The Journal of Value Inquiry
https://doi.org/10.1007/s10790-024-10011-y

When to Fill Responsibility Gaps: A Proposal

Michael Da Silva[1]

Accepted: 12 September 2024

1 Introduction

Consider the following:

> *Crash*: A ship travels on an approved path presenting greater risks than alternatives. It receives routine, if sloppy, maintenance and checks. The crew receives training. Vessel and crew defects contribute to its crashing and sinking. Crew members perish. Traditional accounts of responsibility determine no individual or combination thereof is fully responsible for the quanta of harms produced.[1]
>
> *Oil Spill*: A tanker with a similar background crashes. It does not sink but large holes emit oil into the waters. No one dies in the crash or subsequent good faith attempts at minimizing damage. Several injuries occur. The now-uncontained spill greatly impacts sea life. Oil eventually reaches land, further impacting beaches and land life. Food supplies diminish. Traditional accounts of responsibility reach the same result.[2]
>
> *AWS [Autonomous Weapons System]*: "[A]n autonomous drone bombs a column of enemy soldiers who have indicated their desire to surrender. The drone's commander gave orders to patrol the region and engage legitimate targets. But the drone wrongly identified the surrendering soldiers as legitimate targets" (Himmelreich 2019:273).[3]

[1] *Crash* synthesizes and builds on the sinking of the *Herald* discussed by Pettit 2007 and the Mount Erebus aircraft crash discussed by Collins 2019 in the related collective duty gap context.

[2] I use a tanker here to parallel *Crash*. Wildlife and food issues commonly follow spills, like the Exxon Valdez and non-tanker-based Deepwater Horizon cases. So too do debates as to who, if anyone, is responsible for a spill. Himmelreich 2019:734 discusses Exxon Valdez as a case where corporate liability may need to fill gaps.

[3] This is a variant of 'killer robot' cases that proliferated post-Sparrow 2007.

Michael Da Silva
M.Da-Silva@soton.ac.uk

[1] University of Southampton, Southampton, UK

Published online: 07 October 2024

> *False Positives*: A lung cancer diagnosis Artificial Intelligence (AI) tool meets legal safety and efficacy standards. It is designed for use where expert providers are unavailable and “constructed so that false negative diagnoses are highly improbable … [There is] much less precaution about false positives. Although wrong positive diagnoses are not immediately life-threatening, they can cause great financial, practical and emotional problems. ... [Programmers did] everything possible to prevent false negative diagnoses” (Matthias 2004:177). Significant numbers of false positives occur.[4]

These scenarios model real-world cases and cover issues across diverse domains but share a common concern, namely the presence of a ‘responsibility gap’ whereby there is a mismatch between the amount of responsibility one can attribute on standard models and the amount it is otherwise intuitively appropriate to attribute.[5] There is “a deficit” in the moral “accounting books” (Pettit 2007) arising from how “traditional ways of responsibility ascription are not compatible with our sense of justice” (Matthias 2004:177). Where traditional theories of responsibility cannot account for robust intuitions concerning the amount(s) that should be attributed to someone, many seek means of ‘filling’ the gaps to address corresponding deficits in fundamental interests in having someone to blame, compensate, or otherwise account for harms.

This work examines when one should fill responsibility gaps. Most discussions of responsibility gaps address particular kinds of cases.[6] Many focus on AI and questions of technological capability.[7] But gaps present a general moral phenomenon technology alone cannot deflate. Recent work has begun identifying necessary and sufficient conditions for gaps’ occurrence.[8] Those conditions implicate questions about when to fill gaps. However, the existence of a gap may not justify, let alone require, filling it. Discrete treatment of when and why responsibility gaps should be filled remains comparatively minimal. I further ongoing work on responsibility gaps and when to fill them by proposing a responsibility gap, X, should be filled if (1) X arises in circumstances posing a threat of non-de minimis harm but (2) X is unavoidable or there are reasons to permit X to arise; (3) the harm is addressable and (4) addressing it would fulfill a responsibility-relevant good Y; and (5) there are

[4] I update this real-world case below.

[5] Even authors who state that gaps occur where “no one” (Braham/van Hees 2011; Danaher 2016; Mukerji/Luetge 2014) can be properly held responsible for a state of affairs are primarily concerned with the mismatch between desired and appropriately attributable responsibility, rather than the lack of *one* responsible agent. See my Philosophy Compass article on ’Responsibility Gaps’ for more detail.

[6] On groups, including corporations, see Pettit 2007; List/Pettit 2011; List 2021; Braham/van Hees 2011; Duijf 2018. Smith 2009 discusses similar phenomena. On states, see Lawford-Smith/Collins 2017a (surveying a larger literature). The volume including Köhler et al. 2018 also largely focuses on states. On AWS, see Sparrow 2007; Danaher 2016, 2022; Himmelreich 2019; Swoboda 2017; Zając 2020; Oimann 2023. On cars, see Danaher 2016; Nyholm 2017; de Jong 2020. On AI generally, see Matthias 2004; Champagne/Tonkens 2015; Köhler et al. 2018; Tigard 2021; de Sio/Mecacci 2021; Glavaničová/Pascucci 2022; Tollon 2023. Some accounts address multiple phenomena.

[7] See Tigard 2021’s discussion of attempts to “dissolve” gaps technologically.

[8] See, e.g., Köhler et al. 2018:54; Himmelreich 2019:735. Even these often appear in domain-specific inquiries.

reasons a specifiable person can be called upon to address the harm for Y but (6) no other means of achieving Y. I specify each condition below.

2 Identifying Gaps and What It Means to Fill Them

Responsibility here is a status agents acquire via actions triggering others' permissions to respond in particular ways.[9] Responsibility is traditionally conceived in terms of being fit to blame (Wallace 1994; Nelkin 2016; McKenna 2019; Pereboom 2021) (or, in positive cases, receive praise) for some act or situation.[10] It is sometimes also conceived in terms of being able to demand compensation (List 2021). Responsibility pluralists suggest that it could involve being subject to sanction or reward for one's actions (accountability) or appropriate requests for reasons or justifications (answerability) or being labelled in certain ways (attributability).[11] Responsibility gap scholars discuss 'responsibility' in each of these senses. I accordingly adopt a capacious view here whereby responsibility makes one liable to set forms of treatment and specify *which* forms are appropriate or distinctive of responsibility properly-so-called below.

Responsibility gaps occur where traditional accounts of responsibility intuitively lack "resources to say what should appropriately be said" about cases (Köhler et al. 2018:54).[12] Theoretically-justified responsibility attributions do not match a "pre-theoretical intuition" or "gut reaction" about the responsibility quanta that should be ascribable (or who should have it) (Collins 2019:946). It should at least be intuitively "fitting to hold someone to account" (Köhler et al. 2018:54). This is meant to distinguish gap cases from natural disaster cases where even positing non-human minimal agents cannot establish clear candidates for responsibility.[13] Some believe that mismatches between acceptable and intuitively plausible attributions are themselves problematic (e.g., Matthias 2004; van de Poel 2012). Yet mismatches can be

[9] This is consistent with Braham/van Hees 2011:7's folk sense of responsibility and Shoemaker 2011, 2015's broadly Strawson- and Williams-based approach to responsibility. I do not commit to Shoemaker's responsibility pluralism here but discuss various forms of responsibility invoked in prior work on gaps.

[10] I assume some familiarity with debates on the role of praise here. Recall, e.g., Talbert 2019's summary.

[11] Shoemaker 2011, 2015. See also Burri 2017; Himmelreich 2019.

[12] This is distinct from related problems, like collective duty gaps (e.g., Collins 2017a) and the Problem of Many Hands (e.g., Thompson 1980; Bovens 1998; van de Poel 2012). Some problems, including the Problem of Many Hands, can be described as responsibility gaps. See also the 'retribution gap' (Danaher 2016; Buell 2018). The problems are connected. There are enough parallels to draw from discussions of other gaps. Each case involves responsibility "blurring" (Köhler et al. 2018) such that it is unclear who, if anyone, should be held responsible for an outcome. But the problems are distinct. For another example, one can discuss 'gaps' as 'voids' (Braham/van Hees 2011; Duijf 2018). But the existence of a 'void' is only one of three conditions for a 'gap' in Himmelreich (2019:734)'s leading account. Treating the present phenomenon as distinct appears all-things-considered justified. List and Pettit's identification gap (2011:ch 9) focuses on sustaining a group agent. Again see my Philosophy Compass article on 'Responsibility Gaps' for more detail.

[13] Himmelreich 2019 defines this in terms of actions of "merely minimal agents."

resolved in favour of traditional accounts or intuitions. One may accept that a theory cannot address all cases and accept the persistence of (seeming) responsibility gaps. Others believe responsibility gaps are problematic where and because they permit actions for which no one is accountable (Pettit 2007; Sparrow 2007; Köhler et al. 2018) or make it hard for those harmed to seek redress (List 2021).

While scholars argue about how to characterize the problem and when it arises, responsibility gaps purportedly arise in many morally relevant-domains, including prominent ones concerning government, corporate, and other forms of group agency and AI.[14] The cases above jointly present a general problem: There are (at least apparent) responsibility gaps many want to fill, yet relatively little work on which gaps one should seek to fill and when to do so.[15] This outcome immediately raises two questions: (1) Do gaps exist? and (2) What would it mean to 'fill' them?

With respect to (1), this work primarily examines what to do *if* gaps exist. However, preliminary words on their possibility help motivate the inquiry.[16] Attending to the reasons why persons seek to deny the existence of genuine responsibility gaps offers a fruitful entry point. Consider, for example, Königs (2022)'s claim that gaps are a "philosophical mirage." Per Königs, gaps arise only where no one can be responsible for an autonomous systems' actions or the consequences thereof on traditional accounts (the 'no responsibility' condition) and the absence "is due to the system's autonomy" (the 'autonomy condition'). He also posits a "plausibility constraint" whereby they "arise only in situations in which the negative outcome is not due to carelessness or malice" and suggests most harms in purported gap cases stem from malice or ignorance. Genuine responsibility gaps are thus rare. However, problems plausibly understood as pertaining to 'responsibility gaps' could remain even if Königs's arguments succeed. Königs quickly dismisses a 'demand condition' on which there is a demand to hold somebody to account. Yet robust intuitions in cases above help explain why many theorists, including Pettit and Danaher, believe such 'demands' sometimes call for a response. Their theories offer principled reasons why cases where intuitions can be explained by inapt desires for retribution– such as the *Hurricane* case below –do not exhaust the phenomenon and some accountability deficits are problematic. And gaps in other aspects of responsibility, like answerability, may arise even if accountability-based gaps are illusory.[17] It is worth examining whether and when one should remedy even apparent responsibility attribution mismatches. Attempts to "dissolve" (Danaher 2022) gaps by demonstrating all or most apparent gap cases actually permit one to identify persons with sufficient control to warrant full responsibility when properly described are, in turn, compelling.

[14] Recall note 6.

[15] Plausible exceptions appear throughout this section. Many focus on nearby issues.

[16] As discussed below, if all apparent gaps can be dissolved, this exercise still offers insights into, for example, whether and how to alter theoretical accounts of responsibility to dissolve them and the reasons why gap cases seem problematic.

[17] Similar arguments against gaps in corporate cases (e.g., Moen 2024) accordingly do not defeat this proposal. Königs's conditions are also defined and applied in ways that limit their application to his chosen case types.

But they rest on empirically contentious foundations.[18] The phenomenon plausibly occurs even if one quibbles with cases. And small alterations to cases challenge otherwise seemingly easy resolutions. For instance, many would respond to *False Positives* by positing corporate responsibility.[19] But it is easy to imagine AI being the product of diverse actors working on parts of a project in a garage incubator. The actors may individually fail to meet relevant requirements to a degree that would make them jointly responsible for all attendant harms without qualifying as a group agent that could dissolve the gap on plausible definitions.

These possibilities suffice to motivate this inquiry. One may, of course, still ultimately deny the existence or presence of such gaps in a domain (e.g., Kohler et al. 2018; Himmelreich 2019; Grübler 2011; Tollon 2023) or suggest that they are rare, uninteresting, or not as problematic as claimed (e.g., Braham/Van Hees 2011; Duijf 2018; Moen 2024). However, there is ample room for reasonable disagreement as to the existence and prevalence of responsibility gaps. Case-based and principled reasons to support their existence above/below justify examining what one should do if they exist. And analyses below suggest doing so can help responsibility gap skeptics better understand whether and when to alter general theories to address merely apparent gaps.

This leaves question (2) concerning what it would mean to fill a gap that arises. Some respond to mismatches above by positing entities who can be held responsible, like corporate agents (Pettit 2007:194) or AI (List 2021).[20] Others suggest moral or legal mechanisms can play roles desired by ‘gap-filling’ absent new agents. For example, Collins (2019) suggests some persons should pick up the responsibility ‘slack’ in analogous collective duty gap cases. Still others suggest someone is actually responsible for any harms that accrue in apparent gap cases (e.g., Moen 2024). The call to identify when to ‘fill’ responsibility gaps could simply seek circumstances where gaps are problematic enough to require theory revisions. One could, for instance, alter the control condition on responsibility, remove a requirement that one be able to do otherwise, or lengthen the causal responsibility chain so that more remote actions trigger moral responsibility.

More challenging cases occur when the best theories of responsibility cannot find any individual, collective, or other entity or collection thereof fully responsible for the full quantum of harms even after the kinds of revisions in the prior paragraph have been made. It is at least conceivable that there will be cases in which even a combination of all natural, group, and artificial agents cannot be held responsible under traditional theories and where changing the conditions of full ex-post responsibility is problematic. Some cases above/below plausibly qualify as such hard cases. If robust intuitions remain that someone should nonetheless bear responsibility (in the relevant sense) for the state of affairs, a responsibility gap also remains. One

[18] See e.g., Himmelreich 2019. de Sio/Mecacci 2021:1073-1074 describe this as a form of “deflationism.”

[19] For classic work on corporate responsibility and gaps, see Pettit 2007; List/Pettit 2011.

[20] Per List 2012, where AI cannot be responsible or fill gaps, requiring someone else to accept strict liability for AI-induced harms is a second-best solution.

primary purpose of this work is to identify what should be done when such hard cases arise. However, the conditions for when to fill them below also appear relevant to identifying when gaps are problematic enough to make positing new agents or altering existing responsibility conditions apt. Indeed, one of the conditions explicitly concerns when/how a corporation or AI can/should fill gaps. As with question (1), then, the analysis below is likely to be helpful on at least two readings thereof.

To state this much nonetheless leaves open the question of what it would mean to fill a gap. What is the 'relevant sense' in which someone should bear responsibility for a state of affairs in the prior paragraph? Gap-filling is meant to be a form of holding someone morally responsible, rather than placing someone under a mere legal obligation. It is distinct from final accountability in the sense of being the "person as the one who carries the can, the one who sits at the desk where the buck stops" and forms of regulating conduct whereby we subject persons to certain forms of treatment (e.g., punishment) to guide their future actions (Pettit 2007:173–174). Those interested in gaps believe theories of responsibility should fulfill particular functions and gaps are problematic because applying the best theories of moral responsibility (or legal liability) leaves those unfulfilled. Gap-filling is meant to ensure relevant functions can be properly achieved.

I propose using these functions to define gap-filling. For present purposes, in other words, a gap is 'filled' where someone who is not fully responsible for a state of affairs on leading theories of moral responsibility addresses responsibility-relevant harms within that state to fulfill necessary functions of a theory of moral responsibility (and can be aptly called upon to do so). Gaps are problematic when and because they leave certain functions of responsibility unfulfilled. Gap-filling denotes a moral phenomenon in which particular entities can be aptly called upon to fulfill those functions. Apportioning blame, compensation, and apology are three widely accepted functions that I will initially take as primary.[21] Which functions are important and why remains contentious. The exemplars could alternatively be justified on the basis of desert, social cohesion, etc.[22] A complete account of gap-filling will likely need to make decisions on which functions matter and why. Yet it suffices here if there is broad agreement on some core functions. I am thus initially agnostic as to which functions are essential and why they matter. I follow others in articulating a framework for analysis that require greater elaboration in future work.[23] This move is unproblematic where the framework can work on various articulations of the purposes and a more ecumenical view can help reorient discussions in various debates in which appeals to different purposes are made. It is less problematic still

[21] Each author above/below appeals to one of these functions as a reason for– or constitutive of –gap-filling.

[22] Caruso/Pereboom 2022 helpfully outline options, albeit in service of a view at odds with several moves in the present work. Enoch 2012:124ff appeals to the good of 'valuable relations' to ground an account of 'taking responsibility.' Kiener 2022 offers both agency-based and contractualist accounts of his 'answerability interests.'

[23] Enoch 2012, in fact, only motivates a possible conjunction in responsibility theory. This proves fruitful.

where, as demonstrated below, applying the framework also provides insights into which functions (should) matter when.

At least three options remain as to what it means to state that someone should fulfill the functions. One would 'fill' gaps by finding someone who fill accept responsibility ex-ante for any negative consequences of a given course of conduct. This risks collapsing the distinction between the kind of ex-post responsibility at issue in most gap-filling cases and buck-stopping responsibility. While this alone does not defeat this approach to gap-filling, additional problems arise below. A second approach seeks to find someone that one can appropriately deem responsible ex-ante. If one can find such an entity, this may provide reasons to otherwise alter moral theories of responsibility, nicely connecting to alternative approaches above. Yet the risk of hard cases where no such person can be found may linger. The third, remaining approach states that gaps are filled when someone "takes" responsibility ex-post in Enoch (2012)'s sense of agreeing to do what a fully responsible person would so. The act of agreement is, on this approach, supposed to "make" one responsible ex-post.[24] Others can treat the person as they would have treated a fully responsible agent. Kiener (2022) usefully suggests taking responsibility can address at least some apparent gap cases.[25] Yet Enoch and Kiener rightly note that only particular persons can take responsibility for a state of affairs and thus seek conditions under which identifiable persons can do so.[26] Related questions concern when taking responsibility in this sense is apt and when one can appropriately call on someone to do so. Broader discussions of 'deeming' persons responsible in gap-filling cases also speak to the latter issue.[27] One could deem someone responsible for practical purposes or hold one responsible in theory who would not be responsible on traditional models. Either interpretation implicates concerns below. Consider, for example, List (2021)'s requirement for someone to accept strict liability before AI is allowed on markets. This could be a theoretical requirement of justice or practical solution to theoretical deficiencies in AI ethics. The goal in either case is to find someone who can produce a responsibility-relevant good necessary to address apparent moral failings.

The functional approach to gap-filling here is most intuitive on a view in which gap-filling is a form of taking responsibility. That view captures the sense in which gap-filling is meant to hold someone responsible ex-post while placing the text

[24] On the "backward-looking sense of responsibility" at issue, see, e.g. Enoch 2012:105, 108, 114. His account of taking responsibility for acts within the 'penumbra' of your agency is explicitly (101n6) inspired by Wolf 2000.

[25] This account (582ff) also builds on Owens 2012.

[26] Enoch (2012) requires (1) an act of will to take responsibility for something (2) within the penumbra of one's agency. The penumbra is then contrasted with those acts within your agency (viz., classic intentional action with foreseeable consequences) and those wholly outside of it (e.g., the motion of the planets). Kiener 2022:586n8 contrasts Kiener's own view with Enoch's and offers an AI-specific alternative to Enoch's penumbral account.

[27] Enoch 2012 and Kiener 2022 each discuss contexts in which persons can call on others to take responsibility. Many others above likewise seek someone who can be deemed responsible even where they would not be responsible on classic accounts. Consider, for example, List 2021's aforementioned practical proposal for AI.

squarely within ongoing debates about how to address harms in gap cases and who can be called upon to do so. The following is thus most easily read as specifying conditions under which persons can be aptly called upon to fill gaps by taking responsibility. In such conditions, it is usually apt to state that they should do so. Failure to do so could be an independent wrong (though specifying the conditions under which it would be all-things-considered wrongful is beyond my scope of inquiry).[28] Yet, strictly speaking, the approach at issue only requires that the persons in question fulfill relevant functions. This differs from (at least some) traditional accounts of taking responsibility in at least two ways. First, this work takes the functional deficiencies that gaps produce as central such that the need for someone to, for instance, apologize is of primary value, rather than the quantum of responsibility that necessitates it. And second, those who fill gaps only take responsibility as is necessary to fill those lacks. They do not take full responsibility for all outcomes. Insofar as they take responsibility, in other words, they take only that which has not yet been rightly attributed to others and thus produced that functional deficiency.[29] If, in turn, one finds the notion of 'taking responsibility' independently problematic, the text can alternatively be read as specifying when responsibility gaps are problematic enough to call for a response. The conditions plausibly help identify whether and when theory change is needed to address gaps. Yet the gap-filling at issue again involves particular entities addressing functional lacks in existing theories. It is meant to remain distinct from simply altering one's theory of responsibility, buck-stopping responsibility, and mere legal liability. It also aims to be distinct from other approaches to gaps in Part III.

3 Unsatisfactory Approaches

Issues with existing approaches to this issue help motivate my proposal. Filling all gaps would, for one, be plausible if responsibility gaps are problematic by definition (van de Poel 2012). However, as List (2021:1127) notes, one cannot establish "the conclusion that an entity can be assigned responsibility simply from the premise that such an assignment helps us to avoid" gaps. A case concretizes List's concern:

> *Hurricane*: "[A] hurricane causes a huge amount of damage for which there is little human responsibility."

[28] *Cf.* Enoch 2012 on moral duties to take responsibility in set circumstances (e.g., where your child acts wrongly). Kiener 2022 argues that there are cases where taking responsibility is obligatory. It may not always be so.

[29] Enoch 2012:110, 118-119, etc. suggests offering justifications, excuses, or apologies are paradigmatic ways of taking responsibility. As discussed below, listening to someone's legitimate complaints may be another form. Enoch believes one cannot take the blame for something retroactively (116). If so, a further question arises as to whether one can fulfill a function of blame, such as satiating justified anger (or, indeed, the demand for that apt apology).

Hurricane lacks agents who meet basic causal or epistemic conditions for folk responsibility.[30] It accordingly does not present a genuine responsibility gap even on views that acknowledge some gaps exist.[31] However, other cases with minimal agents also challenge calls to fill all gaps. Consider:

> *Beneficial Health Good*: Another AI-enabled lung cancer detection machine is very effective at early detection, permitting interventions that drastically improve patient life expectancy/quality. It is also more efficient than human care and can operate independently of humans with high efficacy. This produces significant cost and time savings for the healthcare system, providing more funds for other health goods and time for compassionate care. False positives occur but far less often.

This case too is bound to be controversial but exemplifies the idea that circumstances that produce gaps can be highly beneficial. Whether anyone should, let alone must, fill gaps in such circumstances remains an open question. A no-fault compensation scheme would serve many of the same goods as requiring individual entities to fill gaps without making any particular person accept high compensatory costs. It also minimizes scapegoating risks of some accounts below.

A less common response accepts gaps exist and attempts to avoid conditions producing them (e.g., Sparrow 2007). It is also problematic.[32] Avoiding the developments that would lead to *Beneficial Health Good* strikes a problematic bargain that leaves far too many ill. Indeed, Danaher (2022) suggests letting gaps arise can be morally *desirable*. Some may be unavoidable. Consider:

> *Delegation*: Two options for allocating resources in a hospital will each result in one of two patients suffering. Neither will produce death or overwhelming suffering. An AI tool built for rationing makes the decision consistent with reasonableness standards. A majority of healthcare providers would accept the recommendation. Hospital policy directly applies the AI recommendation. One patient does not get needed healthcare.[33]

Suffering for which no one is fully responsible will arise following any choice here. While one could cut funding to eliminate choices in like cases, an intuitively undesirable responsibility attribution is inevitable if anyone is to receive benefits of healthcare funding.

[30] Recall note 9 sources. See also Pettit 2007:175; Köhler et al. 2018:52. Pettit and Copp 2006 also (plausibly) suggest any group/group-like entities must have certain organizational structures, rules, and goals to be agents. Enoch 2012's reference to cases outside one's agency are also relevant here.

[31] Recall note 13, surrounding.

[32] de Sio/Mecacci 2020 call this "fatalism." Tigard 2021 distinguishes 'techno-pessimists' who say we should limit the use of AI due to inability to fill gaps and techno-optimists who will 'bridge' gaps by finding someone responsible. Tigard believes one can always find a responsible party.

[33] This case is mine. Danaher 2022 contrasts 'Delegation' in which we "get someone (or some *thing*) to make tragic choices on our behalf" with 'Illusionism' which denies tragic choices exist and 'Responsibilisation' in which persons just bear the costs for such choices. Danaher argues that Delegation is psychologically attractive, shifts burdens to those better able to bear costs, and can fulfill social benefits by getting people to do what should be done.

Existing attempts at specifying when gaps should be filled do not present sufficiently better results. One may, for instance, take responsibility gaps' constitutive break between philosophically justifiable and intuitive responsibility attributions as a guide. Filling gaps could be beneficial insofar as results better reflect common intuitions. However, the fact that a result is non-intuitive cannot warrant actions fulfilling more intuitive desires in other contexts. It should not here either. Indeed, psychological mechanisms undergirding desires to fill gaps suggest one should reflect on one's intuitions before seeking to fill them. Collins (2019:948) highlights research demonstrating that humans "desire to see *someone* attributed blame for large-scale harmful outcomes" and suggests these base desires produce many apparent gaps. Some mismatches between philosophical reflection and intuition should resolve in favour of intuitions. Intuitions will guide most analyses. Yet the blunt fact of traditional responsibility attribution outcomes producing unintuitive gaps cannot fully justify filling gaps. One should desire to explain real reactive attitudes and fix gaps between considered intuitions and appropriate responsibility attributions. But this is not the only relevant good. The same considerations undermine claims that gaps must be filled to sate reasonable desires for retribution. Bare psychological desires cannot justify retribution. Only intuitions about the need for punishment that withstand reflection should feature in decisions about gap-filling.

One may, finally, more broadly seek to fill gaps to avoid free-standing wrong without appropriate blame. This aesthetically pleasing result might further desires for intuitive results that ensure all harmed parties can hold someone to account. However, ethics rightly admits many aesthetically unpleasing features. We cannot favor aesthetics over truth,[34] particularly where responsibility gap proponents appeal to truth to motivate their views (e.g., List/Pettit 2011:185) and filling gaps requires persons bear costs they would not otherwise face. Ex-hypothesi, the truth is no one bears full responsibility. Requiring they do so for aesthetic reasons is perverse.[35]

If responsibility gaps exist, then, there are good reasons not to fill or avoid all of them or to simply seek the most intuitively plausible result. If the forgoing is correct, the mere existence or possibility of a(n apparent) responsibility gap also may not be sufficiently problematic as to demand theory changes. One may leave gaps in place where, for instance, the theoretical costs of dissolving them are too high. While the following is, again, primarily concerned with what to do if gaps exist, this already makes good on the promise that the present analysis can aid gap skeptics. Future work should examine whether/how considerations below implicate their views.

[34] Compare, e.g., Ross 1920/2002:23.

[35] As an anonymous reviewer notes, excuses are also understood as creating free-standing wrong without appropriate blame. This also counts against the proposed argument for why gaps should always be filled.

4 Conditions for Apt Gap-Filling

The preceding issues call for clarity on when one should fill gaps rather than seeking to avoid them or leaving them in place. I accordingly turn to articulating six conditions for apt gap-filling.

4.1 Condition 1: (Threat of) Non-De Minimis Harms

My first proposed condition is a threat of non-de minimis harms. The concept of a 'de minimis harm' is invoked in criminal and private law to denote trivial harms that do not warrant legal response: producing a de minimis harm does not trigger a need for punitive judgment or compensation for the subject of that harm.[36] I use the phrase here as a threshold marker for the amount of harm necessary to warrant intervention. Below a threshold, seeking to a fill a gap to address harm would be inapt. Any costs associated with gap-filling would not be worthwhile.

Cases above involve major harms. Inabilities to hold anyone accountable or receive compensation appear problematic, helping motivate deviation from what are, here ex-hypothesi, philosophically appropriate responsibility attributions subject to further conditions. But consider:

> *De Minimis*: An AI movie rental pricing tool meets market standards and does not violate any laws in nearly all cases. The tool may rarely err in ways that lead to people paying up to $1 more than the market would normally demand. No one is plausibly responsible for this learning-induced error, which is quickly corrected.

Movie rentals implicate no vital life interests. $1 will not alter the interests for anyone with the disposal income to spend on movie rentals in any case. While the extra $1 constitutes a quantifiable harm, it is not worth filling. The simple financial costs of a system set up provide the $1 back to each person in *De Minimis* will be much more than the $1/person reimbursement. And the costs will expand if one also calls for apologies or other forms of redress or if one acknowledges non-pecuniary costs like the time spent identifying claimants or issuing apologies.

Similar cases are possible if one finds the rental case non-intuitive. Indeed, the movie rental example may undersell the importance of de minimis harm by giving a mistaken impression that there will always be an entity, like a corporate rental chain, who can fairly address all harms. But a variant in which there is no corporate agent capable of easily absorbing the costs makes the risks of high costs for low rewards acute. And, as discussed further below, practical risks of high costs for implementation schemes are far from the only costs of gap-filling. Even cases calling for little practical action present potential costs. Implementing deviations runs theoretical risks of overcompensating by drifting too far from ideals (especially given psychological facts above). The point here is that deviating from philosophically

[36] E.g., Veech/Moon 1947; Inesi 2006; Husak 2011.

appropriate responsibility attribution patterns is not always apt. Costs associated with gap-filling simply are not worth it absent corresponding benefits, which are likely minimal where the risked harms are sufficiently small.

What qualifies as non-de minimis harm will most likely vary across competing accounts of harm, responsibility, and the value of each. I introduced the basic idea of a non-de minimis harm threshold by contrasting small pecuniary harms and harms that impact vital life interests. This provided a useful and intuitive dividing line between cases where gap-filling appears intuitively unnecessary and cases where it appears much more important. Yet there is likely a wide range of non-de minimis harms that do not impact vital life interests. One could even place the dividing line at a point where $1 qualifies as a non-de minimis harm. Cases above/below and clear understandings of the value and costs of filling responsibility gaps still strongly suggest that plausible accounts of responsibility, harms, and gaps are likely to identify a threshold below which the quantum of unaddressed harm is insufficient to trigger any apt call for a response.

Aggregation concerns complicate, but cannot defeat, this condition. If a million persons rent a movie when the error occurs, many would contend that a one-million-dollar harm must be remedied. Class actions plausibly exist to address such cases. Yet granting this does not require rejecting the proposed condition absent strong claims about the relationships between aggregations of value and their parts. It suffices if *some* de minimis harm condition is plausible given the relative size of benefits and burdens at hand. I cannot solve the Sorites Problem here.

This does not mean the threat(s) or harms must be very great or diffuse over many persons. Filling gaps is, all-else-being-equal, also wise in Pettit (2007)'s other classic group agency case:

> *Tenure*: A scholar applies for tenure. Under the tenure process, the applicant must receive 'Excellent' ratings across all relevant criteria to succeed. Tenure committee members do not each rate the applicant 'Excellent' in all categories. Each thus votes against tenure. Yet the members disagree on which criterion the applicant fails to meet. Joint votes of all members for each criterion would lead to applicant success.[37]

I share Pettit's intuition that the applicant cannot fault individual committee members and yet someone should be accountable for a seemingly problematic decision. The responsibility gap plausibly persists and should be filled even where the candidate (who Pettit views as 'borderline' but has strong qualities) quickly finds new work. If one disagrees, *Tenure* still establishes that harms to a single individual's life interests can trigger intuitions that a gap should be filled. A requirement for threats of non-de minimis harm(s) need not require very large or diffuse harms.

[37] See also the related 'doctrinal paradox' (Kornhauser/Sager 1993; Chapman 1998).

4.2 Condition 2: Gap Production is Acceptable or Unavoidable

Gap-filling, again, paradigmatically deviates from leading accounts of justifiable responsibility attributions. Even the most intuitive accounts that understand gap-filling in terms of taking responsibility for actions must accept the controversial possibility that taking responsibility is even possible. And most forms of gap-filling have practical costs above/below. Given these costs, simply avoiding gaps can be desirable, as Matthias (2004) suggests in AWS settings.

Acknowledging gap-filling is costly does not, however entail that all gaps must or should be avoided – or that they can be avoided in the first place. Recall *Beneficial Health Good*: barring AI development to avoid possible gaps at the cost of widespread health benefits can only result from erroneous risk-benefit calculation. Permitting gaps to arise can, indeed, serve many functions. Consider *Crash*. One might seek to avoid its outcome with a better route, training, etc. Yet a variant in which these are provided and a crash remains unattributable to any person remains possible. Anti-capitalists may question whether the ship should have left harbour. But most accept such risks as simple costs of global trade and travel. Similarly, while many will question whether the circumstances producing *Oil Spill* should occur, permitting the gap to arise to ensure access to presently-necessary energy reserves is reasonable. The case for avoiding gaps is weaker still if gaps can be filled by, for example, attributing responsibility to corporate agents or AI, as one may do in some cases at hand. One cannot assume these gaps will be filled here, but even that possibility suggests one need not always simply avoid circumstances that can produce gaps.

There is, of course, a sense in which one 'avoids' a gap by changing one's theory of responsibility to accept non-human agents in those cases. But that is not the sense of avoiding gaps at issue. This condition is instead concerned with claims that one should simply create circumstances in which the apparent mismatch between theoretically apt and robustly intuitive responsibility attributions do not arise by, for instance, limiting trade or forms AI development.

Acknowledging this contrast in the means of 'avoiding' gaps also underlines that other cases may not include acceptable alternatives to permitting gap-producing circumstances. *False Positives* is, for example, constructed such that human providers cannot make relevant diagnoses. Alternatives absent accurate diagnoses are unacceptable. A responsibility gap is, in turn, nearly inevitable in a *Crash* variant where ships are the major mode of trade and transport, everyone does what can be expected of them, and yet a crash produces great harm. Recent shipping issues and their implications for the global economy suggest that such a case reflects a nearby reality.

Permitting harms remains non-ideal: one should avoid them where no countervailing factors override the value of the harm and the harms are actually avoidable. However, one should fill, rather than avoid, responsibility gaps in the (arguably second-best but acceptable and even desirable) cases where non-attributable harms are permissible for other reasons or unavoidable.

One complication for this condition is that many unaddressed harms arise in cases where the gap was avoidable and should not have been permitted to arise.

Cases of this form are likely more common than idealized cases where one has a choice as to whether to permit gaps to arise. Gap-filling should plausibly be apt in some of these more common cases. So, it would a problem if my account could not guide action therein. However, this complication does not defeat the present condition. There is a sense in which gaps in these non-ideal cases are ‘unavoidable’: they have already occurred when one is deciding whether to fill a gap. Conditions 1 and 3-6 then provides useful guidance on when gap-filling is apt in ideal and non-ideal cases. If, in turn, this condition does not apply as directly to cases where a gap has already unjustifiably arisen, that will admittedly make it different from the other conditions. But it will also clarify our understanding of responsibility gaps, their moral value, and when they call for a response. The condition suggests that states of affairs in which questions about whether to fill gaps have already been settled are non-ideal. An understanding of the condition on which one should ideally examine whether it is better to avoid a gap or let it occur highlights that gap-filling in these cases is genuinely non-ideal in the sense that it is inapt relative to simply avoiding otherwise unacceptable gaps’ occurrence. Maintaining this condition helps us better understand responsibility gaps by highlighting that some, but not all, responsibility gaps can or should be permitted to arise and that gap-filling is second-best where they should but cannot be avoided.

4.3 Condition 3: An Addressable Harm

An entity must then be capable of filling a gap. If, for instance, the person harmed cannot be properly compensated, there is no clear reason to attempt to do so on compensation-based grounds. This concern need not be addressed in financial terms alone. If a crash causes me a set amount of harm, a gap will and should remain not only if no one can realistically provide me with funds needed to address that harm (or the actual good of which I have been deprived, as in cases of body parts without extant artificial equivalents). It will and should also remain if no one can fulfill another good that would warrant filling it, like holding someone plausibly accountable. Likewise, if no one can fulfill my justifiable psychological desire to hold someone appropriate to account, this impossibility should matter. Some entity must then be able to address the relevant deficit. If, for example, gap-filling requires financial compensation, an entity must be able to hold/expend funds. Corporate responsibility is partly motivated by corporations’ ability to so-compensate (List 2007; List/Pettit 2011). If, in turn, filling a gap requires someone to face physical punishment, an entity must be capable of experiencing the punishment in a phenomenologically appropriate manner. Corporations are less likely able to do this, requiring another gap-filler.

This state of affairs underscores a further need to operationalize some forms of gap-filling through attributions of individual responsibility that are otherwise inappropriate and potential differences between cases. The just-mentioned cases highlight how desires to fill a gap through one entity may need to be operationalized through further responsibility attributions to other agents. Certain corporate officers are, for example, deemed liable for the corporation’s wrongdoing because corporations cannot experience apt punishments in the phenomenologically relevant way. And where many autonomous AI cannot (yet) hold or expend funds on their own, many consider holding a corporation liable for harms caused by an AI tool desirable (see, e.g., Da Silva 2022).

The need for another entity to operationalize gap-filling admittedly complicates matters in cases like the incubator-based variant of *False Positives*. Whether any technologist in the garage could compensate wronged parties, let alone should do so, is debatable. I elsewhere (*id.*) argue that many intuitions about gap-filling in the AI context are motivated by thoughts about corporate agency such that AI raises at most few unique gap-filling desires. At minimum, the incubator variant highlights how AI can raise challenges finding someone capable of filling the relevant role.

Corporate and AI cases, then, can overlap and yet importantly diverge and differ. This condition suggests they may not always be treated the same. Where, moreover, AI can hold and expend funds, it may do so in ways that make it more like a human person. If so, this will raise the individual responsibility attribution questions about AI analytically prior to any gaps. Some AI may be true individual agents, rather than non-individual agents treated similarly to individuals for the purpose of filling gaps in individual and total responsibility attributions. (As detailed below, though, an entity being capable of filling a gap does not always mean that they should.)[38]

4.4 Condition 4: Furthering a Responsibility-Relevant Moral Good

Filling gaps has costs. Examples include the fact of its deviation from philosophically justified attributions; financial, psychological, and similar costs faced by gap-fillers; implementation costs; and various risks, like giving into psychological desires and unduly placing responsibility on others or (as discussed below) scapegoating particular persons for the sake of filling gaps. They may include negating goods that responsibility gaps can provide in *Delegation*-like cases.

[38] To state that there must be an addressable harm does not require a specific person with a harm-based claim for redress. Enoch 2012:119 highlights cases where negative consequences occur but the extent of the harm is such that the remaining agent who has been harmed and so can directly call for someone to, in his particular example, apologize for the action may lack standing to do so all-things-considered. He further suggests that if all persons are deceased, have no close relatives, etc., there may not be a remaining agent that has been directly harmed. This is fair enough. Yet the present account takes no definitive position on who has standing to call on others to address harms. If it is impossible to address the relevant function since *no one* is available to accept the apology, the gap cannot be filled. But this is just another way of saying that the harm cannot be adequately addressed and so no gap-filling is due. And I suspect cases where there is an apparent harm and nothing can be done are unlikely to be common enough to render gap-filling rare. Enoch believes candidates for the kind of standing necessary to accept “are rather common.” Apology-giving practices can make apologies to those less directly harmed apt. Yet one need not enter into complicated discussions about subjects of harm to acknowledge that gap-filling can address lacks even in cases where harms accrue to the isolated deceased. There are many other types of harms and functions of responsibility that could ground claims even if an apology seems odd there. Calls for someone to face punishment, compensate distant relatives, or apologize at a society level are just three examples of how one could fill specifiable gaps even in cases where direct apologies to the dead are inapt.

These costs are only worth bearing if filling gaps furthers a relevant good. The good in question should relate to the underlying concerns given the nature of the present problem. I again focus here on identifying when we should hold persons responsible *to address problems of responsibility*, rather than other goods we would seek to fulfill through forms of regulation. This condition is accordingly partly analytic. Other approaches' inadequacies provide further motivation. Deeming someone responsible for harms may, for instance, further aforementioned bare psychological desires for retribution. Someone who highly values punishment may find this desirable even in cases like *Tenure*. But that 'good' would not justify imposing punishment and would leave the underlying problem and gap in place. Even cases where punishment may play a role in filling gaps, like *Oil Spill*, do not justify punishments based on psychological desires alone. Rather, an apt call for some kind of redress can only be fulfilled where someone faces punitive sanction. For another example, deeming people responsible for acts might make them more cautious in the future. That good outcome could leave backwards-looking gaps in place.

The present problem, again, concerns mismatches between philosophically justified and robustly intuitive responsibility attributions. The mismatches purportedly raise several responsibility-related issues. Gap-filling is distinctly apt only where it will address one of them. Conditions for justified gap filling otherwise too easily dissolve into a blunt consequentialism. Consequences are always relevant. The benefits in any benefit-risk calculation should be responsibility-relevant. Appeals to the value of ensuring accountability or redress are thus understandable, but desires for accountability and redress prove neither necessary nor sufficient for justified gap-filling. If, for instance, a gap only triggers needs for someone to provide reasons or face sanction, one may fill it to fulfill an answerability- or punishment-based good distinct from (at least some forms of) accountability (including the narrower form above/below).[39] The above already demonstrated that bare desires to, for example, hold someone accountable cannot justify gap-filling.

Even a minimalist version of this condition limits and helps to clarify when any gaps can and should be filled. 'Responsibility-relevant' admits multiple interpretations. van de Poel et al. (2012) define responsibility gaps as morally *problematic* and note that what qualifies as problematic depends on what you think responsibility is for: theories based on fundamental concerns with efficiency, retribution, etc. produce different problems and thus different gaps. One may charge that this condition requires a general theory of the purpose of responsibility or worry that a retribution-based view of responsibility could warrant filling gaps to fill psychological desires. However, the proposed condition's consistency with many views on 'responsibility-relevance' is a feature, not a bug: I seek a general account of when to fill gaps consistent with multiple understandings of responsibility and its aims and functions. This condition establishes a burden, and is thus action-guiding, on many general

[39] Kiener 2022 notes that, typically, "answerability interests are satisfied because human-induced harm normally comes with some people's moral answerability anyway … [but] if AI really creates responsibility gaps, then there is a human-induced harm without moral answerability." Only certain people can fill attendant answerability gaps.

theories of responsibility: whatever one's view, one must identify a responsibility-relevant good and explain how a gap-filling proposal will fill it. The condition may be more plausible when applied using some views rather than others. But this could count in favour of the views on which it is plausible, rather than rendering the condition suspect. For instance, it may simply demonstrate that retribution-based views of responsibility have issues. If so, applying my conditions can fruitfully guide broader conceptual analysis. In either case, the nature of the problem provides principled reasons for this condition's specific response.

My broader argumentative framework then places further restrictions on 'responsibility' and 'responsibility-relevant' that help avoid unduly expansive accounts of each. Responsibility is distinct from regulation, a feature of agents, and only attaches to certain entities for actions. So, no acceptable specification of my proposal can, for instance, seek to fulfill retributivist desires by deeming a non-actor responsible for an outcome or fill gaps to make people act differently.[40] This condition's import is most obvious where compensatory, punitive, and other forms of responsibility are enforced in ways that fundamentally impact persons' life interests, as where moral responsibility produces legal culpability/liability. However, it applies to a wider set of responsibility-relevant practices. Social stigma attached to otherwise-unwarranted blame attributions constitutes a cost that can only justifiably arise when it furthers some end. The condition's implications will depend on the nature of the gap and context in which it arises. For instance, a gap stemming from a failure to provide reasons that is filled to promote answerability can be fulfilled without any direct sanction but requires that the gap-filling circumstances be explainable, leading some to discuss concerns about opaque AI as producing gaps. A retribution-based gap can, in turn, only be filled by physical action. However, this merely highlights the variety of gaps that can occur. All relevant gaps plausibly submit to the basic schema at hand.[41]

[40] This condition admits difficult cases. E.g., Kiener 2022 suggests some persons may be obligated to take on responsibility to respect those harmed in a case or that doing so can beneficially exhibit one's virtue or agency. But his obligatory reasons focus on needs for accountability or answerability that can be accommodated. And Kiener holds that virtue- and agency-based concern can only ground answerability-based gap-filling duties. His conclusions do not exhaust the puzzle. Yet my account is no worse than others here: all general accounts admit borderline cases.

[41] One may challenge the extensional adequacy of even this capacious view. Those interested in 'backward-looking' responsibility attributions (e.g., Matthias 2004; Kohler et al. 2018) are criticized for failing to address other aspects of responsibility (e.g., Collins 2019: 949; Tigard 2021: 599; Glavaničová/Pascucci 2022). Exclusive focus on accountability may miss important concerns. Yet those interested in 'backwards-looking' responsibility gaps have long recognized that responsibility is not exhausted by liability to retributive punishment or even certain reactive attitudes; it also implicates the need to offer explanations (e.g., Matthias 2004: 175). Himmelreich 2019's conditions can be understood in light of concerns with answerability, if not attributability. Kiener 2022 rearticulates the mismatch concern for answerability issues. While 'responsibility gap' could attach to concerns with forward-looking, public accountability, or active responsibility, the problems in each do not clearly combine into one "bigger picture" (contra de Sio/Mecacci 2021). Even pluralists grant that they have different causes. Underlying concerns also qualitatively differ. It would accordingly be surprising if they all submitted to similar solutions. Distinctions above still apply and our gaps admit variety.

As part of this schema, filling responsibility gaps is now only apt where gap-filling will address a mismatch in responsibility attributions and thereby further some responsibility-relevant good. Knowing this is important even if further work is necessary to understand how different forms of responsibility, responsibility gaps, and responsibility-relevant moral goods relate. If, for instance, it is unclear whether one kind of responsibility can fill gaps in another, whether and when we should even attempt to fill a gap remains important. Conditions above/below provide useful guidance.

4.5 Condition 5: A Specifiable Candidate Gap-Filler

If one has a sense of the good gap-filling will further, there is still a need to specify a particular entity or set thereof that not only can but should fill the gap. Gap-filling requires an agent take on the relevant kind of responsibility (or, minimally, fulfill the relevant function of responsibility such as compensating those harmed). The lack of agents explains why *Hurricane* does not create a real gap. What kind of agents should fill which gaps when remains contestable. Whether a state or government should address issues in a *Crash*-style case or a corporation is not only best-placed but ideally suited to fill gaps in *False Positives*-style cases requires further elaboration.

These issues become more complex when we recognize that some gaps can only be filled by particular entities. Corporate or state agency can fulfill some intended ends of gap-filling. Representatives can achieve some goods on their behalf as when a CEO apology accompanies corporate compensation for *Oil Spill*. Yet the corporate agent or its representative cannot address all harms. A new CEO for a company may not be well-placed to provide the reasons for action in a case. A CEO in a variant of *False Positives* using opaque AI may never be able to provide it.

Aforementioned cases where plausible gap-filling by A can only be operationalized through a related agent B make things more complicated still. Recall that corporations as such cannot fulfill all relevant moral goods. Positing corporate agency accordingly cannot fill all gaps. There may be a need for someone to fill the gaps. CEOs are, again, often deemed responsible for corporate actions partly due to perceived needs to address this worry. This is likely because the 'buck stops' with the CEO in many states' corporate laws. Yet the kind of responsibility at issue may not always track such buck-stopping responsibility. Whether it should always do so in gap-filling cases is partly what is at issue. Similarly, mere acceptance of corporate laws cannot establish CEOs should always be accountable for harms in a case, let alone that they should face sanction. And CEOs cannot always provide the reasons for outcomes desired by those harmed.

If few worry about CEOs bearing too much responsibility, further cases where no one is well-placed to operationalize gap-filling roles remain compelling. Recall the incubator case: An inapt person seemingly 'filling' gaps can cause new problems. Further consider another case:

> *Scapegoat 1*: Full responsibility for a *Crash* is placed on a member of middle management with few actual operational duties or other roles. The position

> in the company is mostly symbolic with little pay for little work but complicated role descriptions that are interpreted as providing full responsibility for all crashes.[42]

Some appeals to Silicon Valley's frontier spirit are bald attempts to hide corporate agency. However, the possibility of important developments from non-corporate entities challenge intuitions stemming from anti-capitalist views. As de Jong (2020:733) notes, AI is often developed absent "obvious form of shared collaboration. If, on top of this, (part of) the software, for instance, is developed by several groups who are also not necessarily 'on the same team' either, it will be even more challenging to avoid responsibility-gaps."[43] Consider further:

> *Scapegoat 2*: AI developer teams work at separate locations to produce a tool like that one in *Beneficial Health Good*. Each works on a component independent of others. The eventual result passes relevant safety and efficacy standards. However, an adaptive machine learning development leads to a rash of misdiagnoses, causing harms. No team or member thereof has obvious control over the whole product or another reason for additional responsibility above that provided to others. But part of the AI tool licensing scheme required identification of someone who would be held responsible for errors. A lottery led to one team being responsible for picking out who would be responsible. The team voted for the least socially popular member. This person had little formal role in development, largely assisting. Yet the team's 'egalitarian' structure makes the unpopular lackey an 'equal' member – who now faces significant moral sanction and legal liability.

Intuitions about corporations should not obscure broader truths about which gaps to fill when. Where CEO gap-filling duties are plausible, other cases require specifying other gap-fillers.

This background helps motivate intuitions concerning another case:

> *Maldistribution of Benefits and Burdens:* A team of independent physician-computer programmers develops a machine learning-enabled competitor to the AI tool in *Beneficial Health Good*. It consistently meets safety and efficacy standards and, due to its learning capacities, produces better outcomes over time. It performs so well that the original tool loses its market share and is discontinued. The tool is continually audited to ensure it meets post-market performance standards. Non-negligible false positives produce high health costs for those falsely diagnosed and leads some to undergo unnecessary, sometimes-harmful healthcare procedures. No one is greatly financially or physi-

[42] The 'epistemic fault' Duijf 2018 believes explains most gaps is lacking in many cases here. Assume actor expertise if you worry about epistemic glosses. Duijf grants that gaps can arise from expert actions.

[43] de Jong 2020:733 suggests this challenges vicarious liability-attribution structures too. The structure needed for group agents in note 30 makes this acute by making group agents less common. de Jong also discusses cases where there is a "more particular goal and another person or set of persons is responsible for the means to achieve the goal." These suggest traditional means of using indicia of buck-stopping responsibility cannot resolve all gaps.

> cally harmed by the tool or its consequences. The joint value of the harms of false positives are high. The government will only permit the tool on markets if a member of the development team accepts responsibility for all related harms.

This case could be distinguished from *Scapegoat 2* by the programmer's stronger causal role. But the proposed outcome is unjust even absent scapegoating. Many persons enjoy very large benefits. Corresponding risks fall on a single person who, ex-hypothesi, is not fully responsible for bad outcomes that arise. This distribution of benefits and risks is unjust. No one person should be responsible for risks of a good from which many others derive many benefits (provided that they do not otherwise meet the conitions for full responsibility, which we can assume that they do not where there is a gap). Even if such a distribution were acceptable where risks are low, the large quantum of responsibility and thus amount of compensation owed in *Maldistribution of Benefits and Burdens* remains unjust.

These considerations are not only important in legal contexts where gap-fillers face financial liability or punishment but apply more broadly. An unfair distribution of (especially very large) ill-feeling can greatly impact the lives of those facing such attitudes. And distributions' effects remain germane even on permissive accounts of acceptable distributions of responsibilities within groups. Himmelreich and Lawford-Smith (2019) suggest external justification for group punishment does not always rely on how that punishment is distributed internally among group members. Smith (2009)'s concern with the way in which group punishments are borne by individuals is then at worst mitigated. Smith and others worry that the justification of group punishment depends on the justification of each cost borne by each member. But Himmelreich and Lawford-Smith contend that the effects of punishments can be distributed in proportion to members' contributions, equally among all members, randomly, or at the group's discretion. They defend a permissive and pluralistic approach to the distributions: any plausible account of groups who can be 'punished' permits various distributions (139) and groups may adopt an internal mechanism to plan in advance for how the effects of punishment will be distributed (144). Perhaps 'fair' distributions of burdens are multifarious and do not require identifying of distinct individuals who must fill gaps. Himmelreich and Lawford-Smith, in fact, even accept some scapegoating. They imagine a large corporation calling upon an after-hours cleaner who is not responsible for the contamination of a water supply contaminator to resign and make a large charitable donation on the company's behalf (147). The authors then accept that the cleaner should be able to freely choose to accept liability in exchange for a large compensation package.

Variety in the ways costs can be justifiably distributed does not, however, defeat this condition. I do not share Himmelreich and Lawford-Smith's intuitions about the cleaner. Those authors also notably focus on financial costs and punishments. Their points may not generalize. The cleaner could not fulfill goods related to, for example, apology even if they could donate and resign. More importantly, some distributions would be better than others even if Himmelreich and Lawford-Smith's conclusions generalize beyond punishment contexts. Many relevant cases do not distribute responsibility in the first instance to a collective that could choose how to distribute the 'goods,' limiting that avenue for identifying persons who can plausibly be called upon to take responsibility. Moreover, the need to provide some justification for why this individual should bear the costs remains even on permissive views. Himmelreich and Lawford-Smith grant that distributions can be morally better or worse and justifications of particular distributions will be case- and group-specific (2019: 144-146). Conditions justifying group choices on their proposal may not always obtain and one can say certain choices are preferable where they do. If no one appears apt, gap-filling may not be worthwhile even if it would otherwise be permissible.

Another non-damning objection holds there is no injustice where programmers can choose whether the goods are on markets. Avoiding such choices is preferable, as it would have been in *Delegation*. Matthias (2004:196) notes that society "collectively bears the cost resulting from" conditions not fully attributable to any individual elsewhere, as when storms destroyed the Mars Rover. Rather than avoid gaps where society benefits from a good, shared burdens appear preferable in cases like this. Calls for a no-fault compensation scheme for medical AI are thus plausible (Mahila et al. 2021). But one still must fairly distribute burdens when we let gaps arise.

A variant of *Maldistribution of Benefits and Burdens* in which no one is deemed responsible by the government and a programmer instead chooses to fill the gap may avoid some concerns but leaves open questions about whether and when one can choose to gap-fill. One still must specify an entity who can appropriately take on responsibility without distorting benefits and burdens and yet fulfilling the intended ends of gap-filling. Not all gaps can be filled in the same way or by the same person. Kiener (2022), for example, argues that persons can only voluntarily accept responsibility in the reason-giving sense since other forms of responsibility involve accepting fault, which cannot be accepted on one's own. Some gap-filling is thus impossible. This point too may not generalize outside the accountability context. However, the idea that certain forms of responsibility can only be accepted by particular persons in particular ways remains compelling regardless of one's thoughts on Kiener's proposal. One must determine whether anyone can plausibly be asked to fill a gap before requiring that they do so. Absent some specifiable agent who can and should appropriately fill a gap, unjust distribution risks remain.

If one thinks it is appropriate for a person to bear full responsibility in cases above, then, questions remain as to who should do so. This remains true where initially compelling answers leave key questions open. For instance, group agency and duty cases raise questions concerning whether those with 'steering power' should bear responsibility for filling gaps in group duty-fulfillment or if all members should do so and, if both, whether those with steering powers should bear more than others.[44] The need to address this underscores difficulties here. Either response raises challenges. Individual responsibility attributions already account for steering power. If such power was insufficient to trigger responsibility in the first instance, it is unclear why it should do so here. Focusing on members generally may, in turn, suitably diffuse the burdens but only at the expense of leaving responsibility gaps in place. It is, in short, unlikely that most collections of group members for any case

[44] Compare Zoller 2014; Collins 2017b. See also control-based arguments in Moen 2024.

could fulfill all compensatory duties and many members are ill-placed to further other goods above requiring specific individuals' actions.

Appeals to role responsibilities can then be read as mere specifications of this condition. But those who believe role responsibilities 'dissolve' the present problem by identifying a way of avoiding gaps in advance (e.g., Champagne/Tonkens 2015) may find that they do not provide a simple means of addressing pertinent issues. Fulfilling extant responsibilities will leave gaps in place. For instance, one can fulfill a duty to have backups for and checks on corporate or AI decision-making and still raise gap-producing circumstances.[45] Role responsibilities to accept gaps can then be undesirable. Consider a *Maldistribution of Benefits and Burdens* variant wherein previously-disparate groups incorporate with a common CEO. One person getting a label does not obviously significantly change the moral calculus. They may not even be able to accept responsibility to fill gaps by taking on a role. Per Kiener (2022), where role responsibilities exist, "blame is inappropriate in the absence of fault and, thus, the act of taking responsibility cannot extend liability to blame … [or] blame is appropriate in the presence of fault, but then the act of taking responsibility is redundant since appropriate blame tracks fault and not prior acts of taking responsibility at will." One could, of course, posit a faultless blame-like attitude to partly avoid this dilemma. However, such a response appears ad hoc absent independent reason to believe that it exists. And concerns that role-bound agents may be functionally unable to fill gaps remain even in these best cases. It is more likely that this is again a case where we are discussing who should take responsibility for a set of affairs after the fact. Role responsibilities do not appear to perfectly track this requirement. We are instead focused on a more general question concerning who can and can be appropriately called upon to take responsibility. This post hoc responsibility taking now appears to be distinct from buck-stopping-style ex-ante acceptance of responsibility for possible harms that may follow. This condition states that must be someone who can be appropriately called upon to do the latter work. I have now provided reason to believe that simple appeals to role responsibilities are unlikely to do so.

Those who seek to fill responsibility gaps, then, must identify an entity or set of entities who can and should be asked to fill it. This requires attending to various complications above. Justice demands attending to such concerns if one seeks to fill a gap for the appropriate reasons above.

4.6 Condition 6: No Alternative Means

Finally, one should only fill gaps where morally acceptable alternative means of providing relevant goods, such as accountability or redress, are unavailable. Responsibility gaps stem from philosophically justified responsibility attributions. Deviations, again, have costs. Some costs themselves constitute injustices. Costs can be acceptable to fulfill a responsibility-relevant good. However, one should not bear

[45] Collins 2019b:42. See also Collins 2017b, 2019 on Isaacs 2011.

unnecessary costs. Filling responsibility gaps is, again, second-(or even third-)best. Gap-filling should only occur where justifiable alternatives are unavailable.

This condition does not entail that gaps are presumptively acceptable or desirable. Making particular persons or groups responsible remains a non-ideal solution to problems posed by responsibility gaps. Consider circumstances where one seeks to fill a gap only to ensure people get redress and an insurance-based compensation scheme can provide such redress without holding anyone particularly responsible for the compensated harm, as in a no-fault scheme.[46] One can address the good gap-filling would aim to fulfill here without holding anyone responsible for attendant harms. Pre-theoretical intuitions about how much blame to attribute for harms still would not match the amount provided to individuals or even groups. This would, of course, leave gaps in place (contra e.g., Glavaničová/Pascucci 2022's reading). However, it would solve the underlying problem without raising new problems about fair distributions of responsibility and related benefits and risks. We should seek such alternatives before gap-filling.

This condition underlines independent evaluative reasons to prefer my account to alternatives. When combined with other conditions above, it explains the exceptional nature of gaps and gap-filling in a manner that identifies their unique contribution to morality. This itself explains the pull of views that seek to deny the existence of or otherwise dissolve gaps. Gaps should only be permitted under certain circumstances. Gap-filling will always deviate from our best theories of how to apportion responsibility and so be at best secondary to ideal apportionments and potentially tertiary to finding circumstances where gaps do not arise. Common attempts to avoid or dissolve gaps are thus well-motivated. Gaps and gap-filling remain distinct contributions to normativity and gap-filling addresses harms attendant to right actions. However, we can understand why genuine gaps appear rare and we seldom need to fill them. The proposal also offers a more capacious definition of gaps than those defining them as problematic by definition, providing a clearer target of analysis and broader range of evaluative tools for analyzing cases. Gaps as such can be morally neutral on the account above. One need not deny Danaher's claim that gaps may be desirable by definitional fiat. Gap-filling then always deviates from our best theories of responsibility and so is 'bad' in a relevant sense but can be 'good' or 'bad' all-things-considered. This allows more nuanced analyses of cases that are more likely to fit intuitions.

5 Conclusion

There is ample reason not to avoid or fill all responsibility gaps. The preceding argued for six conditions on when gap-filling is apt and one can be called on to fill a gap. If the conditions lacked a unifying purpose, that would be non-fatal to the account. If the normative realm is messy, the preceding would be better for reflecting that messy reality. Yet I provided independent support for each condition above.

[46] Recall Mahila et al. 2021. See also Hellström 2012 on societal responsibility.

Where the arguments for each connect to the basic structure and purpose of responsibility and responsibility gaps– and explain both why gaps are exceptional and competing intuitions about their value –they do not appear simply piecemeal.

While one may not, in turn, share my intuitions about cases above,[47] neither the existence of gaps nor my conditions rely on casuistry alone. And if gaps do not yet exist, they still present interesting questions about responsibility and how to fulfill its functions. Even some scholars who try to dissolve particular cases admit gaps could arise (Himmelreich 2019). If gaps are rare, they still present a distinct phenomenon demanding scrutiny. Cases above illustrate independently compelling arguments for when they occur and should be filled. And altering many will assuage contrary intuitions. For instance, adaptive machine learning-enabled AI 'learns' on its own and may eventually produce outcomes neither programmers nor regulators could reasonably predict. If no one is fully responsible for the changes, the quantum of harm will produce larger gaps. A *False Negatives* variant involving otherwise well-validated machine learning, like *Beneficial Health Good*, can be compelling.[48] Moreover, even if all apparent gaps can indeed be dissolved and genuine responsibility gaps never arise in the future, attending to reasons why apparent gaps arise and why one may want to address them can help further clarify why we care about responsibility attributions and when we may want to alter our theories.

If gaps exist, in turn, the proposed conditions provide plausible guidance on how to approach particular cases tied to responsibility and its attendant ends, rather than external considerations. The structure of analysis permits one to analyze cases across domains, from corporate responsibility to AI, that feature a host of different forms of responsibility, from classic blameworthiness to explanatory duties. The proposed account's focus on responsibility-specific ends provides a distinct normative framework for analysis. And while some may find the lack of strong statements on the nature and purpose of responsibility in the preceding frustrating, the broader account's consistency with, but non-reliance on, responsibility pluralism ultimately makes it action-guiding not only on pluralist views of responsibility but also on monist views focused on each component. The framework's multifarious nature ensures its action-guidingness even if some criticisms of gaps hit their mark. If, for instance, blame-based gaps cannot exist or be filled, the above still provides guidance for cases involving answerability gaps. And if lingering objections defeat any condition above, knowing why this is so can help address ongoing theoretical and practical issues across several philosophical domains. If nothing else, the preceding is a helpful touchstone for further analysis of responsibility gaps and when to fill them.

[47] Compare Himmelreich 2019 on AWS; Duijf 2018 on *Tenure*.

[48] Positing state action likewise leaves the issue in place. Copp 2006 begins work on related topics with the military strike-induced sinking of an Argentine cruiser during the Falklands War. Yet Copp's description makes it such that no individual can be fully responsible for the strike as decision-making and contributory actions were diffuse. State liability then merely demonstrates that one must identify different kinds of group agents to fill gaps.

Declaration

Conflict of interest None.

References

Bovens, M. 1998. *The Quest for Responsibility*. Cambridge UP.

Braham, M., and M. van Hees. 2011. Responsibility Voids. *Philosophical Quarterly* 61 (242): 6–15.

Buell, S.W. 2018. The Responsibility Gap in Corporate Crime. *Criminal Law and Philosophy* 12 (3): 471–491.

Burri, S. 2017. What is the Moral Problem with Killer Robots? In Strawser, B.J. et al. (eds.), *Who Should Die*: *The Ethics of Killing in War*. Oxford UP.

Caruso, G.D., and Pereboom, D. 2022. *Moral Responsibility Reconsidered*. Cambridge UP.

Champagne, M., and R. Tonkens. 2015. Bridging the Responsibility Gap in Automated Warfare. *Philosophy and Technology* 28 (1): 125–137.

Chapman, B. 1998. More Easily Done than Said. *Oxford Journal of Legal Studies* 18: 293–329.

Collins, S. 2017a. Filling Collective Duty Gaps. *Journal of Philosophy* 114 (11): 573–591.

Collins, S. 2017b. Duties of Group Agents and Group Members. *Journal of Social Philosophy* 48 (1): 38–57.

Collins, S. 2019. Collective Responsibility Gaps. *Journal of Business Ethics* 154 (4): 943–954.

Copp, D. 2006. On the Agency of Certain Collective Entities. *Midwest Studies in Philosophy* 30 (1): 194–221.

Danaher, J. 2016. Robots, Law and the Retribution Gap. *Ethics and Information Technology* 18: 299–309.

Danaher, J. 2022. Tragic Choices and the Virtue of Techno-Responsibility Gaps. *Philosophy and Technology* 35 (2): 26.

de Jong, R. 2020. The Retribution-Gap and Responsibility-Loci Related to Robots and Automated Technologies. *Science and Engineering Ethics* 26: 727–735.

de Sio, F.S., and G. Mecacci. 2021. Four Responsibility Gaps with Artificial Intelligence. *Philosophy and Technology* 34 (4): 1057–1084.

Duijf, H. 2018. Responsibility Voids and Cooperation. *Philosophy of Social Sciences* 48 (4): 434–460.

Enoch, D. 2012. Being Responsible, Taking Responsibility, and Penumbral Agency. In *Luck, Value, and Commitment*, ed. U. Heuer and G. Lang, 95–132. Oxford UP.

Glavaničová D., and Pascucci, M. 2022. Vicarious Liability. Ethics and Information Technology 24:art 28.

Grübler, G. 2011. Beyond the Responsibility Gap. *AI and Society* 26: 377–382.

Hellström, T. 2012. On the Moral Responsibility of Military Robots. *Ethics and Information Technology* 15 (2): 99–107.

Himmelreich, J. 2019. Responsibility for Killer Robots. *Ethical Theory and Moral Practice* 22 (3): 731–747.

Himmelreich, J., and H. Lawford-Smith. 2019. Punishing Groups. *The Monist* 102 (2): 134–150.

Husak, D. 2011. The De Minimis 'Defence' to Criminal Liability. In *Philosophical Perspectives on the Criminal Law*, ed. R.A. Duff and S. Green, 328–351. Oxford UP.

Inesi, A. 2006. A Theory of De Minimis and A Proposal for Its Application in Copyright. *Berkeley Tech. L.J.* 21(2):945–995.

Isaacs, T. 2011. *Moral Responsibility in Collective Contexts*. Oxford UP.

Kiener, M. 2022. Can We Bridge AI's Responsibility Gap at Will? *Ethical Theory and Moral Practice* 25: 575–593.

Köhler, S., et al. 2018. Technologically Blurred Accountability? In *Moral Agency and the Politics of Responsibility*, ed. C. Ulbert, et al., 51–67. Routledge.

Königs, P. 2022. Artificial Intelligence and Responsibility Gaps. *Ethics and Information Technology* 24(3):art 36.

Kornhauser, L.G./Sager, L.A. 1993. The One and the Many. *California L.R.* 81(1):1-59.

Lawford-Smith, H., and S. Collins. 2017. Responsibility for States' Actions. *Philosophy Compass* 12 (11): e12456.

List, C. 2021. Group Agency and Artificial Intelligence. *Philosophy and Technology* 34: 1213–1242.

List, C., and Pettit, P. 2011. *Group Agency*. Oxford UP.

Mahila, G., et al. 2021. Artificial Intelligence and Liability in Medicine. *Milbank Quarterly* 99 (3): 629–647.

Matthias, A. 2004. The Responsibility Gap. *Ethics and Information Technology* 6: 175–183.

McKenna, M. 2019. Basically Deserved Blame and its Value. *Journal of Ethics and Social Philosophy* 15 (3): 255–282.

Moen, L.J.K. 2024. Against Corporate Responsibility. *Journal of Social Philosophy* 55 (1): 45–61.

Mukerji, N., and C. Luetge. 2014. Responsibility, Order Ethics, and Group Agency. *Archiv Für Rechts- und Sozialphilosophie* 100 (2): 176–186.

Nelkin, D.K. 2016. Difficulty and Degrees of Moral Praiseworthiness and Blameworthiness. *Nous* 50 (2): 256–378.

Nyholm, S. 2017. Attributing Agency to Automated Systems. *Science and Engineering Ethics* 24: 1–19.

Oimann, A.-K. 2023. The Responsibility Gap and LAWS. *Philosophy and Technology* 26: 3.

Owens, D. 2012. *Shaping the Normative Landscape*. Oxford UP.

Pereboom, D. 2021. *Wrongdoing and the Moral Emotions*. Oxford UP.

Pettit, P. 2007. Responsibility Incorporated. *Ethics* 117 (2): 171–201.

Ross, W.D. 1920/2002. *The Right and the Good*. Oxford UP.

Shoemaker, D. 2011. Attributability, Answerability, and Accountability. *Ethics* 121: 602–632.

Shoemaker, D. 2015. *Responsibility from the Margins*. Oxford UP.

Da Silva, M. 2022. Autonomous Artificial Intelligence and Liability. *Philosophy and Technology* 35 (2): 1–6.

Smith, T. 2009. Non-Distributive Blameworthiness. *Proceedings of the Aristotelian Society* 109 (1): 31–60.

Sparrow, R. 2007. Killer Robots. *Journal of Applied Philosophy* 24 (1): 62–77.

Swoboda, T. 2017. Autonomous Weapon Systems. In *Philosophy and Theory of Artificial Intelligence 17*, ed. V.C. Müller, 302–213. Springer.

Talbert, M. 2019. Moral Responsibility. *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/moral-responsibility/.

Thompson, D.F. 1980. The Quest for Responsibility. *American Political Science Review* 74 (4): 905–916.

Tigard, D.W. 2021. There Is No Techno-Responsibility Gap. *Philosophy and Technology* 34 (3): 589–607.

Tollon, F. 2023. Responsibility Gaps and the Reactive Attitudes. *AI and Ethics* 3: 295–302.

van de Poel, I., et al. 2012. The Problem of Many Hands. *Science and Engineering Ethics* 18 (1): 49–67.

Veech, M.L./Moon, C.R. 1947. De Minimis Non Curat Lex. *Michigan L.R.* 45(5):537–70.

Wallace, R.J. 1994. *Responsibility and the Moral Sentiments*. Harvard UP.

Wolf, S. 2000. The Moral of Moral Luck. *Philosophic Exchange* 31: 4–19.

Zając, M. 2020. Punishing Robots. *Journal of Military Ethics* 19 (4): 285–291.

Zoller, D.J. 2014. Distributing Collective Moral Responsibility to Group Members. *Journal of Social Philosophy* 45: 478–497.