first and foremost, other human beings, but also (higher) animals, organisations, and even such computer programmes as the word processor which formats my text as I type it.

Many will assume that the ascription of agency to a computer programme is merely metaphorical. This may well be the case, but it does raise the urgent question of what distinguishes between agency which is merely metaphorically ascribed and agency which is 'really' ascribed. If agency were 'real' rather than ascribed, there would be a simple test to distinguish between metaphorical ascription and non-metaphorical ascription. However, if agency is in all cases a matter of 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 – that is, subjective experiences, such as a person's experience of himself doing something – is transformed in the ascription of characteristics which are admittedly not found in reality (in the technical sense) and which are somehow bound to personal experiences, but which are, nevertheless, no longer experiences as such. When people ascribe acts and agency, they do not describe personal experiences; rather, the ascribed acts and agency are found in a world which is basically a world as experienced, a meaningful world.[4]

Causation, to the extent that it is considered to be more than mere regularity, also finds its basis in our experience of the self that manipulates his environment. As Hume points out, our sensory experience cannot provide us with more than a mere regular succession of kinds of events.[5] 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 one event as bringing about another event. A crowing rooster does not bring about the sunrise, but my pushing a vase does bring about its tumbling. In the former case, there may be a regular succession, but there is no causal connection, while in the latter case, there is causation, even though there is no regular connection. (The author does not have a habit of pushing vases.) The experience of bringing about, just as the experience of agency, finds its origin in the experience of the 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 down.

4 The paradox of determinism

The issues concerning 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 made by agents. There may be exceptions, as will be discussed briefly in section 6, such as the absence of capacity control, but these exceptions are exceptional. According to the realist approach, there is no room for either agents who decide what to do or acts that come from such decisions. First, it is difficult to make room for decision-making agents in the realist theory, because it is not at all 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 if not the bodily movements that constitute it? And third, even if decisions to act and acts are given a place in the 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 the conclusion of a discussion of the free will and responsibility in terms of the question whether human

[4] The idea that the world consists of meaningful facts, rather than of facts onto which meaning is projected, is central to my doctoral dissertation (Hage 1987) and, more accessibly, in Hage 2016.

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

behaviour is determined by facts of the past. It says that free will and responsibility cannot exist, regardless of whether human behaviour is determined or not. Apparently, the very fact that free will and responsibility are discussed from the realist perspective makes them disappear, independently of the findings of the realist discussion. I 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 be held responsible only for acts that are the result of his free will. The performance of the act had to have been 'up to this person' exercising his free will. If determinism holds for mental facts and events, a person's will is something that merely happens to him, and not something which he has control over. Consequently, what a person does is not subject to his control either, hence determinism precludes responsibility. Is this correct?

Put simply, determinism holds that all facts and events[6] are necessitated by facts and events from the past on the basis of regular connections. For instance, given the facts that a bar is made of iron, that it was 20 centimetres long prior to being exposed to heat, that it was heated for 5 minutes at a temperature of 500 degrees Centigrade, and that the air pressure was 1050 mBar (and possibly some other relevant fact/s), the bar is now, say, 21 cm long, and this could not have been any different. Given the facts as they are at present 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 a somewhat more distant 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, and this in accordance with physical laws.

Some people may believe that science has proven that determinism is true or that it is at least highly plausible. That is not the case, however. 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. It is probably better to consider determinism to be a paradigm, a kind of preliminary assumption of the natural sciences. We do natural science research on the assumption that all facts can be explained from other facts on the basis of physical laws, and what research largely aims at is finding those laws. Let us suppose, for instance, that there is a domain in which events occur which could not be predicted on the basis of what came before and which appear to be completely random. We cannot find a law (regular connection), but that does not mean that we believe that there is no law, it only means that we have not yet discovered it. Our unwillingness to interpret our failure to find a law as evidence that there is no law signals that we presuppose that all events have a cause, irrespective of whether we have 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 lack (some) relevant information. Whether this strategy is a useful one is something that needs to be established in research, and it may well turn out that it is a good strategy for some domains, but not for all.

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

4.2 Determinism and the mind

At first sight, determinism applies only to physical nature and obviously not to mental phenomena, such as decisions and intentions. If determinism is to be applied to mental processes as well, 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, according to this view, nothing else but the flip side of a particular brain state. The same thing can be described both 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, according to this identity theory, identical to the brain state.[7]

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.[8] A person with a certain brain state will also have a matching mental state, but the mental state does not affect the brain state. The relation between a brain state and the corresponding mental state is a one-way street one, and is comparable to that between light reflecting characteristics of an object and its colour. Whether some object is red or green is determined entirely by the light that this object reflects. The other way round, however, the colour of an object has no influence whatsoever over the light that the object reflects. This colour is merely an 'epiphenomenon', a characteristic added to the object's reflective properties. As epiphenomena of brain states, mental states are determined completely by their underlying brain states.

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

4.3 If determinism is irrelevant

The above argument that determinism excludes the existence of free will presupposes that determinism applies to mental phenomena. This presupposition may be questioned, but if we assume that determinism does not apply to mental processes or states, what would that mean for the possibility of the existence of 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 a cause whatsoever, and that leads to the contraction of muscles and an event which is classified as an act. Would the random nature of the firing of neurons be a reason to ascribe free will to an agent? Randomly firing neurons do not necessarily lead to a conscious phenomenon, such as the will to act to begin with. But, let us 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 have a strong urge to buy an ice cream, completely out of the blue. If he then acted on this urge, would it be a typical exercise of free will? The very opposite seems to be true; the agent seems to be a victim of a will which merely happens to him and which he is certainly not free to either adopt or reject. An uncaused will is not free will.

---

[7] Different variants of the identity theory are discussed in more detail in Rosenthal 1994.

[8] Epiphenomenalism is discussed in more detail in McLaughlin 1994, and in Walter 2009.

4.4 The dilemma

Apparently, we are stuck with a dilemma. Either our will is determined by brain states 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 wants. In the latter case, there is no free will either, because his allegedly free will is something that merely happens to the agent. So, it seems that, on purely logical grounds, free will cannot exist.

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.[9] A possible explanation is that the determinist and the indeterminist arguments share a presupposition which is incorrect. This can be compared with a public prosecutor asking a defendant whether he spent the money he had stolen on a necklace for his girlfriend. It does not matter whether the defendant admits to spending the money, given that the defendant seems to admit that he had stolen the money in the first place, which was, of course, the intention of the prosecutor. An incorrect presupposition can make two seemingly contradictory claims both be false.

Let us hypothesise that the determinist and the indeterminist arguments share an incorrect 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, and when events are purely random, determinism does not apply. This story has no room for a person who, while exercising his free will, intervenes in the regular connections between real facts and events. If one, nevertheless, tries to make room for such an intervening agent, the regular connections as they are without this agent are interrupted and, instead of a free will, randomness appears.

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

5 Compatibilism

Compatibilists keep the realist and the phenomenological approaches to agency separate by assuming that freedom of the will is not something that exists objectively in reality, something to be discovered by science, but is 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 have assigned responsibility to for an act is the one who should take the blame or – more seldom – deserve praise for this act. The usual 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 of the said (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. […]

[9] For logicians: the possibility that a premise is redundant or self-contradictory is ignored.

> 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 to the free will argument: if people attribute agency to a 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.

People who attribute responsibility and free will to agents also determine the grounds on which they do so. Responsibility and free will are not found in a mind-independent reality, but are the outflow of people experiencing themselves both as persons doing things and as free to decide what to do. The standards for determining whether somebody is responsible are set 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 organisations) 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 found in an objective reality, they can theoretically have any content. It is possible to hold an agent responsible for his own doings, to hold parents responsible for what their children did, teachers 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 episode. In short, given the 'right' standard, it is possible to hold anybody responsible for anything. All that is needed is the preferably collective adoption of a standard that makes relevant persons be responsible for relevant acts.

Logically speaking, there is nothing that prevents the adoption of a standard that makes people responsible for acts that they cannot influence at all or for acts that they could not help but perform because they were determined to perform them in the first place. 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 do, we know that we have free will. The social practice in which we attribute 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 our personal experiences as its starting point. But, should we make this assumption? One argument that we should indeed make this assumption was given by Ronald Dworkin (Dworkin 2011: 219-252), one of the more influential defenders of compatibilism. It is worthwhile to take a closer look at his argument, because it provides a nice illustration of the compatibilist fallacy.

6.1 Causal control and capacity control

Dworkin starts his argument with the assumption that we have responsibility only when we are in control of our behaviour. 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 this conclusion, Dworkin distinguishes between two kinds of control. Causal control exists only when a person's decisions are not determined by external forces in the way that determinism holds that all behaviour 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 type of control is capacity control. An agent has capacity control over his acts 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 capacity to form true beliefs about the world and to match his decisions to his normative personality, that is, his settled desires, ambitions and convictions. This capacity control that Dworkin defines comes close to our actual practice of holding people responsible under normal circumstances and of not holding them responsible if certain exceptional circumstances apply. What counts as normal and exceptional in this regard is answered by our social practice of holding people responsible.

Dworkin emphasises, and rightly so, 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 which the best social practice for holding people responsible is. Should we require causal control or should we require capacity control? If we require causal control and if determinism is applied to the mind, we should no longer hold anybody responsible. Our practice of holding people responsible would not make sense then. However, if we merely require capacity control, we can continue our current practice, perhaps do some fine-tuning to get rid of 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, thus, 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 this explicitly. From then on, we should try to determine which kind of control best fits our actual practice. It should not come as a surprise that capacity control fits best with our actual practice, because capacity control was *defined* as the kind of control which 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 fallacious derivation, that is, 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 us having to choose 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, however, mean that his conclusion is false. It only means that the argument that Dworkin offers for the continuation of our actual practice of assigning responsibility does not support its conclusion. It convinces only those who have already been convinced to begin with.

The weakness of Dworkin's argument becomes clearer when we take a look at a similar argument about a practice which most of us do not support: drawing cards to predict the future. Let us suppose that there exists a community in which drawing playing cards to predict the quality of an upcoming marriage has become common practice. The prospective groom drinks a 'predictive potion', a magic formula is uttered, and then the groom draws five playing cards at the most, one by one, from a shuffled deck. The rules are as follows. If, from the five cards drawn, 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 discontinued.

Let us suppose that this practice has existed for some time, when unexpectedly a 'difficult case' arises. The first card drawn by the groom is the Ace of Hearts and the second the Ace of Spades. One interpretation of the rules says that the groom should continue the drawing until he has five cards. Some, however, favour a different interpretation. The Ace of Hearts is the most important red card and, as such, has a clearly predictive power, they say, for a happy marriage. And then the second card is the Ace of Spades, which, had it been the first card drawn, would have been a prediction of a happy marriage anyhow! Such a combination of cards surely indicates that the marriage will be a happy one, and continuing to draw cards is deemed 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, that is, 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 above dispute over the difficult case. According to Dworkin, what we should NOT do is step outside the practice and ask whether the very practice of card drawing to predict the quality of marriage makes sense in the first place. We work within a practice, and we should interpret the practice to determine what the best way to deal with a difficult case arising from it is.

Not many would agree that, in the case of this example, we should take the practice as a whole for granted and argue only from within the practice to find the best solution for the difficult case. Most would say that drawing cards to predict the quality of marriage does not make any sense, and that we would be misguided to argue from the presumption that it does. The proper way to deal with the difficult case is to use it as an opportunity to stop doing what has been nonsensical throughout! Similarly, we should ask whether the very practice of holding people responsible makes sense, and we should not answer this 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. We might then 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 their doings, and that the very practice of holding people responsible makes no sense. Difficult cases, such as those involving individuals who seem accountable only 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 capacity approach

Dworkin's argument for the capacity approach to responsibility may be fallacious, but that does not mean that the capacity approach is wrong. We should, therefore, investigate independently what its virtues are. The underlying assumption of the capacity approach is 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 capacity 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 in which an agent lacks this capacity to have his conduct guided by legal rules. The presence of such circumstances may be a reason not to hold a human agent responsible for his acts.

When a legal rule is violated, the responsibility test allegedly entails investigation of the existence of such special circumstances in a concrete case that took away the agent's general capacity to be guided by the relevant rule. The same point is made more concrete by Dworkin when he assumes that an agent has capacity control over his acts 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 capacity to form true beliefs about the world and to match his decisions to his settled desires, ambitions and convictions.

The capacity approach is used to defend compatibilism, a view that our practice of assigning responsibility is compatible with determinism. At first sight, the capacity approach and determinism seem to be obviously compatible. According to determinism, a human agent who has violated a rule could not have had the capacity to obey the rule. All behaviour is necessitated by regular connections and preceding facts, and therefore his rule violation is also necessitated. He could not have had the capacity not to violate the rule. Since the human agent apparently lacks the capacity to comply with the rule, he should not be held responsible. Accordingly, the capacity approach and determinism lead to the same conclusion: nobody should ever be held responsible for their doings.

Clearly, this is not what adherents of the capacity approach have in mind. They assume that our present practice of holding most human agents responsible for most of their acts is right. To be consistent, they must also assume that most human beings who have violated rules in particular circumstances had the capacity to comply with these rules *under those circumstances*. Such an assumption seems incompatible with determinism, with the following question arising and needing to be addressed: how can compatibilists assume that the actual practice of assigning responsibility can go together with determinism? With the purpose of answering this question, 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 has actually signed her exam because the rules require that she does so, it is obvious that Katarzyna can sign her exam. To put it more generally, if an agent has performed

an act, he had the capacity to do so. However, we are more interested in capacity in cases in which an agent has not done what he had the capacity to do. Let us imagine that Katarzyna has violated the exam rules and has not signed her exam. How can we establish whether she had the capacity to sign her exam?

Capacities – and, more generally speaking, possibilities – are most interesting in cases in which they have not been realised. However, it is notoriously difficult to establish the existence of possibilities, capacities included, in cases in which they have not been realised. To deal with this problem, a thinking tool was constructed: the 'possible worlds' theory.[10] The basic idea underlying the 'possible worlds' theory is that something is necessary when it is the case whatever else may be the case. For instance, regardless of what the other facts may be, every coloured object always has a surface, and regardless of what the other facts may be, number 5 is always greater than number 3. Therefore, every coloured object necessarily has a surface, and 5 is necessarily greater than 3. A different way of expressing that something is the case regardless of everything else also being the case is to say that this something is the case *in all possible worlds*. In all possible worlds, every coloured object has a surface and, in all possible worlds, number 5 is greater than 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 be under different circumstances. In the real world, Bartosz's hair is actually brown, 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 reality he is not, but he might be. 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 signed her exam, but in some other possible world she did not. Therefore, Katarzyna actually signed her exam, but it could have been possible that she did not. This captures the notion of capacity quite well. *We may say that an agent has the capacity to do something if there is a possible world in which the agent does that something.* This would mean that Katarzyna has the capacity to sign her exam if there is a possible world in which she does sign her exam.

7.2 Possible worlds and constraints

We now have a definition of what it means for a person to have a certain capacity, but it may seem that this definition has only replaced one problem, i.e., the nature of capacity, with another problem, i.e., the nature of a possible world. What makes a set of facts a possible world? Here, the notion of constraint plays a role.[11] Not all sets of facts can go together. This is one such 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. This is a constraint on possible worlds. Moreover, this is a logical constraint in this particular case, because it is a matter of logic that a fact and its denial cannot go together. Apart from 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 does not expand once heated. There is no physically possible world, no world that satisfies all the

[10] The idea of the 'possible worlds' theory can be traced back to the German philosopher Leibniz (1646-1716), who, in his *Theodicee*, defines necessity as that which is the case in all possible worlds.

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

physical constraints, in which a metal bar does not expand if heated. And neither is there a physically possible world in which something travels faster than light in vacuum.[12]

We are now in a position to define possible worlds more precisely. A possible world is a world which 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 both 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 that somebody claims that something is possible, it is legitimate to ask relative to which 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. However, is it still possible if we take into account that Bartosz has just dyed his hair brown? This is apparently not the case, and it is worthwhile to consider in more detail why this is not the case. Both with logical and physical necessity (and possibility), this necessity is the result of constraints that consist of laws (regular connections), that is, the laws of logic and of physics respectively. A law expresses a necessary general connection between types of facts, such as the type of fact that something is a metal bar 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 be taken 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 to be constraints, too. One such fact might be that Bartosz has just finished dying his hair brown. Given this 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 that Dobrochna was on departed five minutes ago, it is impossible that she was seen at the railway station one minute ago. More specifically, with regard to 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 but submit the exam without having signed it first, this claim is probably based not only 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 that something. Now, we know that this specification of capacity is still too vague: we also need to specify relative to which set of constraints this capacity exists. The crucial question is the following: 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 the act? Here, I will not attempt to answer this question in the abstract, but will focus merely on the characteristics of individual agents.

It is clear that, in determining the capacities of a particular agent, some personal characteristics of this agent should be taken into account. If we examine the issue by looking at only the laws of physics which are the same for everybody, every agent would have to have the same capacities. This would be an unattractive finding, and to avoid it, we must take

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

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 is illiterate, she does not have the capacity to sign her exam, and she should, most likely, not be held responsible for not signing it.[13] Should we also take into account that the agent might have been highly motivated to violate a norm? Let us suppose that a kidnapper took Katarzyna's baby and demanded that Katarzyna not sign her exam. Almost paralysed by fear that something would happen to her baby, Katarzyna does not sign her exam. Did she have the capacity to sign? Would this be different if Katarzyna was a drug addict who could score only if she did not sign the exam? If we want to distinguish between the latter two cases, would that be a distinction based on a moral judgment about what *ought to* motivate Katarzyna?

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 sense only if not all facts are taken into account as constraints on what is possible. The question then arises which facts should be taken into account, and which facts should not. Capacity becomes a normative issue, the issue of 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 actually do when they ask whether a criminal suspect could have acted differently from the way he actually did. We should understand, however, that if we make capacity a normative notion, we can no longer adduce the capacity of an agent as a reason for holding the agent responsible. What we actually do is 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 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 works, in a sense, from the 'bottom'-up: everything is determined; therefore, agents do not have free will; and, therefore, agents could not have acted differently; and, therefore, agents cannot be held responsible. The 'bottom' of this argument is taken to express a hard fact about the world we live in with the rest following 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 '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, we also attribute to the agent the capacity to have acted differently. These are not, in the compatibilist view, different argument steps that build upon each other, but one single act of assigning meaning to ourselves and the world that surrounds us. This meaning encompasses acts and agency, free will and capacities, responsibility and liability.

[13] This might be different if it was Katarzyna's own fault that she is illiterate, but I will here ignore the possibility of responsibility without capacity.

7.4 Concerning the capacity approach

We have seen that Dworkin's argument for the capacity approach to responsibility rests on a naturalistic fallacy: this is how it is done and, therefore, this is how it should be done. In this section, we take a closer look at the capacity approach to see whether it is attractive for other reasons than those adduced by Dworkin ~~– apart from Dworkin's argument~~. Central to the capacity approach is the assumption that human beings are normally capable of complying with the rules of law which, for them, constitute reasons for action, they are 'reason-responsive', and that, therefore, they should normally be held responsible for norm violations. However, there may be special circumstances in which agents lack the capacity to comply with the applicable rules, and that would be a reason not to hold such agents responsible for possible violations.

The central question is what an agent lacking the capacity to comply with a norm means. If the capacity approach is to lead to results which are different from the determinist approach – which does not hold anybody responsible under any circumstances – it must assume that there are times when agents violate a norm even though they have the capacity to comply with the norm. We have seen that this assumption would cut ice only if, in determining the capacities of an agent, not all facts about the agent are treated as constraints 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 for determining which facts should and which facts should not be treated as constraints on what the agent could and can do. If the choice of facts to be treated as constraints is the outcome of a normative decision-making, it is no longer possible 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 lines: we want to hold the agent responsible for what he did and, therefore, we do not treat the facts which caused him to violate the norm as constraints that define the agent's capacities. For the time being, we may conclude that the capacity approach to holding agents responsible is the outcome of a 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 capacity approach and to the traditional image of man as rational decision-maker. This should not come as a surprise, since the capacity approach is based on this traditional image of man.

8 Conclusion

There are two ways of looking at the world or reality, that is, the phenomenological and the realist way, and if we try to look at acts and agency in both ways at the same time, the result may be paradoxical. This is illustrated by the example of free will and determinism. Working from a realist point of view, determinism either applies or does not apply to agency. However, in either case, there is no room for either freedom of the will or responsibility based on free will. The case is either that, a), acts are determined through the determination of the will, and then there is no free will, or that, b), the will which underlies acts originates in an arbitrary way, and then there is no free will either, or that, c), if arbitrary will is identified with free will, that kind of free will cannot be a basis for personal responsibility.

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

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 an act only if we count it as an act). This is a compatibilist approach. Its most prominent version is the capacity approach adopted by, amongst others, Morse and Dworkin. The capacity approach to acts, agency, free will and responsibility makes these phenomena compatible, by definition, with the facts of a mind-independent reality. Compatibilism is true, but trivially so. The question of whether the social practices which define what we count as acts, as agents, as free will and as responsibility make sense – remains unanswered. Justifying the practice by invoking the practice itself is a variant on the naturalistic fallacy, a variant that we may call the compatibilist fallacy. The question whether the agency practice makes sense is not answered in this article, at least not in detail. In this respect, it only offers 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 a reason other than 'this is what we actually do' does not lead to such a coherent theory.

References

Max R. BENNETT & Peter M.S. HACKER, 2003: *Philosophical Foundations of Neuroscience*. Oxford: Blackwell.

Donald DAVIDSON & Gilbert HARMAN (eds.), 1972: *Semantics of Natural Language* (2nd ed.). Dordrecht: D. Reidel Publishing Company.

René DESCARTES, 1973 [1641]: *Meditations Metaphysiques*. Paris: Larousse.

Ronald DWORKIN, 2011: *Justice for Hedgehogs*. Cambridge (Mass.): Harvard University Press.

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

Jaap HAGE, 2016: Facts and Meaning. How a rich ontology facilitates the understanding of normativity. *The Emergence of Normative Orders*. Eds. Jerzy Stelmach, Bartosz Brozek & Lucasz Kurek. Kraków: Copernicus Press. 13–44.

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

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

Saul A. KRIPKE, 1972: Naming and Necessity. *Semantics of Natural Language* (2nd ed). Eds. Donald Davison & Gilbert Harman. Dordrecht: D. Reidel Publishing Company. 253–355.

Benjamin LIBET, 2011: Do We Have Free Will? *Conscious Will and Responsibility*. Eds. Walter Sinnott-Armstrong & Lynn Nadel. Oxford: Oxford University Press. 1–10.

Brian P. McLAUGHLIN, 1994: Epiphenomenalism. *A Companion to the Philosophy of Mind*. Ed. Samuel Guttenplan. Oxford: Blackwell. 277–288.

Brain P. McLAUGHLIN, Ansgar BECKERMANN & Sven WALTER (eds.), 2009: *The Oxford Handbook of Philosophy of Mind*. Oxford: Clarendon Press.

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

Stephen J. MORSE, 2000: Rationality and responsibility. *Southern California Law Review* 74 (2000): 251–268.

Michael S. PARDO & Dennis PATTERSON, 2013: *Minds, Brains and Law*. Oxford: Oxford University Press.

David M. ROSENTHAL, 1994: Identity Theories. *A Companion to the Philosophy of Mind*. Ed. Samuel Guttenplan. Oxford: Blackwell. 348–355.

Walter SINNOTT-ARMSTRONG & Lynn NADEL (eds.), 2011: *Conscious Will and Responsibility*. Oxford: Oxford University Press.

Jerzy STELMACH, Bartosz BROŻEK & Lukasz KUREK (eds.), 2016: *The Emergence of Normative Orders*. Kraków: Copernicus Press.

Sven WALTER, 2009: Epiphenomenalism. *The Oxford Handbook of Philosophy of Mind*. Eds. Brian P. McLaughlin, Ansgar Beckermann & Sven Walter. Oxford: Clarendon Press. 85–94.

David WOODRUFF SMITH, 2013: Phenomenology. In: *The Stanford Encyclopedia of Philosophy* (Winter 2013 Edition). Edward N. Zalta (ed.). URL: http://plato.stanford.edu/archives/win2013/entries/phenomenology.