**Ronald Bon de Sousa**

**The Myth of Responsibility**

Keywords: free will, mens rea, reactive attitudes, John Searle, status function declarations, Libet, psychopathy

**Abstract:** Judicial pronouncements involve certain performative uses of language. I first sketch a number of interpretations of that notion, including John Searle’s “function status declarations”. A paradox arises from the apparent conflict between the aspiration to objectivity of judicial verdicts, and their performative function. In addition, there are reasons for scepticism about the Kantian notion of free will that seems to be presupposed by ascriptions of responsibility, which makes it look as if that notion is in essential respects a myth. Peter Strawson’s notion of “reactive attitudes” provides a partical solution to the paradox. The potential and the limitations of this approach are discussed in relation to contemporary understanding of the brain peculiarities that characterize psychopathy.

Schlagworte: Wort1, Wort2, Wort3, ...

**Abstract:** (The German translation of the English abstract will be provided by us.)

# Introduction

Judges, juries and legislators have magical powers. Like a spell, their words can, when uttered in the right circumstances, bring new realities into being, produce new truths: they can make it true that doing this is now a *crime*, which was no crime before; or that the decision of a previous court that previously stood is now *quashed*; or that so and so who was heretofore innocent until proven otherwise is now indeed *guilty*. There are differences between these three examples, but each one exemplifies the power of words to bring into being a reality that did not previously exist.

Those magical powers, and the different intuitions they are liable to arouse in different cases, are the topic of this essay. I will begin by sketching a number of different ways in which words make things happen, and then address some antinomies that loom large when we consider some common-sense assumptions about the appropriate conditions for ascribing responsibility. Those assumptions concern the intimate conceptual connection between the concepts of *mens rea*, free-will, and responsibility, and they are likely to be affected by emerging facts about the brain mechanisms involved in making decisions, as well as by philosophical considerations in the light of which they generate some perennial puzzles.

The three concepts just mentioned – *mens rea,* responsibility and free-will – have implications for the causality of action; but legal ascriptions of responsibility also have a *performative* aspect, which is in part detachable from those assumptions about causality. These are what lead to the perception of legal language as having the quasi-magical aspect I have noted. This performative aspect, as I shall shortly explain, admits of several interpretations. Scientific inquiry into mechanisms of will and emotion also muddies the causal picture, and opens the way to other, more overtly pragmatic criteria for ascriptions of responsibility. My aim here will be to suggest some ways in which the paradoxes generated by the apparent tension between the performative function of language and the aspiration to objectivity about responsibility can be disarmed.

To that end, I will begin by sketching some ways in which performative modes of discourse have been interpreted in recent decades by philosophers of language. I will then raise four reasons for scepticism about the Kantian notion of free will that seems to be presupposed by ascriptions of responsibility. Third, I will suggest that we might get some illumination from the notion of “reactive attitudes” articulated by Peter Strawson in an influential article. Strawson contended that we can still, indeed that we still must regard it as rational to respond in certain characteristic emotional ways to other people's actions even if we do not believe in the Kantian presupposition of free-will. The ascription of responsibility, in these cases, will itself constitute or at least derive from a reactive attitude, and in that sense might plausibly be described as a myth. Finally, I will make some remarks about how we could rationally tolerate the practice of ascribing responsibility without requiring the objective existence of free-will in the agent to whom that responsibility is being ascribed. This will imply that labeling the ascription of responsibility as a myth does not invalidate its legitimacy.

# Some uses of language that make things true.

The simplest example of the sort of magical power to bring into existence states that did not previously exist is what John L. Austin called “performatives”. Such uses, of which “I promise” or, in the context of a marriage ceremony, “I do”, are standard examples, are contrasted with “constative” uses of language that consist in making a statement of fact.[[1]](#footnote-1) Basic to Austin’s analysis of the different things we do with words, was his distinction between *locutionary*, *illocutionary,* and *perlocutionary* acts. The locutionary act consists in the utterance of sounds with syntactic and semantic import. The illocutionary act is identified in terms of the point of the utterance. In the relevant sense, the point of an utterance can be the ground for classifying it as “constative”, that is, intended as a statement of fact or logic; “commissive”, as in a promise or undertaking; “interrogative”, if it is used to ask a question; “directive” (requesting or commanding), or simply “expressive”, as in exclamations or expletives. Of those, only constatives are strictly speaking intended to be evaluated as true or false, although other illocutions have conditions of “felicity”, such as sincerity in the case of a promise. Illocutionary force is part of the conventional meaning of an utterance in the circumstances in which it occurs; in this way it contrasts with *perlocutionary* force, which are non-conventional causal consequences, not necessarily determined by the utterance’s illocutionary import. Thus one's utterance may, particularly if it is an interrogative illocution, provoke someone to answer a question, but it may also simply startle someone, or cause them to pass the wine, or to have a heart attack.

The development of the basic idea of performative utterances culminated in John Searle’s much more ambitious *Making the Social World* (Searle 2010). In that work, Searle pointed out that much of our social reality derives from what he called “Status Function Declarations”:

all of human institutional reality ... is created and maintained ... by a single, logico-linguistic operation. That operation is… a Status Function Declaration. The enormous diversity and complexity of human civilization is explained by the fact that that operation is not restricted in subject matter and can be applied over and over in a recursive fashion … to create all of the complex structures of actual human societies.[[2]](#footnote-2)

These “complex structures” owe their existence and their maintenance to the “collective recognition” of the status function declarations. That is what makes it possible for physical objects or people to constitute “a piece of private property, the President of the United States, a twenty-dollar bill, and a professor in the University”, and nearly everything else that is important in our lives. [[3]](#footnote-3)

That collective recognition, in turn, is based on previously recognized “declarations”. Those will have been made by persons endowed by the appropriate status, conferred under conditions that were themselves recognised collectively as giving them the relevant powers. That illustrates their recursive feature: a status function declaration makes it the case that someone is Chief Justice of the United States, and that judge gets to say, under certain circumstances, “you are now the President”; in turn the President can appoint yet others to be a Secretary of Defense, who appoints generals, who are then empowered to give orders to certain other people whose own role has been sanctioned by a declaration, and so forth.

Since the properties and states of affairs created by function status declarations are mind-dependent, it is interesting to compare them with what have been called “response dependent properties”. The central case of a response dependent property is colour, as a “secondary property”; but the category's importance derives from its inclusion of significant moral properties.

The model of colour works like this. Colour is an objective property; it is not merely, as some have occasionally claimed, “in your mind”. Nevertheless, it is not a quality that exists independently of our minds, or rather, of the specific character of our sensory apparatus. Normal humans have three types of cones in their retina, we are “trichromatic”. Our experience of colour results from what our brain computes on the basis of the information captured by those three receptors. Other organisms, such as birds, are tetrachromatic. The colours they see are therefore different from those we are able to see. In every case, however, what any given organism sees depends on its own visual apparatus but is systematically related to the objective reflective properties of the surface contemplated. Colour is therefore best viewed as a dispositional property which is identified by its power to cause a certain kind of sensation, in an organism of a given kind under specific “normal” conditions.

Similarly our emotions, and the moral responses associated with some of them, are typically elicited by certain sorts of things under certain conditions. Inasmuch as the power to cause such responses resides in properties of the object, the latter can be regarded as having a specific “response-dependent” property. Something is labelled *disgusting*, for example, if it has the power to arouse disgust. But not every case in which an object arouses disgust allows us to ascribe that label to that object. Suppose, for example, that someone has been subjected to a systematic programme of Pavlovian associative learning, as a result of which they experienced disgust when presented with a stimulus that to anyone else appears neutral or even attractive. We would then not say that the stimulus in question is disgusting, because it does not standardly have the power to disgust the average person. Much the same can be said about the moral analogue of disgust that can be felt in the presence of certain morally offensive situations. Similarly, something as elusive as love has sometimes been accounted for in terms of a power to cause such feelings; but in that case the claim is considerably weakened by the evident subjectivity of the response, and the notorious difficulty of articulating the conditions under which it is likely to arise. More generally, where any moral properties are concerned, there are difficult questions about what counts as “standard”, “normal”, or “average”. Response dependent properties are therefore tricky to identify. Nevertheless, they afford the prospect of distinguishing “normal responses” to certain things, elicited by properties that can be objectively identified, from merely subjective responses that speak only to the idiosyncrasies of a particular observer.

What then of responsibility? Could it plausibly be regarded as a response-dependent property?[[4]](#footnote-4) Is someone responsible for having committed a crime if and only if that person's action has the power to provoke, in a jury, a verdict of guilty? In view of the slogan that an accused person is to be regarded as innocent unless proven guilty, that idea has its attractions. But it faces an obvious objection, which is that it appears to put the cart before the horse. The presumption of innocence is about how people should be treated; it is not intended literally to imply that there nothing to guilt or innocence that is objectively true, independently of its power to produce a verdict. For when the jury decides that somebody is guilty, surely, the relevant sense of the word ‘decide’ in this case is the sense that is akin to discovery, not to invention. A verdict is a *finding of fact*, not a status declaration that creates a fact that was not there before. If that is the case, then the verdict will be correct only if that objective property is actually present.

The problem was pinpointed by Peter Geach,[[5]](#footnote-5) who pointed out that “ascriptivism”, as he called this sort of view, seems to reduce what common sense would take to be a perfectly objective property to a mere projection of one or more people's attitudes. In order to show why this was a fishy is sort of concept, Geach cites what is supposedly a real situation that occurred in a culture that embraced what we would now call a system of strict liability. Somebody had accidentally fallen off a tree, hit a man under the tree and broke his neck. The victim's brother had demanded a life for a life. The chief of the tribe in question, asked to adjudicate, decided with “Solomonic wisdom” that the culprit should start under the same palm tree, and that the avenger could climb up the tree, fall on top of the offender, and break his neck. The suggestion was declined, and the culprit went free. From this Geach infers that even when it is standard to ascribe responsibility regardless of intention — where it is acceptable to say: “I don't care whether you did it on purpose, I know only that you are responsible for the fact that he died”, people implicitly understand the difference between accidental events and intentional actions.

To see better what is at stake here, it will be helpful to bring in the concept of “direction of fit”. Direction of fit distinguishes between constatives, on the one hand, and commissive or directive uses of language on the other. If you are simply conveying a piece of information, the direction of fit is from language to world. That means that your utterance will be successful or correct if it fits pre-existing conditions in the world. By contrast, in the case of a command or a normative statement, the direction of fit is from world to language: success is determined by whether the world comes into congruence with your utterance rather than the other way round. Status declarations are very peculiar in that they have both directions of fit at the same time. For it is true of them by definition that they actually make true what they declare to be true.

# The Presupposition of Free Will: Four Reasons for Scepticism

The difference between accidental an intentional events appears to presuppose free will. The clearest expression of this presupposition is owed to Kant. For Kant, merely doing what you want is not a sufficient condition for acting of one's free will. That requires that we be capable of absolute origination of action: that we have, in a sense, a will equal to that of God. Free action, Kant writes, presupposes “a kind of causality belonging to living beings in so far as they are rational … efficient independently of foreign causes determining it”.[[6]](#footnote-6) Descartes had a very similar view in *Meditation* IV, when he tells us that God cannot be held responsible for our mistakes. For these arise out of the discrepancy between our finite understanding and our infinite will. About the first, we have surely no cause for complaint, since God owed us nothing in the first place. Still less can we reproach him with granting us infinite will, which is no less than he himself enjoys.[[7]](#footnote-7) On this view, which also presupposes that we believe at will, Descartes claims that our mistakes are due to our jumping to conclusions that are not warranted by our limited understanding but are permitted by the unlimited capacity of our will. Kant does appear to concede that as a matter of fact about the actual world we cannot be sure that we actually have any such capacity: “It must be freely admitted that there is a sort of circle from which it seems impossible to escape.” (ibid., 68) But he insists that we cannot but suppose, or act as if we supposed, that we do have such freewill: the idea of freedom “holds only as a necessary hypothesis of reason in a being that believes itself conscious of a will”.[[8]](#footnote-8)

Let me now briefly describe the four considerations that cast doubt on this presupposition. The first are adduced by David Wegner, who argues that conscious will is an illusion; the second stems from the Libet experiment, to which other essays in this volume have referred; the third is the phenomenon of “ego fatigue”; and the last will consist in some reflections on the nature of psychopathy.

But before I describe some of the difficulties involved in the concept of free will, let me explain why these objections must seem paradoxical. Many a reader will be familiar with Jean-Paul Sartre's *boutade*: “to be free is to be condemned to be free”.[[9]](#footnote-9) In this, Sartre stands firmly in the tradition of Kant and Descartes. We can feel the force of his dictum on the basis of a very simple observation. Suppose I instruct you as follows: in five seconds you must either flex your finger, or refrain from doing so. At the end of five seconds, you will necessarily have made a decision: you cannot escape it. Some of you may have thought that you could outsmart me: you may simply refuse to decide. But even if you do with this, you have decided — if only not to decide. Therefore you cannot simply refrain from deciding altogether. You cannot, for example, claim that as a believer in strict determinism, you will simply wait and see what happens: for that too is a decision, and when the time comes there will be something active rather than simply passive in your scrupulously refraining from doing anything. In this way it seems that we are indeed condemned to be free.

## The Illusion of Conscious Will

Nevertheless, David Wegner has spent the length of a book arguing that choice is an illusion. In the course of his book[[10]](#footnote-10)(Wegner 2002), he presents a number of examples, of which it will be enough to cite one, constructed on the model of a Ouija board. Subjects were given a cursor, with which they could follow a marker on a screen. They were instructed to move the cursor only when required to follow the marker. They must neither cause it to move, nor prevent it from moving. When the board is asked a question, as for example, *What is the name of your loved one?* the marker will apparently move spontaneously towards the first letter of a name.

In fact, of course, people do move the marker, while remaining convinced that they have merely followed its motion. The same phenomenon occurs in demonstrations of ghostly table tipping. More ingeniously, Wegner has found a persuasive way of showing that you can get people to have certain expectations about what is likely to be their own next move; and that they can then be persuaded that they have in fact effected that very motion, even though the device to which they think they are connected has been deactivated. The motion they observed was caused quite independently by the experimenter, and yet the subjects believe they consciously initiated it themselves. In short, Wigner has shown that you can make people think they have acted when they have not, and you can make them think they have not acted when they have.

Perhaps this is not sufficient to show that there is no such thing as free will; but it does seem to show that if free-will indeed exists, then we do not have any reliable way of knowing whether or not we have exercised it. The cover of the British popular science magazine new scientist summed up the situation a few years ago with the headline: “Quantum Underworld: Kiss Goodbye to Free Will!” Actually, quantum theory can have nothing more to do with free will than the ancient atomists' theory of *clinamen.* That was the “swerve” which Lucretius thought broke the otherwise fixedly vertical descent of atoms in the void, causing them, on the one hand, to collide and thus combine to form complex objects, and, on the other hand, to allow for uncaused free actions when the swerving took place in the brain. Quantum theory is an idle loop which people love to bring into this argument. Since nobody understands it, everyone is apt to believe that it holds the key to that other mystery which no one understands, that of free will. But it is idle.

1. **The Libet Experiment**

The second reason for scepticism about free will is the Libet experiment. At first sight, it is simple and compelling. In making a simple decision such as the one earlier illustrated, to flex or not to flex one's finger, there is a moment at which one is aware of making the decision. A moment later, the flexing occurs (or doesn't). But what Libet found is that a certain necessary preparatory phase of activity in the brain actually takes place a fraction of a second before the conscious act of deciding .[[11]](#footnote-11) (In some other similar experiments the delay between the onset of the preparatory phase and the awareness of decision has been seen to stretch as long as seven seconds.[[12]](#footnote-12) What does this mean? It seems undeniable that if acting, consisting in the actual event that begins with brain activity and results in the motion of the finger, comes before deciding to act, that is, before the exercise of will that is supposed to be its determining cause, then it cannot be the case that the decision actually caused the action. The consciousness of the decision, at any rate, must be the effect of something else – whether it be the action itself, or some common cause of both.

Although this seems to be an empirical result, philosophers can defend a claim to priority, because we can get very much the same result on purely analytical grounds. It can be argued that if you mean, by “fully conscious will”, an act of which all the determinants are themselves fully conscious, then you have an incoherent concept.

There are two ways in which we could interpret the notion of a fully conscious free will. On one interpretation attributed above to Descartes and to Kant, the moment of decision is an absolute origin, without any connection to any previously existing state, reason, or causal factor of any kind. It is a moment of pure creation. If that is the case, however, then the link between a free action and any aspect of an agent's character, her reasons for acting, or her motivations, must be excluded a priori. In that case it seems impossible to attribute the allegedly free act to the agent at all. It must be deemed to be purely random in relation to her character, reasons, etc. Surely that is not what the defenders of free will had in mind. On the second interpretation, we could understand a perfectly free decision as one that is determined entirely by factors of which I am fully aware and for which I am also fully responsible. Now either those factors are themselves willed, or they have other determinants. In the first case, I must look further at those other acts of will. Perhaps I will be led to an infinite series of other acts of will. This seems a plain absurdity. But otherwise, some of those acts of will must be due to other sorts of causes, which is the second option. In that case, I can have no awareness and no control over them. They will include genetic and environmental historical factors on neither of which I can claim to have had any influence whatever. Either way it seems I must renounce the idea that I have total control over my own actions.

The moral of this last argument is that we didn't really need the scientific evidence provided by Libet's experiment to warrant much the same conclusion. It's nice to have the science, but philosophers could get there without it. Nevertheless, this experiment has generated some very interesting empirical work on the nature of the mechanisms underlying the process of decision making. Azim Shariff and Jordan Peterson, among others, have suggested that the Libet experiment can be construed to be compatible with our authority over our own actions.[[13]](#footnote-13)(Shariff and Peterson. 2008). The key to this is to distinguish between willing, initiating and execution functions in the production of a deliberate action. Without going into technicalities, the basic idea can be grasped if we consider the experience of complex athletic activity. When you are skiing, for example, you are not concentrating on the tips of your skis on the current turn. You are effecting one turn, while planning the next or perhaps even two or three turns ahead. There are sequential phases involved in their control. In a first phase, the action is planned; it is then delegated to an executive function that will initiate it, and then to an executive that will effect it essentially without further conscious control. The execution of the process is then monitored in what appears to be a moment of decision, but is actually just one result of the cause that consists in the previous planning followed by the initiation of its implementation. But while the apparent moment of decision is not actually the cause of the act's initiation, it shares with it a common cause, and that common cause can be regarded as genuine deliberation.

This way of dealing with the issue suggests that the Libet experiment is not, after all, definitive in its bearing on the issue of free will. Once again, however, philosophy has been there first. David Hume gave us sufficient reasons to think that none of this is really meaningfully relevant to the issue of free will. He pointed out that “Liberty, when opposed to necessity, not to constraint, is the same thing with chance”.[[14]](#footnote-14) A random act can’t be mine. So if you think that in order to save free will you had to reject causation – which appears to have been Kant's position in what he mistakenly thought was an improvement on Hume – that has to be a mistake. For if the opposite of determinism is merely randomness, removing all determination of any kind from the antecedents of my act cannot make it mine. An act cannot be sensibly attributed to me if it is an event determined neither by my character nor by my existing wants. It would seem to follow that I cannot be held responsible for such an event. Once causation is distinguished fromconstraint, we can see that what counts is not whether our behaviour is caused by something other than the will but what kind of cause is involved, and in particular whether the cause in question is internal and external to the self.

To be sure, this leaves open a vast number of questions about what exactly it is that distinguishes the kind of “internal” determination that is compatible with freedom. If my immediate wants are themselves unrelated to my deepest aspirations, to my character, or to my long-term goals, it can be argued that acting on these wants represents a lesser grade of freedom than they would be if they are expressions of the dominating passions of my life. This opens the way for a gradualist conception of free will, allowing us to make a difference between acting on a post-hypnotic suggestion and acting on the basis of careful deliberations based on an educated imagination.

1. **Ego Fatigue**

Roy Baumeister has described a number of experiments to support the hypothesis that acts of will are actually separate events that draw on energy sources, deplete our glucose level, and weaken our capacity to make further efforts.

In one experiment, exerting self-control over eating seems to have depleted subjects’ resources of perseverance: “people who forced themselves to eat radishes instead of chocolates subsequently quit faster on unsolvable puzzles than people who had not had to exert self-control over eating.” In another, “suppressing emotion led to a subsequent drop in performance of solvable anagrams.” In yet another, “an initial task requiring high self-regulation made people more passive.” These results, Baumeister and his colleagues conclude, “suggest that the self's capacity for active volition is limited and that a range of seemingly different, unrelated acts share a common resource.” [[15]](#footnote-15)

Baumeister’s experiments appear to show not only that there is such a thing as free will but that it actually assumes energy in the literal sense. This is attested by the fact that the depleted energy can be restored by an intake of glucose. Thus if you have just made several decisions, it may be a good idea to consume a piece of chocolate before venturing to make another. (Another moral, incidentally, is that the only resolution you should ever make at New Year’s is never to make any New Year’s resolution. That is the only one you are likely to keep.)

So does this vindicate the reality of free will in the sense required by Kant’s presupposition? It does not. On the contrary, it shows that whatever causal factors may be that underwrite the subjective experience of exerting oneself, making an effort, making a decision, and so forth, those factors are as rigidly connected to ordinary physiological causes as any other mental processes. It does, however, open up a crack: for it suggests that when we want to encourage someone to behave in a “responsible” manner, there can be ways in which that person’s behaviour might be influenced without infringing on their sense of autonomy: feed them glucoses at the right moment; don’t expect a long string of successive efforts; allow difficult decisions to be taken on a good night’s sleep and a healthy breakfast – that sort of thing. And those are just the sorts of measures that will seem practically relevant if, when considering the assigning of praise or blame for a specific sort of behaviour, we turn away from moralistic thoughts of blame and retribution to issues of prediction, prevention, deterrence, and rehabilitation—or *PPDR,* as I shall refer to these goals when I return to them in the last section of this essay.

1. **Psychopaths**

Most of us think of psychopaths as evil people, who cause a lot of harm without remorse. But if you take an evolutionary point of view, it might occur to you that psychopathy is in fact none other than a successful evolutionarily stable strategy, based on the existence of frequency-dependent fitness. What does this mean?

Common conceptions of evolution tend to assume that a trait is either adaptive, in which case it will spread and go to fixation, or deleterious, in which case it will tend to be selected out of the population. But this is a mistake. Some traits provide their owners with an advantage only if other members of the same population do not possess the same trait. These are said to confer a degree of *frequency dependent fitness*. The classic example is that of Hawks and Doves, made famous by Dawkins but originally devised by Maynard Smith, provides a vivid illustration. [[16]](#footnote-16)

Imagine two kinds of organisms, Hawks, who fight to the death in any confrontation, and Doves, who can spend some time threatening but retreat rather than risking any dangerous fight. In a group consisting almost entirely of Hawks, Doves will have an advantage and will soon proliferate. That is because Hawks, whenever engaged in a confrontation, have a 50-50 chance of being killed. If most members of the population are Hawks, that means that in nearly every confrontation, two Hawks will be involved, and the result will be one dead Hawk. Doves, on the other hand, may sometimes go hungry, but will not be killed. As their numbers increase, however, Doves will provide adversaries for an increasing proportion of the confrontations with Hawks. They will always lose those, while hawks will always win. When there is a preponderance of Doves, therefore, Hawks will increase in number. If the population consists almost entirely of Doves, Hawks will have an advantage in nearly every confrontation. Depending on the exact value of the gains and losses involved, there will be a specific proportion of Doves and Hawks that will be stable, in that any increase of the number of Hawks will provide an advantage to Doves, and vice versa. That is what is meant by an Evolutionarily Stable Strategy, or “ESS”.

Psychopaths might be rather like Hawks that have found the level of presence in the population that allows them to exist in equilibrium with a majority of non-psychopaths. Clearly, psychopaths are likely to fare well among nice people, while they are not as likely to do so well when surrounded entirely by other psychopaths. The two situations can be construed by analogy with the situation of Hawks among Doves, contrasting with that of Hawks among other Hawks. Psychopaths are similar to everybody else in most ways, but they are differently equipped in certain very specific respects. They fail to respond to the suffering of others with the same empathic discomfort as other people feel in similar circumstances. Further, even the prospect of their own suffering does not arouse the usual anxiety in psychopaths. The prospect of harm to themselves doesn't move them in advance of the event. What sets psychopaths apart from other criminals or anti-social personalities is that their physiological reaction to the prospect of pain is virtually non-existent.[[17]](#footnote-17) If I tell you that in 10 seconds' time your hand will be forced to hover over a flame, you will experience various physiological responses, including a higher level of sweat measurable in terms of galvanic skin response. Psychopaths will not. They will experience the pain as well as anyone else would, but they do not respond to the anticipation of pain. In this way their brains appear to be differently constituted from that of ordinary people.

#  II. Reactive Attitudes

In this case, then, we have conflict of commonsense intuitions. On the one hand, we are readily convinced that in so far as their intelligence and their understanding appear to be normal, psychopaths, like anybody else, ought to be held responsible for what they do. On the other hand, if psychopaths don't really get it, in some important way, then they cannot be held responsible.

Nevertheless, even if we are convinced that there is indeed a reason not to hold psychopaths responsible for the harm they cause, this does not mean that you or I will refrain from experiencing resentment, anger, or other negative emotions in the face of their misdeeds. Such is the premise of Strawson's suggestion about “reactive attitudes”. Even if we are completely convinced that determinism is the correct view, and that there is therefore a real sense in which the psychopath *could not have done otherwise,* we still cannot help reacting with resentment, gratitude, hurt feelings, love, hatred or forgiveness. The responses appear to be inevitably elicited by certain social interactions, but they are independent of any convictions we may have about their underlying presuppositions concerning free-will.[[18]](#footnote-18)

We can think of Strawson's proposal as reconciling the Sartrian paradox that we are compelled to be free with the thesis that randomness is the only alternative to strict determinism. Sartre's paradox is based on the phenomenology of the first person. I cannot help deciding; I alone experience directly the compulsion to be free. I am the only one, in the end, towards whom it is not *possible* to adopt an attitude based on the conviction that strict determinism is true and that free will is irrelevant. In the light of that observation, Strawson is proposing to extend to the second person that presupposition of freedom which makes clear sense only in the first person. Yet on the basis of a sort of analogy based on the first person assumption of freedom, I can project that same freedom onto other agents. Extended to the second person, my attitude says: You *could* have acted differently, therefore you *ought to* have acted differently.

This proposal amounts to introducing an emergent level of social morality. By 'emergent', I mean what most philosophers mean by that word: it refers to a level of description and explanation that is not completely accounted for or predictable on the basis of lower levels of explanation and description. What is emergent here is a kind of *mutuality of responsibility*. This emergent phenomenon occurs inside a social group in which the reactive attitudes are recognised. This only works, of course, if others in the same group adopt the practice of allowing themselves reactive attitudes, and recognise them as legitimate. This is generally true of all myths: they only retain their force if most people at least pretend to believe them.

This leads me to some brief and all too vague vague concluding remarks. Accepting and expecting that others should express reactive attitudes, even if one is not convinced of their underlying presuppositions, is the defining feature of the kind of social contract I am attempting to characterize. This is not a social contract in the traditional senses associated with Rousseau, Locke, or Rawls. It implies no more than the emergent reality that I have tried to sketch. But what seems interesting about this form of social contract is the implication that living with others does not merely entitle us to blame them for what irks us; it also entitles us to claim *a right to be blamed,* even though you might remain somewhat agnostic on the question of the extent to which you are actually capable of an appropriate level of self-control.[[19]](#footnote-19)

If that is the case, what should our attitude be to those people who are simply not capable of entering into the social contract in question? Whether this is the case with the psychopath is a matter of some dispute. Data about the peculiarities of the psychopath's brain have been adduced on both sides of the dispute. On behalf of the view that psychopaths understand perfectly well that what they do is wrong, Adina Roskies has argued that if the psychopath's judgment is intact in every other way when the presence of certain differences in brain response indicates a lack of motivation, it is implausible to claim that the judgment was not “really” endorsed unless accompanied by motivation.[[20]](#footnote-20) But on the other side, Jessie Prinz has pointed out that there *are* certain subtle differences between the moral judgments asserted by psychopaths and those made by others not so diagnosed: in some cases, they actually tend to exaggerate the wrongness of merely conventional prohibitions, in the apparent hope of convincing their interlocutors that they are aware of the gravity of genuine moral prohibitions.[[21]](#footnote-21)

If the true psychopath is indeed incapable of understanding the idea of reciprocity, can we still endorse reactive attitudes as legitimate? It seems to me that we can. For despite the fact that the psychopath’s nature, on that understanding, undermines the point of *retribution*, we can still regard the reactive attitudes as affording us guidance as to the range of acts and responses that will best promote the other aims of the judicial system and its lay analogue in ordinary moral judgment. Recall that I mentioned above four aims for the judicial system that remain legitimate even if retribution, the one most strictly presupposing *mens rea*, is undermined by the skeptical arguments I have adduced. These four aims are those that can be referred to by the acronym, *PPDR,* standing for *predicting* what people are going to do, *preventing* them from doing things which will harm others, *deterring* them from such behaviour, as well as hope of changing them, or *rehabilitating* them to the extent that their nature allows. For those aims to be efficiently pursued, we do need to understand the causal mechanisms that underlie the psychopath's antisocial behaviour. If the widely accepted views expounded in (Hare and Babiak 2006) and (Blair, Mitchell, and Blair 2005) are indeed correct, then it gives us some understanding of ways in which we can and cannot hope to achieve *PPDR* – and particularly the last, the aim of rehabilitation with a view to reinsertion into society. As brain science provides us with increasingly reliable predictions, we may come to see that in the case of psychopaths neither deterrence nor rehabilitation is likely to be achieved. And if, in respect of psychopaths in particular, we are led to renounce the classical presuppositions underlying the notion of *mens rea*, this will pose special problems about how best to respond*.*

# Conclusion

If, as a philosopher, you want to manifest this magical power of declaring someone to be innocent or guilty, and regard the reactive attitudes as rational, regardless of the truth of the presupposition of *mens rea,* then you would do well to remember the admonition of the King in Saint Exupéry’s *Petit Prince*.[[22]](#footnote-22) The King has absolute power over his subjects, extending even to the Sun itself. But he issues only reasonable orders, that take account of facts and circumstances. Thus, he boasts, he would never order a general to change into a bird; neither would he ever command the sun to set when current conditions would make it unreasonable to expect compliance.

What that implies in practice, when the question arises of how to reconcile individual rights to equal treatment under the law, with the needs of *PPDR* and knowledge of the facts about psychopaths, is a question I leave to those who are saddled with more practical responsibilities.

1. Austin, J.L. *How to Do Things with Words*, 2d ed., J. O. Urmson and Marina Sbisà (eds), Cambridge, MA: Harvard University Press, 1975. [↑](#footnote-ref-1)
2. Searle, J.S., *Making the Social World,* Oxford: Oxford University Press, 2010, p. 201. [↑](#footnote-ref-2)
3. Searle, J.S., *Making the Social World,* p. 7. [↑](#footnote-ref-3)
4. I have benefited here from some unpublished work by Christine Tappolet. [↑](#footnote-ref-4)
5. Cf, Geach, P. “Ascriptivism,” *The Philosophical Review* 69 (1960) ,p. 221–25. [↑](#footnote-ref-5)
6. Kant, I. *Fundamental Principles of the Metaphysics of Morals*, trans. Thomas K. Abbot, 1785, Indianapolis, Bobbs Merril, (2008) p. 64. [↑](#footnote-ref-6)
7. Descartes, R., *Meditations on First Philosophy with Selections from the Objections and Replies*, trans. J. Cottingham, R. Stoothoff, and D. Murdoch, reprint, 1641, Cambridge: Cambridge University Press, 1986. [↑](#footnote-ref-7)
8. Kant, *Fundamental Principles,* p. 77. [↑](#footnote-ref-8)
9. Sartre, J-P., *Being and Nothingness, an Essay in Phenomenological Ontology,* trans. H. Barnes. New York, Washington Square (1993), p. 186. [↑](#footnote-ref-9)
10. Wegner, D., *The illusion of Conscious Will.* Cambridge MA, MIT Press, 2002. [↑](#footnote-ref-10)
11. Cf. Libet, B., “Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action,” *Behavioral and Brain Sciences* 8 (1985), pp. 529–66. [↑](#footnote-ref-11)
12. Cf. Soon, C. S., Brass, M., Heinze, H.-J., & Haynes., J.-D.,. “Unconscious determinants of free decisions in the human brain”, *Nature Neuroscience*, 11 (2008), pp. 543-545. [↑](#footnote-ref-12)
13. Cf. Shariff, A. F. and Peterson. J.B., “Anticipatory Consciousness, Libet’s Veto and a Close-Enough Theory of Free Will,” in *Consciousness & Emotion: Agency, Conscious Choice, and Selective Perception.*, R. D. Ellis and N. Newton, (ed), Amsterdam, John Benjamins, 2008, pp. 197–214. [↑](#footnote-ref-13)
14. Cf. Hume, D., *Enquiries: Concerning Human Understanding and Concerning the Principles of Morals*, 3d ed., ed. & introd. by Selby-Bigge, L.D., revised by & notes by P. H. Nidditch, reprint, 1777, Oxford: Oxford University Press, 1975, p. 96. [↑](#footnote-ref-14)
15. Baumeister, R., Bratslavsky, E., Muraven, M., & Tice, D.,”Ego Depletion: Is the Active Self a Limited Resource?” *Journal of Personality and Social Psychology,* vol. 74,(1998), (pp. 1252-1265), p. 1252. [↑](#footnote-ref-15)
16. Cf. Dawkins, R., *The Selfish Gene,* Oxford: Oxford University Press, 1976, and Maynard Smith, J., “Game Theory and the Evolution of Behavior”, *The Behavioral and Brain Sciences*, vol. 7 (1984), pp. 95-126. [↑](#footnote-ref-16)
17. Cf. Blair, J., Mitchell, D., & Blair, K., *The Psychopath: Emotion and the Brain,* Oxford: Blackwell, 2005; Hare, R. and Babiak, P., *Snakes in Suits*, New York: Harper Collins, 2006. [↑](#footnote-ref-17)
18. Cf, Strawson, P. F.”Freedom and resentment”, *Proceedings of the British Academy,*vol.48*,* (1962), pp. 187-211. [↑](#footnote-ref-18)
19. For a defense of the right to be blamed, cf.. Houston, B.,” In Praise of Blame”, *Hypatia* ,vol.7, no. 2 (1992), pp. 128–47. [↑](#footnote-ref-19)
20. Cf. Adina Roskies, “A Case Study of Neuroethics: The Nature of Moral Judgment”, in *Neuroethics: Defining the Issues in Theory, Practice, and Policy,* Judy Illes (ed.), New York: Oxford University Press, 2006, pp. 17–32. [↑](#footnote-ref-20)
21. Cf. Prinz, J., *The Emotional Construction of Morals*, Oxford; New York.: Oxford University Press, 2007, p. 44.

 [↑](#footnote-ref-21)
22. Cf. Saint Exupéry, A. de, *Le Petit Prince,* New York: Reynal & Hitchcock, 1943. [↑](#footnote-ref-22)