1. The philosophical foundations of the insanity defense

1 MODELS OF THE CRIMINAL RESPONSIBILITY OF THE MENTALLY ILL: A FIRST GLANCE

The insanity defense is the standard (adopted in nearly all contemporary legal systems) model of the regulation of criminal responsibility of the criminal offenders who are mentally insane. It is based on the *prima facie* uncontentious assumption that the mentally sane can be held criminally responsible, and the mentally insane cannot. This assumption in fact embraces two different claims: that the concept of mental illness/disorder is fully legitimate, and that mental illness/disorder (in conjunction with its psychological effects, which are certain mental – cognitive and/or volitional – incapacities or at least serious impairments of certain mental capacities) constitute an exculpatory circumstance or at least one mitigating the penalty for committing an offense.

Yet the very fact that this model is dominant (and thus why I call it “Standard”) does not mean that none of its assumptions cannot or have not already been called into doubt. In fact, as already mentioned in the Introduction, one can distinguish four models of the regulation of the criminal responsibility of the mentally insane on purely logical grounds. The distinction is made on the basis of two criteria, namely mental state (mentally sane and mentally insane), and criminal responsibility (admitted, excluded):

(M1 – *Standard*): Mentally sane → Responsible, Mentally insane → Not-responsible
(M2 – *Perverse*): Mentally sane → Not-Responsible, Mentally insane → Responsible
(M3 – *Negative Equalization*): Mentally sane → Not-Responsible, Mentally insane → Not-Responsible
(M4 – *Positive Equalization*): Mentally sane → Responsible, Mentally insane → Responsible

M1 is the already mentioned Standard Model (the insanity defense): It treats the difference between the mentally sane and insane as relevant in the context of criminal responsibility. M2 also treats this difference as relevant but does so
in an idiosyncratic, “perverse” manner: Mental sanity, not mental illness, is an exculpating circumstance. The other two models can be called “equalization models” since they regard the difference as irrelevant (or unfounded): M3 can be called the “Negative Equalization Model” because it assumes that neither the mentally sane nor the mentally insane can be held criminally responsible, whereas M4 can be called the “Positive Equalization Model” because it assumes that both can be held criminally responsible. It is interesting to note that M3 and M4 were endorsed by many philosophers and M4 was in fact adopted in some legal systems (as we shall see later in this chapter). M2 was neither defended philosophically nor adopted in legal systems – so it is distinguished here only for classificatory reasons. But it is worth noticing that one can distinguish a related model that was adopted (albeit rarely) in legal systems, viz. the model in which the mentally sane are held responsible for offenses they have committed, but the mentally insane are held more responsible for the same offenses, i.e., in which mental illness is an aggravating circumstance (I shall say a few more words about this model in Section 4 of this chapter).

The bulk of my analysis will be devoted to examining the philosophical foundations of the above models. I therefore attempt to present the Standard Model in a broader context by depicting other possible models of criminal responsibility of the mentally insane. The importance of providing such a context stems from the fact that the presuppositions of different models of criminal responsibility of the mentally sane have not as yet been presented in a systematic way in the relevant literature. This chapter is aimed at filling in this lacuna, which means that its main goal is analytical. I argue that the models are determined by two factors: the conceptions of criminal responsibility and the accounts of mental illness. However, the task of this chapter is not only analytical; I also provide arguments for the normative claim that the Standard Model is the optimal one (and that thereby it is not by chance that the model is “standard”). The arguments will, at least in part, appeal to the personalist view of human nature presented in the Introduction.

The outline of the chapter is as follows: Section 2 discusses the conceptions of criminal responsibility; Section 3 presents two different accounts of mental illness; Section 4 shows how these conceptions and accounts determine various models of the regulation of criminal responsibility of the mentally insane; Section 5 deals with the process of the so-called deinstitutionalization of mental illness; and Section 6 provides an evaluation of the models.

Before I move on to conduct my further analyses, three important remarks need to be made.

First, as we shall see in Section 2, not all models of the regulation of the criminal responsibility of the mentally insane are determined by both concep-
tions (of criminal responsibility and mental illness); some of the conceptions of criminal responsibility can justify a given model by themselves.

Secondly, the second dimension (an account of mental illness) is only relevant for the so-called “mixed” – psychiatric-psychological – account of mental incompetence (implied in the above description of the Standard Model) and within a “single” – purely psychiatric – account of mental incompetence. It is clear that if one accepts a “single,” purely psychological account (which dispenses with the very concept of mental illness/disorder), then it does not matter which conception of mental illness is assumed. Additionally, it should be noted that the strength of the relevance of a conception of mental illness is greater in the purely psychiatric model than in the mixed one, and how great this strength is in the mixed one will depend on how the relationship between mental illness/disorder and mental incapacities (psychologically defined) is exactly conceived: whether conceptually (in which case it is part of the definition of a mentally insane person that he has suffered, at least at the time of the criminal act, from mental illness/disorder), or, as one may call it, “diagnostically” or “evidentially” (in which case the fact that a person has suffered from a mental disease/disorder is not part of the concept of mental insanity but makes the hypothesis of the person’s being mentally insane more probable). In the former case (which implies that a person who does not suffer from mental illness/disorder cannot be mentally insane) the relevance is stronger than in the latter case (which implies that a person who does not suffer from mental illness can be mentally insane). This first interpretation, postulating the conceptual relation between mental illness/insanity and mental insanity, seems to be dominant in legal doctrine. I shall assume this interpretation throughout my further analyses. I discount here the complication which arises from the fact that, in many penal codes, e.g., in the Polish penal code, mental incompetence can also be caused by mental deficiency/retardation. When one takes into account this complication, one cannot, strictly speaking, say that there is a conceptual/definitional connection between mental illness/disorder and mental insanity, though, of course, one can say that there is such a connection between mental illness/disorder/disturbance or mental deficiency/retardation on the one hand, and mental insanity on the other.

---

1 Pure accounts are much less frequent than mixed ones; for instance, the purely psychiatric model was accepted only for a short time in the American legal doctrine, in the wake of the case of *Durham v United States* in 1954; the so-called Durham rule provided that a defendant is non-responsible if his unlawful act was “the product of mental disease or mental defect.” I shall return to this solution in Chapter 2.

2 The evidential understanding is close to the so-called reductionist thesis, which says that “there is not anything about a mental illness that makes it sufficient to excuse, and that mental illness excuses via ignorance or compulsion” (Reznek 1997: 152).
So, thirdly, I shall not analyze the problem of the moral and legal responsibility of the mentally deficient/retarded. There are, of course, many similarities between the situation of the mentally ill and the mentally deficient/retarded, but the differences seem to be at least equally numerous. For instance, the definitional feature of the mentally retarded is a low IQ (manifesting itself in perception, attention and linguistic deficits, problems with or incapacity of abstract thinking, learning problems, etc.), whereas the mentally ill (e.g., schizophrenics) usually have a normal IQ (as noticed by Ellis and Luckasson (1985: 427): “the irrationality, paranoia, and delusions that indicate mental illness and which are related to criminality are not indicators of mental retardation”). This difference seems to have far-reaching implications, e.g., for the analysis of the cognitive component of the insanity defense: One would have to conduct a separate analysis of this component for the mentally retarded. The volitional component would have to be analyzed separately as well: The mentally retarded are characterized by weakness of will, suggestibility (they are easily influenced by others as they have a strong need for contact with stronger personalities), impulsivity, and emotional immaturity. Furthermore, mental illness usually has a dynamic character, whereas mental retardation is more static (it appears at birth or in early childhood and is in the overwhelming majority of cases permanent), which implies that legal policies addressed to the mentally ill (aimed at their curing) may be inadequate for the mentally retarded. Of course, it may happen that a person is at the same time mentally ill and mentally retarded (about 30 percent of the mentally retarded are mentally insane; cf. Ellis, Luckasson 1985: 425); but, generally speaking, the onset of mental illness is not correlated with the level of intelligence. Another problem is that the mentally retarded are a largely variable group: The concept of mental retardation ranges from borderline (IQ from 70 to 84 on the Wechsler scale) to profound (IQ under 24) – the intermediate levels are mild (55–69), moderate (40–45) and severe (25–39) mental retardation. However, on the other hand, although the concept of mental retardation seems very inclusive, it is much less fuzzy than that of mental illness. All those differences would have to be taken into account in a thorough analysis of the responsibility of the mentally retarded. The analysis might lead, for instance, to the conclusion that the volitional component, which I criticize in Chapter I as part of the insanity defense, should be allowed for in the case of the mentally retarded (especially if their intellectual impairments are profound, and thereby also likely to

3 Although it should be added that “the best modern evidence suggests that the incidence of criminal behavior among people with mental retardation does not greatly exceed the incidence of criminal behavior among the population as a whole” (Ellis, Luckasson 1985: 426).
undermine their volitional capacities); furthermore, the mentally retarded are less likely to abuse the volitional component by invoking it *mala fide* in their defense. The problem of the responsibility of the mentally deficient/retarded would therefore require a separate examination, one which goes beyond the scope of this book.

2 CONCEPTIONS OF CRIMINAL RESPONSIBILITY

I shall use the term “criminal responsibility” in the general sense of accountability for the failure to comply with the norms of criminal law, and by “conceptions of criminal responsibility” I mean theories which determine the conditions under which an agent may be held accountable for this failure. I assume that there is a strict connection (that of a necessary and sufficient condition) between the attribution of responsibility and punishability, i.e., an agent can be punished for an act $a$ if, and only if, he is (criminally) responsible for it. The conceptions of responsibility are therefore at the same time conceptions of punishability.

Since criminal law is, of all the branches of law, the one most closely related to morality, each of the conceptions of criminal responsibility discussed below could also be defended as a conception of moral responsibility (with due allowances for the obvious and philosophically unimportant differences, for instance the type of norms for the violation of which one is responsible: moral or social norms in the case of moral responsibility and norms of criminal law in the case of criminal responsibility). In fact, some of these conceptions were initially proposed as conceptions of moral responsibility. It also seems that a conception of criminal responsibility in a given legal system is usually a reflection of a conception of moral responsibility presupposed in the morality of the society in which the legal system is intended to function (this sociological claim, however, has no consequences whatsoever for my further analyses).

I assume that each at least *prima facie* plausible conception of criminal responsibility must have the following two features: It must provide that a person $P$ can be held criminally responsible for committing an act $a$ only if $P$ did $a$, and that $a$ must be prohibited by the criminal law (let me call these requirements the “basic conditions” of criminal responsibility). Conceptions which fail to satisfy these conditions (e.g., those which *openly* – not just as an unintended and somewhat unlikely consequence of their assumptions – declare that the innocent persons can be held criminally responsible) are omitted in my analysis. The first of these conditions – that for an agent to be responsible for an act he must have at least committed this act – creates the basic link between conceptions of responsibility and the very concept of punishment: One could not speak about punishment but only about social hygiene if this requirement were abolished, and, consequently, if compulsory measures could be used
preventively to agents who, owing to their anti-social tendencies, are regarded as very likely to commit criminal offenses.

The *prima facie* plausible conceptions of criminal responsibility differ, however, in how they describe the “state of the mind” in which $P$ must have been at the time of committing $a$ if $P$ is to be held responsible for $a$. The most general criterion for distinguishing conceptions of criminal responsibility is whether they assume causal determinism and criminal responsibility to be incompatible (the *incompatibilist conceptions*) or compatible (the *compatibilist conceptions*). Thus, for compatibilist conceptions, the question of the truthfulness of causal determinism is irrelevant, whereas it is crucial for the incompatibilist ones, which imply that if causal determinism holds, then no agent can ever be taken to be responsible for what he did (this combination – incompatibilism *plus* affirmation of causal determinism – is called “hard determinism,” in contrast to the combination of “incompatibilism *plus* negation of causal determinism,” usually called “libertarianism”). Hard determinism may have different forms, depending on what version of causal determinism is adopted. But in all its forms it provides a justification of why criminals should not be punished but subject to some therapy. In its arguably most popular sociological form, it states that criminals have themselves been wronged – they are the victims of a badly organized society – and thereby should be subject to some kind of therapy rather than punished; punishing them would be unfair as they are not fully or at all responsible for their actions. This view was defended, e.g., by the Sociological School in criminology or by Karl Menninger (2007) and Charles E. Silberman (1978). I shall present this view in more detail in Section 4.

From among the seven conceptions presented below, conceptions 1–3 are incompatibilist, and conceptions 4–7 are compatibilist. I shall present these conceptions by formulating the necessary conditions of criminal responsibility which they formulate; these conditions, together with the previously mentioned “basic conditions” (that a person $P$ can be held criminally responsible for committing an act $a$ only if $P$ did $a$, and that $a$ was prohibited by the criminal law), constitute necessary and sufficient conditions. An exception should be made for conception 5, which, as we shall see, in fact repeats the “basic conditions.” The presentation of the conceptions will have a critical character, though a more general evaluation of them, appealing to philosophical anthropology, will be made in Section 5.

### 2.1 The Incompatibilist Conceptions

An outline of conceptions 1 and 2 has already been presented in the Introduction. I shall present them now in a more systematic way. I shall not discuss, however, the third conception mentioned in the Introduction, viz. the
“cosmic/God’s eyes” conception of responsibility, since it will not play any role in my considerations pursued in this chapter, which are devoted to the general justification of the insanity defense. However, this conception will play an important role in my analysis (conducted in Chapters 2 and 3) of the two components (the cognitive and the volitional) of the insanity defense.

According to conception 1, P can be held criminally responsible for committing an offense \( a \) (or any other offense) only if P is endowed with free will, i.e., “the ability to do otherwise” (the act \( a \) and the mental intention to commit \( a \) must not therefore be causally determined if P is to be held responsible for \( a \)), and free from external coercion (of the so-called \( \text{vis compulsiva} \) or \( \text{vis absoluta} \)). The phrase “any other offense” is justified because, on this conception, free will is treated as a durable power/feature/faculty: If human beings possess this power, they cannot lose it. According to this conception, P will be held responsible for action \( a \) even if this action, though based on free will, was not fully voluntary, i.e., was not free from both internal and external coercion. Since this conception implies that free will, if human beings possess it, is possessed by all human beings and exhibited also in their non-voluntary actions (unless they are caused by \( \text{vis absoluta} \), in which the decisional component does not appear, or, more controversially, also in \( \text{vis compulsiva} \), where the decisional component is present), it leads to the conclusion that the criminal responsibility of the mentally sane and the mentally insane should be equalized (should be none, if the existence of free will is denied, and should be full, if the existence of free will is affirmed). This conception, therefore, determines by itself (without the necessity of its supplementation by a conception of mental illness) the model of the criminal responsibility of the mentally insane. This conception is highly counterintuitive because it implies that human beings exhibit free will (assuming that it exists) also in not fully voluntary actions, whereas most of those who believe that free will exists seem to assume (though they seldom make this assumption explicit) that it is only manifested in voluntary actions.

This critical remark leads us to conception 2, which says that P can be held criminally responsible for committing offense \( a \) only if P is endowed with free will and his action \( a \) was voluntary (free from not only external but also internal coercion). Unlike conception 1, which implies that the mentally ill can be held criminally responsible (because internal coercion is not an exculpating circumstance), this conception does not by itself determine the problem of the criminal responsibility of the mentally insane but only in conjunction with a given view of mental illness (specifying whether mental illness is a real entity or a fiction).

It is worth mentioning another incompatibilist conception (I shall call it conception 3), which is based on a peculiar understanding of the concept of self-determination. It deserves a closer analysis because if this understanding
were the proper one, it would undermine the personalist view of man assumed in this book. This conception was formulated (by Galen Strawson) as a conception of moral responsibility but I shall present it also as a conception of criminal responsibility.\(^4\) It asserts that \(P\) can be held morally or criminally responsible for committing an offense \(a\) (or any other offense) only if he is \textit{causa sui}, “at least in certain mental respects,” such as “preferences, values, ideals, pro-attitudes,” which determine the way in which we react to various reasons for action (cf. Strawson 1994: 5–7).\(^5\) Strawson argues, however, that no one can be \textit{causa sui} and thus no one (irrespective of whether he is mentally sane or insane) can be held morally and (if we treat his conception also as a conception of criminal responsibility) criminally responsible (the conception of Strawson therefore, \textit{by itself} – without the need to be supplemented by a conception of mental illness – resolves the problem of the responsibility of the mentally insane). The argument of Strawson is simple (he calls it “the Basic Argument”): An agent always chooses an action in accordance with some previous criteria (or “principles of choice” as he calls them); however, these criteria were either not chosen by him in a “reasoned, conscious, intentional fashion” but imposed on him (by his upbringing, genes, etc), in which case he is not \textit{causa sui}, or they were chosen by him; but if they were chosen by him, they must have been chosen in accordance with some earlier criteria, which means that either he is not \textit{causa sui} (because these earlier – second-order – criteria were in some way imposed on him) or, if these higher-order criteria were chosen by him in light of still higher-order ones, then there occurs \textit{regressus ad infinitum} (an infinite series of the agent’s choices), which is impossible. Strawson summarizes his line of argumentation in the following way:

True self-determination is impossible because it requires an actual completion of an infinite series of principles of choice. So true moral responsibility is impossible, because it requires true self-determination … We are what we are, and we cannot be thought to have made ourselves \textit{in such a way} that we can be held to be free in our actions \textit{in such a way} that we can be held to be morally responsible for our actions \textit{in such a way} that any punishment or reward for our actions is ultimately just or fair. Punishments and rewards may seem deeply appropriate or intrinsically “fitting” to us in spite of this argument, and many of the various institutions of punishment and reward in human society appear to be practically indispensable in both their legal and non-legal forms. But if one takes the notion of justice that is central to our intel-

\(^4\) It should be noted that this was not Strawson’s intention, as can be inferred from the quotation invoked further in this text from Strawson’s article; Strawson meant it only as a conception of moral responsibility.

\(^5\) Galen Strawson ascribed this conception, rather controversially, to Kant and Sartre, and also (less controversially) to Nietzsche (as noted by Strawson, Nietzsche rejects the very concept of \textit{causa sui}, e.g., in paragraph 21 of his \textit{Jenseits von Gut und Böse}, and in consequence rejects the very concept of responsibility).
Let me pass to an evaluation of this argument. The gist of this argument amounts to the claim that we must be responsible for who we are in order to be responsible for what we do. But this claim rests on two dubious interconnected assumptions: that the very notion of “who we are” is tenable; and that who we are (“our principles of choice”) determines our choices. The first assumption seems dubious because arguably one can say only about some people that they have a rigid set of principles (values, preferences, etc.) which determine their choices; most people have a rather fluid set of principles, shaped in the course of the very process of their decision-making. The second assumption seems dubious because there may exist the capacity to decide even against one’s own customary “principles of choice” in such a way that the “new principles” of choice are not presupposed but discovered or created – the capacity is (free) will. Thus, even if it were true that we are not responsible for what we are, it would not entail that we act as we do because of who we are: We could still be ultimately responsible for our acts because they are not fully determined by who we are – our capacity of free will underlies this responsibility. Strawson does not consider these objections. But he deals with the objection (similar to the first of these two above) that there may exist “Self” which makes decisions “as it decides because of the way it is. And this returns us to where we started. To be a source of true or ultimate responsibility, Self must be responsible for being the way it is. But this is impossible” (Strawson 1994: 20–21). This counter-objection of Strawson, however, can be criticized on the same two grounds indicated above: The Self need not be identical with a rigid set of principles, and the Self may discover or create principles in the course of making a choice. This kind of Self, which is not causa sui in the strong – Strawsonian – sense, is nonetheless causa sui (in the weaker sense) insofar as it is endowed with free will and thus can always distance itself from the values or principles it accepted in the past (assuming that it did). In other words, Strawson seems to make an (implausible) assumption that in order to be causa sui, one must determine, so to say, the totality of one’s psychological outfit: emotions, preferences, dispositions, principles of choice etc. Yet the capacity to distance oneself from them seems to be sufficient for holding the Self responsible for his actions. In general, it is hard to agree with Strawson that “however self-consciously aware we are, as we deliberate and reason, every act and operation of our mind happens as it does as a result of features for which are ultimately in no way responsible” (1994: 22). The plausibility of
the weaker understanding of *causa sui* is sufficient to profess the personalist picture of human beings.

### 2.2 The Compatibilist Conceptions

The conceptions discussed in this section assume that criminal responsibility is possible even in a fully deterministic universe.

*Conception 4*, one defended by many thinkers, to mention only Aristotle (1947), David Hume (1998), or Kazimierz Twardowski (1983), implies that $P$ can be held criminally responsible for committing an offense $a$ only if his action $a$ was voluntary (free from internal and external coercion). It might be (rightly) objected, however, that my description of this conception is too sketchy as it does not explain satisfactorily why its adherents could assert that an agent who is not endowed with free will but is “merely” free from internal and external coercion can nevertheless be held criminally responsible for his (criminal) acts. One must therefore demonstrate that the lack of coercion allows some other grounds of criminal responsibility (other than free will, which, on this conception, is, *ex hypothesi*, non-existent) to manifest itself.

These grounds may be called generally “an agent’s individuality.” This concept was interpreted in various (although not mutually exclusive) ways by various philosophers: as an agent’s character, as an agent’s “essence,”6 as his “decisional centre,” as his capacity for self-control, as a set of his own beliefs and desires, or as the capacity to respond to moral reasons for and against certain actions (i.e., as being endowed with a reasons-responsive mechanism).

---

6 Aristotle’s term *hekousion*, usually translated as “voluntary,” literally means “flowing from the substance,” “flowing from the essence” (the substance – *ousia* – is the agent himself); this translation shows immediately why one could argue that an agent can be held responsible for an act done non-coercively: this is his *own* act. *flowing from him* as a separate entity/agent (substance), *not accidental* in relation to his own desires and intentions. Aristotle’s definition of voluntary acts is concordant with this etymology: “that of which the moving principle (*arche*) is in the agent himself, he being aware of the particular circumstance” (Aristotle 1947: 1111a23–34). It also follows from this definition that an act may become involuntary in two ways: if it is due to compulsion/coercion (in which case it is *against* the agent’s will) or if it is due to unawareness of specific circumstances of a given situation (in which case it is *independent* of the agent’s will). But Aristotle would arguably deny that being capable of making voluntary acts is a *sufficient* condition of moral responsibility, for also animals and children can have this capacity; it seems that, in his view, an additional condition needs to be fulfilled, viz. having the capacity (so-called *prohairesis*) to make decisions based on a rational reflection. It is worth adding that Aristotle distinguishes also (in Book III, 1 of *Nicomachean Ethics*) a third category of actions – *ouk hekousia* – which can be translated as “non-voluntary” and which may correspond to various automatic bodily movements which cannot be called “actions.”
Depending on how this “ground” is explained, one obtains different variants of this conception. Let me make a short remark on the last variant (based on the mechanism of reason-responsiveness). John Martin Fischer and Mark Ravizza (1998) – its adherents7 – argue that the mechanism consists of two capacities: regular receptivity – the capacity to recognize moral reasons for or against certain actions – and at least weak reactivity – the capacity to translate these reasons into actions. But, as was argued in the Introduction, reason-responsiveness seems to imply free will. So, in my view, this variant of conception 4 collapses into an incompatibilist one.

Conception 5, defended, e.g., by Barbara Wootton (1963), states that $P$ can be held criminally responsible for committing an offense $a$ only if his action caused $a$ (or is identical with $a$); the state of mind of $P$ at the time of committing $a$ (or at any other time) is irrelevant for ascribing him responsibility for $a$. Thus, only actus reus, not mens rea, should be taken into account in ascriptions of criminal responsibility: The prosecution neither has to prove intent or recklessness nor query whether the defendant was mentally sane. According to this conception, all crimes become ones of strict liability. Clearly, this conception by itself implies that both the mentally sane and mentally insane can be held criminally responsible. Evidence regarding mental condition of the agent, including even the evidence concerning his mens rea, should be considered only after the trial process; the trial itself should only be focused on the question of whether the accused “physically” committed the criminal act. It is the experts who, after the trial, should take into account these mental conditions and decide how the perpetrator should be punished. Needless to say, this conception is extremely controversial as it rejects a fundamental assumption of modern criminal law, viz. that the “subjective” side of a crime is equally important as its objective side (the outward act). One can understand, however, why it might have been proposed; Herbert Fingarette contented that “this radical surgery” (as he calls Wootton’s conception) was a

reaction to this miasma of ad hoc legal doctrine and evidentiary confusion …: remove the entire issue of mental disability from the trial process, and leave it to the post-verdict assessment and dispositional decisions by experts; let the behavior, not the mental state or condition of the actor, be at issue in the trial (1976: 242).

But, as Fingarette argued,

such reforms, whatever their procedural efficiencies or inefficiencies, infringe in fundamental ways on the principle that the defendant should have the right to present at his trial a complete defense to an accusation of criminal guilt. Absence

7 This conception is also assumed, e.g., by Stephen J. Morse (2002, 2008b).
According to conception 6, P can be held criminally responsible for committing an offense only if the penalty brought upon him is likely to cause a desired change in his behavior. Thus, as Andrew Eshleman put it, calling this conception “consequentialist,” the agent can be held responsible only if he “exercised a form of control that could be influenced through outward expression of praise and blame in order to promote or curb certain behaviors” (Eshleman 2016). It is worth noticing that conception 6 will lead to more frequent ascriptions of responsibility than conception 4 since the group of people whose actions can be influenced by a penalty will arguably be broader than the group of people whose actions were voluntary, and it will lead to less frequent ascriptions of responsibility than conception 5 (with its modest requirement of only actus reus). Conceptions 5 and 6 are similar in that they do not treat mens rea as a condition of criminal responsibility but conception 6 provides a higher threshold for responsibility ascriptions because, apart from the requirement of actus reus, it also demands that the agent be responsive to sanctions. Conception 6 was also defended by Barbara Wootton (1963), especially in the context of her analysis of the function of punishment (which is in her view purely preventive). Alf Ross, the Danish legal philosopher, summarized Wootton’s view of criminal responsibility in the following way (the quotation seems to merge both conceptions – 5 and 6 – but it is justified in the sense that he focuses on the element which is common for both of them, viz. the rejection of the criterion of mental responsibility):

```
[T]he criterion of mental responsibility (imputability) should simply be dispensed with as a condition for conviction, and the criminological reaction to crime should be arrived at in each individual case without regard to guilt and only with a view to what in the particular instance will offer the best chance of preventing recidivism. The traditional system, which bases punishment on retribution for guilt, should be replaced, according to this view, by a system designed as a means of preventive hygiene …. Wootton’s doctrine amounts to no less than the view that questions of both imputation and imputability be discounted as conditions for convicting a person of an offence, while at the same time they are to be taken into account as
```

8 This is the reason why Fingarette endorsed a traditional account of the insanity defense – he called it the Disability of Mind defense – which “isolates all the elements essential in assessing culpability and places them in the trial; only questions not essential to the issue of culpability are postponed for expert appraisal after the verdict. Thus optimal efficiency within the limits of justice is achievable” (1976: 243).
Ross aptly remarked that if we accept Wootton’s view, then the distinction between penalty as a just retribution and preventive measures becomes blurred. A similar observation was made by Herbert L.A. Hart:

Lady Wootton argues that if the aim of the criminal law is to be the prevention of “socially damaging actions,” not retribution for past wickedness, the conventional doctrine puts *mens rea* “into the wrong place.” *Mens rea* is on her view relevant only after conviction as a guide to what measures should be taken to prevent a recurrence of the forbidden act. She considers it “illogical” if the aim of the criminal law is prevention to make *mens rea* part of the definition of a crime and a necessary condition of the offender’s liability to compulsory measures (Hart 2008: 194; quotations from Wootton 1963: 51–52).

Conception 7, defended, e.g., by Hart (2008), assumes that *mens rea* and *actus reus* of an agent *P* are sufficient (jointly with the “basic condition,” embracing in fact *actus reus*) for holding him criminally responsible even if his (criminal) act was not voluntary (in the sense of freedom both from external and internal coercion). In other words, *mens rea* is preserved as a condition of criminal responsibility but it no longer embraces the requirement of mental normality (sanity): The question of whether the perpetrator is mentally insane should be transferred to the post-trial phase, in which the main role is to be played by medical experts. In fact, Hart treated his own conception as a “less extreme,”

---

9 In Ross’s terminology, imputation is related to the question of the form of the agent’s mental attitude to his act (whether the act was negligent or intentional), whereas imputability concerns the question of mental responsibility/culpability, viz. whether the act can be “imputed” to him.

10 Conceptions 5 and 6 also seem to justify, e.g., the punishing of children since their conduct can assuredly be modified by sanctions. It would seem that some of their adherents (especially those with collectivist predilections, advocating the priority of the community over the individual) are inclined to accept this and other morally dubious consequences. This can be seen, for instance, in the way in which conception 6 was defended by the philosopher Francis Herbert Bradley. In an article advocating eugenics he wrote that the “Chief Good” was “the welfare of the community realized in its members” (1894: 269–270). And, in the same spirit, he argued that punishment should be separated from considerations of guilt and justice, the latter being “but a subordinate and inferior principle. It can hear no appeal from the tribunal of the common welfare” (1894: 276). He added that “once admit that life in this world is an end in itself, and the pure Christian doctrine is at once uprooted. For, measured by that end and standard, individuals have unequal worth, and the value of each individual is but relative, and in no case infinite” (1894: 277); “over its members the right of the moral organism is absolute” (1894: 278).
“moderate” form of the “new doctrine,” i.e., of the conception (of strict liability) proposed by Wootton (conception 5). He described his own theory as replacing

not the whole doctrine of mens rea, but only that part of it which concerns the legal responsibility of the mentally abnormal. In this more moderate form of the theory a mentally normal person would still escape conviction if he acted unintentionally or without some other requisite mental element forming part of the definition of the crime charged. The innovation would be that no form of insanity or mental abnormality would bar a conviction, and this would no longer be investigated before conviction. It would be something to be investigated only after conviction to determine what measures of punishment or treatment would be most efficacious in the particular case (2008: 195–196).

And in a similar vein:

Under this scheme mens rea would continue to be a necessary condition of liability to be investigated and settled before conviction except so far as it relates to mental abnormality. The innovation would be that an accused person would no longer be able to adduce any form of mental abnormality as a bar to conviction. The question of his mental abnormality would under this scheme be investigated only after conviction and would be primarily concerned with his present rather than his past mental state. His past mental state at the time of his crime would only be relevant so far as it provided ancillary evidence of the nature of his abnormality and indicated the appropriate treatment (2008: 205).

Why was Hart reluctant to endorse an “extreme” form of this doctrine (i.e., conception 5 – a conception of strict liability)? He discussed his three “misgivings” about it.\footnote{Note that Hart interprets Wootton’s conception as eliminating responsibility; I treat it in a different way – as a (peculiar) conception of responsibility. This difference in terminology, however, does not have any further consequences.}

First, he asserted that “in a system in which proof of mens rea is no longer a necessary condition for conviction, the occasions for official interference with our lives and for compulsion will be vastly increased” (Hart 2008: 206). We might also be punished, e.g., for accidental harm.

Secondly, Hart objected to treating “what we now call punishment (imprisonment and the like) and compulsory medical treatment … just as alternative forms of social hygiene to be used according to the best estimate of their future effects” (2008: 207). His two objections to “detaching the use of penal
The insanity defense methods” from judgments of responsibility are of a moral and sociological character. The moral objection says that

if we imprison a man who has broken the law in order to deter him and by his example others, we are using him for the benefit of society, and for many people, including myself, this is a step which requires to be justified by (inter alia) the demonstration that the person so treated could have helped doing what he did (2008: 207).

The sociological objection concerns two aspects of punishment which distinguish it from medical treatment and which are likely to be preserved (especially the second one) even if prisons were replaced with “places of safety” (Wootton’s term). The first one is that “unlike medical treatment, we use deterrent punishment to deter not only the individual punished but others by the example of his punishment and the severity of the sentence may be adjusted accordingly” (Hart 2008: 207). The second aspect is that “unlike a medical inspection followed by detention in hospital, conviction by a court followed by a sentence of imprisonment is a public act expressing the odium, if not the hostility, of society for those who break the law” (2008: 207). Should these two aspects of punishment remain, then arguably only those can be punished who could have acted otherwise that they did.

Thirdly, against Wootton’s argument that reference to mens rea should not be introduced into the definition of an offense, Hart argued that “there are some socially harmful activities which are now and should always be treated as criminal offences which can only be identified by reference to intention or some other mental element. Consider the idea of an attempt to commit a crime” (Hart 2008: 208).

Hart summarized his argumentation in the following way:

I do not consider my misgivings on these three points as necessarily insuperable objections to the programme of eliminating responsibility. For the first of them rests on a judgment of the value of individual liberty as compared with an increase in social security from harmful activities, and with this comparative judgment others may disagree. The second misgiving in part involves a belief about the dependence of the efficacy of the criminal law on the publicity and odium at present attached to conviction and sentence and on deterrence by example; psychological and sociological researches may one day show that this belief is false. The third objection may perhaps be surmounted by some ingenuity or compromise since there are many important offences to which it does not apply. Nonetheless I am certain that the questions I have raised should worry advocates of the elimination of responsibility more than they do; and until they have been satisfactorily answered I do not think we should move the whole way into this part of the Brave New World (2008: 207).

However, Hart’s own conception is by no means uncontroversial. Herbert Fingarette, for instance, convincingly argued that this conception exhibits
The philosophical foundations of the insanity defense

much similarity to that of Wootton (of strict liability), as it contravenes “the principle that the defendant should have the right to present at his trial a complete defense to an accusation of criminal guilt” (1976: 242); in consequence, it excludes “by fiat, and counter-intuitively, the possibility that a just determination of culpability can crucially depend upon whether the person’s mind was abnormal at the time of the act” (Fingarette 1974: 293). Furthermore it should be noted that the connection between mens rea and sanity may be so close that their separation may prove to be impossible (arguably, one cannot speak about an agent’s intent, premeditation, deliberation, or malice without assuming his sanity).

Some more general remarks are called for at the end of this section. The incompatibilist conceptions are more prone to the risk of eliminating responsibility than the compatibilist ones since they make criminal responsibility contingent on stronger metaphysical assumptions, whose truthfulness is a matter of dispute, viz. assumptions about the existence of free will or about an agent’s being causa sui. If the metaphysical assumptions of a given conception are rejected, then, on this conception, criminal responsibility proves to be an empty concept: It cannot be ascribed to anyone. For instance, as already mentioned, Galen Strawson claimed that “nothing can be causa sui, and in order to be truly morally responsible for one’s actions one would have to be causa sui, at least in certain crucial mental respects” (Strawson 1994: 21). But if these assumptions are true, they also provide a more solid basis for holding an agent responsible for his actions (i.e., a more solid justification of moral and criminal responsibility). It is also not accidental that they are also usually invoked to justify the retributive theories of punishment. These theories assume that what justifies a punishment is that an agent deserved it by committing an offense. But the concept of desert is only tenable if a sufficiently strong account of agency is assumed. The strong account of agency is provided only by the conceptions based on the assumption of the existence of free will; so these conceptions provide the best justification of retributive justice. Furthermore, the retributive theories of punishment are, arguably, the only ones which do not treat the agent as a means to some other end (e.g., special or general deterrence) but as an end in itself; and, as can be plausibly maintained, only the assumption of free will confers on the agent moral dignity which justifies

---

12 Let me recall that, according to Hart, the question of the defendant’s mental abnormality, since it was to be examined only after conviction, could be concerned only with his present, not past (from the time of committing an offence) mental state (Hart writes “primarily” rather than “only” but it seems to be an understatement: If it were to be examined after conviction, the mental state of the defendant at the time of committing an act would, on Hart’s conception, be irrelevant, at least for the question of culpability). This point of Hart’s conception is particularly controversial.
this kind of treatment (as a fully responsible, autonomous agent, who can be treated seriously by the criminal justice system, i.e., by imposing a punishment on him for what he did just because he did it). One could contend that, on this view, punishment is, in a sense, an end in itself but this would be a misleading way of speaking: Punishment indeed does not serve any future – social or psychological – goals but it serves one basic goal: giving the perpetrator his due as a responsible, autonomous agent.

The compatibilist conceptions do not make any questionable metaphysical assumptions, so they do not create the risk that criminal responsibility will turn out to be an empty concept, but also, arguably, provide a weaker justification of criminal responsibility. It could also be argued that some of those conceptions (5 and 6) should more properly be called “conceptions eliminating responsibility” rather than “conceptions of responsibility.” But the latter term seems to me to be the correct one, at least with reference to conception 5. For even if we dispense with mens rea, we can still speak about, however strongly we may be inclined to reject it, a conception of responsibility. More problematic is conception 6 because of its forward-looking character. Whether we regard it as a conception of responsibility will depend on our semantic associations with the term “responsibility.” Some philosophers may deny the very possibility of treating it as a solely future-oriented notion while others may be ready to do precisely that.13

3 ACCOUNTS OF MENTAL ILLNESS

Let me now turn to the second dimension of models of the criminal responsibility of the mentally insane, viz. the accounts of mental illness. I distinguish two such accounts which I dub the “affirmative” and the “negationist.”

3.1 The Affirmative Account

This account implies that the distinction between mental illness/disorder and mental health is fully legitimate: Its adherents might admit that it may be difficult or impossible to provide a value-neutral conception of mental illness/disorder/health, but will add that it is possible to construct a mixed one combining a value and scientific element. One such conception was proposed by James Wakefield. On his account, mental illness/disorder is an objective phenomenon: It refers to some real dysfunction of a mental mechanism. By

13 For instance, Jaap Hage admits as a possible conception of responsibility the view which allows us to “hold people responsible because of the attractive consequences this has for the future” (2020: 38).
a “real” dysfunction he means one whose recognition does not depend on specific cultural assumptions of the “observer” but is defined without reference to the “observer,” viz. as “the failure of a mental mechanism to perform a natural function for which it was designed by evolution” (Wakefield 1992: 373). The value-neutrality of his account, however, is not complete because, as he notices, not every kind of dysfunction is to be viewed as a disorder. For a dysfunction to be a disorder, it must also be harmful for the agent, and what counts as being harmful depends, at least to some extent, on the norms of the agent’s culture. However, the value component is secondary in the sense that it “selects” disorders from the set of dysfunctions determined by the scientific component. Wakefield’s final definition of a disorder as a harmful dysfunction is as follows:

A condition is a disorder if and only if (a) the condition causes some harm or deprivation of benefit to the person as judged by the standards of the person’s culture (the value criterion), and (b) the condition results from the inability of some internal mechanism to perform its natural function, wherein a natural function is an effect that is part of the evolutionary explanation of the existence and structure of the mechanism (the explanatory criterion) (Wakefield 1992: 384).

This definition is very plausible, although not entirely uncontroversial. The obvious objection that may be raised against it is that it is not always clear what function a given mechanism is supposed to perform. We may therefore not know whether a mechanism has a dysfunction until we know its evolutionary function, and such knowledge, especially in the case of mental mechanisms, may often be difficult to obtain. One could also formulate a different objection to this account – that it does now allow one to distinguish between a mental disorder and mental illness (it is clear that not every mental disorder, e.g., dyslexia, is a mental illness which has to be cured psychiatrically). Yet it should be added that Wakefield’s account does not imply that the concept of mental illness is clear and distinct; it recognizes its vagueness – a number of “borderline” cases, with regard to which it is not clear whether they can be legitimately regarded as mental illnesses or not. For instance, while it is clear that endogenous psychoses, such as schizophrenia, paranoia, or bipolar disorder (manic depression) are mental illnesses, it is less clear whether borderline personality disorder, obsessive-compulsive disorder, psychopathy or pedophilia are such illnesses (the dominant opinion seems to be that they are not). There is no place here to analyze in depth the sources of its vagueness, but it is worth noticing that the uncontroversial cases of mental illnesses seem to be connected with the cognitive defect (being “out of touch with reality,” having delusions), whereas controversies tend to arise mainly in connection with the disorders which affect the volitional or affective side.
3.2 The Negationist Account

The negationist account is connected with the so-called “anti-psychiatry” movement, a term coined by one of its proponents – David Cooper – in his book *Psychiatry* and *Anti-Psychiatry* from 1967. There are three different claims characteristic of this trend, although it must be stressed that not all unorthodox psychiatrists, i.e., anti-psychiatrists, have defended all of them.

The first claim, which can be termed the “sociological,” is the least controversial: It says that many of the practices of orthodox psychiatry (as it functioned before the 1960s), i.e., the typical attitudes and methods of treatment of the mentally ill within the psychiatric system (e.g., electroshocks, lobotomy), cannot be morally accepted as they are inhumane, lacking in respect for the patients’ dignity and rights, based on violence, oppression, the rigidly hierarchical relation between the patient and the doctor. In the more radical versions of this claim, the critique of the practice of orthodox psychiatry becomes part of a more general critique of the system of (capitalist) power, the “doctor” being seen as a representative of this power. The postulate flowing from this claim was that the psychiatric system should be radically changed so as to guarantee that the dignity and human rights of the patients become recognized and respected.

The second claim, more important in the context of my considerations, is that mental illness is a social construct or a myth: The mentally “ill” are not ill in the strict – and proper – sense of an objective dysfunction (only physical illness is an illness in the strict and proper sense, so the argument goes); they just behave in a way which differs from what society regards as “normal”; to put the etiquette of mental illness on them is an insidious attempt at disciplining them and imposing on them the prevailing social norms. On this account, the concept of mental illness is an instrument of stigmatizing socially undesirable conduct and is thereby intentionally used in bad faith by those in power (although it should be added that anti-psychiatrists sometimes also interpreted this concept in a less “unmasking” way, viz. as an expression of sexual, racial, or social prejudices, which may be cherished, or rather, are usually cherished, in an unconscious way). One should also note that the second claim may take a stronger form: viz. that mental “illness,” besides being no illness at all (a weaker form of the claim), is also, or at least may be, a privileged form of cognition, offering an access to some other dimension of

---

14 This inhumanity of institutional psychiatry was especially strongly emphasized by Erving Goffman in his classical book *Asylums* from 1961, which profoundly influenced the “anti-psychiatric” movement. For instance, the works of the Italian psychiatrist Franco Basaglia (1973) have been much influenced by Goffman’s analyses.
The philosophical foundations of the insanity defense

reality. The claim that mental illness is a social construct or a myth is, obviously, fully compatible with the first one, or even implies it: If a mental illness is a myth (regardless of whether regarded erroneously as non-myth or deliberately invented to discipline those who deviate from what society regards as normal), one cannot be uncritical towards the practices and perhaps even the very existence of the institutions of orthodox psychiatry (which, on this view, is based on a fundamental mistake). The second claim in its stronger form was defended, for instance, by Michel Foucault, whereas Thomas Szasz, perhaps the most famous representative of “anti-psychiatry,” defended its weaker form.\[^{15}\] A short presentation of their views may be in order here as it will help better understand their reasons for adopting the critical stance on the traditional account of mental illness.

In his classic book *Histoire de la folie à l’âge classique* from 1961 Michel Foucault maintained that the concept of mental illness is simply one of many ways in which the state tries to control its citizens. According to this account, mental illness, unlike physical illness, is not an objective phenomenon: The deficiencies to which it refers are always socially constructed, distinguished on the basis of subjective ideas and evaluations of those who ascertain them. There exists a certain standard view of happy/healthy/adaptive life in a given society, and anyone who deviates from this standard is considered to be mentally ill. Thus, there exist no strictly scientific, objective grounds for introducing the concept of mental illness. In Foucault’s view, the modern understanding of mental illness – precisely as an illness – is a consequence of the reinterpretation of madness that took place in the Enlightenment and 19th century: The voice of madness – the voice of Unreason – was no longer considered legitimate as it had been earlier in the Middle Ages and in the Renaissance, when reason could be unreasonable and unreason could be reasonable.\[^{16}\] In the Middle

\[^{15}\] Although Szasz preferred to speak about himself as an adherent of “anti-coercion” rather than “anti-psychiatry.”

\[^{16}\] This historical interpretation (but not necessarily its philosophical interpretation) is shared also by many other historians. For instance, George Rosen wrote that in this period (the Middle Ages – especially the waning Middle Ages – and the Renaissance) “madness through its linkage with the revelation of religious truth became a means of achieving knowledge. Madness was a primitive form of revelation, revealing the depths of menace, destruction and evil that lurked beneath the illusory surface of reality. Unreason revealed the unbearable, the things in the world upon which one could not otherwise bear to look” (1968: 155). This passage refers mainly to the Middle Ages, but madness was not totally rejected during the Renaissance either – the mentally ill were still to be considered as “different,” not as dangerous. Let me only mention Erasmus of Rotterdam’s *Moriae encomium* (*In Praise of Folly*), in which the voice of folly is undoubtedly regarded as deserving to be heard: The folly argues that various forms of irrationality are necessary for the functioning of the world (given the precariousness
Ages, various cities may have wanted to dispense at times with the “mad” by means of the famous *Narrenschiff* (ship of fools) – as depicted most famously by Hieronymus Bosch – which left them in other cities or in rural areas, but, in general, the “mad” were tolerated, treated just as somewhat different. But this voice of Unreason was smothered by reason, which, in Foucault’s radical view, is nothing more than the face of power, leading to the disciplining of society, to imposing on people a new social-police order. In the course of this process, the exclusion of the mentally ill from society proceeded in two stages. In the first one, which took place in the Enlightenment, the mentally ill, alongside other socially undesirable groups (e.g., petty criminals, vagrants, beggars, prostitutes, prodigals, the dissolute, libertines, blasphemers), which the emerging, ever stronger modern state wished to separate from the rest of society (as a social burden: as useless to its economy – they were viewed as “idle” – and posing a threat to the social order), were interned in special correctional institutions. The internment – “the Great Confinement,” as Foucault calls it – did not have a medical function but rather an administrative, social, and oppressive one, aimed at controlling those who did not work and who violated common social standards. It is not accidental, in Foucault’s opinion, that the places of internment were located in the former *leprisoria*: The interned were marginalized and stigmatized just as lepers were in the Middle Ages and during the Renaissance – they were perceived as a threat to the social order. The criterion for internment was the mere deviation from social norms; the tolerance of this kind of deviation – large, when folly was regarded as not fully separable from reason – was greatly diminished, all the more so given that “anti-social” people were believed to be in moral error – to have chosen by themselves their
lifestyles. This was an example, on Foucault’s interpretation, of not only physical, but also moral – symbolic – violence directed at excluding from society certain undesirable groups of people. At this stage, the mentally ill were not treated as a separate – sui generis – category (which implies that madness was perceived as being rooted in morality, and in effect as being connected with evil – as multiplying the evil contained in a socially undesirable act). 18

The differential treatment of the mentally ill came about at the second stage – in the 19th century, in the aftermath of the abolition (in 1786–1793) of the internment system (which, besides being non-humanitarian, generated higher costs than economic output) and of the reforms initiated by William Tuke and Philippe Pinel, which gave rise to psychiatric hospitals. These reforms are traditionally interpreted as humanitarian in spirit: as liberating the mentally ill from the ignorance and prejudices of previous ages (when the mentally ill were put together with criminals, and were treated as curious animals often put on display for inquisitive spectators, and, additionally, often held morally responsible for their mental state), and therefore as aimed at providing them with therapeutic help. Yet Foucault disagreed. He wanted to demystify the purportedly humanitarian character of these reforms by pointing out that the mentally ill were not only tolerated in the Middle Ages and the Renaissance but also, at least to some extent, appreciated. So these reforms can only be viewed as progress in comparison with the period of Enlightenment (when people with mental illness were not separated from criminals). Furthermore, these reforms were not liberating as they retained the isolation of the mentally ill (even though the isolation was at the start of this process non-coercive). More importantly, they had, more or less overtly, an ideological goal: The confinement of socially undesirable people (the mentally ill) was not only (or even not above all) to help them but to protect the bourgeois society and its morality; mental asylums were, therefore, an even more efficient instrument serving the moral uniformization of society. Psychiatric practice is, at least on this view, a form of moral tactic aimed at promoting a very specific – dominant, bourgeois – moral lifestyle and suppressing the other lifestyles under the cover of care for the mentally ill. It should be emphasized that, in Foucault’s view, the evolution of social attitude towards the mentally ill was not the result of some series of conscious decisions of a controlling agency; rather, it was

---

18 Yet, as shown by Foucault, the mentally ill were treated in a different way in the legal context – as a separate category: Mental illness was regarded as an excuse. It seems paradoxical that two such different ways of treating the mentally ill could be upheld at the same time. The treatment adopted by the internment institutions implies that the mentally ill person is responsible for his mental state, that this state flows from his (free) will, while the treatment adopted in courts assumed that the mentally ill cannot be blamed for their mental state.
an unintended (though *ex post* predictable and logical) result of the process of emergence of the modern state with its disciplinary ambitions and of a certain type of morality (bourgeois).19

As for Szasz’s view of mental illness, it is in some points similar to that of Foucault’s yet his perspective is different: less historical and sociological and more scientific and philosophical. In his most famous book, *The Myth of Mental Illness* of 1960, Szasz raised the claim that there are no grounds for the belief that mental illness is a scientific category; it is not, unlike any genuine illness (being a scientific category), a physical lesion, a deviation from the anatomical structure, an “abnormal biological condition,” or “abnormal functioning of the body” (1974). Rather than being a scientific category, mental illness is a metaphor or a myth constructed by the psychiatric authorities to support the existing social norms. Thus, in contrast to bodily illnesses, mental “illnesses” are invented, not discovered. This “conspiratorial model of madness”, as it is often called, was aptly characterized in the following way (in the context of schizophrenia): “Schizophrenia is a label which some people pin on other people, under certain social circumstances. It is not an illness, like pneumonia. It is a form of alienation which is out of step with the prevailing state of alienation. It is a social fact and political event” (Siegler, Osmond, Mann 1969: 950). In Szasz’s view, those who are called “mentally ill” can at best be said to suffer from “problems in living” (1974: xi), i.e., from not being capable of adapting themselves to the norms existing in a society or from encountering difficult moral dilemmas which they find themselves unable to resolve (e.g., reconciling the need for intimacy and the need for solitude); 20

---

19 It may be interesting to note that Foucault’s interpretation of the “history of madness” was considerably influenced by his reading of the essay “Van Gogh, The Man Suicided by Society” written by the French writer Antonin Artaud. After a nervous breakdown, Artaud spent nine years in a mental hospital, where he was subjected to severe treatment including 60 electroshocks (cf. Kotowicz 1997: 61). In this essay, which was a scathing critique of orthodox psychiatry, Artaud defended the view that mental illness is not only no illness at all; it is also a privileged way of knowing; as he wrote: “a tainted society has invented psychiatry to defend itself against the investigations of certain superior intellects whose faculties of divination would be troublesome” (quoted from Kotowicz 1997: 62).

20 In Szasz’s opinion, another good example of such a problem is hysteria, of which he wrote as follows:

The typical cases of hysteria cited by Freud … involved a moral conflict – a conflict about what the young women in question wanted to do with themselves. Did they want to prove that they were good daughters by taking care of their sick fathers? Or did they want to become independent of their parents, by having a family of their own, or in some other way? I believe that it was the tension between these conflicting aspirations that was the crucial issue in these cases. The sexual problem – say, of the daughter’s incestuous cravings for her father
accordingly, “psychological interventions are directed at moral, not medical problems (1974: x)” and “while medical diagnoses are the names of genuine diseases, psychiatric diagnoses are stigmatizing labels” (1974: xii), which are “used chiefly to obscure and explain away problems in personal and social relationships” (1974: 182). Szasz has argued that “mentally ill” patients try to solve this and similar problems by playing various games, e.g., the hypochondriac “takes the role of certain medical patients,” whereas the schizophrenic “often takes the role of other, invariably famous, personalities” (1974: 237), and one afflicted with frustration and unhappiness takes the role of a sick person and pretends to suffer from weakness and helplessness. In other words, instead of bravely countering the problems that life inevitably carries with it, they prefer to accept the escapist way of assuming that the problems are not inherent in life itself—that they are in fact a manifestation of a curable mental “illness.” The very idea of mental illness is therefore grounded in the false belief that life itself is free from inevitable conflicts, and that the ideal of personal or social happiness can be easily achieved. It should be stressed that, according to Szasz, many mentally ill “patients” may subjectively not experience a “problem in living”—they may not define themselves as ill, but they are so defined by other people—thus the role of a mentally ill patient is imposed on them against their will. All in all, in Szasz’s opinion, “patients” should be allowed to choose the way they want to be helped by themselves (if they want to be helped at all): “[C]ustodial psychiatry,” as Szasz called it, should replace “institutional – coercive – psychiatry.” The state should not be a “therapeutic state,” i.e., it should not play any role in solving the problems of the mentally suffering; there should be no place for involuntary, psychiatric hospitalization. Consequently, the insanity defense should also be rejected as it implies that the concept of mental illness is a legitimate one. These postulates of Szasz were additionally motivated by his libertarian political convictions. As men-
tioned, Szasz did not maintain that mentally ill people do not need any help; but he believed it should not be the kind of help offered to them in traditional mental hospitals. The “libertarian” reason for this claim is that they undergo in these hospitals an involuntary therapy for an indefinite time; such a form of containment (the so-called “civil commitment”) is, in his view, a flagrant violation of individual liberty, and, additionally, further proof of the claim that mental illness is not like any other illness (the mental patient may not leave the hospital as he wishes, whereas generally the medical patient may do so; only in extremely rare cases – such as contagious disease – is the involuntary hospitalization of non-mental patients allowed; cf. Szasz 1963; 1990). Szasz even went so far as to call involuntary psychiatric interventions “crimes against humanity” (1974: 267), “incarceration in the name of diagnosis” (Szasz and others 1996). As we can see, Szasz’s concern with the individual liberty of psychiatric patients underlies another objection to the insanity defense, viz. the objection that it allows indeterminate containment with respect to people who have committed a crime – the containment which, as he maintained, is, precisely because of its indeterminate character, an even more flagrant violation of individual liberty than jail, a more harsh “punishment” (in the common, not legal sense). Szasz aptly remarked that the insanity defense, which may lead to people being detained for an indefinite time, and civil commitment are “like Siamese twins: you cannot abolish one without killing the other” (Szasz and others 1996). Indeed, the alternative we are faced with has been outlined by Szasz in a clear manner: We must choose between, on the one hand, civil commitment for the mentally ill who did not commit a crime plus no incarceration (but detainment for an indefinite time) for those of them who committed a crime, and, on the other, no civil commitment for the mentally ill who

---

23 To Szasz’s further dismay, the decisions about the containment of the mental patients were made by the medical profession.

24 In response to the objection that the abolition of the insanity defense means that if the mentally ill commit a crime, they will go to jail, Szasz said that it is better for them because jails are less cruel than involuntary hospitalization.

25 Szasz described this first option in a more dramatic way – as involving “imprisonment without trial (involuntary psychiatric hospitalization) and punishment without having been sentenced (psychiatric treatment)” (1990: 560), “imprisoning the innocent and exculpating the guilty” (1980: 120). He also wrote in a more ironic tone:

On Mondays, Wednesday, and Fridays, the psychiatric prevaricators thus go to court to exonerate the guilty: That is called “psychiatric defense.” On Tuesdays, Thursdays, and Saturdays, the same prevaricators go to court to incriminate the innocent: That is called “civil commitment.” The lawmakers, the judges, and the attorneys (for both sides) all shamelessly uses these fakes – which is why each of them is as reluctant to expose and demolish the psychiatric defense of the guilty as he is to expose and demolish the psychiatric incrimination of the innocent (1980: 120).
The philosophical foundations of the insanity defense

...did not commit a crime plus incarceration for those of them who committed a crime. If we assume any of the two claims endorsed by Szasz – that mental illness is a myth or that liberty is the supreme value – we must choose the first option.

Allow me to note that the second – libertarian – objection raised by Szasz against the insanity defense may take a stronger or weaker form, something which Szasz did not distinguish quite clearly. In its weaker form, it says that the insanity defense, even if invoked bona fide, has pernicious consequences as it justifies a flagrant violation of individual freedom. In its stronger form, it says that the insanity defense is prone to be invoked also mala fide – as a pretext to justify holding for an indefinite time in confinement the defendant whom prosecutors and judges know to have no mental illness at all.26 Szasz was arguably endorsing mainly its weaker form, although it seems he also occasionally manifested some sympathies with the stronger one.27

In summary, Szasz put forward two independent but compatible arguments for the rejection of the insanity defense. He claimed that since insanity (mental illness) is a myth, the institution of the insanity defense is groundless: It relies on a non-existent entity. Yet it seems that he attached more weight to the second – “libertarian” – argument, viz. that the insanity defense is a violation of human dignity as it is inextricably connected with the institution of civil commitment (this argument, to be quite strict, is not just compatible with the first one – that mental illness is a myth – but is also strengthened by it). One can also imagine another argument against the insanity defense which does not play an important role in Szasz’s critique, viz. that even if the perpetrator of a crime is insane, he still preserves free will, and thereby can be held respon-

26 The stronger form was defended by, e.g., Joseph Goldstein and Jay Katz, who explicitly wrote that “the insanity defense is not a defense, it is a device for triggering indeterminate restraint” (1963: 868). As they pointed out, if insanity were treated only as “evidence which precludes establishing a crime by leaving in doubt some material element of an offense” (1963: 854), that is, mens rea, then the perpetrator would have to be released (since there would be no crime at all, and so no defense would be needed); but if insanity is an excuse, then the perpetrator is deemed not guilty by reason of insanity, and as a result can be detained for an indeterminate time.

27 This ambiguity in his views on this issue can be discerned, e.g., in the following statement: “The introduction of psychiatric considerations into the administration of the criminal law – for, example, the insanity plea and verdict, diagnoses of mental incompetence to stand trial, and so forth – corrupt the law and victimize the subject on whose behalf they are ostensibly employed” (1974: 267).
sible for his act. We shall return to this argument (which can also be called “libertarian”) in Section 4 of the present chapter.

One point bears stressing: Szasz did not share the political leftist sympathies of many representatives of “anti-psychiatry.” In the article “Anti-Psychiatry: The Paradigm of Plundered Mind,” which Szasz wrote in 1976 for The New Review, we can read: “the anti-psychiatrists are all self-declared socialists, communists, or at least anti-capitalists and collectivists. As the communists seek to raise the poor above the rich, so the anti-psychiatrists seek to raise the ‘insane’ above the ‘sane’” (quoted from Kotowicz 1997: 92). In this article he claimed that the left appeal to the youth who “have nothing to live for, are envious of all those who do, and want to destroy the institutions that give meaning to the lives of ‘normal’ people” (quoted from Kotowicz: 1997: 93). He believed that “the simplest of human truths is that life is an arduous and tragic struggle” (quoted from Kotowicz 1997: 96), with the consequence, as Zbigniew Kotowicz put it, that “the rewards go to those who are competent, patient, modest, silent, and who accept suffering” (1997: 95); Kotowicz commented on this in the following way:

With this in mind Szasz’s campaign to abolish mental hospitals cannot be put in the same bracket as other protests against the inhumanity of the psychiatric system. He

---

28 Even though it only appears marginally in his analyses, it takes (when it appears) a suggestive form:

Today, the role of the physician as curer of the soul is uncontested. There are no more bad people in the world, there are only mentally ill people. The “insanity defense” annuls misbehavior, the sin of yielding to temptation, and tragedy. Lady Macbeth is human not because she is, like all of us a “fallen being”; she is human because she is a mentally ill patient who, like humans, is inherently “healthy”/good unless mental illness makes her “sick”/ill-behaved (2011: 180).

He also wrote in a different paper that his observation that the two paradigmatic practices of psychiatry – involuntary mental hospitalization or civil commitment and the insanity defense or the exculpation of persons guilty of crimes as not guilty by reason of insanity – rests on a philosophically indefensible and morally odious proposition, namely, that unlike the behavior of the sane person, which is governed by free will, the behavior of the insane person is governed by impulses which the subject finds irresistible and for which he or she is, therefore, not responsible (1990: 558–559).

Let us note that this is a very extreme version of the free will doctrine, similar to the one implied by conception 1 of criminal responsibility (though justified in Szasz’s case by his belief that mental illness is a non-existent entity). This conception, as mentioned in the Introduction, is justified by the libertarian but not the personalist picture of man.

29 As is well known, the term “libertarian” has two quite different meanings; in political philosophy it refers to those conceptions which stress the priority of the value of individual liberty over other values, whereas in the philosophy of mind it is used to describe those conceptions which affirm the existence of free will.
is anti-establishment but in the name of the higher law – the law of the market. And so Szasz would want to eliminate all forms of collective response to mental illness. And there is nothing particularly new in this. As long ago as 1961, Enoch Powell, the then Minister of Health, envisaged closing all mental hospitals and setting up a community care system. This has remained the policy of the Conservative Party, although it is becoming manifest that the notion of “community care” is no more than lip service. It is one thing to shut down hospitals and another to genuinely tackle the society’s response to mental illness. This is the difference between the right-wing laissez-faire cost-saving programmes and the Italian experiment, for example (1997: 96).

Finishing my discussion of the second claim of anti-psychiatry (that there is no such thing as mental illness), let me devote a few words to pointing out a paradoxical similarity of this claim with a more popular and more widely accepted claim that the concept of mental illness constitutes a continuum. The continuum view (which arose under the influence of psychoanalysis) amounted to the rejection of the 19th-century account of mental illness as a brain disorder, which some people have and others do not. It resulted in the proliferation of mental illnesses (Attention deficit hyperactivity disorder (ADHD), non-severe depression, adjustment disorders, problems with self-esteem, anxiety, hysteria, and many others), which is visible in the successive editions of the Diagnostic and Statistical Manual of Mental Disorders. The paradoxical similarity between both claims (that nobody is mentally ill or disordered, and that almost anyone can be regarded as mentally ill or disordered) lies in the fact that in both cases the very concept of mental illness loses its distinguishing and classificatory function: The first claim implies that no one can be regarded as mentally ill, whereas the second one, at least in its extreme version, implies that everyone can be regarded as – more or less seriously – mentally ill. As can be easily noticed, the second claim can also be used in the critique of the insanity defense – as leading to possible abuses, or to including all criminals in the category of the “mentally insane” (I shall return to this line of critique later in Section 5 of this chapter).

Let me now pass to a discussion of the third claim that appears within anti-psychiatry – that the mentally ill are entirely or to a large extent responsible for their illness. It was not that widespread among anti-psychiatrists for the simple reason that it remains in tension with the second – principal – claim. The third claim says that mental illness is a reality, not a myth, and the mentally ill are responsible for their illness. The previous (second) claim does not exclude that the mentally “ill” may be responsible for their “illness” but

---

30 Patricia E. Erickson and Steven K. Erickson (2008: 14–17) call such a conception of mental illness (which views it simply a failure of personal responsibility) “moral/punitive.”
this (possible) ascription of responsibility loses moral weight insofar as mental illness is assumed to be a myth – no illness at all. It is worthy of note that each of these claims implies that mental illness cannot be an exculpatory circumstance in criminal law. In discussing the third claim at greater length, I shall focus on Ronald D. Laing’s most famous book *The Divided Self* (from 1960), though this idea can be found also in the works of other anti-psychiatrists, for instance Szasz, who called bodily diseases “happenings” and mental illnesses “actions” (“mental illness is not something a person has, but something he does or is” (Szasz 1974: 167)), thereby suggesting that they are caused by the “ill” agent himself (cf. Szasz 1974: 154).31

Arguably, Laing’s account of mental illness can be interpreted as implying that mental illness is a genuine defect, but one which has been caused by the mentally ill person himself, so that one can speak about his responsibility for his illness. This interpretation can be justified (though not, as we shall see, unequivocally) by two strands in his argumentation.

First, Laing asserted that there is no unbridgeable abyss between sanity and madness: In his view, schizophrenia (on which he was focused) is in fact a development of certain traits already present in the (sane) personality of a schizoid agent, who suffers from “ontological insecurity” which can take three forms: fear of engulfment (of other people as threatening his autonomy), of implosion (of the world “as liable to crash in and obliterate all identity” (1969: 45)), or of petrification (of depersonalization). A person suffering from ontological insecurity has a diminished feeling of himself as a person who is real, autonomous, “spatially co-extensive with his body,” and identical over time. He experiences the division between, on the one hand, his “un-embodied self,” and on the other, the world, his own body, and the “false-self systems” associated with the body (which enable him to function in the world but which he does not treat as expressions of his real self). According to Laing, the transition from the schizoid state to schizophrenia proceeds as follows: The division between the “un-embodied,” “inner” self and the “rest” (body/false-self systems/the world) becomes ever deeper – the false-self systems become more and more extensive, autonomous, and “harassed by compulsive behavioural fragments” (Laing 1969: 142), and the inner self becomes more

---

31 Let me recall that, according to Szasz, the mentally ill “play games” in order to (ineffectively) solve their “problems in living.” This seems to suggest that he would be inclined to hold them responsible for their “mental illness.” But, on the other hand, he argued that the concept of mental illness constitutes a threat to personal responsibility since “for the individual, the notion of mental illness precludes an inquiring attitude toward his conflicts which his ‘symptoms’ at once conceal and reveal. For a society, it precludes regarding individuals as responsible persons and invites, instead, treating them as irresponsible patients” (1974: 262).
and more volatilized, empty (though self-consciousness is enhanced, accompanied by the feeling of being constantly observed by others), unreal, impoverished, mechanical, and finally dead – fragmented; the fragments of the self start to function independently, without knowing what happens in the others (cf. 1969: 129–144), actions are no longer perceived as expressions of the real self. It means that the unity of personality disappears, the border between the self and the world is obliterated, reflective awareness is no longer possible, and memory becomes chaotic. The transition from being schizoid to being schizophrenic therefore consists not only in the further “dissociation of the self from the body” (1969: 129), but also in the destruction, fragmentation of the self itself. The end result is “chaotic nonentity” (1969: 162). Now, in describing the “transition” of the patient from the schizoid to the psychotic state, Laing writes about the individual who is already far on his road to psychosis in that he is “in a world in which, like some nightmare Midas, everything he approaches becomes dead. There are now perhaps only two further possibilities open to him at this stage: 1. He may decide to ‘be himself’ despite everything, or 2. He may attempt to murder his self” (1969: 147). Up to this point one might think that Laing makes the schizoid responsible for his transition to psychosis even at this late stage. But Laing adds that “[b]oth these projects, if carried through, are likely to result in manifest psychosis” (1969: 147). And subsequently:

[S]anity, i.e., outwardly “normal” appearance, dress, behavior, motor and verbal (everything observable), was maintained by a false-self system while the “self” had come to be more and more engaged not in a world of its own but in the world as seen by the self. I am quite sure that a good number of “cures” of psychotics consist in the fact that the patient has decided, for one reason or other, once more to play at being sane (1969: 148).

As we can see, Laing’s account of the transition is not entirely clear: One may interpret it as implying the responsibility of the agent, but this interpretation is by no means uncontroversial. Laing did not say expressly that the schizophrenic patient is responsible for his illness but there seems to be nothing in his account of the transition from the schizoid state to psychosis that would exclude such an interpretation. It is true that he maintained that, at some stage of this transition, there may be no way back to “normality.” It does not have to mean, however, that he believed that the schizoid could not have done so earlier, when the cleft between his “inner” self and his “false-self” system was not so deep.

Secondly, in his account of mental illness, Laing looked at illness from an existential, subjective point of view, as a specific way of life, without invoking any biological factors in his description (his account of the etiology of mental illness is therefore psychogenic). Given this perspective, one could plausibly argue that this description implies that this way of life has been chosen or at
least was not resolutely opposed by the patient. Thus, Laing’s view of mental illness as developed in *The Divided Self* may be interpreted as implying that the mentally ill person is to a large extent responsible for his illness; though it must be admitted that it is not entirely clear whether Laing himself would endorse this interpretation. For even though he assumed the subjective, existential perspective, he did not deny that an illness may have also organic aspects or causes, and that an important part of the etiology of mental illness is played by the parents of the ill (he developed the famous “good-bad-mad” scenario of the development of psychosis, which was to demonstrate that the suppression of the needs of a child may trigger the process of splitting the self into the “inner self” and “false-self” system).

By way of a digression, let me remark that it is also not entirely clear whether and to what extent Laing endorsed the stronger form of the second claim of anti-psychiatry (the claim, let me recall, that mental “illness” is not only no illness at all but also a privileged form of knowing). What is beyond doubt is that his views on mental illness underwent a certain evolution: In his later works he certainly became closer to this claim, although he seems to have only accepted its partial form: that mental illness is a real illness but also a privileged form of cognition. He had already written in *The Divided Self* that “our ‘normal’ ‘adjusted’ state is too often the abdication of ecstasy, the betrayal of our true potentialities, that many of us are only too successful in acquiring a false self to adapt to false realities” (Laing 1969: 12). He developed this line of thought in *The Politics of Experience*, suggesting that mental illness may be a way of counteracting the social pressure to conform to an abnormal society, that the mad may be more sane than the normal, that the mad may have an insight into other dimensions – may become “the hierophant of the sacred” (Laing 1967: 109) (though he did not claim that madness guarantees this kind of insight or that this is the only way of achieving it). Furthermore, as already mentioned, he did not maintain that mental illness is not an illness; as he wrote, “madness need not be all breakdown. It may also be breakthrough. It is potentially liberation and renewal as well as enslavement and existential death” (1967: 110). Laing’s point was therefore that a psychotic breakdown is not only the onset of the disintegration of personality: It may also (but is not guaranteed to) be a way of coming into contact with some other – mystical – reality. Thus, as Zbigniew Kotowicz noted, there is no reason to maintain that Laing was “romanticizing madness” or “encouraging people to go mad as this would enrich their lives” (Kotowicz 1997: 66).32

---

32 Cf. also the following quotation:
At no point did Laing lose sight of the fact that a breakdown is also “enslavement and existential death.” So why should these views provoke such a strong
To sum up, three different claims have been made within anti-psychiatry: (1) that the psychiatric institutions developed within the paradigm of the “orthodox psychiatry” deprive mentally ill patients of their dignity, and violate their rights; (2) that mental illness is not illness at all (the weaker form) and, besides being no illness at all, it is a privileged form of cognition (the stronger form); (3) that mental illness is a real illness, but one for which the patient himself is responsible. As we shall also see more closely in the next section (in this section I mentioned only Szasz’s ideas about abolishing the insanity defense), all three claims have exerted a strong – both beneficial and negative – influence on the psychiatric and penal systems. Especially, claims (2) and (3), though different, may lead (and have led) to similar conclusions as far as the criminal responsibility of the mentally ill is concerned, viz. to the equalization of the responsibility of the mentally ill and the mentally healthy. Claim (2) implies that their responsibility should be the same because there is no such thing as mental illness (though, since on this claim there is also no such thing as “normality,” it may also be invoked to justify the radical postulate of the abolition of all disciplinary institutions, including criminal law). Claim (3) implies that the responsibility should be the same because the mentally ill are themselves responsible for their (real) illness.

4 MODELS OF CRIMINAL RESPONSIBILITY OF THE MENTALLY ILL: A SECOND GLANCE

As already mentioned in section 1, the Standard Model assumes that the distinction between the mentally ill and the mentally sane is sound, and that the mentally sane can be held criminally responsible, whereas the mentally insane cannot. This model is accepted by most contemporary legal systems. For instance, in the Polish penal code the model is expressed by Article 31:33:

§ 1. Whoever, at the time of the commission of a prohibited act, was incapable of recognizing its significance or controlling his conduct because of a mental disease, mental deficiency or other mental disturbance, shall not commit an offence.

§ 2. If at the time of the commission of an offence the ability to recognize the significance of the act or to control one’s conduct was diminished to a significant extent, the court may apply an extraordinary mitigation of the penalty.

reaction? In part this was because Laing’s views were conflated with those of David Cooper, his collaborator in *Reason and Violence*. All of Cooper’s subsequent writing are far more unrestrained than Laing’s. … Cooper really saw the mentally ill as part of the vanguard in the fight against oppression, a position that Laing never adopted” (Kotowicz 1997: 67).

As mentioned, similar regulations are adopted in most other criminal codes.
§ 3. The provisions of § 1 and 2 shall not be applied when the perpetrator has brought himself to a state of insobriety or intoxication, causing the exclusion or reduction of accountability which he has or could have foreseen.

It is clear that this model implies the affirmative account of mental illness. It is less clear which of the conceptions of criminal responsibility are endorsed by those who introduced it or by those who apply it. At any rate, it is compatible with conceptions 2 and 4. When we look at the Standard Model (M1) in the historical perspective, we can discern that this model was also justified in a different way, one somewhat puzzling for the modern reader, viz. by pointing out that *furiosus solo furore punitur* – that the author of a criminal act who suffers from mental insanity is already punished by his madness. To be quite strict, this thesis does not justify M1, since it implies, unlike M1, that the mentally insane criminal deserves punishment, though the punishment is madness itself. However, in practice, the effects of this approach are the same as those of M1: The mentally insane criminal is not punished within the criminal justice system. Let us note that the sentence *furiosus solo furore punitur* echoes Justinian’s *Digesta*. It seems to be based on the belief that mental illness has a supranatural cause. This was a fairly typical belief in antiquity; as noted by George Rosen, “the belief that illness was inflicted by a supernatural power or by an angry deity as a punishment for sin was widespread among the peoples of the ancient world” (1968: 28).

As mentioned at the beginning of this chapter, the *Perverse model* (Mentally sane → Not-responsible, Mentally insane → Responsible) was never adopted in the past. But a related model, in which the mentally sane are responsible, but the mentally insane are more responsible (i.e., mental illness is an aggravating circumstance) was adopted in some medieval legal codes (e.g., in the 7th-century Edict of Rothari, the Lombard king). It was justified by a supranatural conception of mental illness as a punishment for sins (cf. Robinson 1996: 51); on this conception, mental illness of a perpetrator of an illegal act was regarded as additional evidence of his guilt. As we can see, the belief in the supranatural character of madness could support two different conceptions of responsibility of the mentally ill.

In the *Negative Equalization Model* neither the mentally sane nor mentally insane can be held criminally responsible. This model must be accepted by hard determinists, i.e., those who assume a conception which deems free will or being *causa sui* to be a condition of criminal responsibility (conception 1, 2, or 3) and deny their (free will’s or *causa sui’s*) existence. This justification was endorsed, for instance, by the representatives of the so-called Sociological School in criminology, who assumed a deterministic conception of human behavior, including criminal behavior (they stressed the importance of social causes). The most famous of them, Enrico Ferri, an Italian thinker from the
The philosophical foundations of the insanity defense

turn of 19th and 20th century, asserted that the notion of a penalty cannot be legitimately applied given the non-existence of free will; the measures taken by the state against the perpetrators of criminal acts only have a pragmatic justification: They serve difesa sociale – a self-defense of the society against potentially dangerous individuals. This line of thought leads to the conclusion that criminals have themselves been wronged – they are victims of society – and thereby should be subject to some kind of therapy rather than being punished. This view has been defended in modern times, for instance by Karl Menninger (2007), who famously wrote that “all the crimes committed by all the jailed criminals do not equal in total social damage that of the crimes committed against them” (2007: 33), e.g., by their educators using improper child-rearing practices. He argued that a large number of crimes are committed by those who feel desperate, indignant, frustrated, helpless. To punish such perpetrators rather than to make attempts at their socialization is, according to Menninger, a “crime.” Menninger proposed a radical reform of the criminal justice system. Its basic points are the following: Indeterminate sentences should be introduced; “amenability to social control” should replace the concepts of responsibility and culpability; psychiatrists should not be allowed to participate in criminal proceedings since they tend to be instrumentally used – “as a legal weapon” – by lawyers, and the differences in their opinions compromise the psychiatric profession; their opinions – to be expressed only after the resolving of legal issues – should be limited to determining the proper way of resocializing – to finding the proper “penalty” – for the convict.

34 Cf. the following quotation: “The violent destructiveness of the criminal is so often a reflection of the cruel and violent way in which he was treated as a child” (Menninger 2007: 252).

35 The belief in the social causes of crime usually goes hand in hand with a belief in the fundamental goodness of human nature. But with Menninger it is not the case; he rejects the optimistic view of human nature, claiming that we all are potential criminals, that “an instinct for destructiveness” is in all of us (he even goes so far as to claim that this instinct for destructiveness is one of the main reasons for public resistance to humanitarian efforts to reform the criminal justice system). One may query whether this combination of beliefs is consistent.

36 This kind of “don’t blame me – I’m not responsible” attitude was forcefully criticized by Charles J. Sykes (1993): If everything bad we do is the result of “illness” (of our own or of the society), if, in other words, each of us is a victim, then we must dispense with the notion of personal responsibility. This amounts to the medicalization of the concepts of sin and crime. Sykes also argued that mental health professionals are responsible for it.

37 Menninger (2007: 142) made a distinction between penalties and punishment; the former are humanitarian, the latter, e.g., “long-continued torture” of seclusion, “pain inflicted over years for the sake of inflicting pain,” are to be eliminated from the legal system.
Menninger also assumed a peculiar view of mental illness according to which it is a form of “social misadaptation,” for which society itself is responsible (cf. 2007: 160), and from which all people suffer, to a greater or lesser degree, he was against the insanity defense; he also shared Szasz’s fear (even though his leftist political views were radically different from the libertarian right-wing Szasz) that it can be used to detain people unjustly (though Menninger at the same time supported indeterminate sentences, to which Szasz was strongly opposed).

The Positive Equalization Model, assuming that both the mentally sane and mentally insane can be held criminally responsible, can be supported on various grounds. It is implied by conception 5 (objective/absolute responsibility), conception 6 (if it is assumed that the mentally insane are responsive to sanctions), or conception 1 (if it is assumed that free will exists) since, on this conception, even the mentally insane exhibit free will. As we can see, the belief that free will can never be “suspended” may lead to the rejection of the insanity defense.

Furthermore, the Positive Equalization Model is justified by conception 7, conception 2 in conjunction with the negationist account of

---

38 His exact words are as follows: “all people have mental illness of different degrees at different times, and ... sometimes some are much worse, or better” (Menninger 1963: 32).

39 Overall, the differences between Menninger’s and Szasz’s views were much deeper than the similarities; it should therefore come as no surprise that it was Szasz who formulated perhaps the harshest critique of Menninger’s views. Here is an excerpt of it:

According to Menninger, the punishment even of persons guilty of the most heinous crimes, is, quite literally, a crime – whereas their crime is not a crime. Lest anyone think that Menninger is championing liberty for lawbreakers, let me hasten to add that he is not championing liberty for anyone; instead, he seeks to lump the “criminal” together with the rest of humanity – everyone being a fit subject for indefinite psychiatric incarceration at the whim of the psychiatrist” (1980: 113).

Szasz, who believed in personal responsibility and was a true champion of liberty, could not accept Menninger’s position which amounted to the abolition of personal responsibility and punishment, and to the medicalization of law.

40 It is worth noting that an extreme application of this kind of reasoning can be discerned in the witch trials of modern Europe; as Daniel N. Robinson put it:

[D]uring the centuries of witch-hunts, trials, and executions, the insanity defense is virtually absent in trials for witchcraft itself. For all the powers that might be attributed to Satan, the Church remained resolute in denying that the devil could rob persons of their free will. Were this otherwise, the very grounds on which eternity is spent in heaven or hell would be removed. Satan’s power can extend only as far as God allows, and the line is drawn at man’s moral freedom. In this understanding, then, the possessed are not passive victims but willful collaborators (1996: 63).
mental illness and with the acknowledgment of the existence of free will, or *conception 4* in conjunction with the negationist account of mental illness.\textsuperscript{41} Practical arguments have also been invoked in favor of this model, e.g., the fear that the insanity defense will be abused, i.e., treated instrumentally as a tool for shielding guilty offenders from criminal responsibility (the fear may be justifiable if the list of various mental illnesses is extended), or the putative fact that “insanity trials are regarded as the modern-day equivalent of Roman circuses in which psychiatrists, our contemporary gladiators, testify at great public expense to absurd propositions that bring both law and psychiatry into disrepute” (Brooks 1985: 126). This model was very rarely implemented in legal systems yet there are some examples. Perhaps the best known would be Idaho,\textsuperscript{42} where in 1982 the insanity defense was abolished for the first time in the history of the US. The consideration of insanity was eliminated from a criminal trial, which was to be restricted only to ascertaining *actus reus* and *mens rea* (the regulation was, therefore, similar to the one proposed by Hart).\textsuperscript{43}

\textsuperscript{41} It should be noted that the negationist account undermines the very concept of internal coercion.

\textsuperscript{42} Earlier attempts to abolish the insanity defense in the US were made between 1908 and 1928 by three state legislatures (Mississippi, Washington, and Louisiana). For instance, the law abolishing the insanity defense enacted by the Washington legislature in 1909 had the following content:

> It shall be no defense to a person charged with the commission of a crime, that at the time of its commission he was unable, by reason of his insanity, idiocy, or imbecility, to comprehend the nature and quality of the act committed or to understand that it was wrong; or that he was afflicted with a morbid propensity to commit prohibited acts; nor shall any testimony or other proof thereof be admitted in evidence (Washington Criminal Code par. 6, 1909).

Another law – that passed in Mississippi in 1928 – abolished the insanity defense in homicide cases (the defendant, however, could present evidence of his mental state; if the evidence was accepted by the jury, it could find him guilty but insane; in that case, his punishment was life imprisonment rather than – as in the case of mentally sane murderers – death). However, all these laws were struck down by the state supreme courts as unconstitutional (violating the defendant’s due process rights).

\textsuperscript{43} However, the Idaho law admitted an examination of the defendant before a trial; if he was declared to be lacking competence to stand trial, he would be placed in a mental facility and only tried if he regained his ability to participate in his own defense. This, of course, creates the risk that the defendant’s lawyers will try to introduce the insanity defense through the back door; as was noted by one of the prosecuting attorneys quoted by Gilbert Geis and Robert Maier:

> [I]t is probably incorrect to say that Idaho “abolished” the insanity defense. The new tactic of defense attorneys in our jurisdiction is to challenge an individual’s ability to aid his attorney in the preparation of his own defense. If successful in that endeavor the defendant may never have to be tried under our statutory scheme. Thus, our “abolition” has shifted the emphasis from asserting the defense during the trial to pre-trial. Although the standards are different, it
The only possible verdicts are guilty and not guilty; if the defendant is declared guilty but there is reason to believe that his mental state is a relevant factor for the determination of his sentence, then the psychiatric examination is to be conducted in order to ascertain the extent of his mental illness, the prognosis for rehabilitation, the availability of treatment and level of required care (since the defendant is guilty, all forms of care should, strictly speaking, be viewed as a punishment). The motivation for the enactment of this law was complex:

Persons in positions of authority in regard to mental health law reform in Idaho had formed a virtual consensus supporting abolition. That consensus was firmly based on a belief that the insanity plea had been blatantly misused and had failed to provide either satisfactory treatment for offender or adequate community protection against them. Combined with this conviction was the popular view, strongly held by residents of this isolated, highly religious state, that all human beings ought to take personal responsibility for their behavior, that they should not be able to avoid punitive consequences of criminal acts by reliance on either a real or a faked plea of insanity (Geis, Meier 1985: 73).44

After Idaho, several other states also abolished the insanity defense. The differences between their solutions are not deep: While some of them (Idaho, Montana, Kansas, Utah) have no specific insanity defense and the verdicts they admit are only guilty or not guilty (though Montana, Kansas, and Utah, for example, allow evidence of mental disease or defect for purposes of negating mens rea), some others (e.g., Kentucky, Michigan, Alaska, Delaware, Georgia, Illinois, Indiana, New Mexico, South Carolina, South Dakota, Pennsylvania) have some weakened form of this defense because, in addition to the verdict “not guilty by reason of insanity,” they admit verdicts of “guilty but mentally ill” (cf. Erickson, Erickson 2008: 102). A defendant who is declared guilty but mentally ill is held responsible for his act: He is punished (as being guilty) and, while serving his sentence in prison, simultaneously given (as mentally ill) medical treatment. This solution clearly shows that “medical insanity” is

doesn’t appear to be as difficult to have an individual certified not to be able to assist his attorney in preparation of his own defense as it would be to challenge the intent aspect of a particular crime” (1985: 79–80).

Furthermore, in order to determine whether the condition of mens rea is satisfied, some analysis of the mental state is necessary; so the abolition of the insanity defense is not full (though, it is substantial, because the prosecuting attorney does not have to prove that the defendant knew at the moment of committing the prohibited act its normative quality, or that he could control his action).

44 Seeking to determine the causes of this decision, Geis and Maier also highlighted the fact that “Idaho was the first jurisdiction to act in the wake of the mood of public outrage after the acquittal of John W. Hinckley, Jr., on the grounds of insanity, following his attempt on the life of President Reagan” (1985: 73).
something different from legal insanity (a defendant may prove that he was mentally ill at the time of the crime but fail to prove that he was legally insane). The attitude of the US legal system to criminal offenders who are mentally ill deserves, however, a more versatile analysis, one taking into account also a broader, sociological context. I shall present such an analysis in Section 5. I shall finish the following section with a presentation of the Swedish regulation of the criminal responsibility of the mentally ill, which, as we shall see, is similar to the Idaho regulation yet was adopted for different reasons.

Sweden adopted a legal regulation in 1962 according to which if the accused has committed a criminal act in a state of mental insanity the criminal court is obliged to decide that he is _guilty/criminally responsible but mentally insane_ (rather than, as in most other countries, those which assumed the Standard Model, that he is not guilty/not criminally responsible by reason of insanity) and, subsequently, may punish him more leniently or not punish him at all but instead place him in mental hospital.\(^{45}\) As we can read in the Swedish penal code, mental illness cannot release an offender from responsibility – it can only be a mitigating circumstance:

Ch. 29. Section (3) In assessing penal value, the following mitigating circumstances shall be given special consideration in addition to what is prescribed elsewhere, if, in a particular case:

1. the crime was occasioned by the grossly offensive behavior of some other person,
2. the accused, in consequence of a mental disturbance or emotional excitement, or for some other cause, had a markedly diminished capacity to control his actions,
3. the actions of the accused were connected with his manifestly deficient development, experience or capacity for judgment,
4. the crime was occasioned by strong human compassion or
5. the act, without being free from criminal responsibility, was such as is covered by Chapter 24.

The sentence imposed may be less severe than that prescribed for the crime in question if this is called for having regard to the penal value of the crime.

Ch. 30. Section (6) A person who commits a crime under the influence of a serious mental disturbance may not be sentenced to imprisonment. If, in such a case the court also considers that no other sanction should be imposed, the accused shall go free from sanction.

\(^{45}\) Let me recall that a similar solution was proposed (independently of the authors of the Swedish criminal code) by one of the greatest 20th-century legal philosophers, Herbert L.A. Hart (2008). Hart, as already mentioned, opted for transferring the problem of the mental competence of the accused (or, rather, the convict) to the post-trial stage; at the trial stage the judge was to resolve only the issues of _mens rea_ (intent, negligence, or recklessness) and _actus reus_. However, his reasons for adopting this solution, which I have described in Section 2.2, were different from those of the Swedish legislators or from those dominant in the case of the Idaho legislation.
Ch. 31. Section (3) If a person who has committed a crime for which the sanction cannot be limited to a fine, suffers from a serious mental disturbance, the court may commit him for forensic psychiatric care if, having regard to his mental condition and personal circumstances, admission to an institution for psychiatric care combined with deprivation of liberty and other coercive measures is called for.

If the crime has been committed under the influence of a serious mental disturbance, the court may decide that a special release inquiry under the Act on Forensic Psychiatric Care (1991:1129) shall be conducted during the time in care if there is risk for relapse into serious criminality of a serious kind by reason of the mental disturbance.

The court may, in conjunction with a committal to forensic psychiatric care impose another sanction, but not imprisonment or committal for other special care, if this is called for having regard to the previous criminality of the accused or for other special reasons.

Thus, the Swedish solution abolishes the insanity defense but separates criminal responsibility from punishability. Mentally insane offenders are not acquitted: If they have committed a crime then they are declared guilty/culpable/criminally responsible. But it does not mean that they are punishable just like mentally sane offenders: Their illness is taken into account at the stage of determining sanctions. This implies that what is decisive for legal reaction is the mental state of the accused at the time of trial, not at the time of committing an offense. From what has been said so far it follows that the Swedish regulation of criminal responsibility of the mentally insane (just like the regulation from Idaho) does not lead to the full equalization of the legal position of mentally insane and mentally sane offenders.

What reasons stood behind the Swedish regulation? According to the Polish sociologist and journalist Maciej Zaremba (2008), the regulation is a result of an incoherent fusion of two different conceptions of criminal responsibility. On the one hand, it was inspired by the Sociological School of criminology (especially by Enrico Ferri), which, as already mentioned, postulated the abolition of the concept of criminal responsibility on the grounds that since determinism is true, no agent can be deemed guilty, and therefore no punishment is justified (the state can defend against perpetrators of crime but the measures taken by the state cannot be called punishment because they are not the realization of justice but difesa sociale). If this conception were realized consistently, the result would be the Negative Equalization Model. In fact, Ferri’s conception influenced the thought of the Swedish psychiatrist Olof Kinberg, who postulated such a model, eliminating the concepts of guilt and responsibility from criminal law considerations. The state, assisted by psychiatrists, was to isolate the most dangerous individuals, thereby realizing the utilitarian goal of maximizing social utility rather than the goal of retributive justice. This view underlay the project of a penal code (called the “Law of Protection”) proposed...
The philosophical foundations of the insanity defense

by the Swedish Minister of Justice Karl Schlyter in 1956: The project did not avail itself of the concept of punishment, but rather that of “consequences”; the consequences could also be imposed for an indefinite length of time; everyone (the mentally insane and the mentally sane) could face criminal proceedings but no one could be declared guilty. However, this radical but consistent project was ultimately rejected. On the other hand, the Swedish solution was influenced by the commonsense, standard view according to which the mentally insane should be treated differently from the mentally sane; on this view, only the latter deserve to be punished. The eventually adopted compromise solution tried to reconcile these two different conceptions: It assumes that both mentally insane and mentally sane criminals are responsible for their acts but the distinction between them can be introduced at the stage of determining the consequences of their actions. This is indeed paradoxical: The theory of Enrico Ferri implied that no one can be criminally responsible, while the Swedish code, even though it was influenced by this theory, implied that all (including the mentally insane) can be criminally responsible. One might think that this solution, as far as its practical consequences are concerned, does not differ from the Standard Model, but the reality is very different: The Swedish solution has had several non-humanitarian consequences (cf. Zaremba 2008). First, the accused who is mentally insane has to take part in the criminal proceedings. Secondly, if an agent committed the crime in a state of mental incompetence but is healthy during the criminal process, then he will be punished (just like the criminal who committed the crime in a state of mental competence). Thirdly, a large number of mentally ill convicts are punished rather than sent to mental hospitals.

One should add that the deterministic view of human beings, as well as the blurring of the border between the competences of judges and psychiatrists, which underlay this project, were criticized, e.g., by the Social Democrat politician Georg Branting in his article from 1943 “Legal Corruption, or the Fall of the State of Law in Sweden,” and by Professor of Law Ivar Strahl, who was strongly against the rejection of the concept of guilt (cf. Zaremba 2008). One can also speculate that a certain role in advancing this project could have been played by an idea advocated by some legal philosophers from the Scandinavian branch of legal realism (e.g., Alf Ross). This is that “responsibility” is a technical, empty, or functional concept, not carrying any moral weight, and that the notions of “guilt,” “culpability,” or “punishment” are metaphysical fictions and as such should be dispensed with. It should be stressed, however, that Ross himself was in favor of the Standard Model based on conception 4 (cf. Ross 1975).
The criminologists Patricia E. Erickson and Steven K. Erickson have advocated an interesting two-part thesis:

1. that there prevails in the US a negative public sentiment and policy toward the mentally ill [caused by the fact that mental illness came to be regarded] as a failure of individual responsibility rather than an illness requiring a humane orientation … [and] reframing mental illness as a failure of individual responsibility, along with a continuation of the traditional criminal law model, created a shift to a punitive stance toward the mentally ill – hence the criminalization of mental illness (Erickson, Erickson 2008: 8–9);

2. that one of the important causes of this “reframing” of mental illness was the “anti-psychiatric” movement.

As already mentioned, it may seem paradoxical (even though it is by no means illogical) that the “existential” view of mental illness typical for anti-psychiatry could have had non-humanitarian consequences, i.e., could have led to the harsh treatment of the mentally ill, whereas the “biological,” “reductionist” account of mental illness is likely to lead to the humanitarian treatment of the mentally ill. Yet the paradox disappears if we recall that the existential view (unlike the biological one) may imply individual responsibility for mental illness; it may lead to the conclusion that the mentally ill are themselves to blame for their predicament, and that thereby mental illness is to be regarded as a moral failure. If, as has happened in the US, the “existential” account is combined with the retributive theory of punishment, the treatment of the mentally ill is likely to be especially harsh. Thus, even though anti-psychiatry was to a large degree a reaction against the non-humanitarian treatment of the mentally ill, perhaps as an unintended side-effect it may itself have led to the decrease in empathy towards the mentally ill. After all, if they are not mentally ill at all or are responsible for their mental illness, then the important grounds for empathy disappear. Yet it is clear that the process which led to the “punitive stance” towards the mentally ill was caused by many factors, including a genuine concern for the well-being of the mentally ill, not only by the view of mental illness (supported to some extent by “anti-psychiatry”) as a failure of individual responsibility. A crucial stage of this process seems to have been the “deinstitutionalization of mental illness” – “the policy of moving severely mentally ill people out of large state institutions and then closing part or all of
those institutions” (Torrey 1997: 8). This process and its causes will now be described in more detail.

The process began in the 1950s and gained momentum in the 1970s as a result of the critique of “institutions” which, as was already mentioned in Section 3.2., were often indeed brutal, cruel, and dehumanizing. This phenomenon was described suggestively in Ken Kesey’s One Flew over the Cuckoo’s Nest (1962) and Mary Jane Ward’s The Snake Pit (1946), with patients in mental hospitals given cold baths, lobotomies, and electroshock treatment. Mental hospitals therefore not only fulfilled a therapeutic function, but also, as stressed by Erickson and Erickson (2008: 15), a “containment function.” Erickson and Erickson, however, aptly noted that “while the conditions in state mental hospitals rightfully came under attack, the conditions themselves do not explain deinstitutionalization and the movement to community care” (2008: 30). Moral indignation at the conditions prevailing in mental hospitals was therefore the initial cause of the process of the deinstitutionalization of mental illness. But it had to be supplemented by several others for the process to take place.

First, one must stress the role of the cultural revolution of the 1960s (and its “child” – anti-psychiatry), which led to the emphasis of the importance of individual rights, personal autonomy, dignity, anti-paternalism, and the concomitant protest against all – or at least all arbitrary – forms of state coercion. One of its consequences was giving the mentally ill the right to refuse treatment: Involuntary treatment and hospitalization were severely criticized and were only accepted if a mentally ill person could harm himself or other people. As already mentioned, this argument was often raised by representatives of the anti-psychiatry movement, notably by Tomas Szasz, who argued that involuntary hospitalization is crueler than incarceration: The former is for an indeterminate time, the latter determinate.47

47 Yet it should be stressed that Szasz cannot be made co-responsible for the deinstitutionalization process. He was equally strongly opposed to involuntary hospitalization and involuntary dehospitalization; he wrote very critically about it and in a deeply humanitarian spirit:

In the past, thousands of individuals were forcibly incarcerated in mental hospitals, often for life; that was bad enough, but at least many of these unfortunate persons could make the asylum their home. Now the situation is even worse, with thousands of people not only being forcibly incarcerated in mental hospitals, but also forcibly evicted from them as soon as they show any sign of adapting to their new environment. Then the cycle of forcible hospitalization and dehospitalization is repeated over and over again, depriving the “mental patient” of a predictable and stable environment both within and without the insane asylum (1990: 561–562).
Secondly, the process might also have been influenced by the conservative – libertarian but not personalist – conception of human beings, which assumes that they are endowed with free will which can never be “suspended,” and which therefore is active even in agents who suffer from mental illness.

Thirdly, a number of medical experts strongly believed in the efficiency of the pharmacological treatment of mental illness (especially antipsychotic drugs from the 1950s, e.g., Chlorpromazine-Thorazine, introduced in 1954, which, within one year from its introduction, had been given to more than 2 million patients), which could be conducted outside mental hospitals. It was also believed that long-term stays in hospitals “produce[d] institutional behavior and a tendency to chronic illness (Erickson, Erickson 2008: 26)” and that thereby deinstitutionalization served autonomy, liberty, and dignity of mentally ill persons.

There was, however, also, a fourth – economic – cause, connected with the fact that hospitals for the mentally ill were expensive. Edwin F. Torrey plausibly argued in his important 1997 book Out of the Shadows. Confronting America’s Mental Illness that this was the main cause of deinstitutionalization. More specifically, it was the introduction in 1965 of federal welfare programs such as Medicaid and Medicare that accelerated the process. Since the programs did not include state psychiatric hospitals (the states had to pay themselves for their sustenance) the incentive arose for the states to shift the costs and responsibility by closing many of these hospitals and transferring the patients to nursing houses or general hospitals, which were financed by the federal government’s welfare programs. One can therefore justifiably maintain that “federal dollars have contributed substantially to disjointed, uncoordinated care of mentally ill individuals” (Torrey 1997: 105).

As a result of these various causes, the “deinstitutionalization of mental illness” took place, which, in Torrey’s view, “has been a major contributing factor to the mental illness crisis” (1997: 9). Torrey called deinstitutionalization one of the greatest social experiments in the history of the US: While in 1955 there were about 560,000 patients in mental hospitals (from a population of 164 million), in 1994 there were only 72,000 (of a population of 260 million). The difference, as noted by Torrey (1997: 8–9), will prove to be even higher if we take into account the change in population: If the population in 1955 had been at the same level as in 1994, the number of patients in mental hospitals would have totaled around 885,000. It is also worth emphasizing the fact that the most seriously mentally ill were discharged from hospitals (about

His blame may only consist in having contributed (by his works) to creating a climate in which mental illness was no longer regarded as illness.
50–60 percent of those discharged were schizophrenics and about 10–15 percent were manic-depressive or severely depressive patients).

What were the consequences of this “social experiment”? It was believed that the mentally ill would take care of themselves and/or would be taken care of by local communities. Yet this expectation proved to be ungrounded in many cases. The mentally ill who were discharged from hospitals often started to take drugs and abuse alcohol (the result of deinstitutionalization was therefore also detreatmentalization), they frequently refused to participate in pharmacological treatment (because the mentally ill rarely regard themselves as ill), became homeless, and, finally, many of them turned to crime. As a result, there took place what Torrey called the “transinstitutionalization” of the mentally ill from asylums to prisons. Prison proved to be a disastrous alternative to mental hospitals, which is illustrated, for instance, by a high suicide rate among the incarcerated mentally ill (though it should be stressed that transinstitutionalization was, of course, an unintended side-effect of deinstitutionalization). Erickson and Erickson aptly noted that “the proponents of deinstitutionalization … created a climate that was less tolerant of the mentally ill and more willing to use prisons and jails as their places of confinement (2008: 25); and in similar vein: “many of the older commitment laws were predicated on recognizing this difference [between the mentally ill and the mentally healthy], while much of the 1960–70 ‘liberation’ from psychiatry was guided by seeing the mentally ill as the same as everyone else” (2008: 194).

Let me summarize the above description of changes in the situation of the mentally ill: Since the “deinstitutionalization of mental illness” was, in large part, based on unduly optimistic premises (e.g., that the mentally ill can take proper care of themselves), it entangled the mentally ill in criminal behavior, for which they were held responsible, just like the mentally healthy (which was, to some extent at least, a result of the reinterpretation of mental illness as a failure of individual responsibility, and the popularity of the libertarian view of human being); in consequence, “transinstitutionalization” took place.

It is clear that one should not interpret the above analyses one-sidedly – as justifying an overall negative evaluation of the changes in the approach to the mentally ill that took place in the US in the 1960–1970s, and as postulating a return to the state from before the “anti-psychiatry” revolution. The truth,

---

48 This process jibes with the so-called balloon theory, which says that “prison and psychiatric hospital population are inversely correlated: as one rose the other fell – push in one part of a balloon and another part will bulge” (Torrey 1997: 34). One may ask, however, why the juries so often decide to direct the mentally ill to prisons for a fixed period than to mental hospitals. One of the reasons may be their fear that psychiatrists may discretionally decide to release them, accompanied by the conviction that the mentally ill are themselves to blame for their illness.
of course, is much more complex. One should not underestimate the positive effects of deinstitutionalization (which Torrey called “fundamentally a good idea that was carried too far” (1997: 87)) and anti-psychiatry. To give one of many examples, the *Psichiatria Democratica* movement in Italy, founded by the “anti-psychiatrist” Franco Basaglia, led to the adoption of a law in 1978 (the so-called Basaglia’s law) which improved the situation of the mentally ill, protecting their civil rights. One should also remember that “anti-psychiatry” did not engender negative effects by itself but jointly with a number of other causes (e.g., the belief in the efficiency of the pharmacological treatment of the mentally ill), which may have played an even more important role. Finally, it needs to be stressed that deinstitutionalization proved to be beneficial for many, or even the overwhelming majority, of patients; Torrey (1997: 144) admits that 90 percent of mentally ill people were satisfied with this policy, though he adds that a substantial minority found themselves “in terrible conditions” (e.g., in 1994, 150,000 mentally ill people were homeless, and 159,000 incarcerated in jails and prisons (1997: 85)). Accordingly, it is difficult to draw up a balance sheet of the costs and benefits. But the very existence of these negative side-effects fully justifies Torrey’s claim that important corrections to the approach to mental illness need to be introduced. He provides a list of very reasonable recommendations.

First, the idea of the involuntary hospitalization of the mentally ill, including those who are not likely to harm themselves or others, should be rehabilitated since the liberal approach to this issue has had unintended and undesirable consequences.

Secondly, Torrey convincingly argues that we should abandon the continuum/spectrum concept of mental illness (which, as I argued earlier, is structurally similar to the claim that there is not mental illness at all). He postulates that the term “mental illness or disorder” should be reserved for really serious problems – functional brain abnormalities or impairments (e.g., schizophrenia, manic depression, autism, severe forms of depression, panic disorder): “It is necessary to return to the 19th century idea that these illnesses are in a different category from problems of mental health” (1999: 199). They should also, according to Torrey, be conceptually merged with brain disorders called the neurological illnesses (e.g., multiple sclerosis, Alzheimer’s, Parkinson’s). This merger would have two important advantages: It would weaken the stigmatization of the mentally ill (because mental illness would be conceptualized as brain disorder, not as a failure in individual responsibility), and it might also accelerate research on mental illnesses (e.g., neuroscience research on the causes of
schizophrenia would be promoted). Accordingly, instead of devoting energy and resources to the people who suffer from minor mental problems that do not deserve to be regarded as mental illnesses or disorders, we should focus our attention on counteracting real or really serious mental illnesses. In Torrey’s opinion, too much effort has been spent on so-called mental health problems rather than on mental illnesses. Somewhat surprisingly, anti-psychiatry may to some extent be helpful in this context. The anti-psychiatric view may be used to justify cutting the spectrum and making a division between serious or imaginary (or at best non-serious) mental disorders. The latter would be a myth in Szasz’s sense. Clearly, this kind of usage of anti-psychiatry is not quite true to the intentions of its adherents, who were apt to defend the claim that all mental illnesses are not true illnesses. To follow this second recommendation would lead to limiting the grounds for raising the insanity defense. It would, e.g., lead to the abolition of the so-called “syndrome defense,” which mitigates or abolishes criminal responsibility (examples of syndromes invoked within this defense are premenstrual syndrome or battered wife syndrome); this postulate will be additionally supported by considerations devoted to the critique of the volitional component pursued in Chapter 3.

The third recommendation made by Torrey refers only to the context of the US: There should be single-responsibility funding (rather than funding which provides an incentive to the states to shift the responsibility for the mentally ill to the federal government).

Let me summarize. The influence of the cultural revolution of the 1960s on criminal law was multifaceted. The revolution led to the development of the negationist account of mental illness (as a “myth”), which may have induced politicians to adopt the Positive Equalization Model. Yet such a model has only been explicitly (at the legislative level) accepted in some states. And in those states the reasons for adopting was probably different: Legislatures were arguably more influenced by conception 1, with its libertarian (but not personalist) assumption that also the mentally ill are responsible, and by more prag-

---

49 One can also interpret Torrey’s postulate of restricting the concept of mental illness to brain disorders in a philosophical way. The interpretation would imply that one can speak about mental illness only if it leads to the loss of the status of personhood in Max Scheler’s sense. According to Scheler (2014), one of the conditions of personhood is having the spiritual center, which can be viewed as a source of human intentional acts. If we explain someone’s action in causal (objective terms) terms, and not in terms of his intentions, then we imply that this person’s spiritual center is empty. A patient is therefore mentally ill if he lacks this spiritual center. Two other conditions of personhood mentioned by Scheler are the capacity to understand the difference between one’s own and others’ mental states, and the capacity for controlling one’s instinctive impulses. A patient who does not fulfill the first condition is likely to fail to satisfy the remaining two.
matic arguments, e.g., fear that the insanity defense will be abused, leading to exoneration of guilty and deserving agents. The fact that this model was not accepted in many states could have been attributed to some extent to the countervailing idea of this revolution, viz. the idea that criminals themselves have been wronged (that the oppressive “society” is guilty), and that – consequently – the proper reaction to a crime (irrespective of whether committed by the mentally sane or insane) is subjecting criminals to therapy, not putting them in prison, or at least treating them in a humanitarian way. However, strictly speaking, this last idea implies the Negative Equalization Model, according to which both mentally ill and mentally healthy people are supposed to be subject to a therapy rather than punished (this model was never adopted as the official one). Another important aspect of the cultural revolution was a celebration of “transgression,” and even though only few people engaged in the revolution rejected the very idea of crime, many postulated that the number of crimes should be limited. This postulate has been to some extent realized (e.g., there took place the depenalization of homosexuality).

6 EVALUATION OF THE MODELS

Even though the main goal of this chapter was to distinguish various models of the criminal responsibility of the mentally ill, and to identify their philosophical underpinnings, one cannot stop here and refrain from at least attempting an evaluation. In point of fact, my presentation of the various models, especially of the models of criminal responsibility, was critical and not purely descriptive. However, the evaluation made so far must be supplemented by some additional remarks.

6.1 Conceptions of Criminal Responsibility

Let me start from the first dimension of the models – various conceptions of criminal responsibility. I have already argued that the \textit{causa sui} (in its strong form) requirement seems to be untenable, and the “consequentialist” requirement seems to be too weak. The real choice as far as the conditions of responsibility are concerned is between freedom of will \textit{cum} freedom from coercion (\textit{conception 2}), and only freedom from coercion (\textit{conception 4}).

It should be stressed that from the anthropological perspective (the personalist view of human beings) adopted in this book the problem of whether responsibility ascriptions would be justified if the assumption about the existence of free will were to turn out to be false is purely speculative because the central claim of this perspective is that free will exists. So, from this perspective, the adequate conception of responsibility is \textit{conception 2} (which, obviously, provides a stronger justification of responsibility ascriptions than \textit{conception
4). But the question of whether freedom from coercion is sufficient for responsibility ascription is of great philosophical interest. When one analyzes the requirements of criminal responsibility explicitly stated in most contemporary penal codes, one will notice that they include certain perceptual, conative, evaluative, and volitional capacities, but not the metaphysical capacity to act otherwise, i.e., free will. It might therefore seem that penal codes do not treat free will as the condition of criminal responsibility — that they presuppose that freedom from coercion is a sufficiently strong concept of freedom to justify responsibility ascription and punitive practices. Yet drawing this conclusion may be too hasty.\textsuperscript{50} One can argue against it in three different ways. One can assert that a more careful reconstruction of the presuppositions of penal codes, taking account of not only penal codes, but also judges’ opinions and legal doctrine, would reveal that they include also the assumption of free will (one may argue that this assumption is presupposed by the very notion of culpability, or by what is called in the continental legal tradition a “normative conception of guilt”). Furthermore, one can concede that penal codes do not presuppose free will but claim that, if that is the fact, then they are based on a fundamental mistake: They provide a scheme of punishment but this scheme is not sufficiently justified since an agent cannot be held responsible and punished if he could not have acted otherwise, i.e., if he is not endowed with free will. Finally, one can argue that even though freedom from coercion is indeed a presupposition of penal codes and provides sufficient justification of responsibility ascriptions, the justification provided by the free will is much stronger than that provided only by freedom from coercion.

6.2 Accounts of Mental Illness

Let me now pass to an evaluation of the two accounts of mental illness, which constitutes the second dimension of the analysis of the criminal responsibility of the mentally ill. The negationist view seems only partly justified, and, in

\textsuperscript{50} This conclusion was endorsed, e.g., by Ronald Dworkin (2011: 219–252). He attempted to find the best possible interpretation of our practices of holding people responsible, and came to the conclusion that the interpretation that best fits the data and is axiologically attractive is compatibilism: Responsibility is possible also in the deterministic universe on condition that agents possess “capacity control.” Note that the terminology he uses is different from mine: He speaks about “capacity control” rather than freedom from coercion, and about “causal control” rather than free will. But Dworkinian “capacity control” is in fact equivalent to freedom from coercion; an agent has such a control (over an act) if he is conscious of facing and making a decision, when no one else is making that decision through and for him, and when he has the capacities to form true beliefs about the world and to match his decisions to his normative personality (his settled desires, ambitions, and convictions).
addition to that, it is justified in its least controversial part. It is undoubtedly true that certain types of behavior, e.g., masturbation, fellatio, cunnilingus, homosexuality, or sexual promiscuity ("Don Juanism" or "nymphomania") were regarded as mental illnesses (e.g., in the first edition, from 1952, of the *Diagnostic and Statistical Manual of Mental Disorders*) only because they were inconsistent with the then prevailing social norms (all these "disorders" disappeared from the third edition of the *Manual*, from 1980). However, contrary to what Szasz maintained, many other mental illnesses (e.g., schizophrenia) cannot be interpreted in this way. The argument from the lack of evidence for the physical/material basis of, say, schizophrenia, turned out to be the weakest part of anti-psychiatrists’ (especially Szasz’s) argumentation for the thesis that mental illness is a myth. Szasz may have been right to the extent that at the time in which he proposed his theory the physical bases of mental illnesses were not known, so he could be somehow justified in saying that they do not have such bases. But he should have been more open to the possibility that they may be discovered, at least in part, in the future (as eventually happened; today our knowledge about the physical bases of many mental illnesses, though far from being complete, is large enough to justify a peremptory rejection of Szasz’s claim that mental illnesses have no physical bases).

We can also put this criticism in a different way. Szasz may have been right in saying that there is no such thing as mental illness if mental illness is understood in his way – as an “illness” without any neurological/bodily basis; it may be true that all purely mental disturbances can be interpreted as “problems with living.” But the truth is very likely to be, contrary to Szasz’s assertions, that this category does not refer to those mental illnesses (e.g., schizophrenia) which are most often invoked in the context of the insanity defense (most of which would have a neurological/bodily basis). Thus expressed, Szasz’s claim would be close to tautological: There is no mental illness because illness must, by (his) definition, have a neurological/bodily basis, and mental illnesses, by (his) definition, do not have such bases. Of course, it was not tautological *tout court* because his definition was at least to some extent motivated by an empirical hypothesis: that typical mental illnesses do not have such bases. Yet it was precisely this hypothesis which proved to be false and thus his claim was reduced to a tautology.

---

51 It should be noted, however, that the new (more permissive) sexual mores gave rise to new sexual “disorders” or “dysfunctions” (albeit not coined as mental), such as anorgasmia, inhibited sexual desire, sexual aversion, or ejaculatory incompetence (premature ejaculation).

52 It should be added that in one of his latest papers (from 2011) Szasz fully endorsed his position developed 50 years earlier in his *The Myth of Mental Illness.*
Another criticism of this type of theory, provided, among others, by Wakefield (1992), Kendell (2005), and Shorter (2011), concerns his lesion or morphological abnormality account of disorder; as Wakefield noted:

The account consists of two theses: (a) that a lesion (or abnormal bodily structure) is simply a statistical deviation from a typical anatomical structure and (b) that a physical disorder is simply a lesion. First, the idea that a lesion can be directly recognized by its deviant anatomical structure is incorrect. Bodily structures normally vary from person to person, and many normal variations are unusual as any lesion. Moreover, some lesions are statistically nondeviant in a culture, such as atherosclerosis, minor lung irritation, and gum recession in American culture and hookworm and malaria in other cultures. Therefore, recognition of a lesion is not simply a matter of observing anatomical deviance. Second, and more importantly, it is not the existence of a lesion that defines disorder. … A lesion can be a harmless abnormality that is not a disorder, such as when the heart is positioned on the right side of the body but retains functional integrity (Wakefield 1992: 375).

However, though important, this criticism is secondary: Szasz’s main point was that mental illness does not have a material basis, so even if he worked out more carefully the concept of the “material basis” of illness to counter the kind of objection raised by Wakefield, it would still be untenable in the light of contemporary knowledge about mental illnesses. However, as mentioned before, Szasz’s claim that mental illness is a myth is true when it is qualified, i.e., applied only to certain “mental illnesses”; in this sense, as argued before, it may help realize Torrey’s postulate of making a strict division between mental illnesses that are brain disorders and those that are of a different nature (and do not deserve the attention which is paid to them).

Allow me in this place to digress slightly to invoke a more general argument against equalization models of the responsibility of the mentally ill. The argument states that the difference between the mentally ill and the mentally healthy (downplayed in the negationist account) cannot be ignored in our practice of responsibility ascriptions. It can be developed by drawing on Peter Strawson’s (1962) analysis of responsibility which is based on the distinction between various types of moral attitudes. According to Strawson, P can be held morally responsible for committing an offense a if, and only if, the type of act he did (the type including not only outward features of the act but also the mental attitude of the agent) elicits a proper – impersonal – reactive attitude (moral indignation, moral disapproval) in most other agents (so that we can speak about the social practice of reacting in a given way to a given type of offense). R’s reactive attitude is personal when P’s immoral act endangers his personal interests; if it does not (i.e., is harmful for people whose interests are not the object of concern for R), R’s reactive attitude is impersonal/moral. Conversely, an agent is not responsible for an act he did if, and only if, he triggers a specific – objective – attitude of excuse/understanding (or rather explaining).
Here we reach a point of Strawson’s account which is especially important for our analyses, viz. that reactive attitudes (personal and impersonal) are not only sensitive to the positive or negative attitudes manifested in the behavior of others towards us/others but also to whether these attitudes manifest mental sanity or insanity. In the former case we take, as Strawson calls it, a “participant’s view” of their attitudes, in the latter an “objective” view, which is directed at explaining rather than evaluation. This implies that we cannot forbear distinguishing between the responsibility of the mentally sane and the mentally insane. The distinction, as Strawson argues, is deeply rooted in the way our attitudes function, although Strawson does not explain how deeply rooted they are – whether they are changeable or not. He seems to maintain that the rootedness is genuinely deep, since he says the way our attitudes function is part of “the general framework of human life” (Strawson 1962: 6). Furthermore, according to Strawson, our personal reactive attitudes would not change even if it turned out that determinism were true and we knew that it was true; Strawson deems it “practically inconceivable” that “a general theoretical conviction might so change our world that, in it, there were no longer any such things as interpersonal relationships as we normally understand them” (1962: 4). The same applies to impersonal reactive attitudes, though Strawson’s argument for this claim is indirect: He claims that it is harder to imagine the situation in which personal reactive attitudes remain while impersonal reactive attitudes disappear than the situation in which both types of reactive attitudes disappear, since they “stand or lapse together” (1962: 10). Thus, since personal reactive attitudes cannot disappear, impersonal reactive attitudes cannot disappear either. To the question whether it would not be rational to suspend our moral reactions if determinism were true, Strawson replies that “it is useless to ask whether it would not be rational for us to do what it is not in our nature to (be able to) do” (1962: 10). On a superficial reading of Strawson’s paper, one might assert that Strawson’s conception does not have a normative dimension as it takes the social practices of blaming/praising as they are and does not say what they should be like. Accordingly, if the practice were different, e.g., sensitive only to the “outward” action, or if we were mentally constructed in such a way that we would regard the mentally ill as capable of participating in normal relationships, then there would be no reason to introduce the “insanity defense” to a legal system. Yet the crucial point is that Strawson seems to deny that it could be different and so his descriptive analysis becomes normative if

53 As already mentioned, Max Scheler (2014), in his analysis of the concept of a person, argued in a similar fashion: If we stop treating an agent as a person, we intend to explain his action rather than understand it as intentional; we try to know the causes of this action rather than understand its sense.
one assumes the plausible thesis that one cannot require change of what cannot be changed. The consequence of this conception for the problem of criminal responsibility of the mentally insane is straightforward: Given the depth of rootedness of the Standard Model in our social practices, one can suppose that other models, if they were adopted, could not effectively function because they are too much at odds with those practices. This argument seems plausible even if we take into account the fact that, as we have seen in Section 4, non-standard models were implemented in some legal systems. This is all the more plausible if we remember that in those jurisdictions where the insanity defense has officially been abolished, it is often introduced through the back door when determining the accused person’s competence to stand trial, and the question of mental competence is openly invoked at the post-trial stage of determining the punishment.

7 CONCLUSIONS

After this long “tour” through various models of criminal responsibility of the mentally insane, we have returned to the Standard Model as the most cogent one. This model implies the affirmative account of mental illness and is compatible with conceptions 2 and 4 of criminal responsibility (the personalist view of human nature assumed in this book implies, of course, conception 2, based on the idea of free will). The Standard model seems to gain additional support from Peter Strawson’s thesis that the distinction between the responsibility of the sane and that of the insane cannot be abolished because it is “built into” our reactive attitudes (or more precisely: in the way in which our impersonal reactive attitudes function) and cannot easily (or, rather, at all, at least in the long run) be dispensed with. Yet Strawson is less convincing when he talks disparagingly about “the obscure and panicky metaphysics of libertarianism” based on the concept of “contra-causal freedom,” i.e., freedom of will (cf. 1962: 23–25). Contra Strawson, I have argued that the metaphysics of free will is by no means “panicky” but provides a correct view of human nature, a stronger justification of criminal responsibility, and arguably, it is the only view of man which can justify the retributive theory of punishment.