Philos Stud (2021) 178:1361–1379
https://doi.org/10.1007/s11098-020-01479-y

The self-effacing functionality of blame

Matthieu Queloz[1]

Published online: 22 May 2020

Abstract This paper puts forward an account of blame combining two ideas that are usually set up against each other: that blame performs an important function, and that blame is justified by the moral reasons making people blameworthy rather than by its functionality. The paper argues that blame could not have developed in a purely instrumental form, and that its functionality itself demands that its functionality be effaced in favour of non-instrumental reasons for blame—its functionality is *self-effacing*. This notion is sharpened and it is shown how it offers an alternative to instrumentalist or consequentialist accounts of blame which preserves their animating insight while avoiding their weaknesses by recasting that insight in an explanatory role. This not only allows one to do better justice to the authority and autonomy of non-instrumental reasons for blame, but also reveals that autonomy to be a precondition of blame's functionality. Unlike rival accounts, it also avoids the "alienation effect" that renders blame unstable under reflection by undercutting the authority of the moral reasons which enable it to perform its function in the first place. It instead yields a vindicatory explanation that strengthens our confidence in those moral reasons.

Keywords Blame · Responsibility · Self-effacing functionality · Instrumentalism · Alienation · Non-instrumental reasons · Vindicatory explanation · Justification

✉ Matthieu Queloz
matthieu.queloz@wolfson.ox.ac.uk

[1] Wolfson College, University of Oxford, Linton Rd, Oxford OX2 6UD, UK

1 Introduction

Those impressed by the thought that blame is a tool performing an important function in human life have tended to combine this thought with another: that blame is *justified* by its functionality.[1] Instrumentalist or consequentialist proposals involving this *Justification Claim* have a distinguished history, and they are now in the ascendant again.[2] But they continue to meet with great resistance from the majority of philosophers, who are more impressed by the thought that blame is a matter of tracking blameworthiness, and that blame is justified by *moral* reasons—reasons articulating why people *merit* blame—rather than by instrumental reasons.

My aim in this paper is to offer an irenic account that finds a place for the thought that blame performs important functions while divorcing it from the thought that blame is justified by its functionality. Blame is justified by *moral* rather than instrumental reasons, but *this fact itself* is explained by functional demands on the practice: the functionality of blame itself demands that it be justified by reasons other than the instrumental reasons that figure in its explanation; hence, the functionality of blame effaces itself in favour of moral reasons governing its appropriateness, but *for functional reasons*, making the functionality of blame *self-effacing*.

I proceed as follows: in Sect. 2, I provide a *vindicatory explanation* of blame's non-instrumental character: an explanation that is orthogonal to the justification of blame, but that can nevertheless strengthen our confidence in the practice and render us more receptive to the non-instrumental reasons at work in it by showing us why the functionality of the practice itself requires blame to be allocated non-instrumentally. This amounts to the *self-effacing functionality account* of blame, which I sharpen and define more precisely in Sect. 3. In Sect. 4, finally, I show how this self-effacing functionality account differs from other instrumentalist or consequentialist accounts in ways that preserve their strengths while avoiding their weaknesses.

2 Why bare blame does not work

Multiple functions have been attributed to blame: it has been described as a device for protesting, for developing the ability of agents to recognize and respond to moral considerations, and for generating common knowledge about what we owe to each

[1] A number of sophisticated accounts—including Vargas (2008, 2013), McGeer (2013, 2014, 2015), and Jefferson (2019)—have breathed new life into the idea that blame is justified by its functionality (at the level of the act of blame or of the practice). Though much of this paper is devoted to demarcating my position from theirs, this belies the extent to which I have profited from their work.

[2] See Schlick (1939), Nowell-Smith (1948), Smart (1961), Dennett (1984), Bennett (2002), Vargas (2008, 2013), McGeer (2013, 2014, 2015), Miller (2014, 2017), Fricker (2016, forthcoming), Tsai (2017), and Jefferson (2019), among others. Vargas (forthcoming-a) provides a historical overview. In his usage, "instrumentalist" refers to any view highlighting instrumental considerations, while "consequentialist" is reserved for views with a connection to consequentialism in normative ethics. For present purposes, however, the important distinction is that between *justifications* and *explanations* of blame in this vein.

other as a result of wrongdoing.[3] But as proponents of desert-based views have been quick to point out, instrumental considerations seem to be neither here nor there in determining when blame is appropriate—indeed, blame that is guided by instrumental considerations can appear positively disrespectful or manipulative.[4] This feature of blame—let us call it the *non-instrumental character* of blame, the fact that most of the time, the practice of blame is not understood by participants in instrumental terms—is recognized even by the most uncompromising instrumentalists, though they may consider it a flaw inviting revision.[5]

But why does blame have this non-instrumental character? Could there have been a "bare blame"-society which had a nakedly instrumental form of blame? By considering such a "bare blame"-society and identifying just why bare blame must remain ill-equipped to fulfil its function, we can identify the practical pressures driving the emergence of the non-instrumental character of blame. This will indicate good reasons to doubt recent instrumentalist attempts—by Vargas (2008, 2013), McGeer (2013, 2014, 2015), Miller (2014, 2017), and Jefferson (2019), among others—to argue that the instrumental consequences of having a practice of blame are what justifies blame.

The guiding intuition of instrumentalist approaches to blame is that blame is a device by which to regulate the behaviour of those who are out of step with the community's moral sensibility or insufficiently responsive to moral considerations.[6] Of course, not every device that aims to regulate the behaviour of offenders against the prevailing moral sensibility qualifies as a form of blame. Blame is, specifically, an instrument that helps us to regulate behaviour *by* rendering agents more responsive to moral reasons, in particular by explicitly alerting them to, and perhaps sanctioning them for, their past *failures* to respond sufficiently to particular reasons on particular occasions (as opposed to rendering them more responsive to moral reasons through other means, such as positive reinforcement strategies). Our question then is: could there be a stable form of blame that was understood entirely in such instrumental terms?

From the point of view most readily and most comfortably taken up in thinking about blame, that of the blaming party, it is not obvious why bare blame could not work. One *could* resolve to respond to actions betraying insufficient responsiveness to moral reasons with bare blame. Note that an irritated grunt or some other inarticulate expression of feeling would not do the job, for while this would communicate to the recipient that the action was unwelcome, and iterations of this

[3] See Tognazzini and Coates (2018, §1.4), Vargas (forthcoming-a), and Sliwa (forthcoming) for overviews.

[4] This is what McGeer dubs "the anti-regulation concern" (2013, p. 176).

[5] Smart (1961) and Arneson (2003) are examples. Vargas (2013) and McGeer (2015) advocate more qualified forms of revisionism. On revisionism, see McCormick (2017) and Vargas (forthcoming-b).

[6] Despite its dominating presence in philosophy, blame is far from being the unique solution to this problem. There are numerous techniques by which to remedy failures in responsiveness to moral reasons—many far more coercive than blame, as totalitarian regimes and dystopian literature remind us. Some of these techniques, such as the emulation of exemplars, are practiced alongside blame even in societies that deny themselves more coercive resources.

pattern might even succeed in communicating that the action was permanently unwelcome, it could only instil in the recipient a disposition to avoid performing the action in the presence of the blamer ("Tom doesn't seem to like it when I torture kittens; I'd better do it when he is not looking"). It would do nothing to foster a more general recognition of the *moral reason* to act otherwise. Even in its most primitive form, therefore, bare blame would need to involve some articulation of the reason one had to act otherwise.

This may already be thought to present a difficulty for bare blame, since the blamer would necessarily, in the process of blaming, point to a moral reason. But we can still conceive of the bare blamer, not altogether unrealistically, as a *hypocritical moralist*, who remains unmoved by the reason articulated in the act of blaming while nonetheless recognizing that there is an instrumental reason to get *others* to respond to that moral reason.[7] Faced with some agent *B* who disregarded some moral reason *X*, for example, the content of the thoughts of some blamer *A* could be: "Though I don't care at all about *X* myself, things go better if people in general are responsive to *X*, so if, by blaming *B*, I succeed in rendering him more responsive to *X*, I will have been instrumentally justified in blaming *B*." Such a blamer would still be acting in an entirely instrumental spirit, and in the blamer's eyes, the act of blame would be justified solely in terms of its functionality: in line with the Justification Claim, blame would be justified if and when it was effective, and unjustified otherwise.

But what about the recipient of blame? It is here that the instrumental mindset betrays its limits and we see why bare blame cannot do the work required of it: blame cannot normally be effective *unless the recipient thinks it is justified*, and unlike the blamer, the recipient *cannot* coherently understand the justification for blame as purely instrumental, which means that justified blame will need to be justified by something other than instrumental reasons. Since blamers are sometimes the blamed, this entails that blame can only be effective in a community in which each member understands blame (at least some of the time) as justified by something other than its functionality.

Let us examine the argument in more detail. It involves three steps: first, in the basic case, blame cannot be effective unless the recipient thinks it is justified, for when blame is *not* justified from the recipient's point of view, it must remain mere browbeating and cannot hope to become a successful instance of blame.[8] It might intimidate and breed fear or resentment; but it certainly could not elicit remorse, self-reproach, guilt, contrition, penitence, or other markers of *effective* blame, i.e.

[7] The case where the entire community consists of hypocritical moralists raises difficulties of its own, because while we can make practical sense of an isolated amoralist, we cannot make sense of a society of amoralists. It may be possible for an individual to live outside the ethical life, but no human community can get by without some minimal ethical consciousness that stakes claims against self-interest (Williams 2011, pp. 32, 51).

[8] See Williams (1995a, p. 15). One can of course imagine more complicated cases in which blame ends up being justified even though the recipient does not consider it justified, for instance because it initiates an exchange between blamer and recipient that results in the blamer becoming more responsive to a moral reason.

blame resulting in enhanced responsiveness to a moral reason one was previously unresponsive to.

The second step is that since getting the recipient to care about *X* is the success criterion for the act of blame, the recipient cannot remain indifferent to *X* while seeing the act of blaming as justified. The recipient of blame cannot coherently think: "Though I don't care at all about *X* myself, things go better if people in general are responsive to *X*, so since, by blaming me, *A* has rendered me more responsive to *X*, *A* was instrumentally justified in blaming me." For the recipient to come to see the act of blaming as justified, the action that forms the focus of blame must reveal itself to have been accompanied by what Simon Blackburn calls "the hovering 'should'" (2015, p. 222): there needs to be some reason to act otherwise in the offing which the recipient comes to recognize. It is the presence of that hovering "should" that responses such as remorse or self-reproach register. Blame thus cannot be effective without turning recipients' gaze back onto their actions and the moral reasons they proved insufficiently responsive to.

The third step is that once these moral reasons are disclosed to the recipient, they justify not only acting otherwise, but also *being blamed* for failing to do so. If the function of blame is to foster responsiveness to moral reasons in the recipient, the recipient could not take blame to be justified merely on the basis of its functionality, because the thought of blame's functionality would always come either *too early* or *too late*. It would come *too early* if the change that the act of blaming was supposed to effect in the recipient had not yet taken place, because then the act of blaming would not yet have proven functional; and it would come *too late* if the change had taken place but had thereby rendered the recipient sensitive to the moral reason in virtue of which the recipient was in fact blameworthy, because then the act of blaming would already be justified by something other than its functionality. If what blame is an instrument *for* is cultivating responsiveness to moral reasons in the recipient, therefore, it can never be justified without becoming justified by something other than instrumental reasons, for it cannot succeed without alerting the recipient to the moral reasons in virtue of which the recipient is blameworthy, and it cannot alert the recipient to these reasons without rendering the act of blame justified *by* those reasons in the eyes of the recipient.

The key insight brought out by considering bare blame is thus that in order to *be* instrumental in regulating behaviour, the practice of blame must, at least in the eyes of its recipients, be *more* than instrumental, for as long as the participants in the practice of blame think of it in purely instrumental terms, as justified only by its efficacy as a tool, blame will normally fail to be effective. This highlights something that is often missed by accounts which duly acknowledge that blame combines instrumental and non-instrumental aspects, namely the *functional connection* between instrumental and non-instrumental aspects of blame.

The non-instrumental character of blame thus forms what might be called a *precondition of functionality*:

> *Precondition of functionality*:
> Some state of affairs *p* is a precondition of the functionality of something discharging function *F* iff *F* can be discharged only if *p* is the case.

We can then specify the functional connection between the instrumental and the non-instrumental aspects of blame as follows: blame's having a non-instrumental character is a precondition of the functionality of blame as a tool for the cultivation of responsiveness to moral reasons. To grasp this functional connection is to grasp the inherently self-reinforcing nature and practical value of blame's peculiar combination of instrumental and non-instrumental aspects. By grasping it, we not only acknowledge the non-instrumental character of blame, but are also made comfortable with it, because it receives a demystifying and vindicatory explanation: blame turns out to be a practice exhibiting *self-effacing functionality*.[9]

3 Self-effacing functionality

To say that blame is a self-effacingly functional practice is to say that it is functional, but only insofar as it is sustained by reasons that are autonomous, i.e. not conditional on the practice's functionality. As a result, the functionality of blame will be either secondary or entirely absent from the participants' minds as they engage in the practice, but *for functional reasons*. Hence the claim that the functionality of blame is self-effacing.

This claim can be stated more precisely by saying that the practice of blame satisfies the following four conditions[10]:

(a) *Functionality*: the practice is *functional*, i.e. it makes a difference to the lives of those who engage in it which can be shown to be a *useful difference* in light of some of their individual or social needs.

(b) *Autonomy*: the practice is guided and sustained by motives and reasons that are not conditional on its functionality and are in that sense *autonomous*.

(c) *Dependence*: the practice can be functional *only insofar* as it satisfies (b), i.e. it would be ineffective if guided and sustained merely by motives and reasons that were conditional on its functionality.

(d) *Explanatory connection*: the practice fulfils (b) *because of* (c), i.e. there is an explanatory connection between its autonomy and the fact that its functionality depends on its autonomy. This demarcates cases of self-effacing functionality from cases of contingently effaced functionality, in which a practice becomes guided and sustained by autonomous considerations for reasons that have nothing to do with the practice's functionality (an originally instrumentally motivated practice might be harnessed and reinterpreted by a religious movement, for example).

This account of blame as exhibiting the FADE structure (Functionality, Autonomy, Dependence, and an Explanatory connection between the latter two) explains why

[9] It thus stands with blame much as it stands with the practice of valuing the truth intrinsically according to Williams (2002), though Williams does not put it in terms of self-effacing functionality. See Queloz (2018, forthcoming).

[10] Here I draw, with modifications, on Queloz (2018, p. 14).

its functionality is not obvious to participants, having faded from view because that very functionality requires that it efface itself in favour of autonomous considerations. Yet because of (d), the explanatory connection, the functionality is not *contingently* effaced, as it would be if the practice of blame had developed a non-instrumental character for reasons that had nothing to do with its underlying function. Nor does the self-effacing functionality account correspond to the functional dynamics familiar from ideology critique, where awareness of a practice's function is *radically incompatible* with confident engagement in it. On the self-effacing functionality account, blame does not necessarily involve blindness to its function; participants *can* become fully conscious of the functionality of the practice without the insight into the functionality having a destabilizing effect on the practice (a point I return to in the next section).

The self-effacing functionality account of blame is *irenic* because it combines the thought that blame performs a function with the thought that instrumental considerations fail to capture the nature and justification of blame: it preserves the insight that blame serves a function and even extends its reach by exhibiting the non-instrumental aspects of blame as possessing a functionalist rationale; yet it also vindicates those who resist a functionalist understanding of blame by underscoring the poverty of *pure* functionalism: an account which insisted on understanding blame entirely in instrumental terms would find itself at a loss to explain blame's functionality.[11]

4 Upshot and advantages of the self-effacing functionality account

The mark of a truly irenic account, a cynic might say, is that it proves a disappointment to all parties. Each side recognizes enough material in the precariously balanced middle position—either its own or that of the opposing side—to conclude that it must ultimately collapse one way or the other. For the self-effacing functionality account of blame, this takes the form of asking either why the position does not collapse into a purely desert-based view that is oblivious to function, or why it does not collapse into a revisionist instrumentalism which takes the insight into the function to be what really justifies blame and undertakes to overrule desert-based considerations wherever this optimizes functionality.

[11] This provides us with a distinctive way of developing the notoriously elusive ending of P. F. Strawson's "Freedom and Resentment": "It is far from wrong to emphasize the efficacy of all those practices which express or manifest our moral attitudes, in regulating behaviour in ways considered desirable," he writes. "What *is* wrong is to forget that these practices, and their reception, the reactions to them, really *are* expressions of our moral attitudes and not merely devices we calculatingly employ for regulative purposes," for "the very understanding of the kind of efficacy these expressions of our attitudes have turns on our remembering this" (1974, p. 25). If we wish to understand the efficacy of blame, in other words, we must recognize that it is efficacy of a *peculiar kind*, one that we cannot understand unless we recognize that the practice really expresses *moral* and not just instrumental attitudes. This is more than just the suggestion that we need a two-eyed view of blame—a view that duly takes note of both its instrumental and its non-instrumental character. It involves the further idea that we need such a two-eyed view *in order* to understand the peculiar "kind of efficacy" at work in these cases, which suggests that the mark of that efficacy is the *functional connection* between the instrumental and the non-instrumental—a connection which can be more fully brought to light using the notion of self-effacing functionality.

In addressing these concerns, the critical question is how we conceive of the normative *role* and relative *weight* of the instrumental and non-instrumental reasons for blame: do they *explain* or *justify*, and if the latter, *how much* justificatory weight do they possess? While we have so far considered why blame, if understood in instrumental terms, must come to be understood also in non-instrumental terms, little has been said about the relation between these different understandings of blame. If the instrumental considerations are taken to be completely normatively inert, the self-effacing functionality account indeed threatens to collapse into a purely desert-based view. If, on the other hand, instrumental considerations are taken to be so normatively powerful that they overrule non-instrumental considerations and provide the true justification for blame (either at the level of individual acts of blaming or at the level of the practice), the account collapses into the kind of instrumentalist position defended notably by McGeer (2013, 2015) and Vargas (2013, 2015a, b).

Accordingly, my aim in this section is to develop a stable and distinctive account of how, on the self-effacing functionality account, the normative role and relative weight of instrumental and non-instrumental reasons for blame can be understood. This will allow us to contrast it with rival accounts and bring out its comparative strengths. I shall argue that the self-effacing functionality account respects the default *authority* of our non-instrumental reasons for blame by recasting the insight into the functionality of blame in an *explanatory* role; that unlike rival accounts, it does not miss the *functional connection* between instrumental and non-instrumental aspects of blame; and that it avoids what I shall call the *alienation effect* characteristic of accounts committed to the Justification Claim.

4.1 The authority of function-based accounts: explanation versus justification

By itself, the insight into the self-effacing functionality of blame does not fully determine what normative role and weight we should give to it. In thinking about this, it is important to bear in mind a remark of Bernard Williams's, to the effect that when we assess the strengths and weaknesses of some theoretical account of a practice, we ultimately do so by reference to our everyday judgements and our pre-theoretical sense of the life that the theoretical account was supposed to help us to lead, and in many cases, the reasons that a practice immediately provides will often "simply count as stronger than any reason that might be advanced for it" (2011, p. 127).[12] Theoretical explanations or justifications of our everyday practices do not possess more authority simply in virtue of being theoretical or systematic. What authority they may have comes from the way they tie in with what is or should be important to us *already*.

Where the practice of blame is concerned, it likewise seems to me that the reasons that guide and flow from our everyday first-order judgements of blameworthiness *simply count as stronger* than the reasons that guide and flow from our theoretical second-order judgements about those judgements. This is not to

[12] Williams's example is a theoretical account of the concept *person* that is supposed to give us reasons to revise our ethical priorities in our dealings with each other, but ends up seeming less authoritative than the concepts and reasons it is meant to replace (2011, pp. 126–128).

deny that there *can* be theoretical accounts of practices that end up sapping the authority from our first-order reasoning practices (by giving us compelling reasons to think that our adherence to those practices involves some radical form of deception, for instance).[13] But the onus is on the side of theory to show why it should be granted more authority than the reasons we immediately have for thinking that someone does or does not merit blame.

I suspect that appreciation of this fact is at the root of much discomfort with instrumentalist or consequentialist theories, and in particular with the Justification Claim. The moment the functionality of blame is cast in a justificatory role and presented as being what *really* justifies blame—either the act of blame, as in McGeer's (2013, 2015) theory, or the practice of blame, as in Vargas's (2013, 2015a, b)—one is asked to grant instrumental considerations *more authority* than the non-instrumental considerations that guide our everyday judgements of blameworthiness, and to revise one's blaming practices accordingly. But even if one is willing to grant them *some* authority, one understandably feels queasy, in the absence of a fuller account of why the instrumental considerations advanced by a philosophical theory should have such justificatory weight, about granting them *so much* authority over one's life.

For this reason, it seems to me better to recast the insight into the functionality of blame in a different role and to relocate it from the sphere of justification to the sphere of explanation. The practical needs of human beings trying to live together explain why they developed the practice of blame and why it comes to be guided by non-instrumental reasons. This is an account of blame in terms of its consequences, but it is an *explanatory* consequentialism that issues in a non-consequentialist picture of blame. The instrumental considerations it advances remain orthogonal to the business of justifying blame. Like the accounts of McGeer and Vargas, the self-effacing functionality account is function-based, but the insight into the functionality of blame is situated along the axis of explanation, buttressing rather than competing with the reasons at work in the game of justifying blame.[14]

This is not to say that if the insight into the functionality of blame is situated along the axis of explanation rather than justification, it leaves everything as it was; on the contrary, the account of blame as a self-effacingly functional practice is a prime example of what Williams calls a *vindicatory explanation*: an explanation that strengthens our confidence in the practice of blame and in the reasons we take

[13] Under which conditions reasons yielded by a theoretical account have sufficient authority to overrule reasons that are more directly at work in our experience is a complex issue I shall not go into here. Suffice it to say that I do take the authority of first-order reasons to be defeasible.

[14] Note that this approach also remains distinct from evolutionary approaches drawing on Ernst Mayr's distinction between *proximate* and *ultimate* explanations: proximate explanations concern the mechanism by which certain effects are generated and thus answer "How?"-questions, whereas ultimate explanations concern the evolutionary function of having a mechanism with these effects and thus answer "Why?"-questions (Scott-Phillips et al. 2011). But such evolutionary approaches, like other instrumentalist or consequentialist accounts, invite one to see the moral reasons tracked by blame merely as the mechanism by which blame operates, and the regulatory function of blame as the real rationale or justification for having such a mechanism. They thus also subscribe to the Justification Claim which the self-effacing functionality account rejects.

to justify it, because it suggests that this practice, along with the reasons that figure in it, is not just a fetish, a product of deception, or an irrational holdover from the enchanted world, but something that it *makes sense* for us to cultivate, because it provides an important solution to a problem that any human society needs to solve on a continual basis.[15]

Indeed, such an explanation might even be taken to indicate that blame makes particular sense *for us* who live in heterogenous liberal societies. If blame responds to a need for moral alignment, a need to cultivate responsiveness to the same moral reasons, it follows that the more internally diverse societies are, the more there is a need for techniques of this kind, especially in the absence of more draconian ways of responding to that need. In societies that are ethically highly homogenous or that have no qualms about resorting to coercion or force to secure alignment, there may be less of a need for blame. But to the extent that a society is pluralistic and heterogenous and that it denies itself the use of many alternative tools, that society will have a correspondingly greater need for blame.

At the same time, explanatory reflection on the function of blame is also not without a critical edge. As Fricker (2016, forthcoming) has argued, seeing the point of blame also gives us a sense of when blame becomes pointless or even dysfunctional. This is not the same as showing that an instance of blame is unjustified. Nor does it show that the entire practice of blame is unjustified. But it can drain our confidence in particular excrescences of the practice, particularly where they take the form of a *fiat justitia*, *ruat coelum* attitude ("Let there be justice, even should the heavens fall!"). By rendering us more receptive to the thought that blame may not be our best option even where it is fully justified, a reflective understanding of why we go in for blame can inform the way we engage in it even if it does not provide direct input to blame's justificatory logic.

The first important difference between the self-effacing functionality account and accounts subscribing to the Justification Claim is therefore this: for the latter, the instrumental considerations discernible from an external perspective on blame *justify* what we see from the internal perspective, whereas for the self-effacing functionality theorist, authority resides primarily in the non-instrumental considerations to which blame is taken to answer inside the practice. The instrumental considerations visible from the external perspective serve in the first instance not to justify, but to *explain* the way we think as participants in the practice of blame.[16]

[15] See Williams (1999, p. 258; 2002, pp. 36–37, 263; 2009, pp. 198–210; 2014, p. 410). As Williams observes (2002, p. 283n219), his usage of the term is broader than David Wiggins's.

[16] The contrast is even starker if the Justification Claim is coupled with a consequentialism that applies systematically across ethics: whatever authority the non-instrumental reasons for blame possess must then derive from the fact that treating these considerations as reasons appears, at the reflective level, conducive to the production of the states of affairs that are the *real* justifiers. For the self-effacing functionality theorist, by contrast, the considerations that guide us at the first-order level are independently—if defeasibly—authoritative in a way that does not depend on validation by the consequentialist calculus.

4.2 The functional connection

The second difference is that while instrumentalist or consequentialist accounts typically acknowledge that blame involves both instrumental and non-instrumental aspects, they miss the *functional connection* between the two aspects, making it look like a mere contingency of history that blame should display this combination of aspects. By contrast, the insight into the self-effacing functionality of blame is the insight into *why* the justification and the functionality of blame *must* come apart: why it would *not* serve just as well to have an understanding of blame which *looked through* the moral features actually tracked by blame to the benefits of tracking these features.

This functional connection lends succour to the idea that the insight into the functionality of blame is best cast in an explanatory role, because it shows, against the Justification Claim, that there are good functional reasons why the justification for blame not only is, but *should be* a moral rather than an instrumental one. Since blame can only perform its function if recipients in particular see it as justified by moral reasons, the suggestion that its justification could be based instead in its instrumental value then emerges as a conflation of the explanatory and the justificatory which threatens to rob blame of its functionality. It is inherent to the functional demands on blame as a tool by which to cultivate responsiveness to moral reasons that its successful deployment must introduce a new, non-instrumental standard of justification that swamps the explanatorily prior, instrumental justification for blame.

4.3 The alienation effect

The third difference, finally, is that instrumentalist or consequentialist accounts struggle to accommodate a tension between the spirit involved in the practice of blame and the spirit involved in grasping its functionality, whereas the self-effacing functionality account generates no such tension in the first place. In saying that there is a tension in this area, we must work carefully to recover the relevant idea from the rubble of past wars over utilitarianism, naturalism, functionalism, or evolutionary psychology.[17] To bring out the tension at issue, let us first assume that the internal view we normally take of blame as participants in the practice is the following:

> *Internal view*:
> Blame is justified iff someone is blameworthy in virtue of having exhibited insufficient responsiveness to some moral reason.

Given this internal view of blame, there are then various external views of blame with which it might be paired. The one that most plainly generates a tension with the internal view is what might be called the external view of *pure instrumentalism*:

[17] See Stocker (1976), Railton (2003), Sagar (2014), McGeer (2014), Doris (2015), Smyth (2019), and Jefferson (2019). With the exception of Smyth's treatment, however, the tensions addressed all seem to me subtly different from the one at issue here.

> *Pure instrumentalism*:
> Blame is justified iff it produces certain desirable consequences.

Here there clearly is a tension, because the external view of the justification for blame straightforwardly conflicts with the internal view: either blame is justified only insofar as it tracks blameworthiness, or it is justified only insofar as it produces certain consequences, but it cannot be both. One way to resolve this tension is to adopt a ferociously revisionist instrumentalism maintaining that the external view should *displace* the internal one, so that blame comes to be understood in instrumental terms all the way down. But while this would resolve the tension, we saw that there is reason to think that it would do so at the cost of blame's functionality. Some way must therefore be found to accommodate the internal, desert-based view of blame *within* the external, function-based one.

The external views of blame proposed by McGeer and Vargas can be seen as doing precisely that. They are examples of what might be called the external view of *accommodating instrumentalism*:

> *Accommodating instrumentalism*:
> Blame is justified iff it produces certain desirable consequences, but it does this by being understood from the internal perspective as being justified iff someone is blameworthy in virtue of having exhibited insufficient responsiveness to some moral reason.

Though better concealed, there remains a tension here, because from the external perspective, blame is really justified only by its functionality, and while this is taken to justify *treating* blame as justified by something other than its functionality, it does not present blame as *being* justified by something other than its functionality. From the external perspective, the practice of treating blame as justified by something other than its functionality must appear as a mere useful pretence. This is not to say that it also appears this way from the internal perspective; we can grant that the internal view of blame, considered by itself, can draw on enough conceptual and emotional resources to take the thought that blame is non-instrumentally justified beyond mere pretence. The problem arises rather from the *combination* of the internal with the external view, because the external view of the internal justification for blame undercuts it by presenting it as merely instrumental, thus *alienating* one from the normative bindingness of the moral reasons justifying blame from the internal view. This alienation effect is not altogether unlike what the playwright Bertolt Brecht called the *Verfremdungseffekt*, which is often translated as the distancing or alienation effect, but I more particularly have in mind the effect whereby reflection on why we came to think of blame as we do saps the normative authority of the reasons internal to the practice of blame by revealing a deeper justification for that practice.[18]

[18] Smyth speaks in this context of "reflectively induced practical alienation" (2019, p. 193). Brandom also speaks of alienation to refer to the threat that the insight into our own role in instituting norms constitutes for our appreciation of their normative force or bindingness (2019, p. 30). See also Railton (1984) for a differentiation between various specific forms of alienation that do not quite align with the one at issue here. Schacht (1970) traces the term's rich career from Hegel via Marx to the existentialists.

This tension in the instrumentalist or consequentialist account is one that McGeer (2014, §3), for example, takes seriously. The strategy that McGeer recommends to accommodate it follows Philip Pettit's proposal that we should immerse ourselves fully in the internal view of blame while *outsourcing* the control that is supposed to come from external reflection: we should rely on "red lights"—cues from the environment indicating that a situation is abnormal—to prompt ascent to the level of consequentialist reflection only when necessary.[19]

Some of the most forceful doubts over whether this could be done have been voiced by Williams. As Williams points out, an account of blame subscribing to the Justification Claim "most naturally fits a situation in which those who understand the justification, and those whose behaviour is being modified, are not the same people" (1995a, p. 15). It yields a "Government House" account of blame akin to what Williams dubbed Government House utilitarianism (1995b, p. 166), and such a situation must be "inconsistent with ideals of social transparency" (1995a, p. 15).

To sidestep such concerns about transparency, a second possibility—of which the strategy advocated by McGeer is a variant—is to compartmentalize the individual consciousness, relieving the tension using some distinction between theory and practice or reflection and action: keeping one's gaze firmly fixed on considerations of blameworthiness when absorbed in the practice, one would then only consider blame's efficacy off-duty, in the "cool hour" reflection, or only when prompted by "red lights." But Williams made it clear that on his view, such compartmentalizing distinctions possess no real "saving power" (1995b, p. 165). If there is a tension, it will manifest itself even in such a compartmentalized mind in the long run.

A third possibility is to differentiate, not between two sets of people or two stretches of time, but between two *styles of thought* that the same person can engage in simultaneously. But here also, Williams argues, the same tension is bound to manifest itself between the view we take of things in the consequentialist style of thought and the view we take of things in our intuitive responses and dispositions, because both styles of thought have to be part of the same life, which raises the question of how they can be integrated. Merely driving a wedge between internal and external thinking about blame by saying that it happens at different times or in different styles will not suffice to eliminate this tension. It may prevent confusions and conflations, but the question of how these two perspectives *relate* to each other cannot be shirked indefinitely: at the end of the day, the two perspectives have to be integrated into a single life.

The tension stems from the fact that our responses, attitudes, and dispositions are not just black-box mechanisms that can be exhaustively described in terms of their inputs and outputs. As Williams urges in an echo of Strawson, they "constitute a way of seeing the situation," and the problem is that "you cannot combine seeing the situation in that way, *from* the point of view of those dispositions" (2006, p. 80), with seeing it in the consequentialist style, in which the dispositions are merely a means towards the bringing about of some valuable effects or states of affairs. As soon as you see blame not merely as *explained* by its functionality, but as *justified*

[19] See Pettit (2012) and McGeer (2014).

by it—as soon, that is, as you take on board the Justificatory Claim—the effects of blame are presented to practical deliberation in the same normative role as moral reasons, competing with them and compromising their autonomy as justificatory input to the practice of blame. And what are recipients of blame then to think, when they consider whether the blame is justified? The combination of moral and functional justification is bound to unravel under reflection, because taking blame to be justified by its functionality alienates one from the autonomous moral reasons for blame, and alienation from those reasons robs blame of its functionality.

Crucially, the claim is not that you cannot combine seeing the situation from a point of view that focuses on the effects of having certain dispositions and seeing it from a point of view that focuses on the moral reasons visible from those dispositions. That is a combination attempted by any broadly functionalist reflection on moral dispositions, including the self-effacing functionality account of blame.

The claim is rather this: while participants in the practice of blame can unproblematically instrumentalize blame by coming to see the attitudes and dispositions constituting the machinery of blame *as* a machinery, they cannot unproblematically come to see it as *justified* by its effects as a machinery, because the relevant attitudes and dispositions are expressed precisely in seeing blame *as justified by moral reasons*, and the kind of normative bindingness involved in autonomous moral reasons simply does not admit of regulatory supervision by reflection on the efficacy of heeding them. Moral practices like blame resist being understood in instrumental terms, because the dispositions, concerns, and standards they embody demand to be understood independently from instrumental concerns, as making normatively binding claims on us that are characterized by their practical necessity or unconditionality.[20] They are precisely *not* conditional on whether heeding them facilitates the satisfaction of antecedent ends. If, from the external perspective, blame presents itself to me as justified by its consequences, then, by accepting this justification, I lose my sense that blame is unconditionally a matter of blameworthiness. I may continue to accept that blame performs its function if *treated as* justified by moral reasons, but I can no longer genuinely take it to be so justified if, all the while, I keep one eye cocked on the consequences of doing so. Taking blame to be justified by its effects thus alienates me from the moral reasons for blame by undercutting their autonomy and authority.

This holds even if one distinguishes, as Vargas (forthcoming-a) does, between the broad church of instrumentalism and more particular denominations of consequentialism—where the former, unlike the latter, include approaches that see blame as justified by its consequences *without* being committed to the more general idea that anything that has value has it, ultimately, in virtue of its

[20] Smyth (2019, p. 193), drawing on Williams (1981, 1995b, 2002, p. 91), offers the following helpful description: “A settled disposition to avoid violence simply *is* the instinctive refusal to commit acts of violence; once it is regulated by reflection on general consequences, it loses the sense of ‘practical necessity’ which, Williams claimed, accompanies our most basic ethical dispositions. This is what he meant when he described such dispositions as having ‘momentum,’ or, elsewhere, ‘a certain depth or thickness’: phenomenologically, they appear as convictions that a certain behavior must or must not be performed”.

consequences. It is true that the alienation effect becomes even stronger in light of the consequentialist idea that blame's justification in terms of its consequences is the *only* justification there can ultimately be for it, and that whatever authority non-instrumental reasons possess must derive from the fact that treating these considerations as reasons promotes valuable consequences. This consequentialist commitment, one might say, renders the Justification Claim *exhaustive*, displacing any justification for blame that does not derive from its functionality. But even without this consequentialist commitment, the Justification Claim engenders the alienation effect, because the moral reasons justifying blame themselves demand to be understood as justifying it exhaustively.

Miller (2014) has recently defended a consequentialist approach to moral psychology against this objection by appealing to a Strawsonian view of reactive attitudes. On Miller's account, the Williamsian charge overlooks the *resilience* of reactive attitudes. "None of us, not even the utilitarians among us, can avoid experiencing the reactive attitudes in the ordinary way the vast majority of the time" (2014, p. 54), so that occasionally taking a consequentialist view of them, even if it means that we briefly experience them differently (as bare feelings, for instance), will not and cannot prevent us from going back to experiencing them as people ordinarily do. If consequentialists always combined both styles of thought, Miller grants, Williams's charge would be successful (2014, p. 55). But to suppose that this is possible is, on Miller's view, once again to overlook the resilience of our reactive attitudes. We can put them in abeyance only briefly, and cannot avoid experiencing them fully the vast majority of the time.

Even supposing this line of defence to be successful in fending off Williams's charge, it could yield but a Pyrrhic victory, since in establishing the powerlessness of consequentialist thinking to destabilize or otherwise alter our reactive attitudes in any lasting fashion, one also establishes its powerlessness to revise them for the better, thus rendering consequentialist reflection pointless to begin with. In fact, however, the line of defence cannot be successful, since it tries to put a psychological roadblock in the way of a philosophical claim:

> The point is that the thoughts are not stable under reflection; in particular, you cannot think in these terms if at the same time you apply to the process the kind of thorough reflection that this theory itself advocates. That is not a merely psychological claim. It is a philosophical claim, about what is involved in effective and adequate reflection on these particular states of mind. (Williams 2006, p. 80)

In insisting not that the thoughts are unstable, but that they are unstable *under reflection*, and that this is intended as a *philosophical* rather than as a psychological claim, Williams is highlighting that the objection is aimed at the agent's *idealized* process of rational deliberation. As a matter of psychology, our reactive attitudes may be resilient or hard to dislodge, and this may be a blessing or a curse; but Williams's point is that even if these obstacles *could* be overcome and one *were* to "apply to the process the kind of thorough reflection that this theory itself advocates," one would run into the alienation effect and functional instability would ensue, because one would undercut the authority of the very reasons that allow blame to stably perform its function in the first place.

As against this, the self-effacing functionality account of blame offers an external view of blame that can harmoniously and even comfortingly take its place alongside the internal view:

> *Self-effacing functionality account of blame*:
> The explanation for why we think of blame as justified iff someone is blameworthy in virtue of having exhibited insufficient responsiveness to some moral reason is that we need to think this way in order for blame to make a useful difference to human life, but though it can assuage certain doubts about blame, this explanation is orthogonal to the question of whether and when blame is justified.

Here we avoid the alienation effect, because while the external view helps us make sense of why we go on as we do, there is nothing in the external view that conflicts with the view of blame we take from the inside. On the contrary, this external view, unlike the others, offers an account of the relation between the internal and the external view and helps us understand why they differ as they do—the dynamics of self-effacing functionality render intelligible how the non-instrumental reasons for blame relate to the instrumental reasons for blame by explaining the former in terms of the latter.

The two distinctive features of the self-effacing functionality account which allow it to avoid the alienation effect are thus, first, that it helps us *integrate* the inside and the outside perspective into one life, and second, that it casts the insight into blame's functionality in an explanatory rather than a justificatory role, thereby taking care not to harness the non-instrumental reasons to the instrumental reasons for blame. The internal justification for blame then receives explanatory support rather than competition from reflection on the role of blame in human life.

5 Conclusion

My aim in this paper has been to derive, from reflections on why blame could not have developed in a purely instrumental form, an account of blame as a self-effacingly functional practice. This account turns on the idea that blame's functionality presupposes its justification, and therefore its justification cannot be identical with its functionality. In developing that account, I have cast instrumental reasons for blame in an explanatory rather than a justificatory role, thus allowing us to do better justice to the justificatory authority and autonomy of non-instrumental reasons for blame; this is not only truer to the phenomenology of blame, but also goes beyond the mere acknowledgment that there are both instrumental and non-instrumental reasons for blame in highlighting the functional connection between them; and it avoids a problem that other function-based accounts of blame tend to suffer from, of rendering blame unstable under reflection by giving rise to the alienation effect.

The resulting picture does not justify blame as a practice; nor does it justify individual acts of blame. But it does offer a vindicatory explanation of why it makes sense for us—and especially for us—not just to cultivate the practice of blame in some form, but to cultivate it in a form that is guided and justified by moral rather than instrumental reasons. This shows instrumentalists or consequentialists that, even on their own terms, it makes sense to resist understanding the justification for blame in instrumental terms, because on this understanding of what blame tries to achieve, there is simply no logical space for it to succeed. Even from a benefit-minded perspective, therefore, we are vindicated in being bloody-minded rather than benefit-minded about blame.

Acknowledgements I am indebted to Julia Wagner, Damian Cueni, Friedemann Bieber, Ben Waltmann, Miranda Fricker, Markus Wild, Nicholas Smyth, and Simon Blackburn for valuable discussions of these issues. Comments from anonymous reviewers have also greatly improved the manuscript. My work on this paper was supported by grant No. P2BSP1_184477 of the Swiss National Science Foundation.

Bibliography

Arneson, R. J. (2003). The smart theory of moral responsibility and desert. In S. Olsaretti (Ed.), *Desert and justice* (pp. 233–258). Oxford: Oxford University Press.

Bennett, C. (2002). The varieties of retributive experience. *The Philosophical Quarterly, 52*(207), 145–163.

Blackburn, S. (2015). Williams, Smith, and the peculiarity of piacularity. *Journal of the American Philosophical Association, 1*(2), 217–232.

Brandom, R. (2019). *A spirit of trust: A reading of hegel's phenomenology*. Cambridge, MA: Harvard University Press.

Dennett, D. C. (1984). *Elbow room: The varieties of free will worth wanting*. Cambridge, MA: MIT Press.

Doris, J. M. (2015). Doing without (arguing about) desert. *Philosophical Studies, 172*(10), 2625–2634.

Fricker, M. (2016). What's the point of blame? A Paradigm Based Explanation. *Noûs, 50*(1), 165–183.

Fricker, M. (forthcoming). Forgiveness: An ordered pluralism. *Australasian Philosophical Review, 3*(1).

Jefferson, A. (2019). Instrumentalism about moral responsibility revisited. *The Philosophical Quarterly, 69*(276), 555–573.

McCormick, K. (2017). Revisionism. In K. Timpe, M. Griffith, & N. Levy (Eds.), *The Routledge companion to free will* (pp. 109–120). London: Routledge.

McGeer, V. (2013). Civilizing blame. In D. J. Coates & N. A. Tognazzini (Eds.), *Blame: Its nature and norms* (pp. 162–188). New York: Oxford University Press.

McGeer, V. (2014). P. F. Strawson's consequentialism. In D. Shoemaker & N. Tognazzini (Eds.), *Oxford studies in agency and responsibility* (Vol. 2, pp. 64–92). Oxford: Oxford University Press.

McGeer, V. (2015). Building a better theory of responsibility. *Philosophical Studies, 172*(10), 2635–2649.

Miller, D. E. (2014). Reactive attitudes and the Hare-Williams debate: Towards a new consequentialist moral psychology. *The Philosophical Quarterly, 64*(254), 39–59.

Miller, D. E. (2017). 'Freedom and resentment' and consequentialism: Why 'Strawson's point' is not Strawson's point. *Journal of Ethics and Social Philosophy, 8*(2), 1–23.

Nowell-Smith, P. (1948). Freewill and moral responsibility. *Mind, 57*(225), 45–61.

Pettit, P. (2012). The inescapability of consequentialism. In U. Heuer & G. Lang (Eds.), *Luck, value and commitment: Themes from the ethics of Bernard Williams* (pp. 41–70). Oxford: Oxford University Press.

Queloz, M. (2018). Williams's pragmatic genealogy and self-effacing functionality. *Philosophers' Imprint, 18*(17), 1–20.

Queloz, M. (forthcoming). *The practical origins of ideas: Genealogy as conceptual reverse-engineering*. Oxford: Oxford University Press.

Railton, P. (1984). Alienation, consequentialism, and the demands of morality. *Philosophy & Public Affairs, 13*(2), 134–171.

Railton, P. (2003). *Facts, values, and norms: Essays toward a morality of consequence*. Cambridge: Cambridge University Press.

Sagar, P. (2014). Minding the gap: Bernard Williams and David Hume on living an ethical life. *Journal of Moral Philosophy, 11*(5), 615–638.

Schacht, R. (1970). *Alienation*. New York: Doubleday.

Schlick, M. (1939). When is a man responsible? (D. Rynin, Trans.). In *The problems of ethics* (pp. 143–156). New York: Prentice-Hall.

Scott-Phillips, T. C., Dickins, T. E., & West, S. A. (2011). Evolutionary theory and the ultimate-proximate distinction in the human behavioral sciences. *Perspectives on Psychological Science, 6*(1), 38–47.

Sliwa, P. (forthcoming). Reverse-engineering blame. *Philosophical Perspectives*.

Smart, J. J. C. (1961). Free will, praise and blame. *Mind, 70,* 291–306.

Smyth, N. (2019). The inevitability of inauthenticity: Bernard Williams and practical alienation. In S. G. Chappell & M. van Ackeren (Eds.), *Ethics beyond the limits: New essays on Bernard Williams' ethics and the limits of philosophy* (pp. 188–208). London: Routledge.

Stocker, M. (1976). The schizophrenia of modern ethical theories. *Journal of Philosophy, 73*(14), 453–466.

Strawson, P. F. (1974). Freedom and resentment. In *Freedom and resentment and other essays* (pp. 1–28). London: Routledge.

Tognazzini, N., & Coates, D. J. (2018). Blame. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2010 ed.). Stanford: Stanford University.

Tsai, G. (2017). Respect and the efficacy of blame. In D. Shoemaker (Ed.), *Oxford studies in agency and responsibility* (Vol. 4, pp. 248–276). Oxford: Oxford University Press.

Vargas, M. (2008). Moral influence, moral responsibility. In N. Trakakis & D. Cohen (Eds.), *Essays on free will and moral responsibility* (pp. 90–122). Newcastle: Cambridge Scholars Publishing.

Vargas, M. (2013). *Building better beings: A theory of moral responsibility*. Oxford: Oxford University Press.

Vargas, M. (2015a). Desert, responsibility, and justification: A reply to Doris, McGeer, and Robinson. *Philosophical Studies, 172*(10), 2659–2678.

Vargas, M. (2015b). Précis of building better beings: A theory of moral responsibility. *Philosophical Studies, 172*(10), 2621–2623.

Vargas, M. (forthcoming-a). Instrumentalist theories of moral responsibility. In G. D. Caruso (Ed.), *The Oxford handbook of moral responsibility*. Oxford: Oxford University Press.

Vargas, M. (forthcoming-b). Revisionism. In J. Campbell (Ed.), *A companion to free will*. Oxford: Wiley-Blackwell.

Williams, B. (1981). Practical necessity. In *Moral luck* (pp. 124–131). Cambridge: Cambridge University Press.

Williams, B. (Ed.). (1995a). How free does the will need to be? In *Making sense of humanity and other philosophical papers, 1982–1993* (pp. 3–21). Cambridge: Cambridge University Press.

Williams, B. (Ed.). (1995b). The point of view of the universe: Sidgwick and the ambitions of ethics. In *Making sense of humanity and other philosophical papers, 1982–1993* (pp. 153–171). Cambridge: Cambridge University Press.

Williams, B. (1999). Seminar with Bernard Williams. *Ethical Perspectives, 6*(3–4), 243–265.

Williams, B. (2002). *Truth and truthfulness: An essay in genealogy*. Princeton: Princeton University Press.

Williams, B. (2006). The structure of Hare's theory. In A. W. Moore (Ed.), *Philosophy as a humanistic discipline* (pp. 76–85). Princeton: Princeton University Press.

Williams, B. (2009). A mistrustful animal. In A. Voorhoeve (Ed.), *Conversations on ethics* (pp. 195–214). Oxford: Oxford University Press.

Williams, B. (2011). *Ethics and the limits of philosophy*. London: Routledge.

Williams, B. (2014). Why philosophy needs history. In M. Woods (Ed.), *Essays and reviews 1959–2002* (pp. 405–412). Princeton: Princeton University Press.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.