What is punishment?

Frej Klem Thomsen

fkt@dketik.dk

Senior Consultant, Danish Dataethical Council

May 2022 draft

Contents:

What is punishment? 1
1 Introduction 1
2 The six potential conditions of punishment 2
3 What's in a definition? 6
4 Must punishment be in response to wrongdoing? 9
5 Must punishment be of the culpable? 11
6 Must the punisher be an authority? 15
7 Must punishment impose hard consequences? 17
8 Must punishment be intentional? 21
9 Must punishment communicate censure? 25
10 From definitional to justificatory clarity 27

1 Introduction

Punishments, Cesare Beccaria wrote, are "the tangible motives […] enacted against law-breakers [...] to prevent the despotic spirit of every man from resubmerging society's laws into the ancient chaos." (Beccaria 2003 [1764], p.9) The remainder of his classical treatise discusses the justification of punishment, the appropriate forms of punishment, the design of penal institutions, the crimes that ought and ought not be punished, the useful and useless forms of evidence, and many other issues. As to what – *precisely* – punishment is, we get no closer than these eloquent opening remarks. Instead, Beccaria appears to take for granted that we all know of what we speak, when we speak of punishment.

This Enlightenment confidence gave way to modern anxiety in the middle of the 20th century. Contemporary scholars agree that punishment is something that stands in dire need of justification – that it is vitally important exactly why and when punishments are morally permissible and when they are not. It is also widely held that since this is the case, it is crucial to draw the boundary lines for the concept of punishment with some precision. As David Boonin puts it: "[I]f one cannot [identify the properties that make examples cases of punishment], then one cannot satisfactorily determine whether or not a

purported justification of punishment succeeds in justifying punishment or only in justifying something very much like it." (Boonin 2011, p.4)

Defining punishment, however, turns out to be anything but easy, since it involves definitional choices that have prominent and often controversial implications. This chapter explores the ways in which punishment can be defined and the choices required. It argues that contrary to conventional philosophical aspirations there may not be a definition of punishment fit for all purposes, but that we can avoid conceptual confusion and normative malleability by precisely defining the sense of punishment at stake in a given context. It remains important to consider how to define punishment in order to understand and distinguish the diverse ways one might define it and the implications of adopting each.

Immediately below, the chapter introduces what can be called the classical definition of punishment and the difficulties it raises. In order to frame the analysis, the chapter then briefly considers the more abstract topic of what it means to define a concept such as punishment, and what our definitional desiderata might be – that is, what it can mean for one definition of punishment to be superior to another. On the basis of these preliminaries, the chapter considers six central questions that arise when defining punishment:

1) Must punishment be in response to wrongdoing?

2) Must punishment be of the culpable?

3) Must the punisher be an authority?

4) Must punishment impose hard consequences?

5) Must punishment be intentional?

6) Must punishment communicate censure?

For each of these, we shall review arguments for both affirmative and negative responses. A brief final section then summarizes, considers the implications of adopting various definitions, and concludes.

2 The six potential conditions of punishment

What makes defining punishment difficult? A standard dictionary will say roughly that to punish is to make a person suffer because they have done wrong. (E.g. Wiktionary 2022; Hornby 1995) Why, the layperson might well wonder, have generations of scholars expended such efforts thinking about its definition? As a way of motivating and framing the discussion, let us consider a paradigmatic definition of punishment and some of the questions it raises.

H.L.A. Hart presents what is arguably the classical definition of punishment in the scholarly literature when, drawing on work by Antony Flew, he holds that "the standard or central case of punishment [is defined] in terms of five elements": "(i) It must involve pain or other consequences normally considered unpleasant. (ii) It must be for an offence against legal rules. (iii) It must be of an actual or supposed offender for his offence. (iv) It must be intentionally administered by human beings other than the offender. (v) It must be imposed and administered by an authority constituted by a legal system against which the offence is committed."[1] (Hart 2008, p.4-5; cf. Flew 1954).

Although influential, each of the conditions in the Flew-Hart account of punishment has been the subject of debate. Let us separate the conditions, generalize them a little, and review at least a few of the questions they might raise.

The first of the conditions we will label *the hard treatment condition*. Hart's formulation can helpfully be broadened and clarified a little, so that we get the following:

> An act is a punishment of the person punished – the *punishee* – only if it imposes some form of hard consequences on the punishee.

The hard treatment condition must be made still more precise in two respects. First, we must clarify whether imposing consequences means that the punishee actually suffers hard consequences, or whether some weaker modality will suffice, such as Hart's suggestion that they be *ordinarily considered* hard. Second, it must be clarified what it means for a consequence to be hard, i.e. whether this means harmful, unpleasant (as Hart has it) or something else. As regards the first issue, we might suggest that a genuinely harmful consequence should qualify, irrespective of how it is ordinarily considered, and as regards the second, that a genuinely harmful consequence should qualify even if it is not unpleasant. Why not simply say then that the act must harm the punishee? Because, some would argue, there can be cases of harmless or even beneficial punishment, such as the case where sanction leads an offender to reform, and her life goes all-things-considered better for it. We will consider these complications in section seven below.

The second condition we will label *the response condition*, and generalise as follows:

> An act is a punishment only if it is a response to a wrongdoing.

In the context of criminal justice, the wrongdoing may be qualified as a legal offence *qua* Hart's definition, but if punishment can occur in other contexts other forms of wrongdoing may qualify. The crucial issues here are first and foremost what is required for an act to be a response to wrongdoing and second why

[1]Is Hart's definition strictly speaking a definition? The form makes it appear to be, but Hart is cagey enough to say only that it pertains to the "standard or central case", which leaves open the possibility that there are non-standard cases of punishment, which do not meet the criteria. If there are such cases, then it is not a definition of punishment *per se*, but only of a particular subset of cases of punishment, although perhaps the subset of particular interest to the criminal justice theorist.

punishment must be a response to a wrongdoing? As for the first, it seems clear that the act must be in some way causally related to the wrongdoing, but we shall have to say something more precise. As for the latter, it is worth considering whether there could be cases of punishment for acts that are not wrongdoings, or where the response precedes the wrongdoing. We will review these issues in section four below.

The third condition we will label *the culpability condition*:

> An act is a punishment only if the punishee is or is supposed by the punisher to be morally responsible for the wrongdoing at stake.

The most obvious question to ask of this condition is whether our definition should include both actual and supposed responsibility for doing wrong? If what matters is that the punisher believes the punishee to be responsible for wrongdoing, then actual wrongdoing is irrelevant, and vice versa. Why think that both actual and supposed wrongdoing can suffice for punishment, but neither is necessary? A more fundamental question is whether punishment must be of a person in some way responsible for wrongdoing? Can we imagine cases where punishment is in response to wrongdoing but of a person neither actually nor supposedly responsible for the wrongdoing? We consider these possibilities in section five below.

The fourth condition we will label *the intentionality condition*:

> An act is a punishment only if it is intended.

Intentionality is a complex concept, and it is therefore unsurprising that this condition raises many questions. Hart's uncharacteristically ambiguous phrasing signals the difficulty – what about the act *exactly* is supposed to be intentional? Is it merely the act itself, in which case almost anything done outside of sleepwalking will qualify? Must something else, such as the hard treatment also be intended? Finally, why must punishment be intentional in any sense? We will consider these possibilities in section eight below.

It is worth noting that Hart's formulation builds into the condition two substantial further requirements that are not included in the intentionality condition as stated here. The first is that the punisher be different from the punishee. That is, Hart wants to rule out the possibility of an agent inflicting punishment on herself. Call this *the separation condition*:

> An act is a punishment only if the agent (the would-be punisher) is not identical to the person subject to the act (the would-be punishee).

It seems doubtful that separation is genuinely a condition of punishment. Consider:

> **Exile.** A small community punishes certain taboo acts with exile. Having sentenced a number of community members in this way, the elderly magistrate is found to have accidentally violated a taboo herself. In order to preserve respect for the community's institutions and taboos, the magistrate presides over a trial of herself, at which she sentences herself to exile.

It seems reasonable to say that the magistrate in *Exile* punishes herself. If that is the case, then the identity of the punisher and the punishee need not be different, even if they typically will be.[2]

Hart's formulation also includes what we can call *the human agent condition*:

> An act is a punishment only if the agent is a human being.

This is a strange requirement. Presumably, Hart would agree that e.g. human-like aliens, such as the Klingon and Vulcans of the Star Trek universe, would be capable of punishment. It seems more reasonable to require just that the punisher be a person. At this point however, we might ask whether the additional condition is necessary? On plausible interpretations of intentionality, punishment will be restricted to acts carried out by persons simply because it is one of the defining characteristics of such agents that they are capable of acting intentionally. And conversely, if it turned out that there were forms of punishment that did not require intentions, it seems it might well be true of such punishment that it need not be carried out by persons. As such, the hunan agent condition appears to be either implausible or superfluous, and we shall restrict our attention to the intentionality condition.

The fifth condition we will label *the authority condition*:

> An act is a punishment only if the agent is an authority in the relevant context.

We say "an authority in the relevant context" because authority is contextual, and it clearly will not suffice for the agent to be an authority in an unrelated context. It remains open questions for this condition both what it means for the agent to be an authority in the relevant context, and whether we ought to accept the condition. Can we, for example, conceive of punishment carried out by an agent who is not an authority, in the relevant and perhaps no other context? We consider these questions in section six below.

In addition to the five conditions in the classical Flew-Hart definition, we will consider a sixth potential condition. Some contemporary scholars have argued that it is an essential part of punishment that it communicates or expresses something towards the punishee. In Alec Walen's formulation, it is a condition of punishment that "the hard treatment [is] imposed, at least in part, as a way of sending a message of condemnation or censure." (Walen 2021, section 2.1) Call this *the censure condition*:

[2] In fairness, and as previously noted, Hart explicitly defines only the *typical* case of punishment; his fourth condition's inability to apply to atypical cases might therefore be more of a feature and less of a bug.

An act is a punishment only if it communicates censure of the punishee.

One question pertaining to the condition is how to understand the necessary communication of censure. The answer to that question might in turn hinge on what we decide with respect to other conditions, centrally the intentionality and culpability conditions. A second, and more fundamental question, is of course whether we should adopt the condition at all? Can we conceive of punishment that does not communicate censure, perhaps even punishment that is not intended to communicate censure? An important further consideration in that context is what role, if any, the alleged moral attractiveness of the censure condition should play, i.e. the fact that, as some think, communication of censure is morally relevant, perhaps even crucial, to the ethical justification of punishment? We will consider all of these issues in section nine below.

These, then, are the six conditions we will consider in this chapter. However, before we begin our analysis proper, it is worth devoting a moment to readying the tools for the task.

3 What's in a definition?

As we have already seen, defining punishment is no simple task. This difficulty is not a feature particular to the concept of punishment. Conceptual engineering often feels like a particularly unsatisfying game of whack-a-mole, where each attempt at plugging the gaps in an apparently problematic set of definitional conditions reveals two new holes, smug little counter-example rodents metaphorically smirking. In such situations, philosophers are apt to reach for perhaps the most fundamental tool in the philosophical toolbox: abstraction.

Abstraction, in this context, means taking an analytical step back and asking more fundamental questions about the task at hand. Rather than tackling the question "what is punishment?" head-on, one begins by considering what it is to ask what punishment is. Are the difficulties noted above inescapable, for example, as flaws inherent to the task of conceptual analysis, or merely obstacles in a process of gradual refinement? What, for that matter, does it mean to define punishment, and what might be the point of doing so? Answers to such questions form elements of an account of definitions and defining applied to the particular context.

Definitions are themselves definable as sets of individually necessary and jointly sufficient conditions for something to be an example of the concept defined. Definitions generally take clarity and concision as desiderata – a definition is better the easier it is to understand, and the shorter it is – but beyond these common features there are importantly different types of definitions. Broadly speaking, we can distinguish three approaches to defining that serve different purposes: stipulation, lexical fit, and explication. Bearing their differences in mind can help focus the discussion of what punishment is.

Consider first *stipulative* definitions. A stipulative definition serves to concisely state the meaning that the author invests a term with in a particular context. As such, the definition need make no claim to fit. Obviously, deviating too drastically from the meaning the concept takes in other contexts will typically render the stipulation pointless. There would be little point for criminal justice ethicists in discussing 'punishment' in the stipulated sense of "any ice cream of a creamy-white colour". On the other hand, stipulated definitions suitably close to our ordinary understanding can serve to clarify the starting point of the analysis while bracketing definitional difficulties or disagreements.

The second, and perhaps the most common approach to defining a concept such as punishment is to aim for *lexical fit*. This approach combines a particular method with a particular target. The method is the gradual refinement of the definition's set of necessary and jointly sufficient conditions through consideration of counterexamples. The target is "folk intuitions", that is, a widely shared understanding of the concept at stake. (Hansson 2006; cf. Boonin 2011, p.4-5) Defining punishment, on this approach, means to clarify and concisely state what ordinary, competent speakers have in mind when they employ the concept of punishment; the definition succeeds to the extent that it adequately captures this understanding.

The debate on how to define punishment has largely proceeded under the assumptions that a widely shared understanding of the concept can be captured by a set of necessary and jointly sufficient conditions. Arguably, however, this is not the case for all concepts. Some hold that a classical lexical definition can be unobtainable if our shared understanding takes the form of a set of only partially overlapping properties, as in Wittgenstein's famous notion of family resemblance concepts. (Wittgenstein 1991, §66-67) In these cases there is no set of necessary and jointly sufficient conditions that fits how we – the community of competent speakers – understand the concept. One potential solution is to revise the method, e.g. by defining a set of prototypical properties that tend to make something an example of the concept. (Hampton 2006)

A change of target need not be motivated by the perceived failure of a classical definition, however. An *explicative* definition can set a different target, perhaps in an attempt at revising and improving upon established understandings of the concept. In the classical example, defining 'water' as an inorganic chemical compound of two hydrogen atoms and an oxygen atom (i.e. H_2O) would clearly not fit a widely shared understanding prior to the 18th century chemical revolution. And yet, the reference of the term might well have been the same. (Putnam 1973) If asked to identify examples of the substance water, someone living in the 17th century would presumably point to more-or-less the same phenomena as we would today.[3] An *explicative* definition might therefore take not the shared understanding of a concept

[3]There are very deep and complex disagreements in metaphysics and philosophy of language about how to understand the relation of terms, concepts and the world, including about the theory of semantic externalism attributed to Saul Kripke and Hilary Putnam in the context of which the Twin Earth H_2O example is introduced. Exploring these disagreements

but the reference of a term as its target and fit the definition to this in an attempt at isolating relevant properties of the referenced phenomenon that uniquely identifies it. (Rosen 2015) An explicative definition might, for example, be motivated by the way our shared understanding glosses over a significant difference and redraw the boundary lines around the concept to follow this distinction. Why *relevant* properties? Because there are often many different sets of properties that uniquely identify any given phenomenon, most of which do not constitute useful definitions. In yet another classical example, "featherless biped" is not our shared understanding of the concept of a human being and therefore a poor lexical definition (nor was it the shared concept, presumably, for 4th century BCE Athenians, not even with the addition of "with broad, flat nails"). (Läertius, VI 40) But it is not a satisfactory explicative definition either, because even if these properties did uniquely identify the reference of the term 'human', they are not properties we take to be *significantly* human. In criminal justice ethics, an explicative definition might set as its target not our shared understanding, but the practices we recognize as proper referents for the term 'punishment'. It might also include and/or exclude certain conditions to serve a particular purpose significant to the definer, e.g. to better track a morally significant distinction.[4]

Some might worry that revising the definition of punishment to suit our needs risks creating merely apparent solutions to theoretical problems in criminal justice ethics through the use of "definitional stops". (Hart 2008, p.6) If, for example, we define punishment to require that the punishee be guilty then it might seem that we need not worry about the ethics of punishing the innocent – by definition, there can be no such thing. (Cf. Scheid 1980, p.456) As Hart emphasizes, however, it would be a mistake to think that the ethical problem disappears. If we adopted this definition, we would still need to consider the ethics of doing something very similar to punishment to innocent persons. Proponents of restricting the concept of punishment to the guilty might argue, however, that precisely because the distinction is morally significant, it is useful to separate discussion of one from discussion of the other, and that building the distinction into our concept of punishment will help us to do so (we return to this argument in section five below).

In summary, there are several importantly different approaches to defining punishment. Since the approaches serve different purposes, they will be suitable for different contexts. But it will be useful in any attempt at defining punishment to recognize the different approaches, and to be explicit about both the type and purpose of any particular definition. And it will be useful in the following to bear the different approaches in mind when we review arguments for and against defining punishment in different ways.

would take us far beyond the scope of this paper, however, and I believe the more modest points made here should be acceptable (with minor qualifications) even to those who reject semantic externalism.

[4] Cf. Wringe 2019, arguing that there is an important difference between fitting a definition to folk intuitions of a concept and fitting a definition to our best understanding of a phenomenon.

4 Must punishment be in response to wrongdoing?

Let us begin by reviewing perhaps the most fundamental condition of punishment:

> An act is a punishment only if it is a response to a wrongdoing.

Bear in mind that we set aside, as a separate potential condition, the issue of whether punishment must be of the wrongdoer (see section five below). The issue at stake here is simply whether punishment must be in response to wrongdoing. Could there, for example, be punishment when no relevant wrongdoing has occurred? Consider:

> **Sisters**. Habitual troublemaker Pea is caught violating school regulations. Schoolmaster administers 20 strikes with the cane while scolding her for her moral deficiencies. The following year, Pea's sister Pod enters the school. Schoolmaster, sensing that she has a similarly willful character, again administers 20 strikes with the cane while scolding Pod for her moral deficiencies.

Let us suppose for the purposes of illustration that the treatment of Pea meets the conditions of punishment whatever they ultimately turn out to be, e.g. schoolmaster is an authority, she expresses censure, etc. It seems clear then that Pea is punished.[5] The pertinent question is this: can we call the treatment of Pod punishment? Intuitively, it seems strange to say so – whatever the schoolmaster is doing to Pod, it is not punishment – and one obvious explanation is that the schoolmaster is not responding to a wrongdoing.

To properly assess the condition, however, we will need to clarify what it means for punishment to be a response to something. A simple suggestion would be *the causal response condition*:

> An act is a punishment only if the act is caused by a wrongdoing.

Will that work? No – the causal response condition is much too broad. Consider:

> **Downstream**. A judge sentences a person to life-time imprisonment. The sentence is not caused by any recent offence committed. However, several hundred years ago one of the judge's forebears was born as a result of rape.

The sentence in *Downstream* is, of course, caused by the wrongdoing of rape – without that wrongdoing the judge would not have existed, and could not have sentenced anyone – but intuitively the causal link between sentence and wrongdoing is unsatisfactory.[6]

---

[5] *Sisters* is also intended to be a case where both acts are intuitively wrongful, in order to reduce or avoid the impact of a moral difference between the two acts on our intuitions about whether the acts constitute punishment.

[6] This follows straightforwardly if we assume (plausibly) that personal identity depends on genetic background conditions. (Parfit 1984) The judge's maternal forebear might well have had a different child instead, had no rape occurred, but that

Suppose we could solve the above issue by specifying a more appropriate causal link between wrongdoing and punishment. Some might raise a different objection, that the causal response condition is overly demanding. Consider:

> **Gone Girlish.** Amy fakes her own death and plants evidence incriminating her husband Nick. Nick is convicted for her murder and incarcerated.[7]

Has Nick been punished? Here I suspect that intuitions will both differ and waver, but those willing to say that he has must hold that punishment does not require that the wrongdoing at stake has actually been committed. What punishment might seem to require instead is merely that the act be in response to a supposed offence, specifically (i) that the punisher *believes* that an offence has been committed, and (ii) her act is (or is believed by the punisher to be) a response to this offence. (cf. Boonin 2011, p.19) We can label this the *past motivated response condition*:

> An act is a punishment only if the punisher is motivated to perform the act by a wrongdoing that she supposes has been committed.

It would be good to further specify the response condition, i.e. what it means for the past wrongdoing to motivate the act. Roughly, we might say e.g. that it means the punisher takes the past wrongdoing as generating a necessary reason to perform the act. Here, however, I want to focus on a further complication. Specifically, we may want to broaden the scope of the response condition to allow responding to future events. Consider:

> **Pre-crime.** Authorities have excellent grounds to believe that Citizen will commit a serious offence in the near future. Authorities respond not merely by preventing Citizen from being able to commit the offence but by imprisoning her for many years for her would-be offence.[8]

It does not sound particularly odd to say that Citizen is punished in *Pre-crime*. It is also worth noting that we talk about responses in this way in other contexts, as when we respond to learning from the forecast that it is going to rain by cancelling a planned picnic. Perhaps we should therefore adopt the simpler *motivated response condition*:

---

child would not have been the judge's forebear, and would itself have had different children, grandchildren, etc. Furthermore, although the case might then have been tried by a different judge, who could have passed the same sentence, the past wrongdoing remains a cause of the actual act of sentencing.

[7] "-ish" because this is not quite the plot of David Fincher's "Gone Girl"-film, which sees Nick's name cleared prior to imprisonment.

[8] The sci-fi version of this case, from which it takes its name, is Steven Spielberg's 2002 film "Minority Report". Real life scenarios that come close are punishment for the planning of serious crimes, such as kidnappings and terrorist attacks, which are arguably best understood as punishment not merely for the act of planning, but also for the wrongdoings that would have been carried out had offenders not been prevented from doing so.

> An act is a punishment only if the punisher is motivated to perform the act by a supposed wrongdoing.

This version of the condition allows us to accommodate cases like *Pre-crime*. Some will likely still find the condition overly broad, but at least some of these concerns may be resolved when we consider the culpability condition below, which holds (as stated initially) that an act is a punishment only if the punishee is morally responsible for doing wrong.

Before we do so, however, it is worth addressing one last issue for the response condition. All of the above considers the response condition from the point of view of lexical fit, i.e. a definition that intuitively fits what most people understand punishment to mean. Would the condition look different if considered as part of an explicative definition? Is there any reason, e.g. in the shape of a morally significant distinction, to specify the condition in a particular way?

As is likely apparent, the condition touches upon perhaps the most deep-seated division in criminal justice ethics: the difference between forward- and backwards-looking justifications of punishment. Very broadly, forward-looking justifications rely upon reasons in favour of punishment to do with the positive effects of punishing, while backwards-looking justifications rely upon reasons in favour of punishment to do with the actual or counterfactual acts of the punishee prior to punishment. Specifically, mainstream backward-looking justifications can only apply if punishment is a response to wrongdoing. (e.g. Moore 2010; Tadros 2011; von Hirsch 2017; see also chapter(s) xx of this volume) Suppose now someone argued as follows: if one thinks that punishment can mainly or exclusively be justified by backward-looking considerations, then one ought to define punishment such that it is necessarily a response to a wrongdoing, in order for the definition to track this moral distinction. Should we accept this argument?

Not necessarily. After all, how we specify the condition does not in any way affect the moral reasons at stake. Suppose for the sake of argument that backwards-looking justifications are necessary to justify punishment. If we adopt the response condition we will then distinguish between punishment, which will by definition be capable of being justified, and acts similar to punishment in all respects except their failure to satisfy the response condition, which cannot be justified. If we do not adopt the response condition, then we will simply distinguish between justified punishment and punishment that is unjustified for the specific reason of not being a response to wrongdoing (as well, of course, as punishments that are unjustified for other reasons and non-punishment). It is not obvious that the former way of conceptualizing punishment is clearer or more helpful than the latter.

5 Must punishment be of the culpable?

Having reviewed the response condition, let us look next at the closely related culpability condition. Recall that in the Hart-inspired version, this holds that:

> An act is a punishment only if the punishee is or is supposed by the punisher to be morally responsible for the wrongdoing at stake.

There are two obvious questions we can ask of the condition. The first is whether we should accept the biconditional. Why not simply require, as in the response condition, that the punishee is supposed to be morally responsible? And if there are cases where mere supposed moral responsibility is insufficient, then why not require that the punishee be actually morally responsible? The second question is whether culpability – actual or supposed – is genuinely a requirement of punishment? Can we imagine cases of punishment, where the punishee is in no way morally responsible for the wrongdoing at stake?

When we have considered these questions, we will briefly revisit the issue of what impact the specification of the culpability condition has on the justifiability of punishment. We touched upon this previously when discussing the "definitional stop", but the points are worth reiterating here to connect them with the analysis of the condition.

To review the biconditional we will need to compare cases where the former biconditional obtains without the latter obtaining with cases where the latter obtains without the former obtaining. Let us begin with a case of actual but not supposed moral responsibility. Consider:

> **JTP**. Gotham City is plagued by a spree of crimes committed by a new, mysterious villain calling herself "Gettier". Mob boss Falcone takes advantage of the situation by enlisting a friendly judge to frame and convict his business rival, Rachel Eitteg, for Gettier's crimes. Eitteg is sentenced to imprisonment. Unbeknownst to everyone else, Eitteg is actually Gettier.[9]

JTP satisfies the response condition. It also satisfies the culpability condition since Eitteg is *actually* morally responsible for the wrongdoing at stake, even if the punisher (Falcone and/or the judge) does not suppose that she is morally responsible. Set aside the question of whether Eitteg's imprisonment is morally justified – one might reasonably doubt this – the apposite question is whether Eitteg is punished?

If we say no, then the biconditional version of the culpability condition is mistaken; actual moral responsibility cannot suffice for punishment. Perhaps supposed moral responsibility is also required, or perhaps only supposed moral responsibility is necessary. If we say yes, then one of three things are true: either i) only actual moral responsibility is necessary, or ii) the biconditional culpability condition is correct, or iii) no moral responsibility, actual or supposed, is necessary for punishment. To further explore the issue, we will need to look at a case that does the opposite of *JTP*, i.e. one that involves supposed responsibility without actual responsibility. Consider:

[9] The case, of course, draws inspiration from the somewhat similar and justly famous cases discussed by Edmund Gettier in his celebrated 1963 article. (Gettier 1963)

> **Fugitive**. A doctor witnesses the murder of her partner at the hands of a mysterious one-armed man. Unfortunately, all evidence points towards the doctor. Although she attempts to evade police and apprehend the real offender, she is arrested, convicted, and sentenced.

Fugitive is a classical case of wrongful conviction. Note that it satisfies both the response condition and the latter half of the culpability biconditional, i.e. supposed moral responsibility. As such, it seems we should say that the doctor is punished. Is that right? If we say no, then as above supposed moral responsibility is not by itself enough to fill out the set of joint conditions for punishment. If the answer in *JTP* was also no, then one likely wants to replace the original version with *the complete culpability condition*:

> An act is a punishment only if the punishee both is *and* is supposed by the punisher to be morally responsible for the wrongdoing at stake.

If the answer in *Fugitive* is no, but the answer in *JTP* was yes, then one likely prefers an objective culpability condition, according to which only actual moral responsibility matters for punishment.

Suppose instead one says yes – the doctor in *Fugitive* is punished. In that case, if one also answered yes in *JTP*, then either the biconditional is correct or moral responsibility is in no way a condition of punishment. If, on the other hand, one holds that Eitteg is not punished in *JTP*, then one is likely to prefer a subjective culpability condition, which preserves only the requirement that the punishee is supposed to be morally responsible for the wrongdoing at stake.

Clearly, there is already plenty of room for differing intuitions here, but as noted above, if one is willing to say that both *JTP* and *Fugitive* involve punishment, then either the biconditional is correct, or moral responsibility is in no way a condition of punishment. How do we determine whether it is the former or the latter? By reviewing cases where the punishee is neither supposedly nor actually morally responsible.

Two types of case might be thought to show that there can be punishment without moral responsibility: cases of vicarious punishment and cases where an innocent is framed. (cf. Zimmerman 2011, p.2) Consider first:

> **Collective**. A soldier on a training course sneaks out at night to eat at a diner. She is discovered, and the instructor responds by cutting the next day's rations in half for every other member of the soldier's squad.

Are the squad members punished for the wrongdoing of the soldier? I suspect that many, like myself, will have no firm intuition on the matter. However, we should note that at least some of the intuitive pull towards answering in the affirmative might be explainable by the feeling that there is *some* form of punishment involved in *Collective*. The squad members might be punished, at least in part, for failing to prevent the soldier from sneaking out rather than for her sneaking out, if they are collectively responsible

for enforcing the training regulations. Similarly, the squad members will presumably resent the soldier for causing them to lose a half-day's rations and treat her accordingly, in which case imposing costs on the squad members is a way of indirectly punishing the soldier. Although neither of these are vicarious punishments, it is probably difficult to isolate our intuition about the presence of vicarious punishment from such factors.

Perhaps we can get a clearer picture from cases where an innocent person is framed. Consider:

> **McCloskey**. A brutal crime in a small town incites civil unrest that will rapidly escalate to race riots if the offender is not apprehended. The sheriff, unable to find the offender and desperate to prevent the riots, collaborates with the local judge to frame, convict, and sentence an innocent person for the crime.[10]

Note that, unlike *Gone Girlish* discussed above, *McCloskey* satisfies the response condition. The sentence is a response to an actual wrongdoing. This illustrates how the response and culpability conditions can potentially come apart. Furthermore, unlike *Fugitive*, in *McCloskey* the victim is not even supposedly morally responsible, and the case can thus serve to illuminate the question of whether there can be punishment without culpability. Is *McCloskey* a case of punishment, or only of something similar, e.g. simulated punishment? If one holds that it is a case of punishment, then one must reject the culpability condition altogether. Again, I suspect that many will have no firm intuition on the matter, but it bears mentioning that if cases like *McCloskey* are to serve the traditional purpose of grounding an objection to the utilitarian theory of justified punishment, it must be assumed that the cases involve punishment. It would appear, therefore, that at least some scholars have implicitly rejected the culpability condition. (Cf. Scheid 1980, p.459).

If, as it seems to me, there are conflicting and uncertain intuitions on the culpability condition, as illustrated by the cases we have explored above, then there is not a definition of punishment that provides ideal lexical fit. There is rather a certain vagueness and/or disagreement about our shared concept of punishment. Where does that leave us? More specifically, might there be a particular explicative definition worth adopting?

We briefly touched upon this issue when we initially considered the use of "definitional stops". The charge, recall, is that including a culpability condition means that certain intuitively morally problematic cases, such as *McCloskey*, cannot be labelled punishment. If friends of utilitarianism were to respond to the critique grounded in such cases by relying on the claim that these are not cases of punishment, then they would be using a definitional stop. This is justifiably considered a flawed response. It does nothing

[10] This is of course a version of a case famously employed against utilitarian justifications of punishment. (McCloskey 1965) We set aside here the question of what the case might show as regards justifications for punishment, in order to consider only what it might show about how to define punishment in the first place.

to address the substantial point of the critique, which is the claim that consequentialist theories entail that intuitively morally impermissible acts are morally permissible. Whether they happen to be acts of punishment or of something very similar to punishment is in that context irrelevant.[11]

This observation about definitional stops generalizes. The much-discussed problem of punishment is arguably best understood as the question of whether punishment can ever be morally justified, even when it pertains to a person who both is and is supposed by the punisher to be morally responsible for wrongdoing. But even if we defined punishment so narrowly as to include only these cases, we must still contend with the related problem that any system that metes out punishment risks treating and will in fact occasionally treat persons who are *not* actually and/or supposedly morally responsible for wrongdoing in similar ways. (Lippke 2010) If we drop the culpability condition, we can call this the problem of punishing the innocent and distinguish it from the problem of punishing the guilty. But the moral issues do not change, regardless of our definition. Ultimately, therefore, reasons to favour one or the other definition may come down to low-grade pragmatic considerations, such as the fact that it will in a particular context be easiest, e.g. because less convoluted, to discuss an issue using a particular definition.

6 Must the punisher be an authority?

Let us consider now *the authority condition*:

> An act is a punishment only if the agent is an authority in the relevant context.

The condition has some intuitive plausibility. After all, it is certainly true of the paradigmatic examples of punishment that the agent is an authority, specifically a judge in a criminal court is a legal authority invested with powers by the state.

It is worth noting, however, that we regularly speak of punishment in situations that are very different from the paradigmatic example of an offender sentenced to imprisonment by a legal authority. Misbehaving children are punished by their parents, negligent employees are punished by their bosses, and foul-playing athletes are punished by the referee. It is not senseless nor even strange to say that someone punishes someone else in such cases, even if the punisher does not have what we understand as conventional legal authority.

Still, punishment might seem to require *some* form of authority. Parents are authorities relative to their children, bosses to their employees, and referees to competing athletes. Upon consideration, however, this apparent requirement may simply reflect the fact that an imbalance of power is typically a practical prerequisite for punishment. It will ordinarily be difficult to punish someone unless one is an authority,

[11] There are, of course, more promising ways for friends of consequentialism to respond. See e.g. Sprigge 1965; Lyons 1974; Wennberg 1975; see also chapter(s) xx in this volume.

simply because one lacks the physical or psychological means to impose hard consequences on the (ordinarily unwilling) punishee. Are there exceptions to this tendency? Consider:

> **Disgruntled**. A child responds to perceived unfair treatment at the hands of a parent with icy contempt, withholding customary affection, refusing to speak to them or even acknowledge their existence for a week. (Cf. Scheid 1980, p.457)

Certainly, the child might take herself to be punishing the parent, and it does not sound particularly strange to say that she does so. As a matter of lexical fit, therefore, it does not seem obvious that punishment broadly speaking requires authority.

Nonetheless, some may think that specifically *legal* or *criminal* punishment, at least, requires that the punisher possesses a certain form of authority, specifically legal authority granted by the state to impose sanctions for legal offenses. Recalling Hart's definition, let us consider *the legal authority condition*:

> An act is a *criminal* punishment only if the agent is a relevant *legal* authority.

Should we accept the legal authority condition? Perhaps not. Arguably, there are counterexamples such as the following:

> **Lynching**. A convicted offender awaits sentencing for a heinous and very public crime in a small town. One evening, a group of enraged citizens break into the jail, carry off the offender, and hang her in the public square. (cf. Zimmerman p.2)

We set aside here the question of how the actions of the mob differ morally from a court that inflicts capital punishment. The question initially at stake is merely whether it seems odd to say that the lynching is a form of criminal punishment. Offhand, it is not incomprehensible nor even particularly odd to say that the mob punishes the offender. Since that is the case, even criminal punishment may not require authority, legal or otherwise.

Are there other reasons to find the legal authority condition attractive? One suggestion might be that there are morally relevant differences between punishment with and without legal authority, and that these speak in favour of including the legal authority condition. Such a definition would be deliberately revisionist – it would not claim to perfectly fit the meaning we generally ascribe to the term 'criminal punishment', but rather to reserve the term for a particular set of actions, which share morally relevant properties that set them apart from some of the other things we might call punishment. Whether that would make the revision all-things-considered desirable is an open question.

7 Must punishment impose hard consequences?

In the above, we have considered the response, culpability, and authority conditions. Complex as these may be, matters are more tangled still in this section, when we explore the hard treatment condition:

> An act is a punishment of the punishee only if it imposes some form of hard consequences on the punishee.

The fact that a punishment must in some way be hard on the punishee is one of the most prominent features of punishment. The generic statement leaves unclear exactly what this means, however, and it is surprisingly difficult and highly controversial how to make the condition precise. Specifically, there is disagreement on whether consequences must be hard generally speaking, in fact, in expectation, or in intention, as well as on whether being hard on the punishee should mean unpleasant, harmful, or something else.

Let us consider the Hartian approach. Hart, recall, required that consequences be "ordinarily considered unpleasant". (Cf. Wringe 2013) Note first that on the Hartian approach we would need to make precise what it means for something to be "ordinarily considered" one way or another. Considered by who? And by how large a fraction of that group for it to qualify as being *ordinarily* considered in any particular way? Presumably, the answer to these questions would be something like "considered by a majority of persons in the society in which the would-be punishment is imposed", but any such answer would need to be detailed and defended. I say "would be" because Hart's version of the condition faces far more serious difficulties, and we will focus on these.

To illustrate the more serious difficulties, consider:

> **Tourist 01.** A visitor from Singerland is convicted of a minor offence during her stay in traditional farming society Scrutonville. She is sentenced to mandatory participation in the annual harvest-festival, at which a great number of animals are ritually and publicly killed, cooked, and shared by all. Singerland is a progressive, vegan society.[12]

Suppose that the sentence meets whatever other conditions we ultimately believe that punishment must meet, e.g. the act is intentional, communicates censure, etc. Suppose also that participating in the harvest festival, including partaking in the killing, cooking and eating of animals, far from being ordinarily considered unpleasant is ordinarily considered enjoyable in Scrutonville. Finally, suppose plausibly that for the progressive vegan tourist, participating in the killing, cooking and eating of animals is extremely unpleasant.

[12]Peter Singer is, of course, the most famous philosophical defender of animal rights, while Roger Scruton is one of the more notorious philosophical defenders of speciesism. (See e.g. Singer, 1975; Scruton, 1996)

On Hart's account, there is no punishment in *Tourist 01*. The problem, of course, is that intuitively the sentence *does* seem to constitute a punishment, albeit a very unusual and perhaps a fairly mild one, because it is unpleasant for the tourist. A similar difficulty applies to cases where something is ordinarily considered unpleasant but is not experienced as such. (Cf. Hanna 2017) Consider:

> **Tourist 02.** A visitor from Scrutonville is convicted of a minor offence during her stay in Singerland. She is sentenced to a mandatory two-week meat-rich diet.[13]

Again, we can suppose both that the diet would be ordinarily considered unpleasant by Singerland citizens, and that the Scrutonville tourist will enjoy it rather more than the vegan fare otherwise available. In this case, it seems bizarre to insist that the diet constitutes punishment, as we must if we accept Hart's version of the hard treatment condition.

The two cases suggest that a consequence's being ordinarily considered unpleasant is neither necessary (*Tourist 01*) nor sufficient (*Tourist 02*, other conditions being met) for punishment. Given these apparent difficulties, it is tempting to ask why it ought to matter how something *ordinarily* works or whether it is *considered* unpleasant? Why not simply require that the consequences *actually are* unpleasant?

Suppose we adopt a full-blooded actualist interpretation of the hard treatment condition:

> An act is a punishment of the punishee only if it imposes consequences that are actually hard on the punishee.

The actualist interpretation might seem both obvious and attractively simple, but it does have implications that some find problematic. Consider:

> **Home sweet prison**. An offender is sentenced to life-time imprisonment without parole. In prison, she enjoys the physical protection of the guards, receives regular meals, clean clothes, and high-quality health care upon need. She has access to a warm, dry, and comfortable place to sleep, toilet and bathing facilities, as well as decent recreational facilities. All these circumstances compare favourably with her civilian life as a destitute homeless person.

Clearly, many actual prisons are nothing like this case – indeed, the number that are like it may be lamentably low – and perhaps there are or could be societies where even the most desperate are well-off compared to the prisoners in *Home sweet prison*. Nonetheless, it seems entirely possible to imagine that in

[13] Hanna employs a similar case, "Judgment", to criticize Wringe's Hart-like account of hard treatment. However, Hanna takes the case to show that punishment requires intention to harm. *Tourist 02* is deliberately agnostic on the judge's intentions, but the intuitive weirdness remains if we imagine that she intended to harm the tourist. This suggests that a better explanation in both cases is that the tourist is not actually harmed. (Cf. Wringe 2019)

this case, imprisonment could be an improvement for the homeless person. On the actualist interpretation, we therefore cannot say that the homeless person is punished.

If one wants to avoid saying that cases like *Home sweet prison* involve no punishment, one can either move from the actualist interpretation back towards some weaker modality, such as risking or intending hard consequences, or tweak the interpretation of what it means for consequences to be hard. The former solution might say, for instance, that *Home sweet prison* involves punishment because the judge (we can suppose) intends imprisonment to constitute a hard consequence, or because imprisonment is generally a hard consequence. (See Wringe 2013 for discussion of this approach) The trouble for such solutions is that they are apt to reintroduce the difficulties encountered by Hart's version of the condition. For any such interpretation of the condition, there are likely to be both a) cases where the treatment does not meet the condition, that we nonetheless intuitively want to label as punishment because the consequences actually *are* hard, and b) cases where the treatment meets the condition, that we intuitively do not want to label as punishment because the consequences are *not* actually hard.

Suppose instead of changing the interpretation of what it means to impose consequences, we change our interpretation of what it means for them to be hard. We might, for example, require that the punishee suffers:

> An act is a punishment of the punishee only if it imposes suffering on the punishee.

This avoids the *Home sweet prison* challenge, since plausibly the offender is not there made to suffer. However, any solution that relies in this way directly on the experiences of the punishee runs straightforwardly into a different difficulty: the inability to allow punishment in the form of unexperienced harms. Consider:

> **Surprise.** An offender is sentenced to death. The offender is not informed, and the execution is carried out by sedating and then killing her while she sleeps. (cf. Zimmerman 2011, p.3; Boonin 2011, p.6)

There would be several fairly obvious ethical challenges for a criminal justice system that operated as in *Surprise*. The point here, however, is merely that although the offender in no way suffers – not even, we can suppose, fear of impending death – we intuitively want to say that she is punished. An obvious explanation is that she is harmed by the act, because her painless death deprives her of the value her life would have contained had she continued to live.

Should we require that punishment harms the punishee, then? While that approach is tempting, adopting it brings with it a host of issues both similar to those we have seen above and familiar from the broader debate on what it means to harm a person. (See e.g. Holtug 2002; Søbirk Petersen 2014; Hanna 2016) Suppose, for example, that we adopt a counterfactual whole-life welfarist account of harm, where a

person is harmed by an act if that act decreases her whole-life well-being relative to what it would have been, had the act not been performed. Now consider:

> **Something works**. An aspiring career criminal is sentenced to several years in prison. As a result of her imprisonment she is reformed, and upon leaving prison she works diligently to improve her life situation, such that her life as a whole becomes better than it would have been had she not been imprisoned.[14] (Cf. Zimmerman p.4).

Unlike *Home sweet prison*, we need not assume that there is no suffering in *Something Works*. For the offender, her time in prison may be wholly awful – much worse in every relevant respect than her life outside prison. However, as in *Surprise* there are opportunity costs at stake, only this time they attach to non-treatment: had she *not* been incarcerated, her life would have gone all-things-considered worse. So, on the counterfactual welfarist whole-life account of harm, she is not harmed, and therefore not punished. Some think this is implausible – surely, several years of suffering due to incarceration ought to qualify as punishment, even if this has the laudable and perhaps intended effect of reforming the offender? (Zimmerman 2011, p.6; Boonin 2011, p.7; cf. Adler 1991)

In response, one might vary the account of harm at stake in punishment in several ways.[15] One could adopt a time-comparative baseline instead of the counterfactual baseline, such that a person is harmed if the act reduces her well-being relative to what it was prior to the act, or one could adopt a multi-dimensional account of harm, such that a person can be harmed in one dimension, while being benefited in another or across dimensions. In Nathan Hanna's version, referencing work by Ben Bradley, punishment need only inflict "prima facie" harm, where such harm is understood as bringing about something that is intrinsically bad or the loss of something that is intrinsically good. (Hanna 2014; cf. Bradley 2009; Birks 2021; Boonin 2011, p.7; Zimmerman 2011, pp.4-6)

The possibility of punishment harming a person in one dimension while benefitting her all-things-considered follows straightforwardly from value pluralism, where individuals might lose out in terms of one value but gain in terms of another, but it might be accepted even by value monists. In *Something works* prison depresses the offender's level of wellbeing for the duration of her incarceration relative to what it would have been had she not been imprisoned. However, it increases her level of wellbeing for a large part or perhaps even all of the remainder of her life. The net result is a gain in wellbeing, but this, some might say, does not mean that the lower level of wellbeing during her imprisonment is in no sense a harm.

[14] "What works?" was, of course, the title of Robert Martinson's seminal 1974 critique of criminological reform theory, whose conclusion has often been paraphrased as "nothing works". (Martinson, 1974)

[15] It is worth bearing in mind that this need not commit one to a particular general account of harm. It is possible, in theory at least, to say that for an act to constitute punishment, the act must harm the punishee in a particular sense, while holding that harm more generally means something else.

Is there any way to settle the issue of which of the potential versions of the hard treatment condition we should adopt? Not necessarily one that will satisfy all parties. Preferences for different specifications of the hard treatment condition, and over viable definitions of punishment more broadly, are likely to hinge on one's views on the moral significance of different senses of causing harm. For example, those who attribute moral significance primarily or exclusively to actually causing harm might for that reason find themselves attracted to an actualist account of hard treatment. Meanwhile, those who attribute moral significance to intending or risking harm might find a less demanding version of the condition sufficient.

Moral factors might affect our intuitions about the condition in other ways too. If one feels intuitive unease about concluding that *Home sweet prison* involves no punishment it might be due at least in part to the implications of that conclusion. After all, if one denies that there is punishment in the case, but also wants to preserve the possibility of punishing even offenders as badly off as the homeless person in that case, then it seems one might have to resort to draconian measures. Indeed, one might be required to accept the imposition of ever more brutal treatment the worse off an offender already is, reserving the harshest treatment for society's most unfortunate wretches. This anti-egalitarian implication is likely to strike many as unpalatable.

Similarly, some might worry that the hard treatment condition prejudices the debate against certain theories of justified punishment. Might it not, as Boonin puts it, "beg the question against [a position that takes punishment to be morally permissible because it ultimately benefits the offender"? (Boonin 2011, p.7; cf. Adler 1991) Boonin's answer is to adopt the understanding of harm, where it is sufficient that the punishee be harmed in some specific dimension. Another answer could be that it does not because restricting the concept of punishment to situations where the punishee is harmed all-things-considered in no way affects the substance of the theory at stake. That theory must now be understood as the idea that an act, which is in other respects similar to punishment (with the harm condition), but for that fact that it benefits the punishee, is morally permissible, instead of the idea that punishment (without the harm condition) is morally permissible when it benefits the punishee all-things-considered. Drawing the conceptual lines in the two different ways turns out to only entail some moderate rephrasing. This can still count in favour of defining punishment in a particular way. There will often be pragmatic reasons for drawing conceptual lines so as to allow the clearest, least convoluted way of discussing a problem. At the same time, so long as we keep our distinctions clear, and do not equivocate, the crucial issue of what acts are morally permissible and impermissible is entirely unaffected.

8 Must punishment be intentional?

The fifth condition is the intentionality condition:

> An act is a punishment only if it is intended.

As noted initially, intentionality is a complex concept, and it is unsurprising that the condition has been subject to intense debate. An obvious first question is what notion of intentionality is at stake. Is intentionality in the mere sense of acting consciously and willfully sufficient? That would likely rule out too little. Consider:

> **Tudors**. The queen is presented with two documents, one of which pardons and the other of which condemns her treasonous dynastic rival and cousin. After deliberating, she signs a document and hands it to her ministers to be carried out. The next day she discovers to her horror that she signed the wrong document – her cousin was beheaded at dawn.[16]

Did the queen punish her cousin? Note that it should play no role that she does not carry out the physical act of beheading. Had she known what document she was signing we would not have hesitated to say that she had punished her cousin, regardless of whether she took any further part in the proceedings. Note further that her act of signing the document was certainly intentional in the minimal sense that she was carrying out a conscious, willful act. If we feel uneasy about calling the act punishment it is likely because the act was intended to be a pardon. If we are nonetheless willing to say that the queen punished her cousin, we can stay with the intentionality condition in a broad sense, and we might speak then of inadvertent punishment. If we want to say that there is no punishment in *Tudors* we will need a different interpretation of the condition.

One such interpretation that might seem tempting is the notion that *punishment* must be intended, not the mere act. A moment's reflection will show, however, that doing so will lead to a logically troublesome infinite regress: since the intentionality condition is itself part of what punishment means, the punisher must now intend that she intend to punish, which entails that she must intend that she intend that she intend to punish, etc.

We are likely to do better if we focus on intending particular features of punishment. The most obvious candidate is intending that the act impose hard consequences. Call this *the hard intentionality condition*:

> An act is a punishment only if the agent intends for it to impose hard consequences on the punishee.

The hard intentionality condition resurrects the complications we encountered above when reviewing the hard treatment condition. What precisely are the hard consequences that must be intended? The simple answer would be that the act must intend the hard consequences specified in the hard treatment condition, but it is possible to rely on a different sense of hard consequences. We can imagine, for example, a definition of punishment where the hard treatment condition requires only that the punisher

[16] This is, of course, not quite the story of Elizabeth and Mary Queen of Scots.

impose the risk of harm, while the intentionality condition requires that actual harm be intended (or vice versa).

To flesh out the hard intentionality condition, we must say something about what it means for an agent to intend a consequence (hard or otherwise). Here we can rely on a standard account of intentions in moral philosophy, where (roughly) an agent intends an outcome *iff* the agent acts in order to bring about that outcome, either as a goal in itself, or as the means of bringing about some other end that the agent takes as a goal in itself.[17] (Foot 1967; Duff 1982; McMahan 1994; Kamm 1999; see also Tadros 2011, chapter 7)

Having clarified the condition, let us consider the next question: Is intentionality necessary for punishment? Proponents rely on cases like the following:

> **Quarantine Z**. A large, rowdy group, ignoring the bio-hazard signs, break into and hold an impromptu party at an abandoned research facility. In doing so, they expose themselves to a dangerous pathogen. Infected persons become prone to bursts of irrational frenzy, during which they are liable to attack and kill or infect others. Unfortunately, the incubation period is anywhere from a week to a year, and tests for infection are unreliable. For safety reasons, government quarantines the partygoers in the only suitable facility available: a recently decommissioned maximum-security prison. (Cf. Hanna 2008, pp. 127-128; Zimmerman 2011, pp. 9-10)

Quarantine Z is very like conventional punishment. The quarantine is imposed by an authority. Standard forms of quarantine are not responses to wrongdoing and might fail to qualify as punishment for this reason, rather than because they lack the intention to cause harm. (Wringe 2013) However, recklessly exposing yourself to a dangerous and infectious pathogen is plausibly morally wrong, and in this case the quarantine is a response to this wrongdoing. Finally, being locked up in a maximum-security prison is certainly hard treatment.[18] But are the partygoers punished? If we want to say no, then a possible explanation is that the harm the quarantined persons suffer is unintended. Although quarantine is harmful, this harm is neither the goal of quarantining them, nor a means to that goal. The goal of quarantine is only to prevent harm to others, and any harm the quarantined suffer is merely an unfortunate and foreseen side-effect of the quarantine. In conventional punishment on the other hand,

[17] Note that in order to rely on it here we need not believe that the distinction between intending and merely foreseeing is morally significant, a view which faces very powerful objections. (See for example Kagan 1989, chapter 4; Thomson 1999; McIntyre 2001; Nelkin & Rickless 2015; Steinhoff 2018, 2019; for an overview, see FitzPatrick 2012) For the purposes of the present argument we will set that concern aside, and focus on the more limited question of whether punishment requires intention, including whether punishment would require intention *if* the intending/foreseeing distinction was both meaningful and morally significant.

[18] Quarantine cases might also fail to qualify, as Wringe also notes, because they do not meet the censure condition. We reserve treatment of that condition for the section devoted to it below.

harm might be either the goal of hard treatment (e.g. as deserved retribution) or a means to the goal (e.g. deterrence).

Proponents might claim that this analysis shows a further attraction of the intentionality condition. Some scholars of criminal justice ethics believe that we can do something quite similar to what the penal system currently does without *intending* to harm offenders. (Hanna 2014; Boonin 2011; see also chapter xx in this volume) If punishment requires intent to harm, these penal practices are not punishment. Thus, on this account, even if punishment is morally impermissible, non-punitive practices relatively similar to our current penal practices might remain morally permissible.

Would the fact that there is a morally significant distinction between intending and foreseeing harm give us reason to define punishment in this way, so as to allow distinguishing between impermissible punishment and potentially permissible punishment-like practices? It might, but if so, similar to what we have seen for other conditions, it is a fairly pragmatic reason. After all, we can make the exact same point by defining punishment without the intentionality condition and then distinguishing between (impermissible) punishment that intends to harm the punishee and (potentially permissible) punishment that does not.

The fact that the hard intentionality condition makes the punishment dependent on the mental state of the punisher can also be the basis of arguments against the condition. Pragmatically, if punishment requires intention to harm, then it will often be difficult or even impossible to determine whether a particular act is a punishment or not. (cf. Boonin 2011, p.14-15) We cannot, after all, peek inside the heads of other people to find out why they are doing what they do, and they may be unable or unwilling to inform us of their intentions. It may even, as a century of psychological studies of human cognitive biases and the limits of conscious introspection have taught us, be hard for an agent to tell with any reliability why she herself is acting the way she is. Perhaps more importantly, the dependence on mental states has implausible implications in certain cases. Consider:

> **A tale of two convicts**. Offender A is sentenced to three years in prison by judge C; offender B is sentenced to three years in prison by judge D. A and B are alike in all relevant respects, including having committed equally serious offences, and serve their time in the same prison. However, judges C and D differ in one particular respect. C is aware of the wide range of hardships inmates suffer (social, sexual, and recreative deprivation, fear and the risk of harm at the hands of fellow inmates, etc.), and intends these as part of the punishment. D is aware only that prison limits inmates' liberty by restricting their physical access to the rest of society and intends only this hardship. (cf. Kolber 2012)

If we accept the hard intentionality condition, we are forced to say that A and B are punished very differently, specifically that A's punishment is much more severe than B's. This implication could serve as

the basis of a challenge to the idea that there is a significant moral distinction between (intended) punishment and (unintended) hardships suffered as a consequence of punishment. Here, however, we are concerned only with the implications for our conception of punishment. On that issue, it sounds strange to say that A and B receive different punishments, which suggests that hard intentionality is not a condition of punishment.

9 Must punishment communicate censure?

The final element of a definition that we need to review is the censure condition:

> An act is a punishment only if it communicates censure of the punishee. (cf. Feinberg 1965; Duff 2009; Boonin 2011; Zimmerman 2011; Duff & Hoskins 2017; Walen 2021)

Consider first what it means for an act to communicate censure. Roughly, this must be understood as the act sending a message of moral disapprobation. Notably, it is the act of punishment itself that must send this message. In certain cases, the message will be obvious. Scolding might be said to be a form of punishment that consists mainly in verbally expressing moral disapproval. In many other cases, such as a fine or a prison sentence, the message will be communicated non-verbally. The punisher can emphasize the message by verbally supporting and clarifying it, as a judge will typically do at sentencing, but such verbal emphasis is not a substitute for the censure communicated by the act. This need not be problematic – there are many acts of non-verbal communication, including ones that, like rolling eyes and furrowed brows, communicate disapproval – but it does make punishment highly sensitive to context. The meaning of non-verbal communication depends crucially on the social and cultural setting in which it occurs.

Now, for censure to constitute an independent condition of punishment, it must be possible to meet the other conditions without communicating censure. If, for example, responding to a past wrongdoing by imposing hard treatment on the responsible person necessarily in and of itself communicated censure, the censure condition would add nothing to the definition. Is it possible to meet the other conditions without communicating censure? Consider:

> **Ritual.** In order to be eligible for membership in a gang, prospective members are first required to provably commit a serious offence. The leader of the gang responds by initiating them in a ritual that involves branding them with a hot iron, the pain of which is an integral part of the process. (Cf. Zimmerman 2011, p.17)

The case is meant to meet all conditions of punishment except for the fact that the branding, although harmful, does not communicate censure. Intuitively, *Ritual* does not involve punishment. As such, it would seem that censure is a necessary condition of punishment.

To undermine the intuition generated in cases like *Ritual*, we would have to show either that the act does in fact communicate censure, or (more plausibly) that it fails to satisfy some other condition of punishment. Let us grant that the gang leader is a relevant authority, that the act is in response to a wrongdoing and of the culprit, and that the pain is intended (as a means of initiation). That leaves the hard treatment condition. Is the initiation-through-branding a hard consequence? That depends. If one requires only that hard consequences be in some dimension hard, then the answer is likely yes. If on the other hand one requires that the act be all-things-considered harmful, or that the punisher perceive it as such, then the answer might be no. Gang membership might be, or be perceived to be, so valuable that even with branding the initiation is good for the would-be-punishee. And if that is the case, then perhaps our intuition rejects the case not because it lacks censure, but because it does not satisfy hard treatment.[19]

A further argument against the condition is that it might be counter-intuitive in scenarios where the agent is incapable of communicating censure. Consider:

> **The cake is a lie.** In 2092, hostile super-AI has taken control of the world. The AI sets out rules for human behaviour. Violations of the rules are met with an immediate response in the shape of drone-administered electric shocks, the duration and severity of which correspond to the severity of the transgression.[20] (Cf. Hanna 2017)

The AI, we can stipulate, is sentient and capable of forming intentions, but has no moral sensibilities. Its actions therefore do not express moral disapprobation, and we can stipulate that this is understood by all. If the responses communicate anything, it is the threat that violations of the rules will lead to pain. Does the AI punish rule violaters? If we are willing to say yes, then censure is perhaps a typical but not a necessary component of punishment.

At this point, some might raise a familiar argument: even if censure is not necessary for something to be punishment on our standard conception, it might be necessary for punishment to be morally justified. The prominent family of *expressivist* theories of criminal justice ethics has defended variants of the latter view. (E.g. Hampton 1992; von Hirsch 1993, Duff 2001; Glasgow 2015; Wringe 2016) Might there be a moral distinction between acts that do and acts that fail to communicate censure, and might that distinction be worth building into our definition of punishment, by requiring that an act communicate censure in order to be labelled punishment?

In a by now no doubt familiar response, I would suggest that this is doubtful. Even if there were a significant moral distinction between e.g. criminal sanctions that do and do not communicate censure,

[19] Note that, if gang membership is itself disvaluable, as it realistically might well be, then the branding would be unnecessary for the case. Thus, the case presumably involves branding in order to create a scenario that is supposed to work even if gang membership is valuable.
[20] Deceitful promises of imminent cake are prominently offered by GlaDOS, the AI villain of the 2007 hit computer game "Portal".

which is at least debatable, it is not clear that building that distinction into our definition is advantageous.[21] If we do include it, we will speak of potentially justified punishment versus unjustified acts that are similar but fail to communicate censure. If we do not, then we will speak of justified punishment that communicates censure, and unjustified punishment that does not. It is not clear, and might well vary with context, which of these two ways of defining punishment is preferable.

10 From definitional to justificatory clarity

In the above, we have seen that it is possible to doubt the necessity of most of the conditions in the classical Flew-Hart definition of punishment, as well as the more modern censure condition. In many cases, it seems that there may be reasonable disagreement over whether certain conditions do or do not apply. It is possible, of course, that continued reflection may settle matters, and provide a definitive definition. However, it is also possible that such disagreements reflect the fact that different persons employ slightly different conceptions of punishment. This is perfectly compatible with us all speaking meaningfully with each other about punishment. Successful language games require only that there is sufficient overlap in the way we understand the concepts at stake.

There is also another possibility. Forty odd years ago, Don E. Scheid suggested that punishment is what he labelled a "reducible concept". (Scheid 1980, p.461) Although the standard cases of punishment had certain features, he argued, it seems we can speak sensibly of punishment in cases that do not possess all of the features. Today, we would perhaps speak of punishment as a prototypical concept. (Hampton 2006) On that understanding, although paradigmatic examples of punishment may be acts performed by an authority, in response to a wrongdoing, against a wrongdoer, that are both intended to and actually harm the punishee, and which serve to communicate moral disapproval, there may be non-standard punishment that differs in one or more of these dimensions, which still fall under the scope of the concept.

A more important question: does it matter what definition of punishment best fits our shared understanding? Ultimately, if we wanted to resolve which definition achieves the best lexical fit, and the extent to which this definition is shared across persons and groups, philosophers and criminal justice theorists would need to recruit the sociolinguists to study the matter empirically. Deliberations in the armchair will take us only so far. Would such empirical analysis be worth pursuing? Perhaps not.

A central point in our analysis above has been that substantially less may turn on the definition of punishment than has sometimes been assumed. How we define punishment does not (directly) affect the moral permissibility of any act, it merely determines which acts (permissible and impermissible) can be labelled punishment. If we define punishment expansively, then it is likely to include both permissible

[21] For criticism of the view that the communication of censure can provide moral justification, see e.g. Bagaric & Amarasekara 2000; Hanna 2008)

and impermissible acts, and we will need to distinguish between the two. If we define it narrowly, it may include only permissible or only impermissible acts. But in that case, there are likely to be both permissible and impermissible acts very similar to but not labelled punishment.

What is the point of defining punishment, then? Is half a century of philosophical reflection just so much wasted effort? Hardly – defining punishment is important because it can help us sort out the difficult questions of which acts are and are not morally permissible. Recall Boonin's claim, cited in the beginning of this chapter, that if one cannot define punishment, then one cannot determine whether a justification applies to punishment or only to something similar to it. There is an important point to this claim, but it also overstates the issue in two ways. On the one hand, even on a stipulated definition, it may be interesting to establish whether it is possible to justify punishment in that particular sense so long as we bear in mind that this does not guarantee that punishment is justified in any other sense. On the other hand, a lexical definition may help us determine whether punishment in our shared understanding of the concept is justified but does not by itself resolve whether any particular set of practices are justified or unjustified, because some practices may not fit our shared understanding of the concept of punishment. The important point in Boonin's claim is this: precise definitions will help us recognize these limitations, by clearly delineating the scope of any argument for or against a particular conception of punishment being justified. Only with a definition in mind will we know the range of normative arguments for and against the sense of punishment at stake. That precision is worth striving for.

11 References:

Adler, Jacob (1991). *The Urgings of Conscience: A Theory of Punishment*. Philadelphia: Temple University Press.

Bagaric, Mirko & Amarasekara, Kumar (2000). "The Errors of Retributivism", in: *Melbourne University Law Review* vol. 24 (1): pp.124-189.

Beccaria, Cesare (2003 [1764]). *On Crimes and Punishment and Other Writings*. Cambridge: Cambridge University Press.

Birks, David (2020). "Paternalism as Punishment", in: *Utilitas* vol.33 (1): pp.35-52.

Boonin, David (2011). *The Problem of Punishment*. Cambridge: Cambridge University Press.

Bradley, Ben (2009). *Wellbeing and Death*. Oxford: Oxford University Press.

Duff, Antony (1982). "Intention, Responsibility and Double Effect", in: *The Philosophical Quarterly* vol. 32 (126): pp.1-16.

Duff, Antony (2001). *Punishment, Communication and Community.* Oxford: Oxford University Press.

Duff, Antony (2009). "Can We Punish the Perpetrators of Mass Atrocities?", in: Thomas Brudholm & Thomas Cushman (Eds.) *The Religious in Responses to Mass Atrocity – Interdisciplinary Perspectives*. Cambridge: Cambridge University Press.

Duff, Antony & Hoskins, Zachary (2017). "Legal Punishment", in: Edward N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/legal-punishment/

Feinberg, Joel (1965). "The Expressive Function of Punishment", in; *Monist* vol. 49 (3): pp. 397-423.

Feinberg, Joel (1970). “Justice and Personal Desert”, in: *Doing and Deserving*. Princeton: Princeton University Press.
FitzPatrick, William J. (2012). “The Doctrine of Double Effect: Intention and Permissibility”, in: *Philosophy Compass* vol. 7(3): pp.183–196.
Flew, Antony (1954). “The Justification of Punishment”, in: *Philosophy* vol.29 (111): pp.291-307.
Foot, Philippa (1967). “The Problem of Abortion and the Doctrine of the Double Effect”, in; *Oxford Review* vol.5: pp. 5-15.
Gettier, Edmund L. (1963). “Is Justified True Belief Knowledge?”, in: Analysis vol. 23 (6): pp.121-123.
Glasgow, Joshua (2015). “The Expressivist Theory of Punishment Defended”, in: *Law and Philosophy* vol. 34 (6): pp. 601-631.
Gupta, Anil (2021). “Definitions”, in: Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*: https://plato.stanford.edu/archives/win2021/entries/definitions/.
Hampton, James A. (2006). “Concepts as prototypes”, in: Brian H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory:* pp. 79–113. Elsevier Academic Press.
Hampton, Jean (1992). “Correcting Harms versus Righting Wrongs: The Goal of Retribution”, in: UCLA Law Review 39: pp.1659-1702.
Hanna, Nathan (2008). “Say What? A Critique of Expressive Retributivism”, in: *Law and Philosophy* vol. 27 (2): pp. 123-150.
Hanna, Nathan (2014). “Facing the Consequences”, in: *Criminal Law and Philosophy* vol.8 (3): pp.589-604.
Hanna, Nathan (2016). “Harm: Omission, Preemption, Freedom”, in: *Philosophy and Phenomenological Research* vol.93 (2): pp. 251-273.
Hanna, Nathan (2017). “The Nature of Punishment: Reply to Wringe”, in: *Ethical Theory and Moral Practice* vol.20 (5): pp.969-976.
Hansson, Sven Ove (2006). “How to Define: A Tutorial”, in: *Princípios: Revista de la Philosophia* vol.13 (19-20): p. 5-30.
Hart, Herbert L.A (2008). “Prolegomenon to the Principles of Punishment”, in: *Punishment and Responsibility – Essays in the Philosophy of Law*: pp.1-27. Oxford: Oxford University Press.
Holtug, Nils (2002). “The Harm Principle”, in: *Ethical Theory and Moral Practice* vol.5 (4): pp.357-389.
Honderich, Ted (2006). *Punishment – The Supposed Justifications Revisited*. Ann Arbor: Pluto Press.
Hornby, Albert S. (1995) “Punish”, in: Jonathan Crowther (Ed.) *Oxford Advanced Learner’s Dictionary (fifth edition)*. Oxford: Oxford University Press.
Kagan, Shelly (1989). *The Limits of Morality.* Oxford: Clarendon Press.
Kamm, Frances M. (1999). ”Physician-assisted Suicide, the Doctrine of Double Effect, and the Ground of Value”, in: *Ethics* vol.109 (3): pp. 586-605.
Kolber, Adam J. (2012). “Unintentional Punishment”, in: *Legal Theory* vol.18 (1): pp.1-29.
Läertius, Diogenes (1925). *Lives of the Eminent Philosophers* (transl. Robert Drew Hicks). https://en.wikisource.org/wiki/Lives_of_the_Eminent_Philosophers/Book_VI#Diogenes
Lippke, Richard (2010). “Punishing the Guilty, Not Punishing the Innocent”, in: *Journal of Moral Philosophy* vol.7 (4): pp.462-488.
Lyons, William (1974). “Deterrent Theory and Punishment of the Innocent”, in; *Ethics* vol.84 (4): pp. 346-348.
Martinson, Robert (1974). “What works? - questions and answers about prison reform”, in: The Public Interest (Spring): pp.22-54.
McCloskey, Henry J. (1965). “A Non-Utilitarian Approach to Punishment”, in: *Inquiry* vol.8 (1-4): pp.249-263.
McIntyre, Alison (2001). “Doing Away with Double Effect”, in: *Ethics* vol. 111 (2): pp.219-255.
McMahan, Jeff (1994). “Revising the Doctrine of Double Effect”, in; *Journal of Applied Philosophy* vol. 11 (2): pp. 201-212.

Moore, Michael S. (2010). *Placing Blame: A Theory of the Criminal Law*. Oxford: Oxford University Press.
Nelkin, D. K., & Rickless, S. C. (2015). “So Close, Yet So Far: Why Solutions to the Closeness Problem for the Doctrine of Double Effect Fall Short”, in: *Noûs,* vol. 49 (2): pp.376-409.
Parfit, Derek (1984). *Reasons and Persons.* Oxford: Oxford University Press.
Putnam, Hilary (1973). “Meaning and Reference”, in: *The Journal of Philosophy*, Vol.70 (19): pp.699-711.
Rosen, Gideon (2015). "Real definition", in: *Analytic Philosophy* vol.56 (3): pp.189-209.
Scheid, Don E. (1980). “Note on Defining ‘Punishment’”, in: *Canadian Journal of Philosophy* vol.10 (3): pp.453-462.
Scruton, Roger (1996). *Animal rights and wrongs*. London: Metro.
Singer, Peter (1975). *Animal Liberation*. New York: HarperCollins.
Sprigge, Timothy L.S (1965). “A Utilitarian Reply to Dr. McCloskey”, in: *Inquiry* vol.8 (1-4): pp.264-291.
Steinhoff, U. (2018). “The Secret to the Success of the Doctrine of Double Effect: Biased Framing, Inadequate Methodology, and Clever Distractions”, in: *The Journal of Ethics* vol. 22 (3-4): pp.235-263.
Steinhoff, U. (2019). “Wild Goose Chase: Still No Rationales for the Doctrine of Double Effect and Related Principles”, in: *Criminal Law and Philosophy* vol.13 (1): pp.1-25.
Søbirk Petersen, Thomas (2014). “Being Worse Off: But in Comparison with What? On the Baseline Problem and the Harm Problem”, in: *Res Publica* vol.20 (2): pp.199-214.
Tadros, Victor (2011). *The Ends of Harm: the Moral Foundations of Criminal Law*. Oxford: Oxford University Press.
Thomson, Judith J. (1999). “Physician-Assisted Suicide: Two Moral Arguments”, in: *Ethics* vol.109 (3): pp.497-518.
Väyrynen, Pekka (2021). "Thick Ethical Concepts", in: Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy,* https://plato.stanford.edu/archives/spr2021/entries/thick-ethical-concepts/.
Von Hirsch, Andreas (1993). *Censure and Sanctions.* Oxford: Oxford University Press.
von Hirsch, Andreas (2017). *Deserved Criminal Sentences.* Oxford: Hart Publishing.
Walen, Alec (2021). "Retributive Justice", in: Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy:* https://plato.stanford.edu/archives/sum2021/entries/justice-retributive/.
Walker, Nigel (1991). *Why Punish?* Oxford: Oxford University Press.
Wennberg, Robert N. “Act Utilitarianism, Deterrence and the Punishment of the Innocent”, in: *The Personalist* vol. 65 (2): pp.178-194.
Wiktionary (2022). “Punish”, https://en.wiktionary.org/wiki/punish
Williams, Bernard (1985). *Ethics and the Limits of Philosophy*, Cambridge, MA: Harvard University Press.
Wittgenstein, Ludwig (1991 [1953]). *Philosophical Investigations*. Chichester: Blackwell Publishing.
Wringe, Bill (2013). “Must Punishment Be Intended to Cause Suffering?”, in: *Ethical Theory and Moral Practice* vol.16 (4): pp.863-877.
Wringe, Bill (2016). *An Expressive Theory of Punishment*. Houndmills: Palgrave Macmillan.
Wringe, Bill (2019). “Punishment, Jesters and Judges: a Response to Nathan Hanna”, in: *Ethical Theory and Moral Practice* vol.22 (1): pp.3-12.
Zimmerman, Michael J. (2011). *The Immorality of Punishment*. Ontario: Broadview Press.