AI Workslop: The Moral Significance of Withholding Effort

Downloaded from: https://research.chalmers.se, 2026-10-17 03:18 UTC

Citation for the original published paper (version of record):
de Fine Licht, K. (2026). AI Workslop: The Moral Significance of Withholding Effort. Philosophy & Technology, 39(3). http://dx.doi.org/10.1007/s13347-026-01164-8

N.B. When citing this work, cite the original published paper.

research.chalmers.se offers the possibility of retrieving research publications produced at Chalmers University of Technology. It covers all kind of research output: articles, dissertations, conference papers, reports etc. since 2004. research.chalmers.se is administrated and maintained by Chalmers Library

(article starts on next page)

Philosophy & Technology (2026) 39:150
https://doi.org/10.1007/s13347-026-01164-8

RESEARCH ARTICLE

AI Workslop: The Moral Significance of Withholding Effort

Karl de Fine Licht[1]

Received: 2 December 2025 / Accepted: 3 August 2026
© The Author(s) 2026

Abstract
This paper examines the moral status of "AI workslop": AI-generated output that appears adequate but lacks the substance a task requires. Existing research has measured its prevalence and costs and suggested managerial responses, but has not addressed when producing workslop is wrong, permissible, or required. I connect AI workslop to an older problem in workplace ethics: effort-withholding, as in shirking or slacking. Workslop is often perceived as continuous with these practices because it can shift burdens onto others while appearing to spare the sender effort. But the inference from workslop to withheld effort is insecure. Workslop is individuated by output and tool use, not by intention or effort. Its distinctive profile lies in intention-independence, its uncertain relation between output and effort, and the distance AI delegation can place between workers and burdens imposed on others. I argue that responsibility for producing, preventing, and responding to workslop turns on the distribution of control. Managers shape structural conditions, including workloads, deadlines, metrics, AI policies, and training; workers control local choices about whether and how to use AI, how carefully to check its output, and how much burden they pass on. Drawing on harm-based considerations shared across moral traditions, I argue that producing workslop is wrong when it imposes substantial, avoidable, and nonredundant burdens that could reasonably have been prevented; permissible when harms are minor or structurally unavoidable; and, in narrow non-ideal cases, required. The account favors structural reform and worker support over punishment, reserving sanctions for high-stakes settings and repeated deliberate offloading.

Keywords AI workslop · Ethics of artificial intelligence · AI governance · Moral responsibility · Workplace technology

✉ Karl de Fine Licht
karl.definelicht@chalmers.se

[1] Chalmers University of Technology, Vera Sandbergs Allé 8, Gothenburg 411 33, Sweden

Published online: 10 August 2026

1 Introduction

Generative AI (GenAI) can support many types of work and improve efficiency in a wide range of tasks (Dell'Acqua et al., 2023; Brynjolfsson et al., 2025; Jarrahi et al., 2025). It enables rapid text production, assists with routine decision-making, and can help users navigate complex information environments, just to mention a few examples (ibid.). At the same time, its use raises well-known concerns. Models may reproduce or amplify biased patterns in their training data, and their underlying mechanisms many times remain sufficiently opaque that users often cannot determine why a system produced a particular output (Bolukbasi et al., 2016; Buolamwini & Gebru, 2018; Burrell, 2016; Bommasani, 2021; Lipton, 2018). This opacity creates challenges for assigning responsibility when errors occur or when outputs are relied upon in consequential contexts. GenAI can also shift workplace expectations in ways that affect evaluation, autonomy, and workload distribution (Dillon et al., 2025; Mayer et al., 2025; Salari et al., 2025). Taken together, these capabilities and limitations make GenAI both a valuable and a potentially problematic tool.

However, even though these issues have received considerable attention, the ethical dimensions of what has recently been dubbed "AI workslop" remain underexamined. AI workslop is the workplace-specific instance of a broader phenomenon known as "AI slop," now documented across academic publishing, software development, and organizational communication (Baltes et al., 2026; Giray et al., 2026; Niederhoffer et al., 2025). Niederhoffer et al. (2025) define workslop as "AI-generated work content that masquerades as good work, but lacks the substance to meaningfully advance a given task." Existing research has primarily described this practice, documented its prevalence, or proposed operational responses for organizations concerned about its effects. Related empirical work has also examined adjacent phenomena, such as the productivity and quality effects of GenAI on customer-support agents, the risks and burdens experienced by knowledge workers using AI in hybrid settings, and the maintenance burdens created when AI-generated code requires substantial rework (see, e.g., Baltes et al., 2026; Brynjolfsson et al., 2025; Xu et al., 2025). Together, these studies show that GenAI can generate outputs that appear adequate while shifting verification, correction, or interpretive labor onto others. Nevertheless, they do not provide a normative account of when producing such outputs is morally wrong, permissible, or potentially required, nor do they clearly distinguish AI workslop from broader forms of shirking, slacking, or workplace resistance.

In this paper, I develop a role-sensitive account of the moral status of AI workslop. I argue that AI workslop is often perceived as a technologically mediated variant of familiar forms of effort-withholding, such as shirking, slacking, and workplace resistance. That perception is understandable, since workslop can impose burdens on others in ways similar to these older practices. But the inference from poor AI-generated output to withheld effort is often insecure. Unlike ordinary cases of shirking or slacking, AI workslop does not by itself show how much effort the worker invested, what she intended, or whether the failure resulted from negligence, poor judgment, inadequate guidance, workplace pressure, or the limits of the system itself. For this reason, AI workslop requires separate ethical analysis. The account proceeds in three steps. First, I examine how control over the production of workslop is dis-

tributed across managers and workers, contrasting the structural levers available to managers with the more local discretion workers exercise in deciding whether and how to use GenAI and in shaping how their outputs affect others. Second, I assess the moral status of the act itself, that is, when producing workslop is wrong, permissible, or, in rare cases, required. I do so by drawing on de Fine Licht and Folland's (2025) treatment of the harms and benefits of relying on AI in public decision-making, and by focusing on the severity, avoidability, and redundancy of the harms involved. I do not adopt our broader governance framework for public-sector AI; I apply one part of our analysis to a different problem, AI-generated work products in organizational settings. Third, I turn to the separate question of responsibility and blame. I argue that producing workslop is often not blameworthy even when it is wrong, because excusing conditions and unequal control frequently defeat fair attribution. The result is an account of AI workslop as a heterogeneous practice whose moral significance turns on role-specific capacities, workplace pressures, and the distribution of burdens within organizations. This account yields policy recommendations ranging from structural and supportive reforms to, in limited cases, formal sanctions.

This paper is structured as follows. Section 2 defines workslop, puts it in a broader family of effort-withholding behaviors, and explains why it is empirically difficult to detect. Section 3 analyzes the distinct spheres of control of managers and workers in producing and mitigating workslop. Section 4 develops the normative analysis, specifying when workslop is morally wrong, permissible, or obligatory and how this bears on responsibility and blameworthiness. Section 5 derives policy implications, focusing on structural reforms, proportional responses, and the limits of enforcement in non-ideal workplaces. Section 6 concludes and identifies directions for future research on the ethics of AI-mediated labor.

2 Background

AI workslop refers to AI-generated content that appears adequate but lacks the substance needed to meaningfully advance a task (Niederhoffer et al., 2025). It is often perceived as overlapping with a broader family of effort-withholding behaviors, including slacking, shirking, and everyday workplace resistance, in which workers reduce or redirect the effort they put into their work (Kidwell & Bennett, 1993). But AI workslop is not simply a technologically mediated version of these older categories. The differences matter. What sets it apart is not that workslop is concealed while shirking, slacking, and workplace resistance are open. Those older practices are also often undetected. Rather, AI workslop has three distinctive features.

First, and most fundamentally, AI workslop is intention-independent. The classical categories are partly defined by the worker's own engagement with the task. Shirking and everyday workplace resistance involve withholding or redirecting effort (Alchian & Demsetz, 1972; Ackroyd & Thompson, 1999). Social loafing involves a more diffuse, and often unconscious, drop in motivation (Latané et al., 1979; Karau & Williams, 1993). Workslop, by contrast, is not defined in either of these ways. It is defined by the output and by the tool used to produce it, regardless of how much effort or engagement the worker brought to the task. The same artifact may there-

fore come from an earnest but under-resourced worker who invested real effort, or from a strategic free rider who invested none. Effort-withholding may be a common source of workslop, but it is not a defining condition of it. Because effort-withholding accompanies workslop in practice, recipients may sometimes reasonably infer it from the output, and much of the indignation directed at perceived workslop plausibly tracks this inferred motive rather than the artifact itself. But the inference can misfire. The same fluent but hollow output is also consistent with an earnest worker who lacked the time or resources to do better. The anger the concept invites is therefore an unreliable guide to what the producer actually did.

Second, AI-mediated delegation can lower the perceived moral cost of imposing burdens on others through a mechanism that has no close analogue in unmediated effort-withholding. Köbis et al. (2025) found that people were much more willing to have a machine act dishonestly on their behalf when the interface allowed them to induce the behavior without specifying it, that is, by issuing a vague goal rather than an explicit instruction. Across their die-roll studies, honest reporting fell sharply when the interface gave participants greater plausible deniability about what they were asking the machine to do. The mechanism is moral disengagement: the less precisely an agent has to describe the conduct they are delegating, the less they experience it as their own. Their setting concerns deliberate cheating, not the heterogeneous and often unintentional phenomenon of workslop, so the analogy is partial. But the same structural feature is present whenever work is delegated to a generative model through an open-ended prompt. The worker specifies a high-level want, such as "draft the report" or "summarize these findings," and the system fills in substance that the worker never explicitly chose and may never closely inspect. This weakens the perceived link between the worker's intention and the output that lands on others. Burden-shifting can therefore occur without the worker fully registering it as something they did.

Third, AI workslop creates a distinctive problem of identification for structural reasons, not simply because the worker hides it. The problem is not only, or even primarily, that recipients may fail to detect withheld effort. In many cases that is the wrong question, since workslop need not involve withheld effort at all. The more general problem is that generative models can supply, at near-zero cost, the surface markers by which adequacy is ordinarily inferred: fluency, structure, relevant terminology, and a recognizable task format. These features make several judgments difficult at once: whether the work is substantively adequate, whether it is workslop, whether it is AI-generated, and, only further downstream, whether it reflects withheld effort.[1] The difficulty is compounded by the definition of AI workslop itself: because it turns both on output quality and on reliance on GenAI, recipients may have to distinguish it not only from adequate AI-assisted work, but also from ordinary low-quality work. Similar difficulties arise even in mathematics, a discipline whose arguments are meant to be transparent and independently verifiable. The Leiden Declaration on Artificial Intelligence and Mathematics (2026), for example, warns that

[1] For example, studies testing whether instructors can reliably identify AI-generated student work likewise frequently show low accuracy (e.g., Scarfe et al., 2024; Waltzer et al., 2024) which implies that work done by GenAI is hard to detect.

current automated techniques can produce plausible but unreliable, or even incorrect, arguments that are hard to distinguish from correct proofs, including at the level of formalization. Erroneous machine-generated results may therefore accumulate in the literature and propagate as later work builds on them. If fluent but hollow output can evade detection in a field built on proof, the broader epistemic difficulty is unlikely to be confined to "softer" domains. Detectability will vary across domains and tasks, of course, but the general problem remains: AI-generated outputs can be hard to identify and assess reliably in many settings.

Taken together, these three features show that AI workslop, although often perceived as continuous with older forms of effort-withholding, is not reducible to them. It is individuated by its output rather than by motive. It weakens the link between an agent and the burdens they impose. And it is difficult to identify reliably because it can present the surface markers of adequate work while lacking the substance needed to advance the task. This also means that some of the ordinary recommendations given by, for example, Niederhoffer and colleagues, such as encouraging a "pilot mindset," establishing clearer norms for AI use, and providing more explicit organizational guidance, will not always be sufficient to address the underlying problem. The reason is that the worker producing workslop may already be doing the best they can under the circumstances.

Showing that AI workslop differs from older forms of effort-withholding is not yet enough to show that it warrants sustained ethical analysis. Workslop is, however, plausibly a problem for at least two further reasons: it is probably widespread, and it is potentially harmful. Even though AI workslop is difficult to detect with confidence in individual cases, survey-based research provides some indication of its perceived prevalence and effects. In a representative survey of American full-time desk workers, Liebscher et al. (2026) found that 52.5% of respondents reported having sent at least some AI-generated work that they considered low-effort, low-quality, or unhelpful. This figure should not be read as showing that workslop is evenly distributed across workplaces. It captures whether respondents had sent any such content, not how often they did so or how much of the burden they imposed on others. On the recipient side, 37.6% reported receiving such work in the past month. Those recipients estimated that this kind of work cost them roughly 3.3 h per month on average, with most of that time spent reading, evaluating, and reworking it. The burden also appeared to be unevenly distributed within local working networks. Recipients reported receiving workslop from an average of 2.2 colleagues. When this number was set against the average number of colleagues with whom respondents regularly worked, the resulting incidence rate suggested that roughly one in five close colleagues had sent them workslop. Thus, the findings suggest both that sending at least some workslop is relatively common and that the experienced burden may be concentrated among a smaller subset of colleagues in recipients' immediate work environments.

These findings fit within a wider body of evidence on how GenAI reshapes the conditions of work. Synthesizing micro-level evidence across firms, Fregin et al. (2026) document heterogeneous effects of AI adoption on productivity, work, and skills, rather than uniform efficiency gains. Ranganathan and Ye (2026) make the point more sharply: AI integration often intensifies work rather than relieving it, which challenges the assumption that such tools simply free workers' time for higher-

value tasks. Qualitative evidence from adjacent settings supports the same picture of asymmetric burden shifting. In a study of developer discourse about AI-generated contributions in software development, Baltes et al. (2026) document a recurring complaint that generation has become cheap while review has become expensive. Reviewers describe themselves as effectively absorbing the verification work that producers have offloaded. Although their analysis concerns a broader phenomenon than AI workslop as defined here, it independently identifies the same pattern of cost externalization that Niederhoffer and colleagues observe in organizational settings.

3 Controlling the Slop

The production of AI workslop depends on how control is distributed within organizations. Managers and workers contribute to it in different ways. They occupy different roles, face different constraints, and exercise different kinds of discretion. Identifying these role-specific points of control matters for assessing both the moral significance of workslop and the responsibilities that follow from it.[2] The aim is not to list every possible source of influence, but to identify the main levers that matter for the argument. I organize these levers through the upstream/proximate distinction used in Sect. 4. Section 3.1 identifies the structural levers available to managers. Section 3.2 examines the more local forms of control workers exercise in their day-to-day use of GenAI.

3.1 Managers

Managers and other decision-makers occupy positions that give them control over many of the structural conditions under which work is carried out. They decide how tasks are organized, how performance is measured, and which tools and resources are made available. At the same time, there are aspects of work they do not directly control, such as individual workers' skills, personal circumstances, or moment-to-moment choices in using GenAI. Any account of role-specific responsibilities for workslop must therefore begin by clarifying what falls within managerial control and what does not. One obvious way managers shape the production of workslop is through workload and pacing. By deciding how many tasks are assigned, how time is distributed across projects, and what counts as a normal pace of work, they influence whether workers have meaningful room for substantive engagement with their tasks. When output targets are raised or staffing is reduced, reliance on GenAI becomes more likely as workers try to keep up. Managers cannot directly determine how much

[2] The domains identified below are not an arbitrary list. They track the points along the causal pathway from working conditions to downstream burden at which an agent has leverage: managers control the upstream, structural conditions of work (workload and pacing, evaluation criteria, AI norms and procedures, training, and the allocation of review burdens), while workers control the proximate, local production and distribution of output (whether and how to use AI, the quality of what they pass on, disclosure, and the signaling of constraints). These are also the dimensions that recur in the empirical literature on the drivers of workslop and of AI-related strain at work (Niederhoffer et al., 2025; Liebscher et al., 2026; Ranganathan & Ye, 2026).

effort any one worker expends on a given task, but they do shape whether time for substantive work is available at all.

Managers also influence workslop by setting the criteria through which work is evaluated. Performance systems can reward speed, volume, and surface polish, or instead emphasize accuracy, justificatory depth, and task-relevance. Although managers cannot fully determine how every evaluator interprets these standards in practice, they do decide which indicators are formally tracked, rewarded, and sanctioned. In this way, they can make it more or less attractive for workers to rely on GenAI to produce outputs that look adequate while lacking substance. A further point of control lies in the rules and expectations governing GenAI use itself. Managers can decide whether AI tools are integrated into particular workflows, whether AI-assisted drafts must be disclosed or documented, and whether additional verification steps are required before AI-generated material is shared. They cannot monitor every interaction with an AI system, but they can establish policies, review mechanisms, and training practices that structure how such tools are normally used within the organization.

The same is true of competence and support. Managers can provide or withhold training, make guidance on prompting and verification available, and create opportunities for workers to learn how to incorporate AI outputs into task-specific workflows. Crucially, the relevant competence is not only general AI literacy but also domain expertise in the task at hand. A worker who knows a subject well is far better placed to notice where AI output is shallow, wrong, or subtly off, whereas one operating outside their expertise may be unable to tell adequate output from plausible-looking slop. This makes the match between a worker's knowledge and their assignments a further lever of control. By deciding who is asked to do what, and whether they have the background the task requires, managers shape not just general competence but each worker's concrete ability to evaluate and correct AI output on a given task. They cannot guarantee that all workers will reach the same level of competence, but they can determine whether workers are left to rely on GenAI without preparation, are assigned tasks they are not equipped to vet, or are instead given the resources and assignments needed to evaluate and revise AI-generated content responsibly.

Managerial decisions also affect how the burdens generated by workslop are distributed once it occurs. Review and correction do not fall evenly across organizations: responsibility for checking AI-assisted work can be assigned to particular roles, teams, or hierarchical levels, and managers help determine whether these burdens are centralized, diffused, or covertly shifted downward. Although they cannot fully control how often workslop arises, they can shape who is most exposed to its downstream effects and how those effects are managed. Managers are also potential producers of AI workslop. They may use GenAI to draft emails, reports, presentations, or strategic documents that appear polished while lacking substantive analysis or accurate information about the organization's situation. When this happens, the effects are often amplified by hierarchy: vague or misleading AI-generated material from managers can generate confused priorities, unnecessary work, and substantial demands for clarification and repair among subordinates. Managers cannot fully control how their AI-assisted outputs will be interpreted, but they can reasonably foresee that deficiencies in their own use of AI may propagate through the organization rather than be locally absorbed.

Seen in this way, the managerial role is defined by a set of control points over the structural features of work that make AI workslop more or less likely, channel its costs toward particular groups, and sometimes amplify its effects through managerial use of AI itself. These features will matter in the later assessment of when producing workslop is morally permissible or problematic, since they determine which aspects of the phenomenon fall within managerial control and which depend on workers' more local choices. Table 1 summarizes these dimensions of control.

3.2 Workers

Workers occupy positions that may give them control over how they use GenAI in their own tasks and how their outputs affect others, but where they often have limited influence over the broader structures in which they work. They typically cannot set workloads, define performance metrics, or determine which tools are available. Instead, they operate within these constraints and make more local decisions about whether to employ AI, how carefully to evaluate its outputs, and how to present AI-assisted work to others.

One important dimension of worker control concerns whether GenAI is used at all. Sometimes workers can decide whether to rely on GenAI for a given task and, if so, at what stage of the work process. They may use it to draft emails, summarize documents, structure reports, or generate code, or they may complete these tasks without GenAI assistance. Other times, this choice is constrained or even eliminated: where managers encourage or outright mandate the use of GenAI, the worker may retain discretion only over how the tool is used and how carefully its output is checked, not over whether to use it at all. Workers also shape the form and quality of the material that others receive. They may accept AI output largely as it stands, or they may revise, supplement, and partially rewrite it before sharing. They often cannot change the general behavior of the model, but they can influence the final output through

Table 1 Managerial control and its relevance to AI workslop

Domain of managerial control	What managers can influence	Relevance to AI workslop
Workload and pacing	Task allocation, staffing levels, output targets, timelines	High workloads or compressed deadlines increase reliance on GenAI and make workslop more likely.
Evaluation criteria	Performance metrics, reward structures, emphasis on speed vs. substance	Metrics favoring volume or polish incentivize surface-level AI output over substantive work.
AI norms and procedures	Policies for AI use, disclosure expectations, required verification steps	Clear rules can reduce harmful forms of workslop; ambiguous or absent rules create incentives to use AI opaquely.
Training and competence	Access to training, guidance on prompting and verification, support resources	Insufficient competence can lead workers to produce workslop unintentionally due to inability to assess AI output.
Distribution of review/correction tasks	Assignment of quality control responsibilities across teams or hierarchies	Determines who absorbs the burden of identifying and repairing workslop; often shifts costs to particular groups.
Managers' own AI use	Use of AI in reports, directives, strategic documents	Workslop produced by managers has amplified downstream effects, triggering confusion and unnecessary work for subordinates.

prompting, requesting revisions, and deciding how much time to devote to checking and editing. In this way, they determine whether others encounter raw AI text, code, etc., or lightly revised drafts, or more integrated and task-specific material.

A further aspect of worker control concerns transparency. Where no explicit disclosure rules are in place, workers can decide whether to indicate that a text, summary, or analysis was AI-assisted or whether to present it without mentioning the tool. They cannot fully control how such disclosure will be received, especially in settings where AI use is stigmatized or ambiguously regulated, but they do control whether recipients are given information that might shape how they interpret the work. Workers also influence how much interpretive and corrective labor is passed on to others. When they share material that is incomplete, unclear, or only superficially tailored to the task, they leave it to colleagues to infer missing context, check accuracy, and repair errors. They cannot determine exactly how much time recipients will spend on these activities, but they can often anticipate whether their outputs are likely to create substantial additional work downstream. Workers can also sometimes communicate about the conditions under which their work is produced. They can signal when deadlines, task loads, or instructions make it difficult to provide more than AI-assisted drafts, or when they lack the expertise needed to evaluate AI outputs in a particular domain. Such communication may not lead to structural change, but it can still matter by linking the quality and character of work products to the conditions under which they are produced.

Taken together, these features define the sphere in which workers exercise control in relation to AI workslop. Workers do not determine the overall structure of work or the incentives that make workslop more or less likely, but they do shape how GenAI is used in concrete tasks and how its outputs affect others. These local capacities will matter in the later analysis of when producing AI workslop counts as an acceptable use of limited options and when it constitutes a problematic shifting of burdens onto others. Table 2 summarizes these dimensions of worker control.

4 A Normative Framework for Evaluating AI Workslop

This section develops the normative analysis in two stages, corresponding to two distinct questions that the surrounding debate tends to run together. Section 4.1 asks about the moral status of the act, that is, when producing AI workslop is wrong,

Table 2 Summary of worker control in relation to AI workslop

Aspect of worker control	Description
Choice to Use AI	Workers can decide whether to use GenAI for a task, within limits imposed by deadlines, workload, and expectations.
Shaping Output Quality	Workers can revise, supplement, or rewrite AI output, influencing whether colleagues receive raw, lightly edited, or integrated content.
Transparency About AI Use	Workers can disclose or withhold information about AI assistance where no explicit rules require disclosure.
Burden Shifting Through Output	Workers can anticipate whether their outputs impose interpretive or corrective labor on colleagues and adjust accordingly.
Communication of Constraints	Workers can signal when structural constraints limit their ability to produce more than AI-assisted drafts or to verify outputs adequately.

permissible, or required, drawing on a harm-based framework. Section 4.2 asks the separate question of the agent's responsibility and blameworthiness. Keeping the two apart matters because an act can be wrong without its producer being an appropriate target of blame. It is precisely this gap, opened up by structural constraint and unequal control, that generates the policy upshot developed in Sect. 5.

4.1 The Moral Status of AI Workslop

To assess the moral status of AI workslop, we must begin with the harms and benefits it produces and with how those harms and benefits are distributed. When a worker intentionally or non-intentionally (e.g., habitually) submits AI-generated content that appears adequate but lacks the substance needed to advance a task, the result is often not merely a lower-quality output. It can also impose burdens on others: colleagues may need to interpret vague material, correct mistakes, supply missing reasoning, redo parts of the task, or absorb delays and coordination failures caused by the initial output. In some contexts, as we saw in Sect. 2, the relevant burdens extend beyond the organization itself. Clients, citizens, counterparties, and members of the public may also be harmed when AI-generated output is relied on in legal, clinical, safety-critical, or public decision-making settings. These are all harms in the relevant sense: they set back the interests of affected parties (Feinberg, 1984).

In what follows, I draw on Anna Folland's and my (de Fine Licht & Folland, 2025) normative treatment of the harms and benefits of relying on AI in public decision-making, which synthesizes normative principles from a comprehensive mapping of the harm literature. From that synthesis we take four considerations that matter for the moral assessment of harmful conduct: the severity of harm, whether it is avoidable or inevitable, whether it is redundant or nonredundant, and the asymmetry between harm and benefit, along with the limited conditions under which harmful conduct may be morally permissible.[3] I apply these to a different question, the moral status of AI-generated work products in organizational settings. The transfer is warranted because severity, avoidability, redundancy, and the harm–benefit asymmetry are general features of harm assessment rather than artifacts of the public-sector case: Folland and I draw them from the general harm literature, identifying points of convergence across the major moral traditions. These considerations are accordingly not idiosyncratic to that framework. That there is a pro tanto reason against harming others, and that this reason strengthens as the harm grows more severe and as it becomes more readily avoidable, function as prima facie commitments across those traditions: consequentialism treats harm as a setback to welfare that counts against an act; deontology houses it in a duty of non-maleficence that Ross (1930) ranks ahead of beneficence; contractualism would treat a principle licensing such harm as one others could reasonably reject (Scanlon, 1998); and virtue ethics registers it through the vices of callousness and injustice. Nothing in the argument turns on adjudicating

[3] For the main primary sources, see e.g.,: the reason against harming (Algander, 2013: 135; Gardner, 2017: 73–74; Shiffrin, 2012: 361); the proportionality of the reason to severity (Gardner, 2017: 83); avoidability (Gardner, 2017: 85); redundancy (Gardner, 2017: 86); the harm–benefit asymmetry (Bradley, 2013: 39; Shiffrin, 2012); and the conditions under which harm may be permissible—desert, necessity to avert a greater harm, and consent (Shiffrin, 2012: 362).

the differences among these traditions; the considerations are ones that a wide range of moral outlooks have reason to accept as bearing on the permissibility of imposing avoidable burdens on others.[4]

Now, a natural starting point is the general moral reason against harming. If producing AI workslop foreseeably makes others worse off, that counts morally against producing it (Gardner, 2017: 73–74; Shiffrin, 2012: 361). The strength of that reason depends in part on the severity of the harms involved (Gardner, 2017: 83). Minor workslop that costs a colleague a few minutes of clarification or revision generates a relatively weak moral reason against the act. By contrast, workslop that misleads decision-makers, significantly burdens already overstretched coworkers, delays urgent action, or creates serious downstream risks generates a much stronger reason against it. This point applies whether the affected parties are internal coworkers or external parties such as clients, citizens, or patients. In this respect, the moral significance of workslop depends not only on whether it is harmful, but on how serious the relevant harms are and on who is made to bear them.

A second consideration is avoidability (Gardner, 2017: 85). Harm matters more, other things equal, when it could reasonably have been avoided. This is highly relevant to AI workslop. In some cases, a worker could prevent the relevant burden at modest cost: for example, by checking the output more carefully, adding missing context, correcting fabricated claims, or refraining from using AI for a task whose demands exceed what the tool can responsibly support. In such cases, the reason against producing workslop is strengthened, because the burdens imposed on others are not merely unfortunate but reasonably preventable. By contrast, when heavy workloads, unreasonable deadlines, conflicting instructions, lack of training, or organizational pressure make some degree of shallow or incomplete AI-assisted output difficult to avoid, the reason against producing it is weaker. The central issue here is therefore not the mere availability of AI, but the broader structure of working conditions under which people are expected to perform.

A third consideration is redundancy (Gardner, 2017: 86). Some burdens associated with low-quality output would have arisen even without AI use. A worker may lack domain knowledge, receive inadequate instructions, or face task demands that exceed what can realistically be achieved in the time available. In such cases, the incremental harm attributable specifically to AI-assisted workslop may be partly redundant and therefore morally less weighty. By contrast, when AI use introduces new burdens that would not otherwise have arisen — for example, fabricated details, misleading structure, unwarranted confidence, or a polished surface that obscures substantive

[4] The overlap should not be overstated, and a reader may wish to test it consideration by consideration. The harm–benefit asymmetry, though shared by these non-consequentialist traditions and by commonsense morality, and recoverable within rule consequentialism (Hooker, 2000), is not fundamental to classical act consequentialism, which weighs equal harms and benefits symmetrically. Likewise, the discounting of redundant harm is most at home in welfarist and counterfactual-comparative accounts; agent-relative deontological constraints may judge a contribution wrong even where the relevant harm is overdetermined. Familiar moves narrow these gaps, the standard two-level construal of act consequentialism, for instance, can endorse the asymmetry as practical guidance even while denying it is right-making, and parallel manoeuvres are available on the deontological side. But the present argument does not need them. It requires only that the four considerations bear on the permissibility of imposing avoidable burdens in the practical contexts at issue, not that every tradition ground them in the same way.

weakness — the harm is nonredundant and morally more significant. This matters because the wrongness of workslop should not be assessed solely by reference to the final burden experienced by others, but also by whether the agent's use of AI added a distinct and avoidable source of that burden.

A fourth consideration is the asymmetry between harm and benefit (Bradley, 2013: 39; Shiffrin, 2012). AI workslop may produce benefits: it may save time, help a worker cope with competing demands, or prevent serious deterioration in performance across several tasks. But these benefits do not automatically outweigh the harms imposed on others. The moral reason against causing harm is generally stronger than the reason to produce a benefit of equal magnitude. A worker who saves thirty minutes by submitting largely unreviewed AI output, while thereby imposing an hour of corrective labor on a colleague, cannot justify the choice simply by pointing to net efficiency gains. The colleague's burden does not lose moral significance because the sender benefited. Benefits can matter, but their justificatory force depends on what they prevent or enable. A merely private benefit to the sender, such as saving time or avoiding effort, will normally be weak. A benefit that prevents serious failure elsewhere, protects third parties from harm, or allows scarce effort to be redirected to a more urgent task may carry greater weight. The relevant question is therefore not whether producing workslop generates some benefit, but whether the benefit is sufficiently important, and the harm sufficiently unavoidable, to justify imposing the burden.

These considerations help explain when producing AI workslop is morally wrong. It is morally wrong when it imposes substantial, avoidable, and non-redundant burdens on others, and when the agent could reasonably have prevented those burdens without excessive sacrifice. This includes cases in which a worker uses AI in ways that foreseeably shift significant interpretive or corrective labor onto similarly or more burdened colleagues, creates misleading or error-prone outputs that disrupt coordination or decision-making, or externalize costs onto clients, citizens, or other third parties who are not in a position to absorb them. In such cases, workslop is not merely low-quality work. It is a form of wrongful burden-shifting.

At the same time, not all AI workslop is morally wrong. It may be morally permissible when the harms are minor, when they are largely unavoidable under the relevant structural conditions, or when preventing them would require individuals to bear unreasonable costs. Permissibility here is an all-things-considered verdict rather than the absence of any reason against: the considerations just canvassed modulate the weight of the reason against harming rather than switching it on or off. Even permissible workslop may therefore leave a residual reason against it. These conditions include cases in which workers face chronic overload, conflicting instructions, inadequate training, or the absence of feasible alternatives, as well as cases in which teams explicitly agree to treat AI-generated drafts as provisional starting points for shared work. Consent matters here because recipients may willingly accept some additional interpretive labor as part of a cooperative arrangement. Even then, however, the scope of that permission is limited. Consent to receive rough AI-assisted drafts in a low-stakes internal setting does not by itself license the production of misleading AI output in contexts where serious third-party or public interests are at stake.

The most controversial claim concerns obligation. Here the argument must be narrower. The fact that AI workslop benefits a worker or even an organization does not by itself generate an obligation to produce it. At most, such an obligation may arise in rare non-ideal cases where producing limited AI-assisted output is necessary to prevent a greater harm, no feasible less harmful alternative is available, and the resulting output does not expose third parties or the public to grave risks. The relevant harms must be sufficiently serious: for example, serious threats to health, severe burnout, major and foreseeable breakdowns in role performance across multiple tasks, or comparably weighty harm to others that the agent has reason to avert. Ordinary convenience, marginal stress reduction, or routine productivity gains do not suffice. Nor is it enough that producing workslop would help the organization meet ordinary performance goals. The claim, rather, is that where a worker faces a serious and sufficiently weighty harm, where refusal or escalation is not a feasible option, and where limited AI-assisted output is the least harmful available course, producing such output may in some cases be morally required rather than merely permitted, because failing to do so would allow a more serious and avoidable harm to occur when no less harmful option is available. Even this narrower claim is sharply constrained by the requirement that serious harms to clients, citizens, or the public count fully in the analysis. In high-stakes legal, clinical, safety-critical, or public decision-making contexts, these external harms may be too serious for workslop to be permissible, let alone obligatory.

It is worth being explicit about why external harms function as a constraint on the obligation claim rather than as one more harm to be weighed in the balance. The obligation argument runs through a least-harm comparison among the worker's feasible options, and that comparison can license rough AI-assisted output internally in part because colleagues may accept provisional drafts as a cooperative arrangement. This is an instance of consent operating as a condition under which otherwise-objectionable harm becomes permissible (Shiffrin, 2012, p. 362). External parties, such as clients, patients, citizens, and counterparties, are differently situated: they typically cannot consent, cannot readily absorb the resulting costs, and are not participants in any such scheme. Their exposure is therefore not offset by the considerations that make obligation conceivable in the internal case. Once serious external stakes are present, then, the analysis shifts from "possibly obligatory" to "likely impermissible." A clinician whose burnout would be eased by letting an AI system draft clinical notes does not thereby acquire an obligation to produce them, because the risk to patients blocks it. A lawyer who could meet a filing deadline with unverified AI-generated output faces the same barrier where a fabricated citation would imperil the client's case. The very features that make obligation conceivable in the internal case, namely consent, mutual absorption of cost, and contained downstream effects, are precisely what is missing once serious third-party or public interests are at stake.

The upshot is that AI workslop does not have a uniform moral status. Its permissibility depends on the severity, avoidability, and redundancy of the harm it creates, on the benefits it secures, on whether consent is present, on whether costs are mutually absorbed or shifted onto others, on whether downstream effects are contained, and on the availability of less harmful alternatives. In many cases, workslop is morally wrong because it functions as an avoidable transfer of burdens onto others. In other

cases, especially under non-ideal structural conditions, it may be morally permissible. And in a narrow range of cases, where it is the least harmful option for preventing greater harms, it may even be morally required. The next section turns from the moral status of workslop itself to the distinct question of who is responsible or blameworthy for producing it.

4.2 Responsibility and the Limits of Blame

In discussions of AI workslop, there is often a high degree of indignation and blame. It is therefore natural to ask whether these reactive attitudes are justified and, more broadly, how responsibility should bear on the distribution of burdens and on institutional response.[5]

One possible answer is given by basic desert-based views, on which blame is appropriate only if an agent is substantially responsible for conduct that is morally wrong, rather than because blaming them would have good consequences or is otherwise justified on contractualist grounds. Such views require not only that the conduct be wrongful, such as that the worker intentionally or habitually (etc.) distributes AI workslop, but also that the agent satisfy demanding control and epistemic conditions of the kind discussed in the literature on moral responsibility, whether in incompatibilist form (see e.g., Strawson, 1994; Smilansky, 2000; Pereboom, 2001) or compatibilist form (see e.g., Wallace, 1994; Dennett, 2003; Watson, 2004). I will not try to settle here whether responsibility ultimately requires e.g., libertarian free will, or whether the more demanding incompatibilist conditions can ever be met. The question of free will and substantial responsibility arises for any human conduct whatever, and there is no reason to think AI-mediated work is a special case in this respect. Thus, the question will be whether producers of AI workslop fulfill the compatibilist criteria control and epistemic conditions.[6]

At the worker level, some basic control and epistemic conditions will plausibly often be present. Ordinary adult workers typically possess the capacities for practical reasoning, impulse control, and understanding needed for responsibility in the ordinary sense. They may exercise local control over whether and how to use GenAI, how carefully to review its outputs, and how much interpretive or corrective labor to shift onto others, although this control is reduced where GenAI use is formally or informally expected, as increasingly appears to be the case in some workplaces. In that sense, they are not merely passive conduits of organizational pressures.

[5] It should be noted that, even if someone deserves a response for what they have done, it is often difficult to determine exactly what they deserve. This is true both because of the "geometry" of desert itself (Kagan, 2014) and because, even once that issue is settled, it remains unclear what the appropriate response should be. As Scanlon writes, it may be proper to praise someone for their conduct in the course of their work without thinking that this should translate into a higher wage or other economic benefits (Scanlon, 2018: 117—133). I set these complications aside here and assume, for present purposes, that both the geometry of desert and the question of appropriate response are more straightforward than they actually are.

[6] There is also another thorny question regarding how exactly we should spell out the intention, habit (etc.). For me to be blameworthy for an action, and according to some theories, for the action to be morally wrong, I need to act from principles or intentions that for example others could reasonably reject or cannot be universalized (etc.). I will set this complication aside here and assume that many of the relevant cases of harmful AI workslop will include something akin to intention, culpable ignorance, negligence, etc.

When a worker's use of AI foreseeably generates avoidable burdens for others, that fact may count in favor of holding the worker responsible in some form, perhaps through blame, censure, or the assignment of additional remedial or supervisory obligations. At the same time, a wide range of excusing conditions can undermine or mitigate blameworthiness even when these general capacities are present. These conditions divide naturally along the two conditions the desert view already invokes. Some primarily defeat the control condition: time pressure, conflicting demands, coercion, mandated AI use, and other severe structural constraints leave the worker little room to have done otherwise. Others primarily defeat the epistemic condition: misleading instructions and reasonable ignorance may mean that the worker could not reasonably have known that their output would impose the burdens it did, for example because they thought the work looked adequate while lacking the necessary AI literacy skills. The distinction is not always crisp. Misleading instructions can both constrain a worker's options and keep them in the dark. But it is useful, because AI-mediated work sharpens a distinctive instance of each.

The epistemic condition is strained in a distinctive way by AI delegation. The mechanism discussed in Sect. 2 in connection with Köbis et al. (2025) is relevant here as well. Their experiments suggest that delegation to AI can reduce the perceived moral cost of unethical behavior, especially when interfaces allow principals to give vague or open-ended instructions rather than explicitly specify the conduct they want. Their setting concerns deliberate cheating rather than the more heterogeneous phenomenon of workslop, but the mechanism matters for responsibility. AI-mediated delegation can make it harder for workers to recognize and take responsibility for the burden-shifting they enable. By allowing workers to issue general instructions while the system determines much of the resulting output, the technology can create distance between the worker's intention and the burdens imposed on others. This does not remove responsibility, but it may reduce blameworthiness where the worker's contribution is less explicit, less self-consciously endorsed, and therefore less fully attributable to them.

The control condition, in turn, is eroded by the conditions under which work is produced. In their survey of desk workers, Liebscher et al. (2026) found that the *amount* of workslop someone sent was associated, among other factors, with low psychological safety at work. Because the study is cross-sectional, these findings identify associations but do not establish the direction or existence of causal relationships, but the finding bears directly on blameworthiness. Edmondson's (1999) influential work shows that workers' willingness to raise problems, acknowledge limitations, and resist unreasonable demands depends heavily on whether they can do so without fear of interpersonal punishment. Where psychological safety is low, options such as saying "the deadline makes proper checking impossible" are effectively unavailable, and what looks from the outside like culpable burden-shifting may instead reflect a structurally produced inability to communicate the true conditions under which work is being done. This does not eliminate worker responsibility, but it constrains its weight: a worker who could not safely surface concerns about task feasibility or AI suitability cannot fairly be held to the same standard as one who could. Notably, the same authors conclude that, given structural factors such as organizational pressure and low psychological safety, sending workslop "could be an adaptive, rational

response to dysfunctional systems," (Liebscher et al., 2026: 18) an empirical echo of the normative claim defended here.

Nonetheless, it may still be appropriate to treat workers as responsible in a forward-looking sense in cases when workers are not substantially responsible to any greater extent. Desert-based views tend to ask what a worker has done, whether it was morally wrong, and whether they are blameworthy for it; they are in this sense backward-looking, examining the past to determine what should follow now. According to some of these (retributivist) theories it's also good in itself if the person who did the wrong suffer for it. Forward-looking accounts, by contrast, are concerned with what a worker would do in the future, given that we hold them responsible in some way. This distinction between backward-looking and forward-looking responsibility is well established (see e.g., Goodin, 1986; van de Poel et al., 2015).

The harm-based framework of Sect. 4.1 feeds into both, though it plays a different role in each. In the backward-looking scenario, it tells us whether a given instance of workslop was wrong, and thus whether a negative reactive response, such as blame or censure, could in principle be deserved. But establishing wrongness is not sufficient for such a response, since the agent must also satisfy the control and epistemic conditions discussed above. In the cases this paper foregrounds, where low psychological safety or chronic overload excuses backward-looking blame, those conditions are often not met. In the forward-looking sense, the same framework identifies which instances of workslop are harmful enough to be worth preventing, and thus where institutional response is warranted. This remains true independently of whether anyone is blameworthy for past instances. Indeed, blame may often be unjustified not only because the relevant control or epistemic conditions are absent, but also because blame can undermine psychological safety. Forward-looking responsibility gives us a pro tanto reason in favor of intervention where no one is substantially responsible, and therefore blameworthy, for past conduct, but where future individual behavior can nevertheless be improved by holding someone responsible in a forward-looking sense. The two registers can sometimes converge. A worker's controllable past conduct may ground fair attribution backward-lookingly while also being the most effective lever for improving future practice, so that a response such as targeted training is doubly supported. But convergence is not required. Much of the account's force derives from cases where the registers diverge, where blame is inappropriate yet forward-looking response remains justified. The account therefore does not depend on the assumption that the same facts that justify blame must also justify intervention.

The same broad considerations apply to managers, but with an important difference. Managers do not merely act within a structure; they also help create it. They set workloads, define performance metrics, determine whether GenAI use is encouraged or constrained, decide whether workers receive adequate training, and shape how review burdens are distributed. They may also produce workslop themselves, with amplified downstream effects due to hierarchy. Because managers occupy positions of greater structural control, responsibility will often fall more heavily on them than on workers. This is true not only where managers directly produce AI workslop themselves, but also where they sustain the background conditions under which workslop becomes a rational or even unavoidable response. Given the control points described in Sect. 3.1, managers bear responsibility for structural features that make

AI workslop more likely and that channel its costs toward particular groups of workers or third parties. They set workloads, define performance metrics, adopt or reject expansive AI mandates, and decide whether to provide training and guidance on AI use. When these decisions foreseeably create conditions under which workers have little realistic alternative but to rely on GenAI in ways that tend to generate workslop, managers can be responsible for resulting patterns of low-quality output even if they do not produce any particular instance of workslop themselves. In such cases, their responsibility lies primarily in omissions and policy choices rather than in the proximate act of e.g., sending a specific document.

As with workers, there can be excusing or mitigating conditions for managers. They may face constraints from higher-level directives, limited budgets, or incomplete information about how AI tools function in practice. They may also be embedded in competitive environments that reward short-term output over long-term quality. These factors can reduce the extent to which individual managers are appropriate targets of blame or other responsibility-sensitive responses in particular cases. Still, because oversight of workloads, evaluation criteria, AI-related policies, and review structures is part of the managerial role, claims of ignorance or lack of control are often less plausible here than at the worker level, especially where patterns of workslop and their downstream costs are persistent and visible.

For these reasons, managers are likely to be central addressees of remedial expectations even where blame or other reactive responses are inappropriate. The fact that they occupy positions of structural control makes them natural focal points for demands to reorganize work, adjust incentives, revise AI policies, and redistribute the burdens generated by workslop. In backward-looking terms, it may be fair to expect more of managers because more of the relevant conditions were under their control. In forward-looking terms, organizations have especially strong reasons to direct reform efforts toward those actors, including managers, who can change the background conditions that predictably produce harmful forms of workslop. Failing to recognize this asymmetry of control between workers and managers also creates a distinctive risk. Elish (2016) uses the term “moral crumple zone” to describe cases in which ethical blame for unintended outcomes lands on a human even when the outcome lies partly outside that person’s control (p. 40; see also Probst et al., 2026). Although the concept was developed in relation to human-machine systems such as autonomous vehicles and aviation autopilot, the same dynamic can arise in AI-mediated workplace contexts. When workslop is produced under compressed deadlines, ambiguous instructions, inadequate training, or low psychological safety, the worker who sends the deficient output is the most visible target of blame. Managers, by contrast, often remain invisible at the point of failure: their decisions about workload, performance metrics, AI policies, and review structures operate as background conditions rather than proximate causes. The risk, then, is that organizational responses focused exclusively on the proximate sender assign responsibility to the person easiest to identify, rather than to those whose decisions shaped the conditions under which the workslop was produced.

Taken together, these considerations suggest that responsibility for AI workslop should not be understood in a single register. Responsibility-sensitive backward-looking views remind us that it is unfair to hold people responsible for what was not under

their control or what they could not reasonably have been expected to avoid under unjust conditions, while more forward-looking views show why institutional responses may still be justified even backward-looking ones are not. They also help explain why managers will often bear a larger share of responsibility than workers when structural factors under managerial control play a significant role in producing workslop.

Before we move on, it is worth distinguishing this dynamic from the "responsibility gap" as it figures in the ethics of AI. On the original formulation, the unpredictability of learning systems can make it impossible to legitimately attribute culpability to any human agent for outcomes the system mediates (Matthias, 2004). Workslop does not generate a responsibility gap in that strict sense. A human worker often chooses whether to use AI, how carefully to check its output, and whether to pass it on, so identifiable agents remain in control of the morally relevant decisions. The interconnection is nonetheless real, and it operates at a different level. Santoni de Sio and Mecacci (2021) argue that the responsibility gap is not a single problem but a family of problems, including gaps in culpability, moral and public accountability, and active, forward-looking responsibility. They also argue that such gaps can arise from organizational rather than purely technical sources, and even with non-learning systems. On that broader picture, the delegation distance and the "moral crumple zone" dynamic discussed above can produce a practical or perceived decrease of responsibility that mimics a gap even where no culpability gap exists. The risk is not that responsibility disappears, but that organizational and psychological mechanisms obscure where it properly lies. This is precisely the misallocation that the role-sensitive account developed here is meant to correct. The connection is especially close in the case of Santoni de Sio and Mecacci's active-responsibility gap, which concerns the forward-looking dimension foregrounded in this section.

5 Policy Implications

Although the preceding sections establish when and why producing AI workslop is morally wrong, permissible, or even obligatory, these moral assessments do not map straightforwardly onto questions of policy. That an instance of workslop is morally wrong does not by itself imply that it should be prohibited, monitored, or sanctioned. Many morally problematic actions, such as the minor appropriation of office supplies, are not, and according to most people should not be, the subject of institutional regulation. Organizational policy therefore requires an additional line of reasoning, one that considers when a practice is sufficiently harmful, frequent, or structurally significant to warrant formal intervention. Factors such as the seriousness of the harms involved, the feasibility of reliable detection, the proportionality of enforcement mechanisms, structural responsibility, and the risk of collateral damage must all be part of such an assessment. This includes risks such as the erosion of trust or the introduction of intrusive monitoring. The aim of this section is to sketch how the preceding normative analysis might be translated into guidance for organizational policy. The section clarifies when intervention is warranted, when restraint is appropriate, and how institutions can address the structural conditions that give rise to workslop without undermining trust or diminishing the legitimate uses of GenAI.

Several considerations arising from the earlier sections counsel organizational restraint. The most fundamental follows from the responsibility analysis of Sect. 4.2: in many cases, the worker who produces wrongful workslop is not an appropriate target of blame at all. Where low psychological safety, chronic overload, or the delegation distance characteristic of AI-mediated work undercut the control and epistemic conditions for blameworthiness, punitive responses are not merely imprudent but unfair. They sanction people for outcomes they could not, under the prevailing conditions, reasonably have been expected to avoid. Worse, punishment here risks being self-defeating. Because blame and sanction themselves erode psychological safety, and low psychological safety is itself among the conditions that generate workslop (Sect. 4.2), punitive regimes threaten to deepen the very problem they are meant to correct. Detection difficulty compounds this. As Sect. 2 established, AI workslop is hard to identify with confidence, so enforcement regimes premised on catching individual cases are prone to error. The motives and circumstances surrounding any instance are also often epistemically opaque. As Sects. 2 through 4 showed, similar outputs may result from overload, ambiguous instructions, limited competence, principled resistance, or opportunism, and managers seldom have reliable access to these distinctions. Heavy-handed policies, moreover, risk overreach. Strict monitoring, the use of AI detectors, or blanket prohibitions can produce chilling effects, undermining trust, discouraging beneficial experimentation, and pushing workers toward concealment rather than improved practice (Mollick, 2024). Finally, strong sanctions risk falling disproportionately on workers with the least bargaining power, even when the structural responsibility for workslop lies primarily with managerial decisions. This is the very asymmetry of control established in Sects. 3 and 4.2. These considerations together suggest that punitive or surveillance-based approaches are often ill-suited to the phenomenon.

Given the distribution of control described in Sect. 3, organizations should prioritize structural and supportive interventions rather than punitive measures. The first step is to address the structural drivers of workslop. This includes examining workloads, deadlines, and staffing levels in units where workslop appears widespread, and recalibrating performance metrics so that they emphasize substantive quality rather than speed or volume alone. The job characteristics model (Hackman & Oldham, 1976), for example, suggests that workers are more likely to invest substantive effort when they have autonomy over how tasks are completed, understand the purpose of what they are doing, and can see how their work affects others—conditions that fragmented or hyper-quantified work environments often fail to provide. Clarity of norms is also important, but this need not involve criminalizing all instances of workslop. Instead, organizations can offer concrete examples of acceptable and unacceptable uses of GenAI, focusing on clear cases of harmful burden shifting or misleading content.

Goal-setting research has in addition long shown that performance improves when standards are specific and demanding rather than vague or merely aspirational (Locke & Latham, 2002); applied to workslop, this suggests that organizational norms must specify what counts as adequate verification, contextualization, and revision for particular task types, rather than relying on general exhortations to "use AI responsibly." At the same time, investment in competence is often preferable to punishment. Train-

ing programs on critical AI use, verification techniques, and the integration of AI outputs into task-specific workflows, and a general upskilling can reduce the likelihood of harmful workslop and ensure that workers are not left to navigate AI tools without support. Finally, organizations should provide channels through which workers can report unreasonable demands or AI-related pressures without fear of reprisal—a measure that directly targets the low psychological safety that Sect. 4.2 identified as both a driver of workslop and an excusing condition for it. Recognizing that some workslop may function as informal resistance to unjust conditions, policies should not aim to suppress such signals but rather to understand and address their underlying causes.

These supportive measures are, moreover, the natural addressees of the forward-looking responsibility distinguished in Sect. 4.2. Because managers occupy the positions of structural control mapped in Sect. 3.1, they are the appropriate focal points for reform even in the many cases where they are not backward-lookingly blameworthy for any particular instance of workslop. The point of directing organizational effort toward upskilling, workloads, metrics, AI norms, training, and the distribution of review burdens is not to assign fault but to act on the conditions that predictably generate harmful workslop—precisely the forward-looking response that remains warranted when backward-looking blame is excused. This is why the policy weight falls more heavily on managerial structures than on individual workers: not because managers are typically more culpable, but because they control more of what would have to change.

Although many cases of workslop may not warrant punitive responses, there are contexts in which more formal rules or sanctions may be appropriate. These are primarily cases in which workslop occurs in high-stakes environments, such as safety-critical, clinical, or legal settings, where the risks to others are severe and the harms substantial. This tracks the analysis of Sect. 4.1, where the presence of serious third-party or public stakes was what shifted workslop from possibly permissible to likely impermissible. The same external harms that defeat any obligation to produce workslop also provide the strongest grounds for regulating it. Sanctions may also be justified when workers deliberately and repeatedly rely on AI workslop to gain personal advantage, such as securing bonuses or promotions, while shifting significant burdens onto equally constrained colleagues. This is the case in which the excusing conditions emphasized above are, by stipulation, absent. Even in such cases, however, organizations should first examine whether the structural factors under managerial control, including workload, incentives, training, or unclear directives, have contributed to the problem. Sanctions, when used, should be proportionate, transparent, and applied with an awareness of these background conditions. In many instances, graded responses such as feedback, coaching, or reallocation of tasks may address the issue more effectively than immediate punitive measures. Empirical research on feedback effectiveness further suggests that feedback works best when it directs attention to specific task improvements rather than threatening the person (Kluger & DeNisi, 1996). This supports an approach that identifies what was missing in a particular output, such as omitted context, unverified claims, or downstream burdens, rather than treating workslop as a generic failure of diligence.

Given the detection challenges, structural constraints, and moral distinctions analyzed throughout the paper, organizations should acknowledge that a certain amount

of low-stakes workslop is likely to occur and should not always be treated as a matter for enforcement. This is not mere pragmatic tolerance: as Sect. 4.1 argued, workslop may be morally permissible when the harms are minor, largely unavoidable under the relevant structural conditions, consented to as part of a cooperative arrangement, or preventable only at unreasonable individual cost. In such cases, policy should not treat the mere production of workslop as a failure calling for correction or sanction. Like minor personal uses of office resources, trivial instances of workslop do not threaten organizational functioning and should not occupy managerial attention. The appropriate policy aim is therefore not the elimination of workslop, but the reduction of harmful forms and the prevention of systematic burden shifting onto colleagues in similar or more vulnerable positions. Managers should focus on structural patterns rather than attempting to police every marginal instance, redirecting attention toward the organizational conditions that make workslop more or less likely and more or less harmful.

These policy implications outlined above follow directly from the non-ideal, role-sensitive moral analysis developed in this paper. Because responsibility for AI workslop is distributed across roles and shaped by structural conditions that workers cannot fully control, organizational responses must avoid collapsing all instances of workslop into simple failures of diligence. Effective and ethically defensible policy must instead target the structural drivers under managerial control, support workers' capacity to use AI responsibly, and distinguish harmful, avoidable, and nonredundant workslop from the trivial or unavoidable kinds that characterize imperfect workplaces. In this sense, regulating workslop is less a matter of policing individual behavior and more a matter of designing institutions that reduce unjust burdens, preserve workers' agency, and ensure that the benefits of GenAI do not come at the expense of those in similar or more vulnerable positions. For an overview, see Table 3 below.

6 Conclusion

This paper has argued that the moral significance of AI workslop cannot be captured by treating it simply as an individual failure of diligence, competence, or integrity. Whether producing workslop is morally wrong, permissible, or, in a narrower range of cases, required depends on the harms it creates, the alternatives available to the agent, and the structural conditions under which the work is produced. Because GenAI makes it easier to produce outputs that appear adequate while shifting interpretive and corrective labor onto others, the ethics of workslop is inseparable from broader questions about how burdens, discretion, and vulnerability are distributed within social and institutional roles.

On the account developed here, AI workslop is morally wrong when it imposes substantial, avoidable, and nonredundant burdens on others that the agent could reasonably have prevented without excessive sacrifice. It may be permissible when the harms are minor, largely unavoidable, or preventable only at unreasonable cost. In rarer non-ideal cases, limited AI-assisted output may even be required when it is the least harmful means of preventing sufficiently serious harms and does not expose third parties or the public to grave risks. These distinctions also bear on responsibil-

Table 3 Policy Matrix for Organizational Responses to AI Workslop

Condition	Recommended response	Rationale
Minor harm; unavoidable	Structural fixes only (if anything)	Minor severity and limited avoidability both weaken the reason against the act (Sect. 4.1), so the conduct is normally permissible all things considered; and the worker lacks the control that any sanctioning response presupposes (Sect. 4.2).
Minor harm; avoidable	Soft correction, if any (guidance)	Promote improvement without punishment; where the harm is negligible, no response may be needed.
Substantial harm; unavoidable	Structural intervention	The root cause lies in workload, incentives, or constraints under managerial control; this is the forward-looking, managerial-responsibility case of Sect. 4.2.
Substantial harm; avoidable; low-stakes	Graduated response (feedback, coaching, training, supervision)	Corrective action should remain proportionate and target the specific deficiency rather than the person.
Substantial harm; avoidable; high-stakes	Formal sanctions possible, but only where the worker is genuinely blameworthy (excusing conditions absent) and structural contributors have been addressed	Serious external or third-party stakes (Sect. 4.1) can justify stronger measures, but only against agents with the relevant control and no excusing conditions of the kind set out in Sect. 4.2; avoidability alone does not establish blameworthiness.
Substantial harm; avoidable; persistent opportunistic gaming	Sanctions possible regardless of stakes, once structural contributors are addressed	Deliberately and repeatedly producing workslop for personal advantage—securing bonuses or promotions while shifting burdens onto equally constrained colleagues—removes the excusing conditions by stipulation, so blameworthiness is established even where stakes are low.

Note. The matrix covers cases in which workslop is wrong or potentially wrong; it does not address cases in which producing workslop is permissible or required (Sect. 4.1). The rows classify workslop at the level of the act, by the severity, avoidability, and stakes of the harm, so a row indicates at most the type of response that could be appropriate, not that any sanctioning response is yet warranted. "Avoidable" refers to the act-level avoidability of the harm and does not by itself settle whether the producing agent is blameworthy, since excusing conditions (low psychological safety, chronic overload, or the delegation distance characteristic of AI-mediated work) may defeat blameworthiness even for avoidable harm (Sect. 4.2). Accordingly, before any sanctioning response, blameworthiness must be separately established and structural contributors addressed

ity. Because workers and managers occupy different positions of control, and because both act under structural constraints of varying severity, responsibility for workslop cannot be assigned solely by looking at the proximate producer of a deficient output. Judgments of blame must therefore remain sensitive to background conditions, unequal power, and the fact that organizational failures are often mediated through individual action without being reducible to it.

The practical implications follow from this role-sensitive and non-ideal analysis. If workslop arises in part from the organization of work itself, then a response focused primarily on surveillance or punishment risks reproducing the very injustices it claims to correct. Detection is uncertain, motives are many times opaque, and structurally induced cases are likely to be common. More defensible responses therefore begin with institutional reform: adjusting workloads and incentives, clarifying expectations for AI use, investing in workers' capacity to evaluate AI outputs, and creating

avenues through which unreasonable demands can be challenged. Formal sanctions may sometimes be justified, but only where serious and avoidable harms are at stake, and where relevant background conditions have been addressed rather than ignored.

AI workslop, then, is best understood not as a unitary form of misconduct but as a heterogeneous practice whose moral significance varies with context. It is also best understood as a technologically mediated practice often received as a form of workplace effort-withholding or resistance, but with distinctive features: surface adequacy, burden shifting, and an uncertain relation between output, effort, and intention. Seeing it in this way shifts attention from a narrow preoccupation with individual failure to a broader concern with how institutions allocate control, offload costs, and shape the conditions under which people act. As GenAI becomes more deeply embedded in professional life, the central ethical and political question is not simply whether people use this technology at work, but under what conditions AI-mediated labor redistributes burdens unfairly, obscures responsibility, and undermines justifiable forms of cooperation.

Two extensions of the present argument merit further attention. For self-employed workers, the same harm-based framework may still apply, but the distribution of burdens differs because the agent often internalizes more of the immediate costs while still being able to externalize harms onto clients, collaborators, or the public. And in work settings where the value of the task itself is contested, including cases discussed under the label of "bullshit jobs," the significance of withholding effort may be altered because the background justification of the work is itself in dispute. Even there, however, the core issues identified here, avoidable harm, burden shifting, and the distribution of control, remain relevant.

Declaration of generative AI and AI-assisted technologies in the writing process.

During the preparation of this work the author used ChatGPT and Claude to proofread the manuscript. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

Acknowledgements I thank the reviewers, especially Reviewer 2, for their excellent and constructive comments on earlier drafts of this paper. I also thank the associate editor for their valuable input after acceptance.
AI Workslop: The Moral Significance of Withholding Effort.

Authors' Contributions I´m the sole author.

Funding Open access funding provided by Chalmers University of Technology. Not applicable.

Data Availability Not applicable.

Declarations

Ethics Approval Not applicable.

Consent for Publication Yes.

Competing Interests On behalf the corresponding author, there is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Ackroyd, S., & Thompson, P. (1999). *Organizational misbehaviour*. Sage.

Alchian, A. A., & Demsetz, H. (1972). Production, information costs, and economic organization. *American Economic Review, 62*(5), 777–795.

Algander, P. (2013). Harm, benefit, and non-identity [Doctoral dissertation, Uppsala University].

Baltes, S., Cheong, M., & Treude, C. (2026). An endless stream of AI slop: The growing burden of AI-assisted software development. *arXiv preprint arXiv*:260327249.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in Neural Information Processing Systems*, *29*.

Bommasani, R. (2021). On the opportunities and risks of foundation models (arXiv: 2108.07258). *arXiv*. https://doi.org/10.48550/arXiv.2108.07258

Bradley, B. (2013). Asymmetries in benefiting, harming and creating. *The Journal of Ethics*, *17*(1–2), 37–49. https://doi.org/10.1007/s10892-012-9134-6

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *The Quarterly Journal of Economics*, *140*(2), 889–942. https://doi.org/10.1093/qje/qjae044

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77–91). PMLR.

Burrell, J. (2016). How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), Article 2053951715622512. https://doi.org/10.1177/2053951715622512

de Fine Licht, K., & Folland, A. (2025). AI in public decision-making: A philosophical and practical framework for assessing and weighing harm and benefit. *Public Administration*. https://doi.org/10.1111/padm.70029

Dell'Acqua, F., McFowland, E., Mollick, I. I. I., Lifshitz-Assaf, E. R., Kellogg, H., Rajendran, K., Krayer, S., Candelon, L., F., & Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality (Harvard Business School Technology & Operations Mgt. Unit Working Paper No. 24–013). https://doi.org/10.2139/ssrn.4573321

Dennett, D. C. (2003). *Freedom evolves*. Viking.

Dillon, E. W., Jaffe, S., Immorlica, N., & Stanton, C. T. (2025). *Shifting work patterns with GenAI (No. w33795)*. National Bureau of Economic Research.

Edmondson, A. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly*, *44*(2), 350–383. https://doi.org/10.2307/2666999

Elish, M. C. (2016). Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science Technology and Society*, *5*, 40–60. https://doi.org/10.17351/ests2019.260

Feinberg, J. (1984). *Harm to others*. Oxford University Press.

Fregin, M. C., Eijkenboom, D., Özgül-Persyn, P., Pardesi, M., & Rounding, N. (2026). Artificial intelligence in firms: A synthesis of micro-level evidence on productivity, work, and skills (ROA Policy Brief No. 002E). Maastricht University, Research Centre for Education and the Labour Market. https://doi.org/10.26481/umarpb.2026002E

Gardner, M. (2017). On the strength of the reason against harming. *Journal of Moral Philosophy*, *14*(1), 73–92. https://doi.org/10.1163/17455243-46810043

Giray, L., Sevnarayan, K., Maphoto, K. B., & Wider, W. (2026). AI slop in academic publishing: History, characteristics, manifestations, causes, and mitigation strategies. *Internet Reference Services Quarterly*, *30*(2), 221–244. https://doi.org/10.1080/10875301.2026.2637526

Goodin, R. E. (1986). Responsibilities. *The Philosophical Quarterly, 36*(142), 50–56.
Hackman, J. R., & Oldham, G. R. (1976). Motivation through the design of work: Test of a theory. *Organizational Behavior and Human Performance, 16*(2), 250–279. https://doi.org/10.1016/0030-5073(76)90016-7
Hooker, B. (2000). *Ideal code, real world: A rule-consequentialist theory of morality*. Oxford University Press.
Jarrahi, M. H., Li, L., Robinson, A. P., & Meng, S. (2025). Generative AI and the augmentation of information practices in knowledge work. *Behaviour & Information Technology*. https://doi.org/10.1080/0144929X.2025.2551570
Kagan, S. (2014). *The geometry of desert*. Oxford University Press.
Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology*, *65*(4), 681–706. https://doi.org/10.1037/0022-3514.65.4.681
Kidwell, R. E., & Bennett, N. (1993). Employee propensity to withhold effort: A conceptual model to intersect three avenues of research. *Academy of Management Review*, *18*(3), 429–456. https://doi.org/10.2307/258904
Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254
Köbis, N., Rahwan, Z., Rilla, R., Supriyatno, B. I., Bersch, C., Ajaj, T., Bonnefon, J.-F., & Rahwan, I. (2025). Delegation to artificial intelligence can increase dishonest behaviour. *Nature, 646*(8083), 126–134. https://doi.org/10.1038/s41586-025-09505-x
Latané, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology*, *37*(6), 822–832. https://doi.org/10.1037/0022-3514.37.6.822
Leiden Declaration on Artificial Intelligence and Mathematics (2026). https://doi.org/10.5281/zenodo.20302944
Liebscher, A., Lee, A. Y., Rapuano, K., Kellerman, G., Niederhoffer, K. G., & Hancock, J. T. (2026). Workslop: Examining the prevalence, antecedents and consequences of low-quality AI-generated content at work. [Preprint — not yet peer-reviewed]. Data and code at https://osf.io/xh2wn
Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31–57. https://doi.org/10.1145/3236386.3241340
Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist, 57*(9), 705–717. https://doi.org/10.1037/0003-066X.57.9.705
Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1
Mayer, H., Yee, L., Chui, M., Roberts, R. (2025, January 28) *Superagency in the workplace: Empowering people to unlock AI's full potential.* McKinsey & Company. https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/superagency-in-the-workplace-empowering-people-to-unlock-ais-full-potential
Mollick, E. (2024). *Co-intelligence: Living and working with AI*. Portfolio/Penguin.
Niederhoffer, K., Kellerman, G. R., Lee, A., Liebscher, A., Rapuano, K., & Hancock, J. T. (2025). *AI-generated workslop is destroying productivity*. Harvard Business Review.
Pereboom, D. (2001). *Living without free will*. Cambridge University Press.
Probst, et al. (2026). Aligning artificial intelligence with worker health, well-being, and safety: An occupational health psychology agenda. *Occupational Health Science*. https://doi.org/10.1007/s41542-026-00262-5. 10, Article 15.
Ranganathan, A., & Ye, X. M. (2026). AI doesn't reduce work—it intensifies it. *Harvard Business Review*, 9.
Ross, W. D. (1930). *The right and the good*. Oxford University Press.
Salari, N., Beiromvand, M., Hosseinian-Far, A., Habibi, J., Babajani, F., & Mohammadi, M. (2025). Impacts of generative artificial intelligence on the future of labor market: A systematic review. *Computers in Human Behavior Reports, 18*, Article 100652. https://doi.org/10.1016/j.chbr.2025.100652
Santoni de Sio, F., & Mecacci, G. (2021). Four responsibility gaps with artificial intelligence: Why they matter and how to address them. *Philosophy & Technology, 34*(4), 1057–1084. https://doi.org/10.1007/s13347-021-00450-x
Scanlon, T. M. (1998). *What we owe to each other*. Harvard University Press.

Scanlon, T. (2018). *Why does inequality matter?* Oxford University Press.
Scarfe, P., Watcham, K., Clarke, A., & Roesch, E. (2024). A real-world test of artificial intelligence infiltration of a university examinations system: A "Turing test" case study. *PLoS One, 19*(6), Article e0305354. https://doi.org/10.1371/journal.pone.0305354
Shiffrin, S. V. (2012). Harm and its moral significance. *Legal Theory, 18*(3), 357–398. https://doi.org/10.1017/S1352325212000080
Smilansky, S. (2000). *Free will and illusion*. OUP Oxford.
Strawson, G. (1994). The impossibility of moral responsibility. *Philosophical Studies*, *75*(1/2), 5–24. https://doi.org/10.1007/BF00989879
van de Poel, I., Royakkers, L., & Zwart, S. D. (2015). *Moral responsibility and the problem of many hands*. Routledge.
Wallace, R. J. (1994). *Responsibility and the moral sentiments*. Harvard University Press.
Waltzer, T., Pilegard, C., & Heyman, G. D. (2024). Can you spot the bot? Identifying AI-generated writing in college essays. *International Journal for Educational Integrity, 20*(1), Article 11. https://doi.org/10.1007/s40979-024-00158-3
Watson, G. (2004). *Agency and answerability: Selected essays*. Oxford University Press.
Xu, F., Medappa, P. K., Tunc, M. M., Vroegindeweij, M., & Fransoo, J. C. (2025). AI-assisted programming may decrease the productivity of experienced developers by increasing maintenance burden (arXiv: 2510.10165). *arXiv*. https://doi.org/10.48550/arXiv.2510.10165

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