Original article

EXPERIMENTAL PHILOSOPHY

On the Psychological Feasibility of Punishment Theories

Valerij Zisman[1]*

[1] *Max Planck Institute for the Study of Crime, Security and Law, Department of Criminal Law, Freiburg, Germany*

*v.zisman@csl.mpg.de

Abstract

Feasibility arguments in punishment theory hold that the plausibility of a justification of punishment depends in part on its alignment with laypeople's punitive psychology. In this paper, I examine the philosophical significance and empirical support of three such arguments. According to the anti-abolitionist argument, abolishing punishment would have intolerable consequences because the public demands deserved punishment. According to the anti-retributivist argument, retributive institutions undermine themselves, because punishment that does not bring about a change in offenders leaves people dissatisfied and fuels demands for ever harsher sanctions. According to the democratic argument, the criminal law should reflect laypeople's shared intuitions of justice, because this promotes law-abidingness and legitimacy. I argue that the anti-abolitionist argument does not specify when a lack of alignment becomes a problem for a normative theory, and that the evidence does not support its empirical premises. The anti-retributivist and the democratic arguments state their success conditions more clearly, but the available evidence falls short of what they require. I then assess three general objections to feasibility arguments. The expertise objection does not apply to them, whereas the possible malleability of punitive attitudes and the likely small size of feasibility effects may limit their significance. Nonetheless, I argue that punishment theorists are well-advised to take the alignment of their theories with laypeople's psychology into account.

Keywords Punishment, Criminal Justice, Experimental Philosophy, Retributivism, Restorative Justice

1. Introduction

Recent research in both the empirical and the normative work on punishment theory has seen a heightened interest in what we can call the psychological feasibility of punishment theories. In general terms, the psychological feasibility of punishment theories concerns the question whether the normative justifications of punishment that philosophers propose or that a given legal system is designed after are aligned with laypeople's punitive psychology. Alignment here can be specified as a correlation between what the normative system and laypeople see as valuable goals in criminal law, or in terms of what types of sanctions would satisfy the demands of justice of the public.

DOI: 10.18716/ojs/xphi/2026.1.12088

From a philosophical perspective, such feasibility arguments need to clarify why (and exactly how) the alignment of normative theory and laypeople's psychology should matter to normative theory. Does lack of feasibility undermine the plausibility of normative theories; does it only weaken their plausibility; or do feasibility arguments only matter from a pragmatic perspective? In my perception of the literature, such methodological questions are not reflected on in detail, leaving open the question of the relevance of such arguments in particular, and the connection between empirical and normative research in general. This reflection is especially pressing as the relevance of how people think about issues in criminal law (or other areas in philosophy, for that matter) is frequently questioned (for an overview, see Sommers, 2021; Tobia et al., 2025). It is thus worthwhile to not just ask whether the specific feasibility arguments are plausible, but also what their general potential within the philosophical discourse is.

I will proceed as follows. In Section 2, I will start with feasibility arguments as they are presented by critics of abolitionism and restorative justice, specifically by normative theorists such as Husak, as well as empirical researchers such as Carlsmith, Robinson, and Darley. The relevance of feasibility arguments proposed by them is mainly instrumental: the lack of support that the public has for abolitionist or restorative approaches to criminal law should count as a reason not to establish such systems in practice. Whether or not it says anything about the plausibility of the system in an idealized sense remains unclear. In Section 3, I will discuss feasibility arguments with the opposite philosophical target. These attempt to show that retributive systems of punishment are not psychologically feasible. As McGeer and Funk (McGeer & Funk, 2017) argue, establishing systems of retributive punishment leads to non-retributive punishment over time due to features of the punitive psychology of laypeople. While not showing that retribution is implausible as a normative theory, this argument tries to show that a retributive institution is practically self-defeating. In Section 4, I will look at democratic approaches to criminal law theory and how they defend the relevance of the alignment between laypeople's psychology and the norms of the justice system. While these sections focus on specific arguments that have been proposed in the literature, Section 5 focuses on the general philosophical potential of feasibility arguments in criminal law theory. According to the expertise objection (5.1), normative theorists should not care much about such arguments as laypeople do not grasp the correct normative approach to punishment due to their lack of expertise. According to the malleability objection (5.2), philosophers need not pay much attention to the current psychological feasibility of their punishment theories as punishment psychology is malleable. Philosophy should not follow the psychology of laypeople; rather, laypeople's psychology should be educated if at odds with plausible normative theory. Lastly (5.3), I discuss the normative irrelevance of feasibility arguments from a different perspective: even if we grant that psychological feasibility is philosophically interesting in principle, it is too low a bar to take on to be of *much* philosophical relevance. According to the objection, a lot of different theories are feasible in a minimalist sense, so philosophers should not be worried too much about that aspect of their normative theory. While I agree that the philosophical significance of feasibility arguments should not be overrated, I want to emphasize that it is worthwhile to clarify how and why it matters in criminal law theory despite the expertise objection, the malleability objection, and the philosophical irrelevance objection—something not discussed enough in current research.

2. The Feasibility of Abolitionism and Restorative Justice

One way the psychological feasibility of punishment theory is used in current debates is as a pragmatic challenge to specific normative theories. If institutions do not reflect how laypeople think about normative issues, we risk that the public will not cooperate with the demands of such institutions. This, in turn, should be regarded as a reason to reject a normative theory. Punishment, after all, is a very practical research area—and whether or not laypeople respect

the authority of the law, and feel that it is just, might matter a lot to their behavior. The following arguments will rely on the assumption that law-abiding behavior is influenced by how just the institutions are perceived to be, which I discuss in a bit more detail in Section 4, but which will be assumed as correct for this and the next section (for an overview of the justice perception and a defense of the general role of perceived legitimacy, see Robinson & Darley, 1997, section II–III).

Husak (2020), as well as Robinson and colleagues, uses feasibility arguments in order to undermine criminal law abolitionism (Carlsmith et al., 2002, p. 297; Robinson, 2003; Robinson & Darley, 2007, section II). Abolitionism claims that the state should not intentionally harm offenders in response to wrongdoing. The feasibility objection states that such a normative theory goes against what laypeople believe or hold to be normatively correct. Because of that mismatch, the broader public would perceive such a normative system as less legitimate than one that reflects their beliefs and values. This is then taken to be a challenge for the abolitionist approach. Husak, for example, starts with the claim that there is "no groundswell of support among the general public to fundamentally revamp the criminal law as it currently exists" (Husak, 2020, p. 51). Why does Husak expect there to be no support for criminal law abolitionism among the public? Because, according to him,

> [T]he public demands justice—by which I gather they mean some form of punishment. These sociological facts are crucial and cannot be ignored by skeptical reformers. I am doubtful that alternative means to sanction wrongdoers and mollify victims would be adequate to satisfy those who call for justice. If the criminal law were fundamentally altered, all bets would be off as to how the public in general and victims in particular would respond. (Husak, 2020, p. 52)

Why should philosophers (and in this case, abolitionist theorists) care whether or not laypeople agree with the normative theories they painstakingly crafted in the towering heights of their reclusive ivory tower? For Husak and other critics of abolitionism and restorative justice, there is purely instrumental value in feasibility arguments (Husak, 2020, p. 53). Paying attention to the psychological feasibility of criminal law as a whole or specific punishment types or motives will likely "produce a wealth of further goods, such as an increased willingness to cooperate and conform to law" (Husak, 2020, p. 53).

Before getting to the question whether or not the empirical evidence supports Husak's claim, I want to raise the more general methodological question to what extent philosophers should care about such arguments—as this is not clarified by the respective arguments. After all, one radical way to deal with this objection is to simply acknowledge the empirical case but reject its normative relevance entirely. If abolitionism, for example, is the morally correct approach to punishment—because state punishment is taken to be an unjustifiable intentional infliction of harm (Boonin, 2008)—then what laypeople expect from the criminal law should not matter for the normative evaluation of the argument brought forward by abolitionists. Or, at the very least, abolitionism might very well be worth the costs in psychological feasibility that come with it. In the same way that it would be right to go against the institution of slavery despite (let us assume for the sake of analogy) widespread public support for slavery, we should not institutionalize a morally fraught system of punishment simply because the public demands it. A sound philosophical theory is immune (at least to some extent) to such practical concerns of its alignment with public opinion.

There are at least some examples in punishment theory that would suggest such a resistance to the relevance of feasibility arguments. Obviously, Kant is well known to have said that "if justice goes, there is no longer any value in human being's living on the earth" (Kant, 1797/1996, p. 473 [6:332]). More recently, Morris similarly argued that an institution of punishment that is not focused on giving offenders their just deserts first and foremost fails to respect the dignity of moral agents (Morris, 1968).

While pure retributivist approaches to criminal punishment might indeed have no room for feasibility arguments and their instrumental concerns to play a role, many retributivists both in the past and today have some room for instrumental concerns to play a role in the justification of the amount or type of punishment. Husak, a retributivist himself, acknowledges the need to have non-retributive justifications for the institution of punishment to justify the costs involved in pursuing intrinsically valuable aims, such as retributive justice (Husak, 2016; Moore, 2010, p. 186 similarly acknowledges other values despite desert).

Therefore, given the views many retributivists hold, there is some space for feasibility arguments in criminal law to latch onto existing debates even without accepting a purely forward-looking approach to criminal punishment. The crucial question, then, is how much weight such arguments should carry and how exactly they impact the plausibility of normative theories.

Unfortunately, feasibility arguments have not yet clearly specified when the lack of alignment between normative theory and laypeople's psychology becomes a problem for normative theories—and it is unclear whether there is hope for any clear criteria of relevance in principle. Estimating when exactly the consequences should make us worry about an otherwise plausible normative theory is no easy task.

What are the relevant criteria offered in the current debate? In the case of the anti-abolitionist argument, we get a criterion: failure of institutions. Husak, for example, claims that "all bets would be off" (Husak, 2020, p. 52) if we were to ignore the justice intuitions of the public. While 'all bets are off' does not provide a clear criterion, I think it is fair to assume that Husak has in mind consequences that philosophers should be quite wary of, such as a potential collapse of the institutions as such. If implementing an otherwise plausible normative theory would result in societal collapse, so the argument would have to go, then that normative theory should be rejected.

Similarly, Robinson and Darley argue that when philosophers ignore laypeople's retributive psychology, bad things are going to happen:

> We must face the reality that human beings will demand justice for serious wrongdoing, and that the absence of a system that allows for the imposition of deserved punishment would produce intolerable consequences. (Robinson & Darley, 2007, p. 18)

Again, while it is clear that the argument is supposed to count against the normative theory that is being criticized, we only get a vague criterion ('intolerable' consequences) that would need to be spelled out in more detail to have clear argumentative force. We are not told what these intolerable consequences are exactly, and how severe these might be and would have to be to count as an objection against an otherwise plausible normative theory. In a different paper, Carlsmith and Darley similarly argue that "[t]he 'rule of law' is threatened when the rules of law violate citizen intuition" (Carlsmith & Darley, 2008, p. 235). How severe the threat is, and at what point normative theorists should worry about it, again remains unspecified.

The reason I insist on this lack of specificity is not simply to point out flaws in the arguments of anti-abolitionists. It is to reflect on the needs of such arguments in general to clarify—if possible at all—what would need to be the case for their arguments to have philosophical relevance. In my estimation, 'all bets would be off', 'intolerable', and 'threat to the rule of law' as descriptions of potential consequences of implementing a normative theory are too vague, and therefore philosophically unhelpful (even regardless of whether the claims are empirically correct) as long as there is no argument for when such consequences matter, normatively speaking.

While there will likely still be disagreements among philosophers about how much instrumental detriment is enough to make us suspicious of specific theories of criminal punishment (a 10%, 20%, or even 50% reduction in law-abidingness?), such clarifications nonetheless need to be made, and their relevance needs to be argued for. We have seen that even many retributivists

allow for such instrumental considerations to play a role (Husak, 2016, pp. 50–51; Moore, 2010, p. 186)—but to know how much of a role, and for feasibility arguments to work, this needs specification.[1]

Because of the lack of specificity, it is hard to analyze whether these arguments are convincing on empirical grounds. Anti-abolitionists point to a plethora of research showing that people engage in second- and third-party punishment (for an overview of this case, see Husak, 2020, section 9), with some evidence supporting the claim that punishment (of some sort) is a universal human trait (Curry et al., 2019, 2022). But while it is undoubtedly true that people engage in punishment and find *some* form of sanctioning satisfying, the details are more complicated than anti-abolitionists appear to think. First, while practices of sanctioning indeed appear to be universal, it is less clear that people are pure retributivists when it comes to punishment (Zisman & Rehren, 2025), which at least would weaken Husak's and Robinson et al.'s claim that people demand retributive justice. It is also less clear that punishment as the intentional infliction of harm on offenders is superior in satisfying both observers and victims of wrongdoing compared to restorative sanctions, such as different forms of restitution (van Doorn et al., 2018; for a more optimistic assessment of some form of abolitionism, see Zisman, 2023). Therefore, because the dangers of abolitionism are not specified by feasibility arguments, and because the empirical case against abolitionism is not as strong as retributivists appear to hold, anti-abolitionist feasibility arguments are not currently convincing.

3. The Feasibility of Retributive Justice

Ironically enough, a feasibility argument structurally similar to the one retributivists advance has also been made against retributive justice itself. Not only did McGeer and Funk argue that there is little reason to suspect that, in Husak's words, 'all bets would be off' if we were to steer away from (strictly retributive) punishment, but they also argued that retributive institutions would likely undermine themselves in practice. They write that

> [O]ur model suggests that the rhetoric of 'just deserts,' far from ensuring 'proportionality' in response to crime, may, under certain circumstances, produce a powerful ratcheting-up effect in people's estimation of how much punishment is, in fact, deserved. (McGeer & Funk, 2017, p. 539)

The basic idea of their argument, in broad terms, is that what laypeople really care about when punishing—that is, what they derive most justice-satisfaction from—is to see offenders change their behavior after their wrongdoing (or at least promise to do so). Retributive punishment alone, in the form of incarceration or otherwise harsh treatment of offenders, tends to not satisfy the supposed justice demands of victims or the public. In fact, Funk and other social psychologists have picked up work from the anti-abolitionists (Carlsmith et al., 2008), which found that people who engage in retributive punishment did not feel more satisfied after doing so. This led to a line of research which now suggests that what gives people justice-satisfaction when punishing offenders or seeing offenders being punished is not seeing just desert being given to offenders but rather prompting changes in the behavior of offenders (Aharoni et al., 2022; Funk et al., 2014; Gollwitzer & Denzler, 2009; Gollwitzer et al., 2011; Nahmias & Aharoni, 2017).

According to the anti-retributivist feasibility argument, we might expect people who live in punitive systems built on retributive normative frameworks to react with demands for harsher punishments to a felt dissatisfaction with the retributive practice (assuming that dissatisfaction leads to such punitive demands). This would then form a vicious feedback loop. More

[1] Feasibility arguments work more seamlessly in the context of consequentialist approaches to punishment. In the context of such theories, if a certain punishment scheme or rationale leads to more law-abiding behavior, this would automatically speak in favor of adopting such a scheme or rationale, all else being equal.

punishment does not lead to more justice satisfaction if it lacks proper responses from offenders that indicate a change of their future behavior—and due to a lack of justice satisfaction, the public might demand even more punitive reactions to wrongdoing. Therefore, one might argue, the institutionalization of a purely retributive system of punishment that is normatively blind to the behavioral changes of offenders is psychologically infeasible. Because the sense of justice of victims and the public at large would remain unsatisfied, such a system would lead to ever harsher punishment, ultimately departing from its own demands of measured and proportionate responses to crime.

Such an argument is also familiar from discussions in normative ethics. In the context of utilitarianism, some have argued that implementing decision-making practices (or institutions) according to act utilitarianism might lead to less utility than institutionalized rule-based approaches to utilitarianism—also due to quirks of human psychology, such as mistakes in the calculation of how much utility an act brings about (for a critical discussion, see Singer, 1972). In that sense, institutionalizing act utilitarianism and retributivism might similarly be pragmatically self-defeating.

Caring about justice and proportionality might thus require us to shift away from narrowly retributive understandings of criminal punishment, and require institutions to allow for alternative types of sanctioning (those that make sure that offenders change their behavior in the future) and for a loosening of the concept of proportionality (if proportionate punishment is not the best method of making offenders change their behavior in the future).

In the context of the anti-retributivist feasibility argument, the criterion for when the feasibility of punishment theory should matter to the normative theory in question is clearer. When implementing systems of retributive punishment leads to disproportionate punishment over time due to features of laypeople's psychology, then this should be relevant to retributivists themselves. In this case, retributivists do not have the chance to say that they do not care for laypeople's psychology and the consequences of not paying attention to it, because making sure that proportionate punishment is given to offenders is exactly the point of implementing the institution of retributive punishment.

To be fair, there is potentially some wiggle room for retributivists because the concept "proportionality" allows for a range of specific punishment practices. So, they might counter that a certain increase in punitiveness over time after implementing retributive practices would not necessarily undermine the retributive rationale for punishment, as even an increase over time might still count as proportional. After all, one might make the argument that both the US and Norway have proportional responses to murder, even though these are significantly different from each other. But this response cannot take just any shape, because otherwise the concept of "proportionality" would become empty, which is something that retributivists themselves cannot accept.

This version of the feasibility argument succeeds more clearly in showing how the psychology of laypeople matters to the plausibility of normative theories. The main problem, however, is its own plausibility. McGeer and Funk present data suggesting that people are not satisfied with merely proportional punishment of offenders but rather require communicative elements from the offender that involve promises of changes in behavior by the offender—and this communicative and behavioral dimension of punishment satisfaction has also been found by other studies, as mentioned above. But this alone falls short of showing that there will be a ratcheting-up effect which undermines the retributive principle in practice. The lack of justice satisfaction might not be severe enough to drive people to demand ever-increasing punishment severity over time. Or institutions might implement other ways than punishment to make people satisfied with the justice system, such as procedural justice mechanisms. It would also be hard to establish that the increase in punishment would be due to the lack in justice satisfaction alone, rather than any other changes in society.

Therefore, while the relevance of such an argument seems clear, the current data at best hint at the potential of an anti-retributive feasibility argument. As it stands now, it falls short of providing the evidence necessary to make a complete case against retributivism.

Because of the shortcomings of the anti-abolitionist and the anti-retributivist arguments, a potential solution would be to argue that the psychology of laypeople supports a punishment system that is in between these extremes.[2] Indeed there is research suggesting that restorative justice and retributive justice might work best when combined, and used for different types of crimes (Gromet & Darley, 2006, 2009; Wenzel et al., 2008). While a normative theory that accounts for the plurality of punishment motives has a better chance of surviving empirical scrutiny, it also comes with challenges, especially theoretical ones. While some theories of justice allow for a mix of retributivism and restorative justice (Daly, 2002, 2005), abolitionists, for example, would not. A normative theory that wants to account for laypeople's psychology therefore has some reason to embrace the plurality of punishment motives. Theoretical and normative considerations, however, may give us pause, since monistic theories do not easily accommodate a pluralistic psychology.

4. Feasibility and Democratic Legitimacy

Another area in which arguments rely on the alignment of laypeople's attitudes and normative theory (or specific policy proposals) is democratic justifications of criminal justice.[3] Democratic theorists offer broadly two types of reasons why the law should reflect (some of) laypeople's punitive psychology.

The first type of reason is closely aligned with the rationale covered in the previous two sections: the law should reflect the attitudes of laypeople (or at least the people that the law applies to) because doing so will make it more likely that people adhere to the law. Robinson and colleagues have made this argument most forcefully in the literature, claiming that when the law reflects laypeople's attitudes, this leads to laypeople perceiving the law as more just, which in turn makes it more likely that people adhere to the law (Robinson, 2017; Robinson & Darley, 1997).

The second type of reason for alignment is non-instrumental. That value is typically spelled out in terms of legitimacy. A criminal law that pays attention to what its citizens value, and what they want punishment to look like, is more legitimate than one that ignores such attitudes. I will focus on instrumental arguments first, and at the end of this section comment on non-instrumental arguments.

But for either of these arguments, the first step when using them in criminal law theory—as with the arguments discussed in the previous sections—has to be to specify exactly in what ways laypeople's attitudes matter, and which attitudes specifically matter. There is a range of options for both tasks. Some democratic theorists are at the more extreme end, demanding that we design most of the criminal law in line with the attitudes of the relevant public (Kleinfeld, 2016, 2022). At the other end of the spectrum, one could hold merely that practitioners and scholars should be aware of the attitudes of the public, without making any specific commitments to reflecting these in the law.

While this section cannot do justice to all the approaches to democratic criminal law theory (for an overview, see Ryberg & Roberts, 2014), I want to clarify the challenges of using such arguments in criminal law theory with the example of Robinson and colleagues, as they are most explicit in their instrumental argument in defense of democratic criminal law theory.

[2] Thanks to a reviewer for pointing this out.

[3] Thanks to a reviewer for the nudge to add this section about democratic criminal law.

Robinson et al.'s argument proceeds in two rough steps (Robinson & Darley, 1997). The normative part of the argument says that the criminal law should value those policies (and theories) that make people more likely to adhere to the law, all else being equal. The empirical part argues that laypeople are more likely to adhere to the law when the law reflects their views, as this makes them see these laws as more just. Perceiving laws as just, in turn, correlates with a stronger motivation to adhere to said laws.

Which views of the public matter, then? Robinson et al. argue that laypeople endorse 'empirical desert', which entails that the law should criminalize core wrongdoings (such as physical aggression, property damage, and deceit), and that it should punish wrongdoings in proportion to their blameworthiness. This, according to Robinson et al., does not come with agreement on cardinal punishment levels, but only with an endorsement of relative punishment severity (ordinal proportionality). That is to say, laypeople do not agree on how severe specific punishments should be in absolute terms, but they do agree on how different types of wrongdoing should be ranked relative to each other in terms of how severely they should be punished.

The reason why it is valuable to use Robinson et al. as an example is that they are very clear on these criteria. This allows readers to offer direct rebuttals, and especially a critical discussion of the empirical basis of the claim. I will turn to this now, but I want to stress that I think this is exactly the right way to use feasibility arguments, that is, by clarifying the criteria for the argument's success, as Robinson et al. do.

Let us accept the normative claim that theory and practice should value improving law-abidingness. The two central empirical claims here are that laypeople hold the view of empirical desert, and that reflecting empirical desert in the law will lead to more law-abidingness. There are reasons to doubt both empirical claims.

First, people only (if at all) agree on ordinal proportionality under specific circumstances, where, for example, political and other types of biases are controlled for (Kolber, 2009, p. 441ff.). Agreement on ordinal proportionality rankings is thus limited. This is fine, proponents of empirical desert could counter, because the relevance of laypeople's attitudes should be limited to their unbiased opinions, not those that are biased (Robinson, 2008, p. 149).

While this might be a plausible move judging from a normative perspective, it raises an issue that Robinson and colleagues do not deal with adequately. If we restrict the relevant attitudes to the 'true' (unbiased) attitudes, this means that the argument needs to show that perceptions of justice and law-abidingness correlate with these true views, rather than the biased ones. So what drives laypeople's perception of the law being just?

Robinson and colleagues claim that there is a connection between judgments of ordinal proportionality, perceptions of the law doing justice, and self-reported intention to abide by the law. The first problem is that Robinson et al.'s own studies in that regard fail to establish that connection, as their manipulation depends on a mismatch between absolute punishment judgments of laypeople and the law, rather than merely a mismatch of ordinal proportionality judgments (Robinson et al., 2010, especially Studies 2a and 2b). Also, there is some doubt whether the link between laypeople's attitudes and law-abidingness is as strong as Robinson et al. suggest. Many of their studies depend on reported willingness to cooperate with the law, not actual behavior, and other people have in general cast doubt on whether laypeople value the law being applied in accordance with the same values that they have (but see Engelmann et al., 2024). De Keijser, for example, found evidence that laypeople sometimes think that it is a good thing that courts keep their distance from laypeople's attitudes and values (de Keijser, 2014, pp. 110–111). While this is not a direct counter to the idea that legitimacy perceptions lead to law-abiding behavior, it at least suggests that unaligned values do not necessarily lead to dissatisfaction with parts of the justice system.

While there are thus many doubts about Robinson et al.'s specific democratic version of a feasibility argument, it remains philosophically interesting, provided it can satisfy certain conditions. To succeed, the argument would need to accomplish three tasks: First, it must show that promoting law-abidingness matters from a normative perspective—a claim that many legal scholars in the debate would likely accept. Second, it must specify which aspects of lay psychology should be reflected in law. This task is more difficult, as there are *prima facie* normative reasons not to pay attention to all of laypeople's attitudes. Third, and perhaps most difficult, it must establish that laypeople actually hold the views that are taken to be relevant to criminal law. None of this is easy to achieve, but an argument that succeeds would be a philosophically relevant version of a democratically inspired feasibility argument.

Up to this point in the discussion, the focus has been mainly on instrumental arguments for taking laypeople's attitudes seriously in criminal law theory, which is to be expected with arguments on the psychological feasibility of punishment theories, as feasibility can be measured against law-abiding or theory-congruent behavior. But alignment arguments need not be restricted to instrumental concerns. In fact, democratic approaches also lend themselves to non-instrumental arguments for some form of alignment between the law and laypeople's psychology. From a democratic perspective, alignment between laypeople and the law might be valuable because it imbues the law with legitimacy (regardless of whether that would make people more or less likely to adhere to the law).

Such legitimacy concerns can take at least two shapes. First, legitimacy can come about by the law reflecting the actual views of the people, as established by surveying them in some form. Second, legitimacy can come about by allowing people to directly impact the content of the law, for example, by voting on policies directly.

The crucial empirical challenge for the first type of legitimacy argument is again to establish which attitudes are relevant for legitimacy, and that people actually agree on a set of views. Robinson et al. tried to solve this problem by keeping the set of views on which people need to agree rather minimal (criminalization of core crimes and ordinal proportionality). For legitimacy arguments, the problem I discussed above regarding the link between the minimalist account and law-abiding behavior does not apply, as legitimacy is primarily non-instrumental. Still, you might worry that agreement on such a minimal set of base principles does not give the law much legitimacy. After all, ordinal proportionality is silent on the worst types of punishment a state can use—and lay attitudes certainly are not.

The second type of legitimacy argument avoids this problem, as it is procedural rather than based on specific content. As long as people have ways of meaningfully engaging in the process of establishing the relevant principles and values of the criminal law, the relevant criterion is fulfilled. Here, one potential problem is that alignment stops being the relevant measure of legitimacy. Whether or not the outcome of the participatory process then also reflects what laypeople want the law to be is an open question—and at least for feasibility arguments a mismatch should be worrisome, as it would cast doubt on the relevance of such legitimacy arguments.

5. Objections to Feasibility Arguments in Punishment Theory

In the previous sections I have aimed to establish how feasibility arguments could in principle be used in normative debates, and whether any of the existing ones are plausible. I have argued that philosophers have until now not been precise enough regarding how exactly the empirical dimension of the feasibility arguments would impact the normative theory that is being criticized, and whether the data actually support the critique. I have argued that there are clear cases in which feasibility arguments become relevant in philosophy. Nonetheless, I have pointed out that neither the anti-abolitionist nor the anti-retributivist feasibility argument is supported

by empirical evidence in such a way that it would succeed in undermining the normative theory it targets. Lastly, I have argued that democratic feasibility arguments have done well to specify the empirical and normative success conditions they rely on, but have failed on the empirical side. Thus, while powerful in principle, current feasibility arguments do not succeed.

Still, feasibility arguments could be dismissed on more general grounds. This is what I will discuss in this section, aiming to defend the relevance of feasibility arguments in punishment theory while acknowledging some limitations.

5.1 Expertise Objection: What Laypeople Think Does Not Matter

Expertise objections are familiar territory in experimental philosophy (Horvath & Wiegmann, 2016; Weinberg et al., 2010; Williamson, 2011). In the context of feasibility arguments, the objection states that while some empirically minded philosophers might be interested in how laypeople think about various aspects of the law, this ought not to be of any concern to the armchair normative theorists, as the correct moral theory or theory of punishment is true independently of how laypeople think about punishment. This might be because laypeople are biased, or because they simply lack specific training necessary for proper reasoning in an expert field, which philosophers and legal theorists do have.

While there might be something to say in favor of the expertise defense in (experimental) philosophy more generally, it fails to apply directly to feasibility arguments in punishment theory—or at least to those presented in this paper. The gist of feasibility arguments, as discussed in the previous sections, is that when normative theories and descriptive accounts of punishment psychology diverge too far, we risk creating institutions that do not function well or promoting conflict with the law. Such a warning, if correct, is not affected by the claim that laypeople get the law wrong. It might be very unfortunate that we should bow to the wrong attitudes of the public, but if we are to accept the relevance of feasibility, then the quality of laypeople's judgments does not matter directly. To put it crudely: the best normative punishment theory is of no use if implementing it leads to chaos and lawless behavior. On the other hand, a normatively deficient theory of punishment might still be the better option if it ensures that people at least adhere to the (morally) non-optimal legal rules.

This defense against the objection will not impress the persistent ideal theorist who would argue that the plausibility of a normative theory has nothing to do with how laypeople think about the law. And they would need to hold that the potential (moral) costs of implementing that system are normatively irrelevant. In that regard, the defense of feasibility arguments is limited, as it presupposes a minimal normative significance to the pragmatic consequences of implementing normative theory into practice. While this is a limitation, my perception of the debate in criminal law theory is that researchers (experts in punishment theory) do care about the costs and consequences of their proposed normative system. Therefore, feasibility arguments as discussed in this paper should not be dismissed based on the expertise objection.

5.2 The Malleability of Laypeople's Punitive Attitudes

A tacit assumption for the success of feasibility arguments is the claim that the punitive psychology of laypeople is to some extent set in stone—otherwise, feasibility arguments do not have much normative force. For instance, if the anti-abolitionist from Section 2 claims that retributive justice desires of the public cannot be ignored on pain of terrible consequences, this assumes that laypeople cannot simply get used to non-punitive systems over time and accept them as just.

So, for feasibility arguments to have any bite, punitive psychology cannot be too malleable—if it were very malleable, normative theorists would only need to figure out the correct normative theory, implement it in practice, and wait until people get used to the institution that is built on the correct normative framework. Even if there were some conflict at first, over time, feasibility would not be a worry for normative theorists. Feasibility arguments would then be philosophically uninteresting.

How malleable is our punitive psychology? If we think about how different institutions of criminal law were historically and are now cross-culturally designed, we might expect punitive psychology to be quite malleable. Humans have lived in systems with gruesome punishments and comparatively mild punishments, as well as institutions that criminalize very different types of actions. Even now, there are institutions with quite gruesome prison conditions and those with significantly less terrible prison conditions. On the face of it, it is unclear whether law-abidingness depends a lot on such details of the justice institutions.

To my knowledge, there is no clear data on how malleable punitive psychology is. It is therefore difficult or perhaps even impossible to assess the limits of a minimally feasible justice system. While it is likely a minimal condition of justice institutions that there is some form of law enforcement, and thus coercion, the details of what the coercion can look like can and do vary a lot. There might be minimal conditions such as ordinal proportionality, as we have seen in Section 4, but even here there is some evidence that, independent of sanction severity, procedural justice perceptions can have an important impact on overall justice satisfaction (Casper et al., 1988). Such a finding adds to the claim that the specific punishment satisfaction is somewhat flexible, as long as the wrongness is scaled coherently with the punishment and procedural justice is maintained. But it also sets limits on the malleability of justice satisfaction if procedural justice turns out to play a role in all societies and institutional arrangements.

On the other hand, sometimes a single criminalization decision can be enough to lead or motivate citizens to rebel against the rule of law, as can be seen when citizens protest for voting rights and the decriminalization of certain actions. As far as I can see, we have no solid basis for judging how much we can experiment with criminal justice institutions before the rule of law is threatened, as Carlsmith and Darley worried (and if, indeed, criminal law would be the cause of such societal distress in the first place, rather than economic or other problems within a society).

We are familiar with this problem from other debates in philosophy and law—where researchers also disagree about the malleability of laypeople's psychology. In debates on free will, blame, and responsibility, some compatibilists in the tradition of Strawson (including Strawson himself according to a predominant interpretation) argue that reactive attitudes (and by extension maybe punishment) are fundamental human attitudes we can perhaps temper but not get rid of if we are to treat each other as moral agents (Strawson, 1962). Others, however, do suggest ways of living without ideas of free will, reactive attitudes, or punishment (Pereboom, 2001). Again, I do not see the debate settled on empirical grounds in one or the other direction. There are certainly some ways of living that suggest life without reactive attitudes might be possible (and by extension, the acceptance of institutions that do not promote these), but it is of course an open question whether this would be possible across the whole population, or at least among enough people that their dissatisfaction with a non-punitive justice system would not lead them to rebel against it.

While there is clearly some malleability to our punitive psychology—and correspondingly, the range to which people will be willing to accept justice systems modeled on different normative theories—we cannot say for certain how far this malleability can be taken. What does this mean for philosophers interested in feasibility arguments in punishment theory? First, if future research suggests that people can accept vastly different justice institutions, including those which do not punish at all and those which punish harshly, then feasibility arguments might lose their philosophical significance to a large degree. If there are no long-term detrimental

effects due to lack of alignment between normative theory and laypeople's psychology because laypeople simply get used to different systems, then worries about the current feasibility of certain normative theories can be largely ignored.

This might be bad news for feasibility arguments, but for the empirically minded legal and moral philosopher this is still good news. Whether and to what extent laypeople's psychology is malleable is an empirical question, after all, so there is good reason for normative theorists to pay attention to the work of their colleagues in other fields. While it might turn out in the future that feasibility arguments should not be taken too seriously, we first have to do the empirical work to establish this conclusion.

5.3 Practical Irrelevance of Feasibility Arguments

I want to close with a third way of rejecting feasibility arguments in punishment theory—one that might be implied by the discussion in Sections 2 and 3 of this paper. In these sections, I specifically looked at two currently prominent types of feasibility arguments that point in two very different directions: the rejection of abolitionism and the rejection of retributivism. For both conclusions, I argued that the evidence is not quite there to support either side—and that strong evidence for either side would be necessary for the arguments to bring about their quite interesting conclusions regarding the criminal law and criminal punishment.

If extraordinary evidence would be necessary to support either side of the argument, then a realistic expectation might be that none of the arguments succeed and that therefore feasibility arguments—even though in principle relevant for a philosopher thinking about the justification of punishment—are never supported by the evidence in such a way that would allow us to draw interesting normative conclusions. Rather, the most successful feasibility arguments would then only show that one system of punishment would lead to slightly more law-abidingness or justice-satisfaction than another system of punishment. But depending on the degree to which this is the case, such results might become philosophically less significant—and armchair philosophers might have more ground to reject the relevance of such arguments.

I discussed this limitation in Section 2: philosophers might choose to simply ignore the philosophical relevance of weak (but realistic) feasibility arguments if their otherwise plausible normative theory is psychologically less feasible than a theory-competitor. If retributive institutions of criminal punishment are slightly less feasible in the sense that they would lead to less justice satisfaction and thus less compliance with the law than systems of punishment built on other normative frameworks or that suggest different types of sanctions—the persistent retributivist can simply ignore such arguments and proceed happily in the knowledge that their proposed justification is the normatively more plausible one. A small cost in feasibility and corresponding instrumentality might be an easy bullet to bite. Small differences in feasibility might only matter to pure consequentialists who aim to maximize law-abidingness (assuming that justice satisfaction is the only relevant metric in that regard—which it likely is not).

But even with this limitation in mind, I would maintain that it is worth paying attention to feasibility arguments in criminal law scholarship from the philosophical perspective. First, as I have argued, it is still quite an open question—empirically speaking—how feasible different punishment theories are in practice. While the arguments I discussed in this paper do indeed require strong evidential support—such evidence might be provided in the future. Second, even if only weak versions of feasibility arguments turn out to be plausible, that is, feasibility arguments that succeed in showing that some systems of punishment fit better with laypeople's psychology than others and are thus—all things being equal—preferable on instrumental grounds, such arguments should carry at least some weight to all normative arguments that allow room for instrumental considerations, which many current justifications of criminal punishment do. Feasibility might not be the first concern of armchair philosophers, but it should

cross their minds at some point in the analysis of their theories. Acknowledging this also invites more theorizing about when exactly feasibility should matter to a given theory.

6. Conclusion

In this paper, I have examined the potential of feasibility arguments in punishment theory. The core idea of feasibility arguments is to support or criticize a normative justification of criminal punishment based in part on its alignment with laypeople's punitive psychology.

I have discussed three feasibility arguments in detail—one that poses an objection to abolitionism in criminal law theory, one that poses an objection to retributivism, and one that borrows from democratic theory to defend the relevance of laypeople's attitudes in criminal law theory. For all of them, I have discussed how much philosophers ought to pay attention to such arguments in their research. As all these arguments can be interpreted as instrumental, one necessary condition of relevance is that instrumentality matters (to some degree) in criminal law theory. For many theories this condition is granted, but for some it might not be—and in such cases feasibility arguments do not directly become relevant to philosophical debates.

Another such limitation is that the existing feasibility arguments make strong empirical assumptions regarding the punitive psychology of laypeople. The anti-restorative and anti-abolitionist cases require laypeople's psychology to be so punitive that an institution without proportionate punishment would collapse due to a lack of justice satisfaction among the public and a corresponding decline in norm compliance. The anti-retributive case would require that the implementation of retributive punishment, due to laypeople's psychology, necessarily leads to disproportionate punishment over time, therefore undermining retributive justice in practice. While both are interesting arguments, the evidence does not appear to support their strong conclusions. Interestingly, the democratic argument proposed by Robinson et al. makes weaker assumptions as it claims only that laypeople universally agree on some core criminalization issues and ordinal proportionality. While there is some doubt about whether laypeople agree even on that, the main challenge for this argument is to show that with such a minimal account of the necessary content of the law, the empirical link to law-abidingness still exists.

Despite these limitations, I have (in part) defended the relevance of feasibility arguments in punishment theory against more fundamental objections in Section 5. While feasibility arguments are unlikely to resolve any normative debates in punishment theory for good, I have argued that philosophers and legal theorists spend their time well when testing their preferred normative theory against the empirical research on how laypeople think about punishment and different forms of sanctions. This is even more true for policy makers.

It should also be noted that psychological feasibility is only one of the relevant feasibility concerns in assessing the overall plausibility of normative punishment theories. As a reviewer of this paper has correctly pointed out, economic and political feasibility concerns also play an important role. Different normative theories have different implementation costs. Furthermore, punishment theories should not only be psychologically feasible but also promote other values of the state or criminal justice system. A theory that is psychologically more aligned with how laypeople think about punishment is not necessarily the best suited to combating discrimination, for example. Future research or a more comprehensive justification of a given normative punishment theory thus also has to weigh psychological feasibility against other relevant normative concerns.[4]

[4] I am thankful for helpful feedback from participants at a workshop at the Center for Interdisciplinary Research (ZiF) in Bielefeld, Germany, and at Jagiellonian University in Kraków, Poland, as well as from the anonymous reviewers and the editors of *Experimental Philosophy*.

7. References

Aharoni, E., Simpson, D., Nahmias, E., & Gollwitzer, M. (2022). A painful message: Testing the effects of suffering and understanding on punishment judgments. *Zeitschrift für Psychologie*, *230*(2), 138–151. https://doi.org/10.1027/2151-2604/a000460

Boonin, D. (2008). *The problem of punishment*. Cambridge University Press.

Carlsmith, K. M., & Darley, J. M. (2008). Psychological aspects of retributive justice. *Advances in Experimental Social Psychology*, *40*, 193–236. https://doi.org/10.1016/s0065-2601(07)00004-4

Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology*, *83*(2), 284–299. https://doi.org/10.1037/0022-3514.83.2.284

Carlsmith, K. M., Wilson, T. D., & Gilbert, D. T. (2008). The paradoxical consequences of revenge. *Journal of Personality and Social Psychology*, *95*(6), 1316–1324. https://doi.org/10.1037/a0012165

Casper, J. D., Tyler, T., & Fisher, B. (1988). Procedural justice in felony cases. *Law & Society Review*, *22*(3), 483–507. https://doi.org/10.2307/3053626

Curry, O. S., Alfano, M., Brandt, M. J., & Pelican, C. (2022). Moral molecules: Morality as a combinatorial system. *Review of Philosophy and Psychology*, (13), 1039–1058. https://doi.org/10.1007/s13164-021-00540-x

Curry, O. S., Mullins, D. A., & Whitehouse, H. (2019). Is it good to cooperate? Testing the theory of morality-as-cooperation in 60 societies. *Current Anthropology*, *60*(1), 47–69. https://doi.org/10.1086/701478

Daly, K. (2002). Restorative justice: The real story. *Punishment & Society*, *4*(1), 55–79. https://doi.org/10.1177/14624740222228464

Daly, K. (2005). Restorative versus retributive justice. *Criminal Justice Matters*, *60*(1), 28–37. https://doi.org/10.1080/09627250508553613

de Keijser, J. W. (2014). Penal theory and popular opinion: The deficiencies of direct engagement. In J. Ryberg & J. V. Roberts (Eds.), *Popular punishment: On the normative significance of public opinion* (pp. 101–118). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199941377.003.0006

Engelmann, N., Almeida, G. da F. C. F., Oliveira de Sousa, F., Prochownik, K., Hannikainen, I. R., Struchiner, N., & Magen, S. (2024). Apply the laws, if they are good: Moral evaluations linearly predict whether judges should enforce the law. *Cognitive Science*, *48*(10), e70001. https://doi.org/10.1111/cogs.70001

Funk, F., McGeer, V., & Gollwitzer, M. (2014). Get the message: Punishment is satisfying if the transgressor responds to its communicative intent. *Personality and Social Psychology Bulletin*, *40*(8), 986–997. https://doi.org/10.1177/0146167214533130

Gollwitzer, M., & Denzler, M. (2009). What makes revenge sweet: Seeing the offender suffer or delivering a message? *Journal of Experimental Social Psychology*, *45*(4), 840–844. https://doi.org/10.1016/j.jesp.2009.03.001

Gollwitzer, M., Meder, M., & Schmitt, M. (2011). What gives victims satisfaction when they seek revenge? *European Journal of Social Psychology*, *41*(3), 364–374. https://doi.org/10.1002/ejsp.782

Gromet, D. M., & Darley, J. M. (2006). Restoration and retribution: How including retributive components affects the acceptability of restorative justice procedures. *Social Justice Research*, *19*(4), 395–432. https://doi.org/10.1007/s11211-006-0023-7

Gromet, D. M., & Darley, J. M. (2009). Retributive and restorative justice: Importance of crime severity and shared identity in people's justice responses. *Australian Journal of Psychology*, *61*(1), 50–57. https://doi.org/10.1080/00049530802607662

Horvath, J., & Wiegmann, A. (2016). Intuitive expertise and intuitions about knowledge. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *173*(10), 2701–2726. https://doi.org/10.1007/s11098-016-0627-1

Husak, D. (2016). What do criminals deserve? In K. K. Ferzan & S. J. Morse (Eds.), *Legal, moral, and metaphysical truths: The philosophy of Michael S. Moore* (pp. 49–62). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198703242.003.0004

Husak, D. (2020). The price of criminal law skepticism: Ten functions of the criminal law. *New Criminal Law Review*, *23*(1), 27–59. https://doi.org/10.1525/nclr.2020.23.1.27

Kant, I. (1996). The metaphysics of morals (M. J. Gregor, Trans.). In M. J. Gregor (Ed.), *Practical philosophy* (pp. 353–604). Cambridge University Press. https://doi.org/10.1017/CBO9780511813306.013 (Original work published 1797)

Kleinfeld, J. (2016). Reconstructivism: The place of criminal law in ethical life. *Harvard Law Review*, *129*(6), 1485–1565.

Kleinfeld, J. (2022). *By the people: Restoring democracy in criminal justice*. SSRN. https://doi.org/10.2139/ssrn.4107451

Kolber, A. (2009). How to improve empirical desert. *Brooklyn Law Review*, *75*(2), 433–461.

McGeer, V., & Funk, F. (2017). Are 'optimistic' theories of criminal justice psychologically feasible? The probative case of civic republicanism. *Criminal Law and Philosophy*, *11*(3), 523–544. https://doi.org/10.1007/s11572-015-9381-2

Moore, M. S. (2010). *Placing blame*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199599493.001.0001

Morris, H. (1968). Persons and punishment. *Monist*, *52*(4), 475–501. https://doi.org/10.5840/monist196852436

Nahmias, E., & Aharoni, E. (2017). Communicative theories of punishment and the impact of apology. In C. W. Surprenant (Ed.), *Rethinking punishment in the era of mass incarceration* (pp. 144–161). Routledge. https://doi.org/10.4324/9781315170602-9

Pereboom, D. (2001). *Living without free will*. Cambridge University Press. https://doi.org/10.1017/cbo9780511498824

Robinson, P. H. (2003). The virtues of restorative processes, the vices of "restorative justice." *Utah Law Review*, *2003*(1), 375–388. https://doi.org/10.63140/p9i_4f6xw4

Robinson, P. H. (2008). Competing conceptions of modern desert: Vengeful, deontological, and empirical. *The Cambridge Law Journal*, *67*(1), 145–175. https://doi.org/10.1017/S00081973 0800010X

Robinson, P. H. (2017). Democratizing criminal law: Feasibility, utility, and the challenge of social change. *Northwestern University Law Review*, *111*(6), 1565–1596.

Robinson, P. H., & Darley, J. M. (1997). The utility of desert. *Northwestern University Law Review*, *91*(2), 453–499.

Robinson, P. H., & Darley, J. M. (2007). Intuitions of justice: Implications for criminal law and justice policy. *Southern California Law Review*, *81*(1), 1–68.

Robinson, P. H., Goodwin, G. P., & Reisig, M. D. (2010). The disutility of injustice. *New York University Law Review*, *85*(6), 1940–2033.

Ryberg, J., & Roberts, J. V. (2014). Exploring the normative significance of public opinion for state punishment. In J. Ryberg & J. V. Roberts (Eds.), *Popular punishment: On the normative significance of public opinion* (pp. 1–13). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199941377.003.0001

Singer, P. (1972). Is act-utilitarianism self-defeating? *The Philosophical Review*, *81*(1), 94–104. https://doi.org/10.2307/2184228

Sommers, R. (2021). Experimental jurisprudence. *Science*, *373*(6553), 394–395. https://doi.org/10.1126/science.abf0711

Strawson, P. F. (1962). Freedom and resentment. *Proceedings of the British Academy*, *48*, 1–25.

Tobia, K., Almeida, G., Prochownik, K., & Hannikainen, I. (2025). Experimental jurisprudence. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Summer 2025). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2025/entries/experimental-jurisprudence/

van Doorn, J., Zeelenberg, M., & Breugelmans, S. M. (2018). An exploration of third parties' preference for compensation over punishment: Six experimental demonstrations. *Theory and Decision*, *85*(3), 333–351. https://doi.org/10.1007/s11238-018-9665-9

Weinberg, J. M., Gonnerman, C., Buckner, C., & Alexander, J. (2010). Are philosophers expert intuiters? *Philosophical Psychology*, *23*(3), 331–355. https://doi.org/10.1080/09515089.2010.490944

Wenzel, M., Okimoto, T. G., Feather, N. T., & Platow, M. J. (2008). Retributive and restorative justice. *Law and Human Behavior*, *32*(5), 375–389. https://doi.org/10.1007/s10979-007-9116-6

Williamson, T. (2011). Philosophical expertise and the burden of proof. *Metaphilosophy*, *42*(3), 215–229. https://doi.org/10.1111/j.1467-9973.2011.01685.x

Zisman, V. (2023). *Criminal law without punishment: How our society might benefit from abolishing punitive sanctions*. De Gruyter. https://doi.org/10.1515/9783111027821

Zisman, V., & Rehren, P. (2025). The problems of empirically-informed arguments for and against retributivism. *Review of Philosophy and Psychology*, *16*(3), 805–831. https://doi.org/10.1007/s13164-024-00753-w