

Beneficiary-Based Responsibility for AI Training Labour: Lessons from Kenya's Gig Economy and Beyond

Ruoyi Lu^{1,2}, Brenda Maina³

¹ Blavatnik School of Government, University of Oxford

² Information Society Law Center, University of Milan

³ Center for AI and Digital Policy

ruoyi.lu@chch.ox.ac.uk, wbrendamaina@gmail.com

Abstract

AI systems are built from large volumes of outsourced human labour, including data annotation, content moderation, and reinforcement learning feedback. Much of this work is performed in low-and middle-income countries under conditions of low pay, insecure contracts, psychological harm, and weak legal protection. Existing frameworks in labour law, business, human rights, and investment law struggle to address these harms because they rely on forms of legal or contractual proximity that outsourced AI training chains are designed to weaken. This paper argues that responsibility for AI training labour should be grounded not only in formal employment or contractual ties, but also in benefit and structural influence. Drawing on the Beneficiary-Pays Principle from climate justice, it develops a conditional framework under which obligations arise where harms are foreseeable, firms exercise structural influence over labour conditions, and benefits are systematic and durable. Building on this structural diagnosis, it proposes a two-layer institutional response: extending due diligence to digital labour chains and establishing an AI training levy to support social protection and remedy for affected workers.

Introduction

Across the global AI economy, the labour that makes general-purpose systems usable is increasingly outsourced, hidden, and precarious (Gray and Suri 2019; Altenried 2022). Kenya has emerged as a key hub in this infrastructure and as a jurisdiction in which courts and lawmakers are beginning to confront outsourced digital labour, where a growing number of young workers in Nairobi and other cities perform content moderation and data annotation for major technology companies, often under harsh conditions and with little formal protection (Pidoux et al. 2025; Aapti

Institute 2026). This pattern forms part of a broader African data labour ecosystem in which annotation, content moderation, and reinforcement learning from human feedback are supplied through intermediary firms to predominantly North American and European AI clients (Pidoux et al. 2025; Graham, Hjorth, and Lehdonvirta 2017). In 2019, for example, the business process outsourcing (BPO) firm Sama began providing content moderation and AI data labelling services for Meta from a delivery center in Nairobi (Perrigo 2023). Moderators reported spending hours each day reviewing violent, sexually abusive, and otherwise traumatic material, working to strict productivity targets and with limited access to mental health support (Perrigo 2023). In 2023, more than 100 former workers brought a case in the Kenyan Employment and Labour Relations Court against both Sama and Meta, alleging forced labour, union-busting, irregular pay, and severe psychological harm.¹ Meta argued that, because it had no direct employment relationship or corporate presence in Kenya, it could not be sued there. The High Court rejected that argument in 2023, and the Court of Appeal later dismissed Meta's jurisdictional appeals, allowing the litigation to proceed on the basis that questions about Meta's effective authority over the virtual workplace required full adjudication.²

The subsequent trajectory of the case illustrates the structural issues at stake. In April 2026, Sama announced that it would terminate 1,108 jobs in Nairobi after Meta declined to renew its contract, reportedly giving workers around six days' notice.³ Investigations by civil society groups characterise this as a form of corporate accountability evasion. The contractual architecture was

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹ Meta Platforms, Inc & 2 others v Motaung & 186 others; Kenya National Human Rights and Equality Commission & 14 others

(Interested Parties) (Civil Appeal E232 & E445 of 2023 (Consolidated)) [2024] KECA 1262 (KLR).

² *Ibid.*

³ Hürlimann, Corinne; and Weston, Phoebe. 2026. Sama to Cut 1,108 Jobs in Kenya After Meta Ends Contract. The Guardian, April 2026.

designed to ensure that the global platform could rapidly exit without bearing the legal or financial consequences of mass layoffs. At the same time, recent investigations and mapping studies suggest that this pattern extends well beyond the Kenyan cases. A 2026 SOMO investigation found that Meta, Amazon, Google, Microsoft, and Nvidia collectively rely on at least 30 intermediary companies for AI data work globally, while shaping labour conditions through pricing pressure, productivity demands, and other commercial constraints that travel through the outsourcing chain (SOMO 2026). Research mapping Africa's AI training labour supply chain similarly finds that data annotation, content moderation, and reinforcement learning from human feedback now occur in 43 out of 55 African countries, supplied by at least 17 intermediary firms to predominantly North American and European clients (Pidoux et al. 2025). Workers typically hold short-term contracts, earn below national wage averages, and have little or no access to collective bargaining or effective grievance mechanisms. High youth unemployment in Kenya and the wider region makes this labour model attractive to both governments and firms, and difficult for workers to refuse.

These developments raise a basic question about responsibility in AI value chains. The firms that design model architectures, set the performance and safety requirements, and capture the bulk of the commercial value are rarely the same entities that formally employ training workers or appear on their contracts. Existing legal frameworks in labour law, business human rights, and investment law all condition responsibility on some form of relational proximity between duty-bearer and worker, such as an employment tie, a direct contractual relationship, or an investor–state link. Yet the outsourcing architectures used in AI training are structured precisely to erode that proximity while preserving flows of value and control. The result is a structural decoupling of benefit and legal responsibility. The actors with the greatest power to change working conditions, and the greatest stake in the continued availability of cheap training labour, are also the actors least exposed to existing accountability mechanisms.

This paper addresses that gap. It asks how responsibility should be allocated when AI training labour is organised through cross-border chains in which lead firms derive systematic benefits and exercise significant influence over working conditions, but stand at several removes from the workers who bear the harms. The central claim is that responsibility for AI training labour should be grounded in benefit and structural influence, instead of solely in formal employment or contractual proximity. Drawing on the Beneficiary-Pays Principle (BPP) from climate justice, the paper adapts BPP to the context of AI labour and develops a conditional beneficiary-based framework for assigning obligations (Caney 2010; Page 2012). Under this approach, obligations arise when harms are foreseeable, when lead

firms materially shape labour conditions, and when they derive systematic and durable benefits from outsourced training architectures.

Methodologically, the paper combines doctrinal analysis of Kenyan labour and data protection law and recent litigation over outsourced digital labour with a comparative examination of relevant regional materials and emerging EU due diligence and AI regulation. It uses these legal materials, together with recent empirical work on AI training labour, as the foundation for institutional design. The argument proceeds in three steps. First, it offers a structural diagnosis of the AI training accountability gap, identifying a common defect across labour law, business, human rights, and investment law in their reliance on relational proximity. Second, it develops and defends a conditionalized version of BPP tailored to AI training chains, addressing key objections and clarifying the status of the principle as a basis for legal development. Third, it sets out a two-layer governance model including extended digital value-chain due diligence and an AI Training Levy that operationalises beneficiary-based responsibility in practice. In this respect, the paper sits between existing literatures of both empirical work on ghost work and data labour, which has richly documented the organisation and harms of AI labour, business and human rights scholarship, which has pushed supply-chain accountability forward but has not fully addressed the structural mismatch between existing legal triggers and outsourced digital labour chains (Gray and Suri 2019; Casilli 2021; Roberts 2019; Altenried 2022; Tubaro and Casilli 2019). The aim is to offer a concrete, normatively grounded starting point for rethinking how responsibility is assigned along AI training value chains.

The Outsourcing Veil and the Limits of Proximity-Based Responsibility

How AI Training Labour Is Structured

Contemporary AI systems depend on large volumes of human labour. Data workers label images and text, clean and validate datasets, review model outputs, moderate harmful content, and provide reinforcement learning feedback that helps make models more usable and commercially deployable (Gray and Suri 2019; AI Now Institute and Share Lab 2018; Tubaro and Casilli 2019; Miceli and Posada 2022). Yet this labour is rarely organised through direct employment by the firms that ultimately benefit most from it. Instead, AI training labour is typically sourced through multi-tier intermediary structures, including business process outsourcing firms, crowdwork platforms, and local subcontractors, which connect global technology companies to workers in low-and middle-income countries through layered contractual arrangements (Weil 2014; Gray and Suri

2019; Casilli 2021; Altenried 2022; Lehdonvirta 2022; Ekbj and Nardi 2017).

The issue of this structure is that it distributes value and vulnerability unevenly. Commercial value, strategic control, and informational power tend to flow upward to lead firms, while labour risk, insecurity, and legal exposure are pushed downward onto intermediaries and workers. Most data workers do not contract directly with, and often do not meaningfully interact with, the major AI firms whose systems they help build. One reason for this arrangement is straightforwardly economic. It is far cheaper to source training labour in Nairobi, Kampala, or Antananarivo than in New York, London, or Strasbourg, especially where high unemployment, weaker labour protections, and limited bargaining power make workers easier to discipline and replace. The result is a labour regime marked by low pay, short-term contracts, opaque management, limited social protection, and exposure to psychologically harmful material, particularly in content moderation and related tasks (Perrigo 2023; Gray and Suri 2019; Roberts 2019; Aapti Institute 2026).

Kenya illustrates the broader pattern. It has become a major hub for AI training labour because of its large English-speaking workforce, expanding digital infrastructure, and high levels of youth unemployment (Pidoux et al. 2025; Graham, Hjorth, and Lehdonvirta 2017). At the same time, these conditions have made workers especially vulnerable to outsourcing arrangements that combine economic dependence with weak legal protection. As reflected and recognised in the Sama litigation, exploitation does not solely occur in individual cases. The current organisation of AI training labour reproduces a familiar political economy in which commercially valuable functions are retained by firms in capital-rich jurisdictions, while labour-intensive and harmful work is displaced onto workers in lower-income settings with fewer effective protections (Mezzadri 2017; AI Now Institute and Share Lab 2018). In that sense, outsourced AI training labour is one of the ways in which global inequalities are built into the contemporary AI value chain.

The Outsourcing Veil: Concept and Mechanism

The governance problem identified in the previous subsection can be understood through what this paper calls the outsourcing veil: an organisational architecture that

systematically separates decision-making power and commercial benefit from formal employer status, and in doing so insulates lead firms from labour accountability. In a conventional employment model, the entity that directs work, sets conditions, and benefits from output is also the entity that appears as the employer in law. In outsourced AI training chains, these roles are divided. A worker's formal employer is a BPO firm, digital platform, or local subcontractor, while the entity that sets key performance indicators, content standards, and pricing, and that ultimately monetises model outputs, sits several contractual steps away. In jurisdictions such as Kenya, for example, the category of "employee" remains legally tied to a person working for wages or salary under a contract of service (The National Council for Law Reporting 2007, 2022a, 2022b, 2022c, 2023; International Labour Organization 2006), and core labour protections continue to track that formal status rather than the broader chain of actors involved in organising work (The National Council for Law Reporting 2007, 2022a, 2022b, 2022c, 2023).⁴

Existing legal concepts do not quite capture this pattern. Company law's corporate veil doctrine is concerned with how liability is distributed among companies within a group, and debates about piercing the corporate veil focus on misuse of the corporate form. Franchising models deal with brand and quality control across legally separate entities, rather than with long chains of intermediaries linking a distant client to a dispersed workforce. Debates about misclassification ask whether a particular worker should be treated as an employee or an independent contractor in a single relationship. Courts have responded by refining tests of control, integration, mutuality of obligations, and economic dependence in order to classify individual relationships (International Labour Organization 2006; Bogg et al. 2020), but they still apply these tests within a bilateral frame and thus struggle to account for complex, multi-tier outsourcing structures.⁵ Kenyan courts, for instance, generally treat outsourcing as a lawful practice and classify outsourced workers as employees of the outsourcing company rather than of the client enterprise, which tends to widen the gap between the firm that is formally responsible in law and the firm that exercises strategic control over the work.⁶

In AI training, this architecture has become increasingly sophisticated. Lead firms contract with global delivery centres and specialised vendors that in turn manage local workforces, often combining on-site delivery centres with

⁴ Employment Act 2007 s 2; Employment and Labour Relations Court Act 2022 s 2; Labour Relations Act 2022 s 2; Labour Institutions Act 2023 s 3; Occupational Safety and Health Act 2022 s 2.

⁵ *Yewens v Noakes* (1880) 6 QBD 530 (Court of Appeal); *Ready Mixed Concrete (South East) Ltd v Minister of Pensions and National Insurance* [1968] 1 All ER 433; *Christine Adot Lopeyio*

v Wycliffe Mwathi Pere [2013] eKLR; *Geoffrey Makana Asanyo v Nakuru Water and Sanitation Services Co & 6 others* [2014] eKLR; *Stanley Ominde Khainga v Nairobi Hospital* [2018] eKLR.

⁶ *Kenya Airways Ltd v Aviation & Allied Workers Union Kenya & 3 others* [2014] eKLR (Court of Appeal); *Abyssinia Iron & Steel Ltd v Kenya Engineering Workers Union* [2016] eKLR (Court of Appeal).

remote, platform-mediated work. Outsourcing promises access to specialised skills, rapid scaling, budget control, and a formally “elastic” workforce, while allowing lead firms to present themselves as several steps removed from day-to-day labour management. At the same time, governments and clients are beginning to demand assurances about worker well-being, transparency, auditability, and bias mitigation. Providers who can demonstrate “good AI governance” practices increasingly market that fact as a competitive advantage. But without structural changes to how responsibility is assigned, these developments leave the basic veil intact. Firms can showcase governance commitments while the risks of low pay, contract insecurity, and psychological harm remain concentrated at the bottom of the chain.

The outsourcing veil extends and hardens what Weil described as the “fissured workplace,” in which core firms shed direct employment in favour of contracting and franchising (Weil 2014; Altenried 2022). In AI training, fissuring operates across borders, through legally autonomous intermediaries in multiple jurisdictions, and in a domain where the work is digital rather than physical (Weil 2014; Casilli 2021). This combination makes it particularly effective at frustrating frameworks that tie responsibility to formal employment or territorial presence. This system-level design underpins the structural decoupling of benefit and responsibility that the rest of the paper seeks to address.

Why Three Major Frameworks Share the Same Structural Defect

The accountability gap created by the outsourcing veil cannot be simply mitigated by enhancing the level of enforcement. It reflects a more basic limitation shared by the main legal frameworks through which responsibility for AI training labour might otherwise be assigned, that each relies on some form of relational proximity between the duty-bearer and the affected worker. In different ways, labour law, business and human rights frameworks, and investment law all presuppose a legally cognisable relationship that can anchor responsibility. AI training supply chains are structured in ways that weaken or sever precisely those relationships.

Labour law illustrates the problem most directly. Its protections are ordinarily triggered by the existence of an employment relationship, or at least by a sufficiently close functional equivalent (International Labour Organization 2006; Bogg et al. 2020). Tests based on control, integration, economic dependence, or contractual form all assume that the relevant firm stands in a recognisable employment position vis-à-vis the worker. But in AI training chains, the lead platform or model developer is typically not the formal employer. Instead, workers are hired through BPO firms,

digital labour platforms, or local subcontractors that stand between the worker and the commercial beneficiary of the labour (International Labour Organization 2021; Lehdonvirta 2022). As a result, labour law obligations attach most readily to the intermediary that contracts with the worker, rather than to the firm that structures the workflow, sets the commercial terms, and captures most of the value.

Business and human rights frameworks appear, at first glance, to move beyond this limitation. Instruments such as the UN Guiding Principles on Business and Human Rights and the EU Corporate Sustainability Due Diligence Directive extend attention beyond direct employment ties by referring to “business relationships” and “chains of activities.” (United Nations Digital Library 2011; European Union 2024; Choudhury 2018) Yet these frameworks still depend on a notion of proximity in the value chain, and in the context of AI training labour that notion remains unsettled. Digital and intangible labour supplied through dispersed platform and subcontracting arrangements is not yet clearly or consistently treated as part of the regulated chain (Casilli 2021; Tubaro and Casilli 2019; Joseph and Kyriakakis 2023). This leaves firms significant room to argue that the workers who label data, moderate content, or generate reinforcement feedback fall outside the part of the value chain for which they bear clear due diligence obligations.

Investment law is even less well suited to addressing the problem. Its central relationship is not between firm and worker, but between investor and host state (Sornarajah 2021). The point of departure for most investment treaties and investor–state dispute settlement mechanisms is the protection of capital against public interference. It was not designed as a vehicle for allocating labour obligations to firms operating across borders, and most investment treaties and investor–state dispute settlement mechanisms are silent on the position of workers in outsourced production chains (Sornarajah 2021; The ETO Consortium 2013). For that reason, investment law does little to close the accountability gap in AI training supply chains. If anything, it can reinforce the imbalance by furnishing legal protections for investors without creating corresponding pathways through which workers can hold lead firms responsible for labour harms associated with outsourced production.

What links these three frameworks, then, is the way each depends on a recognised legal relationship as the trigger for responsibility. Labour law looks for an employment tie. Business and human rights law looks for a sufficiently proximate business relationship or value-chain connection. Investment law looks for an investor–state nexus. The outsourcing veil is effective because it is organised to disrupt or dilute each of these connecting points while preserving the underlying flows of benefit, control, and commercial dependence. The outsourcing veil thus becomes a structural problem where existing frameworks fail due to

the fact that AI training labour is organised in ways that place its primary beneficiaries at a legal distance from the workers on whom the system depends.

A Beneficiary-Based Responsibility Principle for AI Training Labour

Intellectual Genealogy: From Polluter-Pays to Beneficiary-Pays

The beneficiary-pays approach developed in this paper grows out of a familiar idea in environmental governance: those who generate harmful social costs should bear the costs of preventing, mitigating, or remedying them (United Nations Digital Library 1992; Sands et al. 2018). The polluter-pays principle works best where the actor causing the harm and the actor profiting from the underlying activity are the same (Caney 2010; Page 2012; Birnie, Boyle, and Redgwell 2009; Schwartz 2015). In those settings, liability rules and levies can internalise externalities by directing responsibility to the source of the harm. But this logic becomes less effective when causation and benefit come apart, as they often do in contexts of historical injustice, transnational production, and structurally unequal exchange.

It is in that gap that the beneficiary-pays principle becomes normatively attractive. In climate justice theory, the principle has been used to argue that those who benefit from unjust or harmful structures may bear obligations even where they are not the direct or proximate cause of the injury (Caney 2010; Page 2012; Gosseries 2004; Shue 1999; Meyer and Sanklecha 2017). The strongest versions of the argument arise in debates over historical emissions and inherited advantage, that present actors continue to enjoy gains produced by past carbon-intensive development, even though they did not themselves produce the original emissions (Caney 2010; Gosseries 2004; Miller 1998; Pogge 2005). On this view, responsibility can arise from continuing to hold and build on advantages produced by harmful structures, even where direct causation lies in the past (Shue 2020; Truccone-Borgogno 2023; Huseby 2015).

That move is especially relevant to AI training labour. Here too, the firms that capture the greatest commercial benefit are often not the firms that directly employ the workers who bear the risks. The AI case goes beyond a straightforward analogy with climate justice and, in one respect, presents an even stronger normative claim. In many climate cases, beneficiaries are passive inheritors of an already existing structure. In AI training, by contrast, lead firms actively design, maintain, and refine the outsourcing arrangements from which they benefit. They determine sourcing models, pricing pressure, workflow architectures, performance metrics, and platform standards, while relying

on intermediaries to absorb the labour consequences. The beneficiary is therefore a downstream recipient of unjust advantage and, at the same time, an active participant in organising the structure that generates it.

This is the key reason for moving from polluter-pays to beneficiary-pays in the AI context. A strictly causation-based model asks which entity directly inflicted the harm or stood in the closest legal relationship to the worker. But where outsourcing chains are deliberately structured to separate benefit from formal responsibility, that question is too narrow. A beneficiary-based approach instead follows the direction of value extraction. It asks which firms systematically derive commercial advantage from a labour regime whose harms are foreseeable, whose organisation they materially influence, and whose costs are otherwise pushed onto workers, intermediaries, and weakly protected jurisdictions.

The claim, however, should not be overstated. This paper does not argue that beneficiary-pays is already a settled rule of positive international law for AI labour governance. Rather, it argues that beneficiary-pays provides a more appropriate normative foundation for assigning responsibility in cases where traditional employer-based and proximity-based frameworks predictably fail. In that sense, the principle is best understood as an emerging justificatory basis for legal development, which is intellectually rooted in environmental and climate justice debates, but adapted here to the political economy of outsourced AI production.

Conditionalized Application: When Does BPP Apply in AI Training?

Applied directly from climate justice, the Beneficiary-Pays Principle could be read to generate obligations for any actor that benefits from AI training. In the context of complex, multi-firm AI value chains, that formulation is too broad to be operational and too vulnerable to familiar objections about overreach, diffuse responsibility, and “guilt by association.” To make BPP usable in law and policy, this paper therefore adopts a conditionalized version: beneficiary-based obligations arise only when a set of more demanding criteria are jointly satisfied.

The first condition is foreseeability. The beneficiary must know, or have reasonable grounds to know, that the outsourcing structure through which it organises AI training is likely to generate specific kinds of labour harm. In the AI training context, these harms are well documented including psychological injury from sustained exposure to violent and abusive content, wage suppression driven by pricing pressure on intermediaries, structural insecurity from short-term and highly contingent contracts, and gaps in social protection (Gray and Suri 2019; Perrigo 2023; Tubaro and Casilli 2019; Roberts 2019; Aapti Institute 2026; SOMO 2026). Public reporting on the Kenyan content moderation

cases, investigative work by civil society organisations, and litigation such as the Motaung proceedings together make it implausible for large AI developers to claim ignorance of these risks. For any firm operating at scale in this space, the harms associated with current outsourcing models are now plainly foreseeable.

The second condition is structural influence. The beneficiary must materially shape the organisation and conditions of the labour process, even where the workers are formally hired and managed by intermediaries. In AI training chains, this influence is exercised through contract pricing and volume commitments, key performance indicators, content and quality standards, platform design choices, and audit requirements imposed on intermediaries. Lead firms may not supervise workers directly, but they often determine the contours of the work, including the tasks to be performed, the time pressure for the tasks, the level of exposure to harmful content, and minimum economic terms. Structural influence therefore marks out firms that help organise the outsourcing architecture itself, rather than firms that simply purchase an off-the-shelf service. (Weil 2014; Altenried 2022; SOMO 2026).

The third condition is systematic and durable benefit. The benefit must arise from a stable organisational arrangement designed to generate recurring value over time, in contrast to a one-off or incidental interaction. Commercial AI systems are developed and improved through training pipelines that run for years and underpin multiple product generations (Gray and Suri 2019; Miceli and Posada 2022). Firms deploy models built on that pipeline into markets where they generate significant and ongoing revenue, reputational advantage, or strategic positional gains. In that setting, it is not plausible to treat the relationship between beneficiary and labour as accidental or marginal, given that the outsourcing structure is part of the firm's core production system, and the benefits it derives from the system are systematic and durable.

Only when foreseeability, structural influence, and systematic and durable benefit coincide does the conditionalized BPP proposed here generate obligations of prevention, mitigation, and partial compensation. This framing narrows the scope of the principle in ways that matter for practice. It excludes incidental and downstream beneficiaries, end users of AI systems, and small research actors whose use of AI training labour is neither structurally shaping nor commercially central. It focuses instead on firms that actively design and sustain outsourcing architectures, know or ought to know the harms those architectures produce, and derive enduring commercial advantage from them.

Responding to the Strongest Objections

A beneficiary-based framework for AI training labour invites several predictable challenges. Some are conceptual, questioning whether benefiting from an unjust structure can ground obligations at all. Others are institutional, arguing that existing doctrines already provide the necessary tools or that beneficiary-based responsibility cannot be allocated in a principled way. The account developed here aims to show that neither worry is ultimately compelling.

A first objection is the “innocent beneficiary” worry. On this view, even if large AI firms profit from outsourced training labour, they do not directly inflict harm and therefore should not bear special obligations. That concern has more force in contexts where beneficiaries inherit advantages passively from distant historical processes. It is less convincing here because in AI training chains, major firms actively select, design, and maintain the outsourcing structures that generate those benefits, instead of simply being the receiver of benefits. The conditionalized BPP developed in this paper further narrows the target class. Obligations arise only where harms are foreseeable and where firms exercise structural influence over labour conditions. The framework is therefore directed at active structural beneficiaries, not at passive downstream users.

A second objection is that existing doctrines ought to suffice. If the core problem is that firms exercise control without accepting responsibility, why not rely on joint employer status, enterprise liability, or expanded supply-chain due diligence? In some settings, those tools are indeed valuable, and nothing in this argument denies their relevance. The difficulty is that they do not fully address the structural gap identified earlier. Joint employer and enterprise doctrines generally require showing a high degree of control within a single legal system, and that threshold is hard to meet when labour is fragmented across multiple intermediaries and jurisdictions (International Labour Organization 2006; Bogg et al. 2020; Sornarajah 2021). Due diligence regimes broaden responsibility but still depend on contested notions of value-chain proximity and do not, by themselves, explain why firms that systematically benefit from outsourced training labour should bear redistributive obligations (United Nations Digital Library 2011; Choudhury 2018; Joseph and Kyriakakis 2023). In this sense, BPP matters in this context because it supplies a normative basis for extending and coordinating existing tools once proximity-based approaches become unstable, partial, or too case-specific to address the architecture as a whole.

A third objection concerns multiple beneficiaries. AI systems are often developed, trained, deployed, and monetised through constellations of firms, raising the question of “who owes what to whom.” This is a serious concern, but it is not unique to AI or to beneficiary-based

reasoning. Legal and policy systems routinely allocate burdens in contexts of shared causation and diffuse benefit using proportionality rules, pooled mechanisms, and proxy measures. The present framework does not require exact moral accounting of each worker's marginal contribution to each firm's revenue. It requires a principled basis for assigning partial responsibilities among actors that derive systematic benefits and exercise differing degrees of structural influence, which is a task that later sections address through disclosure requirements and a levy mechanism.

A final objection is that BPP might be overly expansive, sweeping in actors whose connection to AI training labour is too remote or whose activities are socially valuable, such as open-source developers, academic researchers, or smaller firms that do not themselves organise large training pipelines. The conditional structure of the framework is designed precisely to prevent that outcome. Beneficiary-based obligations arise only when benefit is systematic and durable, when harms are foreseeable, and when the actor exerts material structural influence over the labour process. These conditions exclude many downstream users and non-commercial actors, and focus instead on large commercial entities that help sustain and shape outsourced training architectures.

Taken together, these objections sharpen rather than undermine the proposed framework. They highlight the need to distinguish active structural beneficiaries from passive ones, to clarify the relationship between BPP and existing legal doctrines, and to show that shared benefit can be handled without demanding impossible precision. The conditionalized version of BPP advanced here is structured to meet those requirements, while leaving room for more detailed institutional design in the sections that follow.

The Status of the Principle

The Beneficiary-Pays Principle is used here as an emerging normative basis for assigning responsibility in AI training labour, rather than as an already established rule of positive international law. It is introduced to address cases in which existing employer-based and proximity-based frameworks fail to reach the firms that derive the greatest benefit from outsourced training labour. The principle is therefore offered as a basis for legal and institutional development under conditions in which benefit and formal legal obligation are structurally decoupled.

This move is familiar in transnational regulation. Frameworks that later acquire legal force often begin as softer normative standards or coordination projects, as illustrated by the trajectory from the UN Guiding Principles on Business and Human Rights to more formal due diligence regimes, and by OECD-led coordination in international tax governance (United Nations Digital

Library 2011; OECD 2021; Deva and Bilchitz 2013; Joseph and Kyriakakis 2023; European Union 2024). These examples show that principles often crystallise before their legal form is fully specified. For present purposes, beneficiary-pays is best understood as a justificatory principle for institutional design, rather than a ready-made rule for immediate application.

Operationalising Beneficiary Responsibility: Due Diligence and a Training Levy

Why Two Layers Are Needed

No single governance tool can address the full accountability problem created by outsourced AI training labour. The gap has both behavioural and distributive dimensions. On the one hand, lead firms structure training supply chains in ways that obscure responsibility and externalise risks onto intermediaries and workers. On the other hand, those same firms continue to derive substantial and durable benefits from labour arrangements whose costs are only partially borne within the chain. Extended due diligence speaks mainly to the first dimension, which works *ex ante* by requiring firms to identify, disclose, and mitigate risks in their training labour chains before those risks crystallise into harm. Properly designed, it can make outsourcing arrangements more visible, narrow the space for plausible deniability, and oblige lead firms to take responsibility for the labour conditions embedded in their systems. But due diligence alone remains a framework of behavioural obligation. A firm can comply formally while workers still remain underpaid, exposed to severe psychological risks, or excluded from social protection.

A levy speaks to the second dimension. It works *ex post*, after benefits have been generated, by requiring firms to contribute financially to the protection and repair of the labour force on which AI systems depend. This matters because many of the harms at issue are not exhausted by future risk prevention. For example, workers have already contributed to training pipelines that underpin commercially valuable systems, and are often under conditions of insecurity and weak legal protection. A redistributive mechanism is therefore needed both to address future harms and to internalise some of the social costs of value already extracted. The two layers are therefore complementary.

Layer One: Extending Due Diligence to Digital Labour Chains

The first layer of the proposed framework is an extension of due diligence obligations to digital labour chains involved in AI training. The basic claim is straightforward. There is no principled reason to protect workers who generate value in physical supply chains while excluding workers who

generate comparable value in digital training pipelines (International Labour Organization 2017; The ETO Consortium 2013; Deva and Bilchitz 2017). In both settings, lead firms organise production through intermediaries, benefit from cost reduction and flexibility, and distance themselves from the workers who bear the risks. The fact that the product of labour is labelled data, moderation output, or reinforcement feedback rather than a physical component does not make the underlying accountability problem any less real. If anything, the relative invisibility of digital labour makes the case for explicit governance stronger (Miceli and Posada 2022; Tubaro, Le Ludec, and Casilli 2020).

The most plausible legal pathway for such an extension lies in the emerging due diligence framework in Europe. The Corporate Sustainability Due Diligence Directive already moves beyond direct employment ties by requiring firms to address adverse impacts arising in their own operations, subsidiaries, and chains of activities (European Union 2024). On a narrow reading, digital labour supplied through dispersed platforms and subcontractors may appear peripheral to that framework. On a purposive reading, however, the case for inclusion is stronger. Human annotation, content moderation, and reinforcement learning feedback are not ancillary to AI production; they are part of the development of the service itself (Gray and Suri 2019; AI Now Institute and Share Lab 2018; Casilli and Posada 2019). The OECD Due Diligence Guidance for Responsible AI, published in 2026, reinforces this direction by explicitly applying responsible business conduct standards to actors across the AI value chain, including those supplying inputs to AI development (OECD 2026). This does not mean that existing law already resolves the issue. It means that there is now a credible interpretive and institutional basis for treating digital training labour as a proper object of due diligence rather than as a residual or optional concern.

If due diligence is extended in this way, the practical implications are relatively clear. Lead AI firms should be required, at a minimum, to map the labour architecture through which training work is sourced, disclose the jurisdictions and intermediaries involved, assess risks associated with pay, contract insecurity, psychological harm, and social protection gaps, and take reasonable steps to mitigate those risks (OECD 2023; OECD 2026; International Labour Organization 2017). They should also be required to establish monitoring and remedy mechanisms that do not depend on workers first proving the existence of a direct employment relationship. The point of this first layer is not to collapse all intermediated labour into formal employment, nor to make lead firms guarantors against every harm in complex value chains. It is to remove the current opacity that allows firms to benefit from digital labour while treating its conditions as someone else's problem. In that sense, extending due diligence to digital

labour chains is the minimum institutional response required to make outsourced AI training labour visible, governable, and attributable to the firms that depend on it.

Layer Two: An AI Training Levy

The second layer of the proposed framework is an AI training levy. If extended due diligence is meant to make outsourced training labour visible and governable, a levy is meant to address the continuing extraction of value from labour whose social costs remain only weakly internalised. This matters because the harms identified in earlier sections are not limited to future risks that better oversight might prevent. Workers have already contributed to training pipelines that support commercially valuable AI systems, often under conditions of insecurity, low pay, psychological strain, and limited access to remedy. A redistributive mechanism is therefore needed to ensure that firms bear at least part of the costs of sustaining and repairing the labour force from which they have derived durable benefits.

The basic logic of such a levy is familiar from other areas of regulation. The aim is not to punish AI development as such, but to internalise part of the social cost of a production model that currently allows lead firms to externalise those costs onto workers, intermediaries, and under-resourced states. In operational terms, the levy would attach to AI-related commercial revenue and be calibrated, at least in part, by information disclosed through the due diligence layer about the location of the training labour and the degree of reliance of the firms. This would not produce perfect precision, nor is perfect precision required. As with other schemes that allocate burdens under conditions of shared benefit and incomplete information, the point is to create a principled mechanism for partial contribution rather than an exact accounting of each worker's marginal role in the value of a given model (Caney 2010; Page 2012; Gosseries 2004; Shue 1999). Properly designed, such a levy could fund social protection, mental health support, severance backstops, and other forms of remedy for workers whose labour underpins AI systems while remaining structurally undervalued.

A levy of this kind would face obvious implementation challenges. If adopted only at the national level, it could encourage firms to shift training activities across jurisdictions in search of weaker obligations (Sornarajah 2021; OECD 2021). For that reason, its most plausible pathway is to go beyond a standalone domestic measure to part of a broader coordinated approach, initially among jurisdictions with sufficient regulatory capacity and market leverage (Joseph and Kyriakakis 2023; Deva and Bilchitz 2017; OECD 2023). Even then, the proposal should be understood modestly. A levy would not eliminate exploitation, nor would it substitute for labour law reform or stronger due diligence obligations. Its function is narrower

but still important. It is meant to ensure that large commercial AI firms do not keep profiting from outsourced training labour while shifting the costs of protecting and supporting workers onto everyone else.

Conclusion

Contemporary debates on AI governance have paid sustained attention to safety, bias, transparency, and downstream social impacts, but they have devoted far less attention to the labour arrangements through which AI systems are built and maintained. The Kenyan moderation and data-labelling cases, together with recent evidence on the wider geography of AI training labour, show that this omission is no longer sustainable. Outsourced AI training labour is part of the infrastructure through which commercially valuable systems are produced. Yet the firms that benefit most directly from that labour are often insulated from responsibility by the very organisational structures through which the labour is sourced.

This paper has argued that this is a structural problem produced by a mismatch between existing legal triggers of responsibility and the way AI training labour is organised across borders. Labour law, business and human rights frameworks, and investment law each rely on some form of recognised proximity between duty-bearer and affected worker (United Nations Digital Library 2011; Sornarajah 2021; International Labour Organization 2006). The outsourcing veil works by weakening or severing those links while preserving flows of value and control. For that reason, responsibility in this context cannot be grounded only in formal employment or contractual ties. It must also take account of who benefits from the labour and who helps structure the conditions under which it is performed (Caney 2010; Page 2012; Gosseries 2004; Shue 2020).

On that basis, the paper has proposed a conditionalised beneficiary-based framework for AI training labour. Under this view, obligations arise where harms are foreseeable, where firms exercise structural influence over labour conditions, and where they derive systematic and durable benefits from outsourced training architectures. It has then suggested a two-layer institutional response: extending due diligence to digital labour chains, and establishing an AI training levy capable of directing part of the value extracted from training labour toward social protection, remedy, and workforce support. The point of this proposal is to show that a more plausible and normatively defensible allocation of responsibility becomes possible once the analysis moves beyond proximity alone.

Several limitations remain. The argument developed here is primarily normative and institutional rather than predictive, and it does not model how firms would adjust their sourcing strategies in response to a levy or expanded

due diligence obligations. The threshold question of which firms should bear the full weight of these obligations also requires further specification, particularly in multi-firm development ecosystems. And while the role of human labour in AI training may change as automation and synthetic data expand, this does not eliminate the central problem addressed here: existing AI systems have already been built through labour arrangements from which powerful firms continue to benefit. Future work should therefore examine more closely how digital labour disclosures could be operationalised, how redistributive mechanisms might be calibrated across jurisdictions, and how affected workers themselves understand the forms of protection and recognition they most urgently need (Pidoux et al. 2025; Tubaro, Le Ludec, and Casilli 2020; Aapti Institute 2026).

References

- Aapti Institute. 2026. Invisible Workers, Visible Harms: Perils and Precarities of AI Labour. <https://aapti.in/blog/invisible-workers-visible-harm-perils-and-precarities-of-ai-labour/>. Accessed:2026-08-02.
- AI Now Institute., and Share Lab. 2018. Anatomy of an AI System: The Amazon Echo as an Anatomical Map of Human Labor, Data and Planetary Resources. <https://anatomyof.ai/>. Accessed: 2026-08-02.
- Altenried, M. 2022. *The Digital Factory: The Human Labor of Automation*. Chicago: The University of Chicago Press.
- Birnie, P. W.; Boyle, A. E.; and Redgwell, C. 2009. *International Law and the Environment*, 3rd ed. Oxford, New York: Oxford University Press.
- Bogg, A.; Collins, J.; Freedland, M.; and Herring, J. 2020. *Criminality at Work*. Oxford: Oxford University Press.
- Caney, S. 2010. Climate Change and the Duties of the Advantaged. *Critical Review of International Social and Political Philosophy* 13(1): 203–228. doi.org/10.1080/13698230903326331.
- Casilli, A. A., 2021. Waiting for Robots: The Ever-Elusive Myth of Automation and the Global Exploitation of Digital Labor. *Sociologias* 23(57): 112–133. doi.org/10.1590/15174522-114092.
- Casilli, A. A., and Posada, J. 2019. The Platformization of Labor and Society. In *Society and the Internet: How Networks of Information and Communication Are Changing Our Lives*, 2nd ed., edited by Graham, M., and Dutton, W. H., 293–306. Oxford: Oxford University Press.
- Choudhury, B. 2018. Balancing Soft and Hard Law for Business and Human Rights. *International and Comparative Law Quarterly* 67(4): 961–986. doi.org/10.1017/S0020589318000155.
- Deva, S., and Bilchitz, D., 2013. *Human Rights Obligations of Business: Beyond the Corporate Responsibility to Respect?* Cambridge: Cambridge University Press.
- Deva, S., and Bilchitz, D., 2017. *Building a Treaty on Business and Human Rights: Context and Contours*. Cambridge: Cambridge University Press.

- Ekbia, H. R., and Nardi, B. A. 2017. *Heteromation, and Other Stories of Computing and Capitalism*. Cambridge, MA: The MIT Press.
- Europe Union. 2024. Directive (EU) 2024/1760 of the European Parliament and of the Council of 13 June 2024 on corporate sustainability due diligence and amending Directive (EU) 2019/1937 and Regulation (EU) 2023/2859 (Text with EEA relevance). <https://eur-lex.europa.eu/eli/dir/2024/1760/oj/eng>. Accessed: 2026-08-02.
- Gosseries, A. 2004. Historical Emissions and Free-Riding. *Ethical Perspectives* 11(1): 36–60. doi.org/10.2143/ep.11.1.504779.
- Graham, M.; Hjorth, I.; and Lehdonvirta, V. 2017. Digital Labour and Development: Impacts of Global Digital Labour Platforms and the Gig Economy on Worker Livelihoods. *Transfer: European Review of Labour and Research* 23(2): 135–162. doi.org/10.1177/1024258916687250.
- Gray, M. L., and Suri, S. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Boston, New York: Houghton Mifflin Harcourt.
- Huseby, R. 2015. Should the Beneficiaries Pay? *Politics, Philosophy and Economics* 14(2): 209–225. doi.org/10.1177/1470594X13506366.
- International Labour Organization. 2006. Employment Relationship Recommendation, 2006 (No. 198). https://normlex.ilo.org/dyn/nrmlx_en/f?p=NORMLEXPUB:12100:0::NO::P12100_INSTRUMENT_ID:312535. Accessed: 2026-08-02.
- International Labour Organization. 2017. Tripartite Declaration of Principles Concerning Multinational Enterprises and Social Policy. https://normlex.ilo.org/dyn/nrmlx_en/f?p=1000:62:0::NO::P62_LIST_ENTRIE_ID,P62_LANG_CODE:2453910,fr:NO. Accessed: 2026-08-02.
- International Labour Organization. 2021. World Employment and Social Outlook: The Role of Digital Labour Platforms in Transforming the World of Work. https://labordoc.ilo.org/discovery/fulldisplay/alma995218610802676/41ILO_INST:41ILO_V2. Accessed: 2026-08-02.
- Joseph, S., and Kyriakakis, J. 2023. From Soft Law to Hard Law in Business and Human Rights and the Challenge of Corporate Power. *Leiden Journal of International Law* 36(2): 335–361. doi.org/10.1017/S0922156522000826.
- Lehdonvirta, V. 2022. *Cloud Empires: How Digital Platforms Are Overtaking the State and How We Can Regain Control*. Cambridge, MA: MIT Press.
- Mezzadri, A. 2017. *The Sweatshop Regime: Labouring Bodies, Exploitation, and Garments Made in India*. Cambridge: Cambridge University Press.
- Meyer, L. H., and Sanklecha, P., 2017. *Climate Justice and Historical Emissions*. Cambridge: Cambridge University Press.
- Miceli, M., and Posada, J. 2022. The Data-Production Dispositif. *Proceedings of the ACM on Human-Computer Interaction* 6(CSCW2): 1–37. doi.org/10.1145/3555561.
- Miller, R. W. 1998. Cosmopolitan Respect and Patriotic Concern. *Philosophy and Public Affairs* 27(3): 202–224. doi.org/10.1111/j.1088-4963.1998.tb00068.x.
- OECD. 2021. Tax Challenges Arising from the Digitalisation of the Economy – Global Anti-Base Erosion Model Rules (Pillar Two). <https://doi.org/10.1787/9afd6856-en>. Accessed: 2026-08-02.
- OECD. 2023. OECD Guidelines for Multinational Enterprises on Responsible Business Conduct. doi.org/10.1787/81f92357-en. Accessed: 2026-08-02.
- OECD. 2026. OECD Due Diligence Guidance for Responsible AI. <https://doi.org/10.1787/41671712-en>. Accessed: 2026-08-02.
- Page, E. A. 2012. Give It Up for Climate Change: A Defence of the Beneficiary Pays Principle. *International Theory* 4(2): 300–330. doi.org/10.1017/S175297191200005X.
- Perrigo, B. 2023. Exclusive: OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less Toxic. *TIME*, <https://time.com/6247678/openai-chatgpt-kenya-workers/>.
- Pidoux, J.; Tighanimine, M.; Kypraiou, S.; Kgomo, S.; Malgwi, K. I.; Mathenge, R. M.; Okinyi, M.; and Oyange, J. 2025. Mapping Data Labour Supply Chain in Africa in an Era of Digital Apartheid: A Struggle for Recognition. arXiv preprint. arXiv:2512.04269. <https://arxiv.org/abs/2512.04269> [cs.CY].
- Pogge, T. 2005. World Poverty and Human Rights. *Ethics & International Affairs* 19(1):1–7. doi.org/10.1111/j.1747-7093.2005.tb00484.x.
- Roberts, S. T. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven, CT: Yale University Press.
- Sands, P.; Peel, J.; Fabr. A.; and MacKenzie R. 2018. *Principles of International Environmental Law*, 4th ed. Cambridge: Cambridge University Press.
- Schwartz, P. 2015. Principle 16: The Polluter-Pays Principle. In *The Rio Declaration on Environment and Development: A Commentary*, edited by Viñuales, J. E., 429–450. Oxford: Oxford University Press.
- Shue, H. 2020. *Basic Rights: Subsistence, Affluence, and U.S. Foreign Policy: 40th Anniversary Edition*. Princeton, NJ: Princeton University Press.
- Shue, H. 1999. Global Environment and International Inequality. *International Affairs* 75(3): 531–545. doi.org/10.1111/1468-2346.00092
- SOMO. 2026. Big Tech Sets Unfair Terms and Conditions for AI Data Workers Globally. <https://www.somo.nl/big-tech-sets-unfair-terms-and-conditions-for-ai-data-workers-globally/>. Accessed 2026-08-02.
- Sornarajah, M. 2021. *The International Law on Foreign Investment*, 5th ed. Cambridge: Cambridge University Press.
- The ETO Consortium. 2013. Maastricht Principles on Extraterritorial Obligations of States in the Area of Economic, Social and Cultural Rights. Heidelberg: FIAN International.
- The National Council for Law Reporting, Republic of Kenya. 2007. Employment Act, Cap. 226. <https://new.kenyalaw.org/akn/ke/act/2007/11/eng@2024-04-26>. Accessed: 2026-08-02.
- The National Council for Law Reporting, Republic of Kenya. 2022. Employment and Labour Relations Court Act, Cap. 8E. <https://new.kenyalaw.org/akn/ke/act/2011/20/eng@2022-12-31>. Accessed: 2026-08-02.
- The National Council for Law Reporting, Republic of Kenya. 2022. Labour Relations Act, Cap. 233.

<https://new.kenyalaw.org/akn/ke/act/2007/14/eng@2022-12-31>. Accessed: 2026-08-02.

The National Council for Law Reporting, Republic of Kenya. 2022. Occupational Safety and Health Act, Cap. 236A.

<https://new.kenyalaw.org/akn/ke/act/2007/15/eng@2022-12-31>. Accessed: 2026-08-02.

The National Council for Law Reporting, Republic of Kenya. 2023. Labour Institutions Act, Cap. 234.

<https://new.kenyalaw.org/akn/ke/act/2007/12/eng@2023-12-11>. Accessed: 2026-08-02.

Trucone-Borgogno, S. 2023. Climate Justice and the Duty of Restitution. *Moral Philosophy and Politics* 10(1): 203–224. doi.org/10.1515/mopp-2021-0071.

Tubaro, P., and Casilli, A. A. 2019. Micro-Work, Artificial Intelligence and the Automotive Industry. *Journal of Industrial and Business Economics* 46: 333–345. doi.org/10.1007/s40812-019-00121-1.

Tubaro, P.; Le Ludec, C.; and Casilli, A. A. 2020. Counting “Micro-Workers”: Societal and Methodological Challenges Around New Forms of Labour. *Work Organisation, Labour and Globalisation* 14(1): 67–82. doi.org/10.13169/workorglaboglob.14.1.0067.

United Nations Digital Library. 1992. Rio Declaration on Environment and Development. <https://digitallibrary.un.org/record/144444>. Accessed: 2026-08-02.

United Nations Digital Library. 2011. Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework. <https://digitallibrary.un.org/record/720245?v=pdf>. Accessed: 2026-08-02.

Weil, D. 2014. *The Fissured Workplace: Why Work Became So Bad for So Many and What Can Be Done to Improve It*. Cambridge, MA: Harvard University Press.