

Note for the Colloquium: This is a chapter in a book several chapters of which have been considered in Colloquia in past years. I realized the need for this chapter after a discussion of Jürgen Habermas's views about free will last year.

Responsibility Without Freedom

1.

A threat hangs over my argument. I have been writing freely about moral responsibility ignoring a view popular among philosophers that there is no such thing. People are responsible for their acts only when they are in control of what they do – only, in the standard philosophical jargon, when they have and act out of a free will. You are not responsible for the injury when someone else pushes you into a blind beggar, for instance, or when a hypnotist makes you steal from the beggar's cup. Many philosophers – and many other people – believe that this apparently innocent observation is in fact wholly destructive of morality. They press what we might call the “no free will” challenge in the following form. “People are never actually in control of their own behavior, even when it seems to them that they are. Their will is never free because their behavior is always caused by some combination of physical or biological laws and environmental influences acting on their brains, some combination of forces and events entirely beyond their control. Moral responsibility is therefore an illusion, and morality itself also an illusion because if people are never responsible for what they do then it is senseless to suppose that they are morally required to act in one way rather than another.” This challenge to moral responsibility is quite independent of the issue I discussed in Chapter *n*: whether what I called the causal impact thesis is sound. The challenge would be at least equally as threatening

if we accepted that thesis and believed that the truth about value has or can have a causal impact on our minds.

The threatening question whether people have free will and hence can be responsible for their actions is one of the few philosophical questions that have escaped the textbooks and entered popular literature and imagination. The philosophical literature is in itself vast. It divides into two rival positions. “Compatibilists” insist that even if physical and biological laws do determine all our behavior, we are still morally responsible, at least normally, for our acts. They insist, as they often put the point, that free will is consistent with determinism. “Non-compatibilists” deny this: they accept that if all behavior is determined in that way morality is an illusion. Some of them – optimistic non-compatibilists – hold that morality nevertheless survives the challenge because, they believe, our behavior is not fully determined by forces beyond our control. But other non-compatibilists are pessimistic. They insist that morality does not survive because all our behavior is indeed fully determined by forces that are, in the appropriate sense, beyond our control.

I have postponed taking up the issue until this late point in the book because I believe that three themes in the argument so far are pertinent to understanding and responding to the no-free-will challenge. These are, first, the independence of reasoning about ethics, morality and value from scientific and metaphysical truth, second, the interpretive character of proper reasoning about ethics and morality, and, third, the capital importance of being able to ground our moral judgments, and in particular our judgments about moral responsibility, in ethics: in our sense of the importance of living our own lives well. A word about the pertinence of each of these themes.

That value and science are distinct and independent orders of inquiry means that no assortment of physical, psychological, biological or social facts can dictate, just on their own, that anyone is or is not morally responsible for any particular behavior on any particular occasion. That latter judgment is an ethical or a moral judgment and no argument for such a judgment can succeed unless it includes, explicitly or implicitly, another pertinent value judgment. The issues of free will and moral responsibility are often twinned in philosophical argument: many compatibilists and non-compatibilists propose accounts of what genuine freedom of action must be like and draw their

diverse judgments about responsibility from those accounts. That twinning has, I believe, dulled philosophers' sense of the difference between metaphysics and value. We would do better to separate the ethical and moral question from any question – if there is any sensible question – about whether we have wills and if so whether those wills are in some sense free. Perhaps the ethical question will turn out, for some reason, to require us to face some metaphysical issue. But we should begin with no such assumption: the problem of responsibility is big enough on its own.

Our earlier conclusions about the nature of reasoning about value are therefore directly in point now. The world of value is the world of interpretation: ethics and morality are best understood as interpretive genres. We can defend our convictions about how to live well and how we must treat others only by embedding these convictions in a larger structure of value from which these convictions draw and to which they contribute argumentative support. We must therefore test particular theories of moral responsibility, in this chapter, by seeing how well they stand up to the rest of what we think about how to live. We cannot accept any account of responsibility, no matter how compelling it seems on its own, if we cannot accept its extended ethical or moral consequences. In particular, we must resist the appeal of any claim about the conditions of responsibility that is offered as axiomatic: there are no axioms in an interpretive structure. It is sometimes presented as obvious, for instance, that a person cannot be morally responsible for any action if it was entirely predictable, given full knowledge of the physical and chemical structure of his brain, that he would act in that way, unless he was also responsible for his brain being that way. We cannot take any value judgment of that kind for granted: we must see whether we can accept it once we see what else it would require us to accept.

Classical discussions of responsibility begin in a moral issue. Is it right to punish or chastise someone for an injury he has inflicted when he was drunk? Or hallucinating, or suffering from some other mental disorder? Or if he had an unfortunate upbringing or acted under duress? Would it be fair to jail someone who acted under one or another of these disabilities? If not, why not? I propose to begin differently: by asking how and why people normally hold *themselves* responsible for what they have done, and why, in some circumstances, they should not do so. That tack brings the strategy of this chapter into line with the general strategy of the book, which tries to draw

morality, as one department of value, out of the best understanding of ethics – the best understanding, that is, of how to live well. In this instance the strategy allows us to concentrate on something important that the more usual strategy tempts us to ignore. When we begin in the first rather than the third person, we pay more attention to how it feels to be confronted with a decision. We pay more attention, in particular, to the impossibility of deciding without taking responsibility for how one has decided.

2.

Responsibility is a tricky concept; we use “responsibility” and “responsible” in many different and easily confused senses. The long discussion of Chapter *n* focused on virtue responsibility. That is the kind of responsibility we have in mind when we say that someone is a responsible or irresponsible person, or that he behaved responsibly or irresponsibly in acting as he did. (He acted responsibly in refusing the bribe. He would have been irresponsible had he accepted it.) We make distinctions among different forms or modes of virtue responsibility: we distinguish intellectual, practical and moral responsibility. A scientist who does not check his calculations lacks intellectual responsibility; a writer who does not back up his files lacks practical responsibility; a voter who votes for a candidate because he finds her sexy lacks moral responsibility. A politician who sends his nation to war on plainly inadequate intelligence is irresponsible in all three ways.

We distinguish virtue responsibility, first, from causal responsibility. A person is causally responsible for an event, we might say, if some act of his is part of the causal chain that best explains the occurrence of that event. I would be causally responsible for an injury to the blind beggar if I had collided with him absent-mindedly or while drunk or deranged or even just accidentally. But not when someone else has pushed me into him because then no act of mine has contributed to the injury. We distinguish both virtue and causal responsibility from assignment responsibility. Someone has assignment responsibility for some matter if it is his duty to attend to or look after it. The last person to leave a room, we might say, is responsible for turning off the lights and the sergeant is responsible for his platoon. Virtue, causal and assignment responsibility are in turn different from consequence responsibility, which is liability

for the results of some act or event. I have consequence responsibility for the damage I cause by my careless driving; an employer may have consequence responsibility for any damage his employees cause. These are all, finally, to be distinguished from judgmental responsibility. Someone has judgmental responsibility for some act if it is appropriate to criticize his act against a normative standard: a standard of how one ought to behave. I have judgmental responsibility for my decision to walk past the beggar giving him nothing, but not for the harm when someone else pushes me into him.

These different senses of responsibility are plainly interrelated. Judgmental responsibility is not necessary for causal responsibility – someone who is insane is causally but perhaps not judgmentally responsible for taking coins from the beggar’s cup. But judgmental responsibility is necessary for responsibility in the other senses. Only acts for which someone is judgmentally responsible can properly figure in the overall judgment whether he is virtue responsible, for example, or has acted contrary to his assignment responsibility. True, people sometimes have consequence responsibility for the acts of others – as I said, employers may have consequence responsibility for the damage their employees cause – but that consequence responsibility depends on their having judgmental responsibility for some act of their own: setting employees to work, for instance.

In any case, it is judgmental responsibility that the challenge we are now concerned with calls directly into question. The pessimistic non-compatibilists insist that, for exactly the reason that it is never appropriate to attribute praise or blame to someone who is pushed into a beggar, it is also never appropriate to attribute responsibility to him when he takes coins from the cup or walks past the beggar without giving. He ignores the beggar because his genes or background conspire with neural physics to push him past the beggar in that case too. He is no more in control of his conduct in the second case than in the first. So no one ever has judgmental responsibility for anything. Can that be right?

3.

It is important to notice, right from the start, that we can’t actually believe it. I do not mean only that we would find it difficult to believe it in the way we might find it

difficult to believe that a lover has betrayed us or that slavery is morally permissible. Nor do I mean only what Peter Strawson showed: that the way we relate to and respond to other people is so comprehensively built around the reactive emotions of blame and resentment – emotions that depend on the assumption of responsibility – that it would take an unimaginably seismic change in our lives fully to accommodate the idea that these emotions are misplaced.¹ If we could be convinced, intellectually, that responsibility is an illusion, then we might describe Peter Strawson’s argument as revealing the practical difficulties – perhaps insurmountable practical difficulties – that we would face in trying to incorporate our new belief into the way we live. But our situation is different: we cannot be convinced even intellectually that we are not judgmentally responsible for our actions, because there is no way we can live without asking, as we decide what to do, which decision would be in some way best, which means that we cannot live without judging our behavior as we act. We cannot even imagine behaving, let alone actually behave, in a way that would be consistent with the belief that it is never appropriate to judge and criticize our own behavior. A sense of judgmental responsibility is not an optional add-on for people as they act; they cannot act without that sense.

Suppose you really believe that free will is an illusion. You are convinced, after you pass the beggar by, that you were forever destined to pass him by. Or, after you have given him money, that your conscious decision to hand him a dollar or two played no effective role in your actually doing that. Nevertheless, as you approach him, you cannot repeal either the thought or the fact that you have a choice to make. You cannot lift yourself above yourself just to watch how you choose. You must choose. You might pause, frozen in your tracks, to see what will happen. But then nothing will happen, and even then you have chosen to stop and eventually you will choose to do something else. You cannot choose, moreover, except in particularly banal matters, without supposing that there is a better and a worse choice for you to make; you cannot choose, that is, without supposing that your choice is a matter appropriate for criticism. Of course this need not be moral or even ethical criticism: it rarely is. You may criticize yourself on what you take to be purely instrumental grounds – can

¹ Strawson, *Freedom and Resentment*.

you afford to give to every beggar you confront? But you will still be holding your choice to a standard not simply choosing as if the choice were a tic or a cough. After you choose you might be able to treat your choice that way: you might insist, even to yourself, that you are not to blame and have nothing to regret in having ignored the beggar because you were destined to do it. But the threatened conclusion, that you never have judgmental responsibility, claims more than that. It claims that your decision, like your cough, is immune from judgment from the start. That is what you cannot believe.

I make this claim in the first person: no one can doubt his own judgmental responsibility. Defenses of non-compatibilism are usually constructed, as I said, in the third person; it seems possible for one person to suppose, even before the fact, that another person's actions are determined by forces beyond that person's control and that in consequence it would be wrong to attribute judgmental responsibility to him. But if I cannot believe that I myself lack judgmental responsibility even though I accept that my own actions are determined, I have no ground for supposing that anyone else lacks judgmental responsibility just because his actions are determined. In fact, moreover, the third-person case is actually a first-person case as well, because it ends in a conclusion about how it would be proper or fair for me to treat someone else. Some criminologists insist that we must abandon traditional criminal law, with its apparatus of guilt and punishment, and substitute only therapeutic treatment because, they declare, people are never responsible for what they do and so it is unfair to blame and punish them. These criminologists contradict themselves: if no one ever has judgmental responsibility then officials who treat accused criminals as responsible for their actions are not responsible for their own actions, and it is therefore wrong to accuse them of acting unfairly. Of course it is then also wrong of me to accuse those criminologists, as I just have, of acting wrongly in accusing the officials of acting wrongly, because the criminologists are not responsible either. And wrong of me to accuse myself of accusing them wrongly because I am not responsible either. And so on. This recursive nonsense shows, even if nothing else does, that we cannot believe the proposition on which it hinges, which is that we all lack responsibility for anything. If we can't believe it, then we'd better not believe it. We mustn't hold it true. That injunction shows the real threat of the no-free-will

argument. If we accept it we end in incoherence. We find ourselves forced to believe what we know is false.

4.

I must be clearer about what I take the scientific challenge to be. Deciding to do something and then doing it feels a certain way. I see the beggar; I wonder whether to give him something. I quickly rehearse reasons for and against doing that. He looks hungry, I won't miss a dollar or two, he'll spend it on drugs, I gave at the office. I decide against giving; I walk past. It certainly feels to me that I could have decided to give him something. – after all, I've given to other beggars – and that if I had decided to give I would have given. That is, we might say, a crude sketch of how the process feels from the inside. I shall try to provide a more complex account of how it feels to make a decision later in this chapter, but that crude sketch will do for now.

Non-compatibilist philosophers – both optimists and pessimists – see two different causal claims embedded in that crude account. The first is a claim about the origin of my decision not to give. I must be assuming, at least unreflectively, that I myself initiated the causal chain that ended in that decision. The causal buck stops here. If I thought it was fixed in advance that I would not give – if the combination of my genes and my environment made this inevitable – it would have been silly, a pointless waste of time, for me to deliberate. In deliberating, that is, I must be assuming that my decision was not settled by physical and biological forces in advance, that the question whether I would give was unsettled until I, of my own free will, decided not to give. The second causal claim they find is a claim about the effect of my decision. I must think that my decision caused various physical events to take place in my brain, nerves and muscles: the physical events that constituted my act of walking past the beggar without giving. If my decision made no difference – if these physical events would have taken place, produced by physical and biological forces, no matter what was going on in my mind – then once again my deliberation would have been pointless. In deliberating, I must be assuming that my decisions have causal potency.

I shall call these two causal claims “hydraulic” to distinguish them from a different kind of causal claim that it would be more natural for me to make if asked why I walked past the beggar. Earlier in the book I described a grand divide between two

worlds: the order of nature and the order of interpretation. I might say, if asked why I ignored the beggar, that I was worried he would use any money I gave him for drugs. That is a kind of causal explanation – philosophers sometimes call it a teleological causal explanation – but it belongs to the order of interpretation. The causal claims that non-compatibilists see in the internal account belong to the order of nature. They are hydraulic claims about what in the flow of past events caused my mental state when I decided to ignore the beggar – nothing in the past did that – and what caused the flow of current along my nerves to my muscles – my decision did that.

The first of these two hydraulic claims is contradicted by a scientific hypothesis usually called determinism and the second by the different scientific hypothesis often called epiphenomenalism. Determinism, in its now most fashionable form, begins in the assumption that all mental activity, including decisions, are correlated to brain states: no one feels that he is making a particular decision unless his brain is in a particular (and in principle identifiable) state of electrical excitement. But brain states, the argument continues, are physical phenomena that do not arise spontaneously but must be caused by something whose existence is prior to that state. Even if we accept that an act of will (whatever that is) can cause brain states to change, that act of will must itself be correlated to a different brain state that must in turn be caused by something prior to it. So all changes in the brain that are correlated with mental activity must have causal antecedents that are not themselves pieces of mental activity but are rather states of the world external to the organism. It is therefore nonsense to suppose that people have the power to make decisions that are not fully determined by something external to them and prior to their decision.

Determinism contradicts the first of the two hydraulic claims I described, but not the second. It only insists that any decisions that cause behavior are themselves caused by something else: that the causal chain that ends in action cannot start in mental life. Epiphenomenalism, however, denies the second claim. It denies that mental events even figure in the causal chain that ends in movements of nerve and muscle: it supposes that the internal sense of having decided to do something is only a side-effect of the physical and biological events that have actually produced the behavior decided upon. Certain now famous experiments at least illustrate this proposition, though they hardly demonstrate that it is true. An experimental subject is asked

spontaneously to raise whichever of his hands he wishes: scans indicate that the brain activity that ends in his raising one hand begins before the brain activity that constitutes awareness of which hand he will raise. The experimenters conclude that his decision is not the cause of his raising his right hand, but only another effect of whatever did make him raise his right hand. Epiphenomenalism supposes that *all* conscious decision is a side-effect rather than a cause. It supposes, for instance, that the series of physical events that culminated in my typing the last word in this sentence began before I actually decided which word to type assiduously. They began while I was still, or so I thought, hesitating over my choice of words. If every conscious decision is only a side-effect, then whatever part of me forms that decision, whether we call it my “will” or by some other name, can hardly be in charge of what happens. It is only the fraud of Oz, pulling levers and plumbing steam to no effect whatsoever.

I assume that both determinism and epiphenomenalism may be true. My description of these theories is presumably defective, so I assume that each may true with any defects in my account cured. I am not competent to judge either of them as scientific theories. I believe that neither has been demonstrated to be true or been shown to be false. Everything is possible. Every Tuesday’s New York Times brings fresh surprises about brain geography, physics and chemistry, about potent alleles on neglected chromosomes, and about the interrelations among all these and our mental life. Every dinner party brings fresh speculation about the sexual reasoning of baboons, the moral lives of chimpanzees, the reptilian brain at the core of your brain and the neo-Darwinian explanation of the trolley problem. Our grandchildren had better be ready for anything. What follows when we assume that determinism and epiphenomenalism are actually true?

5.

Deliberate behavior has an internal life: there is a way it feels to act deliberately. We intend to do something and we do it. There is a moment of final decision, the moment when a die is cast, the moment when the decision to act merges with the action decided on. That internal sense of deliberate action is at the core of our familiar ideas about judgmental responsibility. It marks the indispensable distinction between acting

and being acted upon: between pushing and being pushed. We think that we are judgmentally responsible for what we do but not for what happens to us: for driving too fast but not for being hit by lightning. Our more complex ideas about responsibility depend on refinements in these crude ideas. We distinguish the normal occasions in which people decide to act not just from those in which they are acted upon but also from those when they act under the control of someone else, as in hypnosis or higher-tech forms of mind control, or when they are victims of certain forms of mental deficiency or illness. We say, in the former case, that it was not their decision, but rather the decision of their mind-controller and in the latter case that though it was their decision they ought not to be held responsible for it because they lacked some capacity essential to responsibility. We have in mind, as part of that capacity, some minimal ability to form true beliefs about the world, about the mental states of other people, and about the likely consequences of what they do. Someone who is unable to grasp the fact that guns can injure people is not responsible when he shoots someone. We also have in mind, as part of that capacity, the ability to make decisions that fit what we might call the agent's normative personality: his desires, preferences, convictions, attachments, loyalties and self-image. Genuine decisions, we think, are purposive, and someone who cannot match his final decisions to any of his desires, plans, convictions or attachments is incapable of responsible action.

This system of ideas about responsibility – the “responsibility system,” we might call it – is very widely shared stated as abstractly as I just have. But much of it becomes controversial when specified in greater detail: we disagree, for example, about whether someone who is unable to resist impulses stemming from blind rage, even when these impulses contradict all of his more reflective purposes, or someone who is forced to act against his convictions by threats of grievous harm if he does not, or someone whose sense of right and wrong has been warped by watching violence on television, is judgmentally responsible for his acts. A theory of responsibility must either reject this popular scheme for attributing and withholding praise and blame or justify that scheme; if it offers a justification it must draw on that justification to take sides in these controversies. It can do this only by placing the responsibility system in a larger context of ethical and moral values in order to attribute some point and value to the system as a whole. We cannot decide whether the alleged scientific challenge I

described – the doctrines of determinism and epiphenomenalism – are really threatening to the responsibility system until we have grounded that system in that way. The contest between the compatibilist and non-compatibilist position is a contest in normative theory: it cannot be decided until we are clearer about the moral and ethical presuppositions of our general opinions about judgmental responsibility.

I have found little extended attention to the normative issues in the literature of the so-called free will problem, however. Most writers seem to assume that the ethical and moral basis for the responsibility system is obvious: it lies in the principle that people should be held responsible for behavior that they can control, but only for such behavior.² If my pulling the trigger was for some reason not under my control – if I pulled it because someone else passed a current into my brain that produced a reflex pulling, for instance – then it would be unfair to hold me responsible for the shooting. But the principle that declares control necessary for responsibility can be interpreted in different ways. One interpretation – I shall call this the hydraulic control principle – is plainly vulnerable to the scientific challenge. It holds that someone is in control of his behavior, in a way sufficient to attract judgmental responsibility, only if the two causal claims I described above are true: only if his decision is not caused by prior events beyond his control and only if that decision figures in the hydraulic causal chain that ends in action. A second interpretation – the creative control principle – is not so plainly vulnerable. It holds that people are judgmentally responsible for their decisions if they have the two capacities I described: the normal capacity to form beliefs based on evidence and argument and to make the decisions that are called for by their normative personality. Each of these two different principles purports to interpret – to fit and justify – the general structure of the familiar responsibility system. If we find the first of these interpretations persuasive we must accept non-

² Historically another formulation of the root ethical and moral assumption was popular: people are judgmentally responsible for some act only if they could have acted otherwise than as they did. It now seems widely accepted, however, that this assumption is not credible. Harry Frankfurt constructed hypothetical cases that make it seem implausible. If a scientist has implanted a device in my brain that can force me to decide to kill a colleague, and intends to exercise that latent power unless he knows that I am about to decide to kill that colleague myself, I am nevertheless responsible I do make that decision myself, even though I could not have decided otherwise.

compatibilism and, on the assumption that determinism and epiphenomenalism are both true, we must conclude that judgmental responsibility is indeed a myth. If we reject the first principle, however, and find the second compelling, then we will find ourselves compatibilists instead. Or so I shall argue.

Most prominent non-compatibilists, I believe, assume that some principle like the hydraulic control principle provides the best justification of the responsibility system. The most prominent compatibilists, including Hobbes, Collins and Hume as well as Harry Frankfurt and other modern compatibilists, have the creative control principle, or something very like it, in mind. However that historical claim is no part of my argument, and I shall not pursue it. Moreover the interest of these two principles, and the choice between them, is not exhausted by their bearing on the classical dispute about the possibility of judgmental responsibility. For if we do not reject the very idea of responsibility, either because we are compatibilists or because we are optimistic non-compatibilists, then we must confront the controversies within the responsibility system I just mentioned. Which of the two principles we take to justify the responsibility system will be decisive for many of these controversies.

The two principles differ dramatically in what they take to be ethically and morally crucial in the phenomenon of decision – its psycho-biological genesis and consequence or its felt character just as a decision. The hydraulic principle makes everything turn on what caused the decision and what flowed from it as historical events. The creative principle ignores these hydraulic questions and makes decisive the phenomenal fact of decision itself together with the place of that decision in the larger cognitive and normative structure of the agent's personality. I pause to speculate in the Just-So style once again. This disagreement about what is ethically and morally significant in a person's decisions signals a deeper difference in conceptions of human dignity. The Enlightenment project I mentioned in the Introduction supposes that our self-respect as distinctive animals depends on being able to see our mental life as in some way exempt from the causal order of nature. Otherwise we are each just an ordinary part of that natural order, pushed and pulled along with the other crude clumps of matter. We feel free – we think we can decide as we wish and that we have responsibility for how we decide – but that freedom may be

only an illusion. The brute physical and biological forces that shaped our brains may have inoculated them with that illusion as an evolutionary benefit or a bad joke.

When God was in charge we inherited dignity from Him. Perhaps he gave us free will as a miraculous act of grace. Or, if he denied it, at least our pre-destination was decreed by a supreme intelligence who had made us in His image not a barren mechanics. The deism of the Enlightenment blocked that escape even as the dramatic success of its physics magnified the threat. Two sources of dignity might be thought to remain. First, we might find that our decisions are after all in some way independent of the transactions of the physical and biological world. That hope is captured in the hydraulic control principle and threatened by determinism and epiphenomenalism. Earnest philosophers cling to the hope, in the face of the threat, with Kantian metaphysics or various forms of dualism. Or, second, we might think that the undeniable phenomenal world of challenge – of lives to lead and thousands of unscripted decisions to make – itself gives us all the dignity we need or should crave. Nature may know what we will decide but we don't and so we must struggle to choose. Nothing else we know of in the universe faces that challenge or has that opportunity to create value on such a platform. We might read the long existentialist tradition in philosophy as built on that second view of our dignity: that undeniable and important sense in which our existence precedes our essence.

That is the end of my speculative pause. Now we must consider the normative case for two principles I distinguished, in turn.

6.

We might begin our inspection of the hydraulic control principle by focusing on the second requirement of hydraulic control: that decisions be causally potent. Suppose epiphenomenalism in the dramatic form I described earlier is true. Everything you do is initiated in your nervous and muscular system before you take the decision to do it. Your decisions, from the simplest to the most complex and far-reaching, are only part of an after-the-fact documentary film playing on the screen of your mind: what you do causes your sense of having decided to do it, rather than the other way around. The hypothesis is of course amazing. But what can it have to do with judgmental responsibility?

Responsibility is an ethical or moral matter: it attaches to final decisions whether or not these are causally effective. We might say: someone who decides to injure someone else, but whose decision is only epiphenomenal, is guilty merely of an attempt. He is trying with all his heart to do something bad. But he fails because his decision is not the cause of what happens. He wants to kill his rival, he decides to do so, the gun he is holding fires, the rival dies. But it wasn't he who killed him; it was (we might say) only his programmed reptilian brain. So what? At least in this kind of case, an attempted murder is morally as bad as a successful murder.

Lawyers like to invent cases like this one: A puts arsenic in B's coffee, intending to kill him, but just as B is about to drink, C shoots him dead. A is not guilty of murder, of course, but only of attempted murder. Nevertheless A is morally as much at fault as if he were a murderer; that is the assumption that makes the lawyers' question – why should A be punished less severely than C? – difficult to answer. Lawyers discover – or invent – policy or procedural reasons to explain why attempted murder should be punished less severely than murder. We want to encourage people to change their minds at the last moment; we can't be sure that A wouldn't have warned B just before he sipped the coffee. But these reasons of policy have no application here. So why shouldn't we say that the person who tries to kill his rival and fails because his decision is not the cause of the rival's death but only an epiphenomenal consequence of muscular behavior is nevertheless morally culpable? He is judgmentally responsible for having tried.

I agree that this comparison between the action of a single person and of two distinct people is strange. It is strange to treat a person and his reptilian brain as separate actors, the way we treat A and C in the lawyers' imagined case. But that artificial bifurcation of a person is exactly what the hydraulic control principle itself relies on. We normally treat people as whole people: the same person who has a mind also has a brain, nerves and muscles, and a person's acting involves all of these. The hydraulic control principle separates mind from body, personifies part of mind as an agent called the will, and then asks whether that agent actually causes the body it inhabits to act in a certain way, or whether it is only a fraud pulling levers disconnected from anything. If we accept this picture, however, we must, for the purposes of moral and ethical criticism, hold the person-within-the-person responsible for what he has tried

to do. Kant said that nothing is really good except a good will. If we were persuaded of epiphenomenalism, we would add that nothing is really responsible except a purposive will.

So there is little to be said for the second requirement of the hydraulic control principle. Now turn now to the first requirement. I am considering firing an employee. I investigate his performance and failures, I look into the consequences for his family, I calculate my chances of finding someone better, I reflect on whether loyalty or fairness requires giving him another chance. Then I fire him. Why should it matter, to my responsibility for that very deliberate decision, whether or not I was destined to take just that decision by events beginning in the spectral origin of the universe? Here is a way to focus that question. Assume for a moment that determinism is only *generally* true. Many of my decisions are predetermined by the long chain of historical events; my will – my self within myself – has no power to step out of these timeless causal chains. But – wonder of wonders – on some occasions this is not so. On these occasions my will does have power to originate decisions out of nothing. I then have the ability to act other than as someone with full knowledge of these causal chains would predict. So sometimes my decisions are only the causal upshot of external forces and sometimes they really do originate in me.

Of course the difference is invisible to me. Nothing in a person's experience can reveal to him whether his action is determined by external forces or originates in some uncaused act of will. However we make sense of that distinction, if indeed we can, the experience would be the same for the agent. So though I can believe that some but only some of my decisions are determined, I can have no idea, as I act, whether my decision to fire my employee is free or determined. Now also assume that instruments exist through which other people can detect my occasional sporadic originating spasms of will. They can connect me to scanning instruments that will allow them to decide, after I have acted, whether any particular action was determined or not. The hydraulic control principle holds that they would be right to praise or blame me for my decision if their instruments revealed that, as it happens, that particular decision was an original, causeless act, but wrong to praise or blame me if they revealed that it was determined from the beginning of time. That seems crazy. How could the crucial question whether I am responsible for some act depend

on information that is in principle and wholly inaccessible to me as I act? Something that can make absolutely no difference to the wisdom or stupidity of my act, its roots of motivation in my personality, its pre-meditation, the degree of passion with which I act, the pressures I was under as I acted, or anything else that we normally count pertinent in praising or criticizing people for what they have done? It seems perverse to make something so fundamental and important turn on random singularities that have absolutely no further consequences beyond their own occurrence. But if it would be perverse to apply the hydraulic control principle in these imagined circumstances, then it would be perverse to apply it at all, even if we assume that decisions are either always or never uncaused.

Making causal originality a condition of responsibility seems even stranger when we notice that none of what we might call the ingredients of rational decision display that kind of originality. We make decisions based on our beliefs and values but no one supposes that he can choose his beliefs or his values by an act of uncaused will. On the contrary, another form of responsibility I described – intellectual responsibility – depends on our *not* having that power. The laws of physics, among other things, fix how the world is and if we are rational those laws therefore fix how we think it is. It would be silly to think that we would have more judgmental responsibility for our acts if that were not so: if we had spontaneous thoughts about geography, physics and cosmology, or if we could whimsically decide for ourselves which beliefs on these matters would take root in our minds. If we didn't think that some combination of the state of the world and the state of our nervous apparatus produced our beliefs, we would have to count those beliefs as random visitations, and that randomness would provide an *excuse*, a reason for denying responsibility, not a reason for insisting on it.

Nor do people choose their values: their tastes, preferences, convictions, allegiances and the rest of their normative personality. I argued in Chapter *n* that our moral convictions are not caused by moral truth, that the causal impact hypothesis is false. If it were true, however, then our convictions would of course then be caused by something outside us – moral fact – not an originating will inside. If it is false, as I claim, then any competent causal explanation of our convictions must lie in the personal history I described in that chapter, which means that a complete explanation would include not only facts about my genes, family, culture, and environment but

also the causes of these: it would include the laws of nature and the history of the universe. This is even more evidently true of our tastes, desires and preferences. We cannot create these from nothing by some wondrous act of will. Yes, to some degree people are able to influence their preferences and convictions. We struggle to like caviar or sky diving, or to become better people by enrolling in churches or extension philosophy courses. But we do this only because we have other convictions or preferences or tastes we did not choose. People try to train themselves to like caviar or skiing because for a variety of reasons they desire to be the kind of people who do like them, and they did not choose to have the latter desire. They join churches or self-help groups to acquire or strengthen convictions they already want to have. In Chapter *n* I described what I called a rationality project: this requires people to try to work their various convictions, including their sense of authentic conviction, into a coherent and integrated whole. But these efforts at integrity respond to still deeper aspirations that we do not originate by any act of will either, and they are also sadly and inevitably frustrated, at least to some degree, by what we find we just cannot believe.

If the materials out of which my decisions emerge are inevitably determined by events beyond my control, why should it impair my responsibility that my actual decision is also determined by those same events? If I am rational, any seer who knew my beliefs, convictions, desires and tastes in the most marvelous and absolute detail, and had an incredible computer at this disposal, could predict my decisions with absolute accuracy: with greater accuracy, perhaps, than weather forecasters can ever hope to achieve. Given my beliefs about my employee's crimes, performance and value to the organization, my desires for an honest and efficient organization, my convictions about my assignment responsibility to my clients and my other employees, my positive taste for confrontation, and all the other features of my personality, it was inevitable that I would fire him. It would therefore be natural to say that whatever complicated personal history and state of the world caused me to have those beliefs, desires, preferences, tastes and convictions also caused me to make that decision.

Suppose I insist that if this is really true – if my decision was really inevitable – it follows that I was not in control of my behavior and that I am not responsible for my

act. It would be wrong to praise or criticize me for it, or for me later to take satisfaction or feel regret in it. I would only be responsible for firing the employee if, in spite of the conclusive case I recognized for firing him, I might not have fired him – if it was not inevitable that I fire him. Only, that is, if I could have ignored my own beliefs, desires and convictions and acted contrary to what these required. This seems at least initially paradoxical. Since in these circumstances it would have been irrational for me not to fire this employee, you would not think I was judgmentally responsible had I done so. On the contrary, you would look for some external pressure or mental instability that prevented me from acting as rationality would require: you would, in short, have supposed that I was not responsible for my decision not to fire. So if someone is in full control of his action, then he is the kind of person whose behavior is entirely predictable given absolutely full knowledge of the beliefs and values he did not choose. If his behavior is not predictable, given that knowledge, then he is not in control. Only something alien, like a disease, can then account for his behavior. So the hydraulic control principle seems to make someone responsible only when he is not responsible.

We might try to escape this alleged dilemma in the following way. We might say: someone is in hydraulic control of his behavior only when two conditions are met. First, it is possible for him to act entirely contrary to what all his beliefs, convictions and tastes would make it rational for him to do. Second, nevertheless he does act exactly as those beliefs, convictions and tastes require. However, it is hard to see why, if the second condition is met, the first condition is ethically or morally important at all. Imagine two people. Mother Teresa is through-and-through good. She has deep moral and religious convictions and it would be psychologically impossible for her to violate these. She finds any even mildly selfish act so repugnant that she is psychologically incapable of acting selfishly. Sister Teresa has the same convictions, and almost always behaves just as Mother Teresa does. But Sister Teresa has a streak of madness in her makeup and on some occasions, inexplicably, she steals small change from the church contribution box. On the latter occasions, Sister Teresa is not responsible for her acts: she acts out of madness. But the fact that she is capable of such acts shows that, even when she acts as selflessly as Mother Teresa, it is possible that she might act selfishly instead. So, according to our new hypothesis,

Sister Teresa deserves praise almost always, though not of course for her occasional mad lapses. Mother Teresa, on the other hand, never deserves praise because we can be satisfied, through careful psychological examination, that she is absolutely incapable of selfish behavior. Once again, that seems crazy.

Perhaps, when we consider the thesis that no one is judgmentally responsible unless he has the power to act against all his preferences and convictions, we should have in mind not psychological possibility but something else that may or may not hold at some deeper level. We might call this metaphysical possibility. You might think that Mother Teresa was metaphysically even though not psychologically capable of acting selfishly, and that it is that metaphysical possibility to act selfishly that makes it appropriate to praise her when she acts selflessly. Some philosophers offer analyses of metaphysical freedom that might fit this case. They say that people are really free when they act according to their reflective convictions about how they should act, or when they act consistently with maxims they can consistently legislate for everyone including themselves, or when they act as God would wish them to. But none of these theories is in point now, because we are not discussing metaphysics. We are not considering how we should define the conditions under which it would be right to say that, whatever their psychological limitations, people remain at some more fundamental level free. We are considering an ethical and moral claim: that it is wrong to blame or praise someone if factors beyond his control have made what he has done inevitable. If we accept that claim at all we must accept it as holding in all cases when, for whatever reason, someone is in fact incapable of deciding to act contrary to what he judges best overall. It would make no moral or ethical sense to exempt people when their acts are metaphysically inevitable but not when they are psychologically inevitable. We must then accept that Mother Teresa is not responsible for her wonderful acts of charity. Nothing turns, incidentally, on her goodness. We must take the same attitude to someone – Koba the Dread – who is psychologically incapable of acting against his wholly absorbing and (as we judge them) evil ambitions. We must judge it inappropriate to blame him for his horrible career.

I have yet to consider what some will think the strongest case for the hydraulic control principle. The popular responsibility system I described makes exceptions to judgmental responsibility. We are not responsible when someone physically forces

our body or manipulates our mind through hypnosis or electrical intervention. That is understandable; these are not our acts. But we are also not responsible when we are small children or seriously mentally ill. It might seem an important strength of the hydraulic control principle that it identifies and justifies all these exceptions. Indeed the familiar threatening argument that I described earlier begins with that claim. Pessimistic non-compatibilists argue that if we accept that mentally ill criminals should be excused because they are not responsible, we must for that reason accept that no one is ever responsible because everyone is actually in the same position. People who are mentally ill are not in control of their behavior, but neither are people whose actions are caused entirely by events beyond their control.

The structure of that familiar argument is important. It is addressed to people who think that they and other people are normally responsible for what they do, but who also assume that children and the mentally ill, among other people, are not responsible. It aims to show them that they already accept the hydraulic control principle. “You assume,” it tells them, “that there are crucial differences between your normal situation and that of children and the mentally ill. The hydraulic control principle captures what you must take the crucial difference to be. You think that in the exceptional cases people’s decisions are caused by events beyond their control, while in the normal cases people’s decisions originate the causal chain that ends in action. We now show you, by demonstrating the truth of determinism, that people’s decisions are never original in that way but are always the product of events wholly beyond their control.” The strategy assumes that the distinction people see between normal and exceptional cases is best explained as a difference in causal roles: that decisions in exceptional cases but not in normal cases are determined by events beyond the agent’s control. But the strategy fails because that cannot be what ordinary people think. They do assume that they are responsible for their decisions, and that children and the mentally ill are not, but the hydraulic control principle cannot be, for them, what justifies that distinction.

Consider, first, young children. Senior citizens make decisions that give effect to their desires and preferences, given their beliefs. We have no reason to think that young children, who certainly do make decisions, make them in any other way. We therefore have no basis for ascribing a different internal agency or hydraulics of decision to

them. I imagined, a page or so ago, that the hydraulic principle might be thought to require a capacity for perversity as a condition of responsibility. Few parents would deny that capacity to their young children. Whatever view we take about the freedom of an adult will must therefore hold for a young child as well. But of course there is a difference: it is the difference the other principle I mentioned, the creative control principle, picks out. Young children have a defective capacity, judged by normal adult standards, to form correct beliefs about what the world is like, and hence about the consequence, prudence and morality of their having what they want and doing what they want to do. They are often ignorant of “the nature and quality” of their acts. It is these incapacities, not any assumption about the causal pedigree of their decisions, that strikes people as requiring that children be relieved of some or all judgmental responsibility.

Now consider someone suffering from a serious mental disease: he thinks himself Napoleon or God and he also thinks that his status as such entitles or even requires him to kill and steal. He lacks the normal capacity to form beliefs about facts that are guided by facts. He is crazy and the familiar responsibility system holds him exempt from judgmental responsibility for that reason. But there is no reason to suppose that his decisions have either less or more initiating power than they would have had if he were not crazy. Like normal people, he acts in a way that is fully predictable given a full knowledge of his beliefs and normative personality. Ordinary people who adopt the responsibility system could have no reason to suppose that the decisions of a person crazy in that way have any less causal independence or originality than their own. True, they might find it natural to say that a crazy person’s disease has made him kill, which might suggest something special about the causal pedigree of his decisions. But that is only a figure of speech. Taken literally it is absurd; the disease, un-personified, is not capable of that kind of action. We speak more accurately when we say that the disease has distorted its victim’s judgment. But then, once again, we are invoking the creative not the hydraulic control principle to justify the exception.

Now consider a different form of mental disease: someone who though he is possessed of normal powers to form true beliefs and though he is committed to unexceptional moral, ethical and prudential convictions, is nevertheless unable to square his actual decisions with those convictions. Instances range from psychopaths

– the killer who begs society to catch and stop him before he kills again – to the physiological or psychological addict – the smoker or shooter or alcoholic or compulsive hand-washer who is desperate to stop but cannot. I distinguish these unfortunate people from people who have been hypnotized into behavior they would reject or whose minds are manipulated by a villain with a thought-control ray gun. I do not know what it feels like to be hypnotized, and no one knows what it feels like to have his thoughts zapped into being. I shall assume, however, that victims in these latter cases do not make what I called final decisions: real, felt decisions that merge into the actions the decisions contemplate. Their behavior is like a cough or other production of their autonomic nervous system. They do not act and so their behavior raises no question of judgmental responsibility. (If I am wrong, then their cases raise the same problem as those of the ill people I do discuss.) I do suppose, however, that psychopaths and addicts make final decisions: to kill or to light or shoot up. So we can sensibly ask whether it would make sense for ordinary people, who think that they themselves have judgmental responsibility for their acts, to excuse psychopaths or addicts from such responsibility because of some perceived difference in the etiology of their own and the latter's decisions.

We ordinary people, who believe that we are responsible for what we do but that psychopaths and addicts are not, concede that we ourselves are sometimes unable to overcome temptations of various sorts: we sometimes decide to do what our reflective values condemn as imprudent or wrong. We might or might not deliberate much; we might or might not struggle. But temptation wins: we say: "Just this once," or, "The hell with it," and we order another steak pommès frites. We do not think that on these occasions we have been hypnotized or zapped; we do not think our wills have been robbed of their ordinary originating power. We think, on the contrary, that our wills are to blame: we say we have been weak-willed and we resolve not to sin again. We count the occasion as showing not a conquest of our minds by some alien force but a failure of our mind's ordinary capacity to organize and direct our reflective convictions. We can find no reason, in this account of our own lapses, to think that an addict's situation is an entirely different matter rather than only a difference in degree. We have no basis for supposing that some alien force has usurped the role of the addict's will either. We may say that since he yields even though at some level he

knows that the result will be disastrous, he is very much weaker than we are. He is in fact incapable, we might say, of controlling his immediate impulses; perhaps, in the moment of acting, he is even incapable of understanding his peril. But then we are not assuming that something about the causal history of mental events distinguishes his case from ours. We count the difference between us and him as one of capability and therefore of degree. That latter explanation does not invoke the hydraulic control principle; it makes no assumption, either way, about determinism or epiphenomenalism.

I'll summarize this part of the argument. The hydraulic control principle is popular among philosophers as an interpretation of the more basic idea that people are not judgmentally responsible for their acts when they are not in control of what they do. But that principle is an ethical and a moral judgment, it must be assessed as such, and it finds no support in those departments of value. It is contradicted by familiar principles and assumptions: that people are responsible when they attempt harm, even when the attempt is unsuccessful, for example. It seems arbitrary because when we imagine that the test might be met in some ordinary circumstances but not in others, the difference it makes in those circumstances seems wholly inconsequential. It seems pointless since the ingredients of decision – desires, tastes, convictions and the rest – are in any case not chosen or under an agent's control, so that independence from external causes could only mean freedom to be irrational. It seems unhelpful because it fails to explain why people who believe they normally have judgmental responsibility nevertheless suppose that they and others lack that responsibility in exceptional circumstances.

7.

The creative control principle is much less ambitious and much more a matter of common sense than the hydraulic control principle. Someone is in creative control of his action when at the time of acting he has the two capacities I mentioned in my initial description of the responsibility system. He must have a minimal capacity to form pertinent beliefs about the world in which he acts, beliefs that respond to genuine evidence. He must also have a minimal capacity to match his decisions to his full normative personality – the full set of what he identifies as in some way good or

desirable or appropriate for him to have or do. People in fact have these two capacities to very different degrees. A brilliant scientist is better at forming true beliefs about the physical world than I am, and someone less impulsive is better at conforming his decisions to what he actually thinks good to have or do. The principle supposes a threshold level of these capacities, and much of the controversy among lawyers and laymen about when it is proper to hold people responsible for their behavior is controversy about where the threshold should be set. I shall take as my initial examples of failed creative control those instances in which the failure is egregious and undeniable. An idiot cannot form a large enough stock of stable true beliefs about the world to make his life safe let alone profitable; he lacks the minimum level of the first capacity. Someone with serious frontal lobe brain injury may be wholly unable to avoid aggressive and violent behavior even though nothing he thinks or wants or approves recommends that behavior. The creative control principle holds that the idiot and the victim of serious brain damage are not judgmentally responsible for the decisions that manifest these incapacities.

That principle offers to justify the central features of the responsibility system I described: it explains why people normally have judgmental responsibility for the decisions they make and also why, in the exceptional cases I described, they do not. But we need to justify the principle itself. Why, if it does it not matter whether our decisions are made inevitable by events long ago, is it nevertheless crucial whether we now have the capacities I described? If a decision is inevitable anyway, why should it matter what capacities were exhibited when the decision was finally made? Once again, as on many other occasions in this book, we must distinguish that question of justification from different questions of explanation. The creative control principle, I believe, is very widely accepted; it provides the core of the familiar responsibility system. If so, what explains that popularity and persistence? That is a question of psychological, social and perhaps biological explanation: neo-Darwinians would no doubt have an answer to it on hand, or could easily manufacture one. Our question, however, is one of justification not historical explanation.

We cannot rule out in advance a consequentialist justification that brings explanation and justification closer together. A utilitarian might suppose, for example, that the widespread acceptance of the creative control principle contributes to the general

welfare and is justified for that reason. The consequentialist case for the creative control principle is at least more plausible than any consequentialist case could be for the hydraulic control principle. It is fatal to the supposed utility of the latter principle that no one can apply it, either to himself in retrospect or to others, without knowledge of mind-body interaction that is in principle unavailable. The creative control principle on the contrary could readily be applied, at least in fairly clear cases, with knowledge readily available to anyone considering his own past behavior or the behavior of others. It is also sensitive to improvements in medical knowledge as any principle underlying responsibility that is justified on consequentialist grounds should be. The history of the criminal law in developed countries shows the impact of those improvements on the development of the insanity defense, for instance. On reflection, however, we cannot accept a utilitarian or any other familiar form of consequentialist justification for the creative control principle. I argued, in a Chapter *n*, that we have no reason to assume that we each have, as part of our basic personal responsibility to realize value in our own lives, any general duty to improve the overall welfare of humanity as a whole or any particular part of it.

The strongest case for the creative control principle, I believe, is a more fundamental one. It draws on the foundational principles of ethics that we have been exploring throughout this book. [I now summarize claims I explained and defended earlier in the book.] Each of us has an inescapable ethical responsibility to try to make something valuable of his life. That responsibility yields two projects. Each of us must try, first, to live well and also, second, to make his life a good one. The former, adverbial, project is the more fundamental of the two and it may require and therefore justify compromising the second. Only the former project directly implicates judgmental responsibility. How good a life you have depends not just on what you do yourself but also on what happens to you. But how well you live depends only on what you do yourself.³ Living well means identifying, even if inarticulately, standards

³ The distinction between having a good life and living well is important in many areas of ethics and morality. For example, it allows us easily to account for the familiar and largely unrelated phenomenon now often called “moral luck.” (See Nagel and Williams.) People often and sensibly feel great personal remorse for terrible events in their lives for which they have no fault. The school-bus driver who drives impeccably but crashes nevertheless, killing many of the children in his care, may believe that his life has been ruined by the tragedy. He regrets

of success and then creating a life that is structured by those standards. It means making of a life not just a chronology but a narrative woven around values of desire, ambition, character, taste, loyalties and ideals. No one creates a narrative of perfect integrity: we all act out of character, as we put it, sometimes. Many people's lives, judged as narratives, are picaresque or even shambles. – Hubbard's "one damned thing after another" or Millay's "one damn thing over and over." But shambolic lives are not lived well, no matter how full of worldly success they turn out to be.

We construct our personal narratives through what I called final decisions: those final decisions that are merged into and that we cannot pry loose from our actions. I have already emphasized that we cannot not make such decisions. We cannot test determinism or epiphenomenalism by waiting to see whether our nerves and muscles will act on their own without the mental component of that final decision. It is those final decisions that are the raw materials out of which we construct our life's narrative. It would make no sense for me to deny, as I write a paragraph or end a love affair, that this is my act or that it should count in my own or other people's judgment about my successes or failures. I can make no other sense of my life going well or badly except to suppose that this is a matter of what I have decided to do. Taking responsibility for a decision is just to accept that it counts, and if I am aware that I am leading a life I cannot act without supposing that each final decision does count. Others, and I myself later, may judge that some particular decision does not after all count, that I should not be held responsible for something I have done. But once I am conscious of leading a life, I cannot think this as I act. Almost all of these final decisions – those we think trivial – are made unreflectively, of course. But even these are assessable in retrospect if they turn out to be consequential. That retrospective

the children's deaths, as anyone would, but also and independently regrets that it was he who drove the bus that killed them. He has indeed, as he knows, had a worse life in consequence. But he has not lived his life less well: he should not have the different kind of corroding despair that someone feels who acknowledges his fault for a tragedy. Someone who has done great harm while seriously mentally ill is in the same position. His life has been ruined by his illness, but it would be wrong to say that in consequence he has lived that life badly .

assessment of all our consequential decisions will ask: have we decided as well as we should? How have we stood up, in the only way we can stand up, to our mortality?

We can exempt certain decisions from counting in any overall ethical assessment; we can do this for other people as they act and in retrospect for ourselves. Does every decision we have made count, even in retrospect, in judging what narrative our life exhibits? The decisions we made while children? While ill or under extreme pressure of some sort? A theory of responsibility is a response to these questions. It offers a screening filter and we must judge any account of responsibility, at least in the first instance, by asking how well it performs in that role. We must design such a filter with an eye to the overall ethical project and to the human situation as we understand it. We must not exclude so much that we have made the project of living well either pointless or impossible. We cannot, for example, exclude every decision we took when guided by a conviction or desire we did not choose. That would exclude everything – it would make the project of living a life impossible. In much of this book I argue that we live well only if we make our important decisions only after reflecting on, and attempting to find integrity in, our ambitions and convictions. But it is not a matter of choice for us which ambitions and convictions survive that process. I must take judgmental responsibility for my normative personality even though I did not choose that personality, if I am to respect my more basic assignment responsibility to live well. It is the imperative of that latter responsibility, not any metaphysical freedom, that designs our responsibility system.

We can, however, adopt a much less dramatic screen. In a variety of quotidian contexts we distinguish between someone doing a job badly and not being able to do it at all. We do not hold a blind person accountable for his reading deficiency. The creative control principle offers a similar distinction at a more abstract level. It requires that we count, in assessing how well someone has lived, only those decisions that he has made when he had threshold levels of the two capacities that principle names. I asked for a justification of the creative control principle. If it seems plausible that we should not hold someone to have failed in his ethical responsibility to live well unless he has those particular capacities to a minimal degree, than that fact supplies the justification of the creative control principle that we seek.

Dogs, we believe, can have good or bad lives: they suffer pain and are often mistreated. Not every dog has a dog's life: some have lives that other dogs might envy. But dogs cannot live well or badly. People normally can, because they normally have the two capacities the principle cites. Creating a life requires reacting to the environment in which that life is lived; a person cannot sensibly be treated, or in retrospect treat himself, as creating a life unless he can form beliefs about the world that are largely responsive to how the world is. An idiot or someone who thinks he is Napoleon or that pigs can fly lacks that minimal ability. Philosophers sometimes imagine that they are only disembodied brains in a nutrient vat, pervasively deceived by a master intelligence into thinking that they are embodied organisms living on planet Earth. If that is true, then they are not leading lives. If we all assume, as we all must, that we are not brains in a vat, then almost all of us have the epistemic capacity we need for most of our lives. But from time to time some of us lack or lose that normal ability in one way or another, and then our responsibility for what we do is called into question.

The second capacity the creative control principle requires is regulative. Someone cannot lead a life if he is not capable of forming a normative personality – a stable system of desires, preferences, tastes, convictions, attachments, loyalties, ideals and the rest – and making decisions that exhibit that personality. Of course, as I said, everyone acts out of character from time to time – seized by a whim or impulse, perhaps. And people's normative personalities change over time, sometimes dramatically. A sybarite may turn into an ascetic or – though this is rarer – the other way around. But if someone's behavior cannot be interpreted by himself or others, even from time to time, as revealing any particular personality, any coherent ordering of tastes, desires, ambitions or convictions, any ground for ascribing selfishness or selflessness, industry or laziness or anything else except randomness to the course of decisions that make up the course of his life, he is just enduring a life not creating one. His life may be a bad one – or, perhaps, a good one – but he has done nothing that can be judged a success or even, I think, a failure in living. Harry Frankfurt imagines human beings – he denies them the title of people – who lack that power altogether, either because they have no normative personality or because they cannot bend their decisions to that personality. He calls these creatures wantons, and says

that wantons lack the kind of freedom we must have in mind when we speak of a free will.

Of course I am not responsible for the damage when someone else pushes me into a blind beggar. I make no decision so the question of my epistemic or regulative capacities to make competent decisions does not arise. Perhaps, in spite of my earlier assumption to the contrary, I do make final decisions when I am hypnotized or when a fantastic scientist manipulates electrodes implanted in my brain. In any event, these events might – they probably would – make my life worse. But they would not count in judging how well I have lived. If I am to respond to the challenge of living well, I must have the capacity to match my behavior to my sense of what living well would mean. It does not impair that capacity that this sense has inevitably been molded by forces gathered in my personal history including my genetic history. These forces have shaped my personality, but they do not impair my capacity to make purposive decisions by matching those decisions to the personality they have shaped. It obviously impairs that capacity when others have taken over my decision-making capacity to serve their own ends. That usurpation disconnects my decision from my personality so that it is at best an accident when these match. It is therefore sensible that when I ask how well I have lived I distinguish between what I did when I had ample capacity to put my own desires and convictions into practice and what I did when I lacked that capacity. I take responsibility only for the former.

For the same reason, the creative control principle requires me also to disregard, in retrospect, what I have done when suffering from serious cognitive impairment or mental disease. These conditions diminish or destroy the capacities on which judgmental responsibility depends. A beginning infant does not make decisions at all; a very young child does, but he does not have the cognitive or critical ability needed to match his decisions to any self-consciously recognized desires. Mental illness may savage either or both of the judgmental capacities in anyone; indeed serious loss of either might be a defining condition of mental disease. The history of the insanity defense debate that I shall briefly describe later in this chapter shows a pendulum swing between a strict doctrine that emphasizes loss of epistemic capacity and a more generous doctrine that also makes regulative capacity critical.

The creative control principle functions, of course, as a moral as well as an ethical principle. In that role it plays no direct part in anyone's judgment of how well he or anyone else has led his life: instead it serves as a threshold condition for blame and sanction. We must therefore ask what justification we have for exporting the principle from the ethical to the moral arena in that way.⁴ It is a central demand of self-respect, I said in Chapter *n*, that we must not only take personal responsibility for making something of our own lives but must also treat the principle that requires this as an objective principle of value. This means recognizing and respecting the same responsibility in others. That requirement cannot be met – we cannot be treating the principle of personal responsibility as having an objective standing – unless we understand personal responsibility to have the same character and dimension for everyone – the same character and dimension in morality, that is, as it has in ethics. Suppose I rely on the creative control principle in criticizing myself: in deciding whether it is appropriate to feel shame or guilt or only deep regret for some decision I wish I had not taken. I hold myself responsible unless I am satisfied that I lacked some capacity essential to creative control when I took that decision. What justification could I then have for using a different – stricter or more lenient – standard for judging the guilt of someone else? For deciding whether it is appropriate, other conditions being met, to punish him in some way, or appropriate only to sympathize with him? That would mean my judging and treating him as I think he ought not to judge and treat himself. It would be an act of disrespect to him. Since it would deny that the principle of personal responsibility is an objective one, it would also be a failure of self-respect.

We have already met a dramatic form of that incoherence. Some criminologists say that since science has shown that no one has free will it would be wrong to punish anyone for anything. We should treat those we now style criminals medically rather

⁴ I must take care, here as elsewhere, to guard against being understood to mean that the moral use of the responsibility system is subordinate to or dependant on its ethical role. I have emphasized the ethical role of the creative control principle because I believe it is easier to see its importance from the first person perspective. But any interpretive argument for the truth of the principle in ethics depends on its making independent sense in morality as well. The history of the criminal law across civilized nations seems to me to demonstrate that it does; I amplify that statement briefly at the end of this chapter.

than criminally, hoping to reprogram rather than to punish them. This judgment supposes that “we” have responsibility that other people lack, that we can judge ourselves to act unfairly and therefore wrongly while we can only judge everyone else to act dangerously or inconveniently. Most people have a strong negative reaction to the proposal that outlaws should be treated medically rather than punished criminally. They think that this would de-humanize outlaws. They sense, I believe, that this proposal fails the sovereign requirement that we treat responsibility in others as we cannot help but treat responsibility in ourselves.

8.

The hydraulic control principle fails the test I set: it finds no support among our other moral and ethical convictions. The creative control principle passes that test. It serves as the foundation for the root ethical enterprise of making value through our lives and so it also fits the structure of morality that flows from and into that enterprise. Now we must consider whether the latter principle makes judgmental responsibility hostage to science in the way the former one does. If we embrace the creative control principle, can we also accept determinism or epiphenomenalism without sinking into the incoherence of denying that we have judgmental responsibility for anything?

That principle makes responsibility turn, not on the hydraulic causes or consequences of a decision, but on the drama of decision itself. It treats the struggle of decision as the proper theatre of responsibility but makes no assumptions about how the stage on which the drama unfolds came to be arranged. Of course, in judging the merits or demerits of our final decisions, we and others pay great attention to the consequences that we foresee, or ought to foresee, of acting as we decide to act. But that attention presupposes no causal efficacy. It presupposes only what logicians call material implication: that if I decide to pull the trigger, someone will die through an action of mine; if I do not he will not. I can know the truth of those conditionals from my experience without making any assumption about the hydraulic force of my decision on the muscles that pull my trigger finger back. The conditionals are consistent, that is, with epiphenomenalism, even though they are also consistent, of course, with denying epiphenomenalism.

The principle does make exceptions for what it treats as pathological cases: it conditions responsibility on the capacities of the agent. But again these are not causal conditions. The principle makes capacities crucial to responsibility not because normal people have wills that are in charge while a child or an idiot or a madman does not, but because it sets conditions on responsibility with an eye to the overall ethical judgment whether an agent has created value or disvalue in his life by the way he has lived it. It declares that that overall assignment responsibility is in play only when a person is capable of pursuing the assignment. A toddler or idiot or madman makes decisions and presumably makes them with some sense of responsibility for them. But he should reject judgmental responsibility for those decisions later, when he grows or if he recovers, and the rest of us should reject them now. We think -- and the toddler, at least, will later come to think -- that it would be right not to count those decisions in deciding how well he has lived.

I conclude that the creative control principle makes judgmental responsibility compatible with determinism and epiphenomenalism. If we accept that principle as the ethical basis for our responsibility system, we can await the latest exhilarating discoveries about the geography and electro-dynamics of our brains with boundless curiosity but with no terror. However that claim will hardly pass unchallenged, and I must now consider certain challenges. The most primitive is simple disbelief. How can it *not* matter, to whether someone is to blame for what he does, whether he was made to do it by forces beyond his control? Galen Strawson puts that objection this way: we need an account of judgmental responsibility that explains why, at least in the eyes of many people, God is justified in sending some people to Hell while others escape to Heaven. Any account that meets that condition, he insists, must accept that "To be truly morally responsible for what you do you must be truly responsible for the way you are -- at least in crucial mental respects."⁵ In other words, judgmental responsibility for some decision requires causal responsibility for the beliefs, desires and convictions out of which the decision is made. The argument of this chapter so far is an answer to that claim. Since we can find no support for the hydraulic control

⁵ Galen Strawson, *The Impossibility of Mental Responsibility*, *Philosophical Studies* 75: 5-24, 13.

principle in ethics or morality, causal responsibility is neither a necessary nor a sufficient condition of judgmental responsibility. The creative control principle holds that a final decision, on its own, can be a sufficient basis for condemnation or punishment. Final decisions merit Hell if anything does.

But how can it be right to blame or punish someone for what he could not help doing? That question, we now see, is too crude because it neglects the crucial distinction between what someone does without deciding to do it and what he does after deciding to do it. If he cannot help himself because he is pushed or constrained by some physical force, or because his brain has been taken over by someone else, then it is unfair to punish him because punishment is appropriate only for decisions. If he cannot help the way he has decided, because the history of the universe conspired to give him beliefs and desires that made his decision inevitable, then it is permissible to punish him, if other necessary conditions are met, because, again, punishment is appropriate for decisions. Galen Strawson thinks these cases are on a par, morally and ethically speaking. But they are not. Our inescapable overall ethical responsibility to live well is not engaged when we are pushed or constrained or our brains are manipulated. That is the distinction that both Hobbes and Hume, as well as many contemporary compatibilists, rely on. They say that it is sufficient for responsibility if the agent was not impeded by some external force in doing what he decided to do. If he is not impeded, they say, then he could and would have acted differently if he had chosen to. It is said, in reply, that some people – addicts, for instance – are plainly unable to choose otherwise than as they have chosen, and that if determinism is true that is the situation we are all in. So if addicts are excused from responsibility, and determinism is true, we must all be excused as well. That response makes the mistake I have been trying to expose. We excuse an addict on grounds of competence; we do not excuse him because we think the biological mechanisms of his decisions are different from our own.

Now consider a more complex objection: that if determinism or epiphenomenalism is true, then the creative control story I told rests on illusion. What illusion? I conceded, when I described how it feels to make a decision, that we feel as we ponder that we could decide either way. But that feeling is sufficiently vindicated by what Hobbes and Hume suppose: we feel that we can decide as we finally think best. Determinism

does not contradict that sense: it claims rather that what we finally think best has already been determined, though in a way that denies us all access to the content of the determination. No doubt many people believe that determinism and epiphenomenalism are both wrong: in fact absurd. They believe that it has not already been decided what they will think best; that this is a matter of their spontaneous manufacture here and now. But whether that further thought is coherent or not, it plays no part in the story I told.

Perhaps the illusion is a deeper one. I described the role a responsibility system built on creative rather than hydraulic control plays in people's efforts to live well. I distinguished having a good life from living well, and said that the latter goal presupposes the capacities the creative control principle recognizes. It might now be said that the project of living well makes sense only for people whose decisions are not fixed in advance by external forces over which they have no control. I said that if we were, unknown to us, only brains in a vat, we would not be leading lives at all. If determinism is true, why aren't we all in a similar position? We are manipulated by outside forces just as we would be if we were only floating brains. But we are not in that position. The floating brains are in complete ignorance of their situation; they have no way to discover it. They wholly lack the capacity to form beliefs based on evidence. Most of us have that capacity in ample degree; indeed we are now supposing that we have the capacity even to discover that all our decisions are determined by ancient events. We are not in either complete or terminal ignorance.

Am I missing the point? The objection may be more straightforward: it is an illusion to suppose that anyone can attract credit or blame for doing something well or badly if it was not his doing at all, but rather the work of forces beyond his control. The bare phenomenon of a decision can't attract these judgments; the decision must be the agent's original achievement. We wouldn't credit an artist with any success or failure if his hand created a painting because it was attached to a mechanical device like a signature machine manipulated by someone else. So we shouldn't claim credit for living a life well through our decisions if it was determined long ago, by cosmic forces, how we would decide. This is only the old objection we have been considering in different forms for some pages. But with a helpful twist, because it switches focus from judgmental responsibility for particular decisions to creative credit or discredit

for a life as a whole. I said that the decisions a person makes, considered as phenomena in themselves without regard to their hydraulic ancestry or consequence, fix how well he has lived. Does this make sense?

The painter begins on a giant canvas. He dreams and imagines. He sketches, draws, paints, rubs out, paints over, despairs, drinks, returns, stands back, paints feverishly, stands back, sighs, lights up. He is done; his canvas is exhibited, we adore it and we celebrate him. Then a guru in the Arctic Circle calls a press conference. He unveils an exact replica of the great painting. Newly sophisticated carbon dating proves that it was created an hour before our artist began his own work. The guru explains that he has a painting machine directed by a powerful computer at whose disposal he has placed an exact description of every event since the beginning of time, including information about the artist's various abilities, his convictions about greatness in art and his beliefs about the tastes of rich collectors. We are amazed. But does the revelation lead us to value the artist's efforts or achievement less? Of course not. We valued what the artist did, before the press conference, because we admired the way he made the many thousands of decisions that ended in the wonderful picture. None of that has changed; our amazing discovery cannot have cheapened the worth of a single one of those decisions. It would be incoherent to admire him less now.

Suppose the guru, instead of predicting the picture, actually made it. He didn't tie the painter's arm to a mechanical device; instead he transmitted radio signals that moved the molecules of the artist's cerebellum in such a way that artist's arm moved as the guru dictated. We wouldn't give the artist credit then, of course, any more than we would if the guru had used the mechanical device. But now suppose that the radio signals also made the artist think that the thousands of decisions he was making were his own decisions though they were not. He thought, as he painted, that he was making his own painting not someone else's. But he was wrong. Making artistic decisions yourself means bringing to bear your own sense of the various aesthetic values in play and your own skill in exhibiting those values in a concrete work. That is why the creative control principle makes some level of the second, regulative, capacity essential to responsibility. And that is why someone else's painting through you is different from your painting yourself even if, in the latter case, your aesthetic values and skills were predestined by nature to take exactly the form they do take.

Our artist might be brainwashed, we are assuming, into thinking that it is his own artistic genius that is now displayed on the canvas before him. I imagined that a hypnotized patient might be in that position. But when he learns that the canvas actually signals the artistic skills of someone else, and his own only by accident if at all, he will abandon all pride – or shame – in what he has done.⁶

One more challenge. It might be said that if determinism or epiphenomenalism is true, people never have the capacities the creative control principle assumes they normally do have because these capacities require the kind of causal originality or power that one or the other of these hypotheses would make impossible. But the capacities require no such thing. The first of these is the capacity to form true beliefs about the physical world and the mental states of other people. It does not damage that capacity that people's beliefs about the external world are caused by events beyond their control; on the contrary, as I said, it is exactly that fact that endows people with that capacity. Nor can it damage that capacity that their final decisions do not enter into causal relations with their nerves and muscles; that fact, if it is a fact, is completely irrelevant to the existence of the first capacity. The second capacity is regulative: the principle assumes that people can normally make final decisions that can be understood as serving their desires and convictions in the light of their beliefs. That is an assumption about the character not the etiology or causal consequence of final decisions. People have that capacity if their decisions have the stipulated character whether or not they were fated to have that character. A fast car, whose behavior is certainly determined by events beyond its control, nevertheless has the capacity to exceed the speed limit.

⁶ We can turn this screw through more turns of fantasy. We imagine that the guru didn't radio discrete hand movements to the artist's brain but rather implanted the more general tastes – a sense of the artistic possibilities of abstract expressionism, perhaps – to which the artist responded. Or – a more difficult case still – the more concrete insight that this genre might be exploited brilliantly by swinging leaking paint cans over a prone canvas. We can in this way manufacture hard cases for any judgment about the artist's responsibility for what his hand has made. These fantasy cases are hard, however, because we imagine two decision-makers rather than one, and the facts make it unclear whose values and skills a particular decision should be understood as exhibiting. That complication is absent when it is nature, rather than an Arctic guru, that has shaped an artist's skills, taste and judgment.

9.

The choice between the hydraulic and the creative control principles is important, I said, for reasons that go beyond the old free-will controversy. It is important to the much more practical controversies I mentioned, among people who accept the general structure of the responsibility system, about its application to particular cases. If we think that people are responsible only when their actions in fact flow from a spontaneous, uncaused act of will, then we think the determinative question in these practical controversies is a psycho-biological one. When someone claims that he committed his criminal act in a blind rage, or when overcome by an irresistible impulse, or under extreme duress of some kind, or because he grew up disadvantaged in a ghetto, or because he had watched too much violence on television, we would ask: were these forces or influences strong enough, in the circumstances, so that they usurped his will's normal causal role, like a drunken sailor pushing the helmsman aside and taking the wheel? So that it was not his will but rather an overwhelming surge of sexual jealousy that provided the efficient cause of his muscles contracting around the trigger? I do not understand that question, and I doubt that many of the citizens, lawyers and judges who would have to answer it understand it either. Perhaps the popularity of the hydraulic control principle among philosophers has contributed to the confusion that marks this area of the criminal law.

If we reject that requirement of responsibility, however, in favor of the different principle I have tried to defend, then we pose a different question. We must ask instead: did the accused lack one or the other of the pertinent capacities to such a degree that it is inappropriate to ascribe responsibility to him? That question calls for two judgments: an interpretive judgment about his behavior and an ethical and moral judgment that reasonable people will make differently. It is therefore often a difficult question but not, I think, a mysterious one.

People who must try to answer it – jurors after hearing volumes of testimony, perhaps – will have different opinions about the interpretive issue. They will disagree, for instance, about whether the defendant's general behavior showed an admiration for violence as a virtue, as part of his self-image, so that his violent act on this occasion confirmed rather than contradicted his general capacity to suit his action to his

convictions. They will also disagree about the more evidently normative issue – about what level of incapacity is sufficient to let someone off the responsibility hook. We admire people who at least begin to answer that question introspectively. Would I regard myself as responsible, in retrospect, if I had the kind of incapacity the defendant’s act revealed? That is the spirit of the attractive thought, “There but for the grace of God go I.”

Sadly, neither common experience nor the history of the insanity defense suggests that many people do trace their judgments of others to hypothetical self-criticism. Outrage is a more frequent spur. When the public has been particularly anxious for vengeance after some crime, judges and legislators have responded by cutting back the coverage of the insanity defense. The M’Naughten Rule, named after the woodcutter who killed Peel’s secretary while trying to kill the prime minister himself, shrunk the defense to allow only the first, cognitive, capacity to count, and stipulated that only a particularly low level of even that capacity could excuse. Over many decades most American states moved from that strict rule to a more forgiving one that permitted the accused to argue that he was confronted with an irresistible impulse. But asking juries to judge the appropriate level of the second, regulative capacity proved unwieldy, and the results often seemed too permissive to many scholars as well as to the general public. The argument, made to a Florida court, that the defendant lacked the necessary regulative capacity because he had watched too much television, seemed a *reductio ad absurdum* that called the standard itself into question.

Many states have now adopted a different approach based on a PLI recommendation: the insanity defense is available only when the defendant is shown to exhibit a mental disease. That rule by no means eliminates the need for judgment that different lawyers, judges and jurors make differently. But it does change the focus from discrete event to general capacity. This has evidentiary advantages: it is easier to judge whether a defendant has shown a general incapacity, manifested in other ways, rather than just a single-shot temporary incapacity exhausted in the crime it is alleged to excuse. Requiring a showing of mental disease also reduces the vagueness of the defense: the label “disease,” even if not a medical term of art, is itself a classification. We do not regard someone as suffering from mental disease if his cognitive and

regulative capacities fall only somewhat short of what we take to be normal. They must be low indeed.

Understanding the responsibility system as based on the creative control principle is also helpful in distinguishing a plea of incapacity that might justify denying judgmental responsibility from other kinds of excuse that must find some other definition and defense. Duress plainly falls into the latter category. If someone obeys an order to kill but only because he is threatened with death himself if he does not obey, he cannot claim that he lacks either of the pertinent capacities to any degree. He obeys because he understands his situation accurately, and because he is able to conform his decision to his reflective judgment of what is best for him. If he kills, his act might or might not be excusable. It would clearly be excusable if he was ordered, under the same threat, to commit some minor crime: stealing handkerchiefs for Fagin, for instance. Perhaps it is excusable even when the crime he commits is grave; even perhaps when it is capital. But there is no place here for any suspicion that he lacks responsibility for whatever decision he makes. He is responsible, so the question is rather whether he made the decision that the law requires him to make. Torture, at least in extreme forms, is different. The villain who threatens torture hopes to change his victim's options just as the villain does who threatens death. Someone facing torture remains responsible for his choice whether to obey to avoid it. But when the torture begins, the torturer's aim is different: it is to create so much pain that the victim does lose his capacity to decide whether to yield or not. The torturer aims to reduce his victim to a screaming animal who no longer is able to reason in that way. He aims, that is, to extinguish his victim's responsibility.

Further questions arise if the general account of judgmental responsibility I have defended is right. It is common to say, for instance, that people who grow up in ghetto poverty are less responsible for any anti-social behavior than are people from more privileged backgrounds. I suppose it would be possible for someone to think, if he accepted the hydraulic picture of a will normally in control, that poverty or other disadvantage can usurp the will's place and cancel responsibility in that way. That makes as much sense to me as any other use of the hydraulic will picture. But on the different view of responsibility I defended the claim of diminished responsibility seems problematic. Ghetto survivors are no less capable of forming accurate views

about the world or of matching their decisions to their desires or convictions than people from more comfortable social strata are. I leave this question dangling in case the paper has not yet supplied sufficient material for a long discussion. If we hold people responsible for what they do in service of their normative personality even though they did not choose the personality they have, can we find any justification for the impulse to think that people who grew up in disadvantage are less responsible for their behavior? Is mitigation of blame part of a compensation package for prior injustice? Or do we have some other reason for supposing that though a person's convictions are never of his own choice, the particular circumstances in which these convictions were formed nevertheless matters for judgmental responsibility?