Ethics in Progress

17/1

Ethics in Progress

Vol 17, No 1. 2026.
Peer-reviewed online journal.

A multidisciplinary forum for research bridging between academic ethics and social practice.

EDITORIAL TEAM

We are grateful to Dr. Stephen Dersley for his linguistic assistance in working on this issue of *Ethics in Progress*.

Cover graphic: Marcello Bacciarelli, „Allegory of Justice – Themis", photo by P. Ligier (with permission of the National Museum in Warsaw)

Vol. 17 (2026) No. 1

Punishment in Question:

Responsibility, Structural Injustice, and Social Transformation

Table of contents:

Ethics in Progress (ISSN 2084-9257). Vol. 17 (2026). No. 1,
Art. #1, pp. 4-19. DOI:10.14746/eip.2026.1.1

Beyond the Traditional Grammar of Punishment: An Introduction

Chiara Magni

(Paris Lodron Universität Salzburg, Salzburg, Austria/Università degli Studi Roma Tre, Rome, Italy; chiara.magni123@gmail.com)
ORCID: 0009-0008-4399-7716

Anna De Giuli

(Università degli Studi di Milano, Milan, Italy; anna.degiuli@unimi.it)
ORCID: 0000-0003-1074-2895

Giulia Battistoni

(Universität Münster, Münster, Germany/Università degli Studi di Verona, Verona, Italy; giulia.battistoni.90@gmail.com)
ORCID: 0000-0002-5391-8590

Abstract: While modern criminal law has largely been structured around established notions of individual responsibility, blame, punishment, and reparation, some of the most pressing challenges facing contemporary societies increasingly expose the limits of these categories. This introductory contribution presents the themes and common concerns that animate the essays collected in this thematic issue, namely the growing dissatisfaction with traditional penal paradigms and the search for alternative conceptions of punishment, responsibility, and justice capable of addressing structural injustice, environmental harm, social conflict, and emerging forms of accountability. At the same time, the article introduces the issue's central problem: the tension between individual accountability and the systemic conditions within which wrongdoing and harm take shape. Situating these debates within a broader critical tradition that has questioned the relationship between law, violence, and justice, the contribution examines the opportunities and limits of criminal law as an instrument of social change and explores the extent to which contemporary legal institutions are capable of responding to harms that exceed traditional punitive frameworks.

Keywords: Punishment; criminal law; individual responsibility; collective responsibility; structural injustice; social transformation; restorative justice; transformative justice.

I. The Rationalization of Punishment: Between Legal Justification and Critical Deconstruction

Punishment is among the oldest institutions through which societies have sought to respond to wrongdoing.[1] Yet the history of punishment is also the history of its rationalization,

1 Although the article was jointly conceived, Part I was written by Chiara Magni, Part II by Anna De Giuli, and Part III by Giulia Battistoni.

which aims at circumscribing the field of inquiry to the specifically juridical concept of penalty. Modern legal thought has progressively sought to distinguish punishment from other forms of coercion and violence, narrowing its scope to a specifically juridical framework governed by principles of legality, proportionality, and procedural fairness. Within this model, accountability presupposes a relatively stable architecture: a wrongful act is attributable to an identifiable agent, responsibility is assigned according to established criteria concerning both the objective aspect of the act and the subjective condition of the agent, and an appropriate response is formulated in terms of sanction, compensation, or condemnation by an impartial third party. This framework continues to structure a wide range of contemporary legal institutions.

It may be noted that the long rationalisation process of punitive action coincides:

(1) first and foremost, with the attempt to legitimise the gap between just punishment and legal punishment, an endeavour that has taken different forms over time. Theories seeking to provide a justification for punishment can generally be grouped into three broad categories: (i) retributive theories; (ii) utilitarian theories, traditionally associated with the Enlightenment tradition in criminal law; and (iii) mixed theories. More recently, however, Anglo-American scholarship has challenged the classical foundations of sentencing through the development of "normative justifications" – whether moral (Murphy 2014) or political (Kelly 2018; Caruso 2021). Closely related to this rethinking is the need to conceive punitive practices in a way that also takes into account the structural conditions and inequalities that shape both criminal behaviour and the attribution of responsibility;

(2) secondly, with the gradual transition from "big punishments" to "small disciplines": the abandonment of exemplary punishments in favour of confinement, ranging from imprisonment to probation, electronic monitoring, and home detention.

Punishment has thus been conceived as a form of *justified* violence. Whether understood through retributive, utilitarian, or mixed theories, it presupposes the legitimacy of inflicting suffering, deprivation, or constraint in the name of a normative order, giving rise to competing accounts of desert, deterrence, rehabilitation, prevention, or social defence. Then, the peculiar link between justice and law is pointed out. It is precisely this presumed connection that has attracted the sustained attention of some of the most influential critical thinkers, who have sought to deconstruct it. Nietzsche, for instance, famously argued that punishment originated in relations of debt and creditorhood rather than in a concern for justice, thereby revealing the historical contingency and violent genealogy of concepts such as guilt, responsibility, and desert:

> The equivalence is provided by the fact that in place of an advantage that pays directly for the injury (...) the creditor is granted a kind of pleasure as repayment and compensation – the pleasure of being allowed to vent his power uninhibitedly on someone powerless, the thrill "*de faire le mal pour le plaisir de le faire*," the enjoyment of violating (...). By means of "punishment" of the debtor the creditor partakes of a *master's right*: finally he too arrives for once at the elevating feeling of being allowed to despise and abuse a creature as something "beneath him" – or

> at least, in case the actual punishing authority, the execution of punishment has already transferred to "the authorities," he is able to see the creature despised and abused. The compensation therefore consists in a court order and right to practice cruelty (Nietzsche 2014, 253–254).

Another paradigmatic author in this respect is Walter Benjamin, who famously challenged the assumption that law and justice are interchangeable concepts, arguing that every legal order remains inseparable from both law-founding (*die rechtsetzende Gewalt*) and law-preserving violence (*die rechtserhaltende Gewalt*). Jacques Derrida, in turn, distinguished law as a historically situated and therefore deconstructible institution from justice as an experience of the impossible. Yet this impossibility does not amount to a negation of justice; rather, it is precisely in its fundamentally unattainable character – which grounds the deconstructive nature of law – that justice is most fully honored and finds its strongest defense (Derrida 1992). For all these authors, the crucial issue is how to make sense of a form of violence that is simultaneously constitutive of law *and* responsible for its preservation. The latter aspect was articulated most clearly by Kant, for whom

> Right should not be conceived as made up of two elements namely an obligation in accordance with a law and an authorization of him who by his choice puts another under obligation to coerce him to fulfill it. Instead, one can locate the concept of Right directly in the possibility of connecting universal reciprocal coercion with the freedom of everyone. (…) Right and authorization to use coercion therefore mean one and the same thing (Kant 1991, § E, 57–58).

Yet the claim that violence is also constitutive of law is the truly unsettling one, since it reveals that the legal order cannot be understood solely as a rational mechanism for restraining violence, but must also engage with the fact that violence is inscribed in the very moment of the legal order's foundation. This appears to compromise any attempt to secure the autonomy and purity of law as a distinct rational sphere. In his recent book *De la violence à l'institution. Philosophie juridique de Hegel*, Jean-François Kervégan (Kervégan 2025) reflects precisely on this issue, showing that Hegel himself was far from ignoring the problem of the relationship between violence and law, including in its constitutive dimension. The German term *Gewalt* itself denotes both violence and legitimate authority – a polysemy that seems to capture its intrinsic ambivalence. Kervégan argues, however, that the Hegelian conception of power, understood as *Macht*, differs fundamentally from violence by virtue of the mediations through which it operates, mediations that allow it to be comprehended within the sphere of rationality. This is reflected precisely in the most familiar juridical regulation of the use of force, namely punitive power: for Hegel, the second coercion [*Zwang*] should not be understood as a form of violence externally imposed by society upon a recalcitrant empirical will, but rather as an internal requirement that the free will conform to its own concept – to the point that punishment is described as a "right *posited in the criminal himself*" (Hegel 2008, § 100, 102). From this perspective, whereas Benjamin conceives law in terms of the reproduction of founding violence, Kervégan, following Hegel, emphasises its role in the protection and actualisation of freedom, while nonetheless acknowledging its constitutive impurity. The juridical sphere, though necessary for overcoming the stage of anomic violence, does not naively erase every trace of it. Violence remains, as Kervégan writes, "present at the margins of law, both before and beyond the normative order it institutes" (Kervégan 2025, 49, our trans.). In the Hegelian view, the legal order is indeed anchored in that which radically negates it; yet this does not function as a principle of legitimacy, but rather constitutes an ineliminable moment of its genesis – one that is merely phenomenological rather than foundational.

For Derrida, by contrast, the origin of law is neither legal nor illegal: it belongs to a "mystical" dimension (Derrida 1992, 14) in which the very categories of legitimacy and illegitimacy have not yet been constituted and which exceeds the opposition between founded and unfounded. The origin of law is not, therefore, an anomic act identified as such with respect to a normative horizon, but rather an act that wholly exceeds any preliminary articulation of such dimensions. For this reason, law is, in the Derridean sense, essentially deconstructible: not because it rests merely on contingent or revisable normative layers, but because its ultimate foundation is itself without foundation, exposed to the groundlessness that inhabits every instituting act. Yet such a deconstructible nature does not simply undermine law; rather, it opens the very space in which justice can be conceived as something irreducible to law itself. It is in this sense that deconstruction is not external to justice but coincides with it: "deconstruction is justice" (Derrida 1992, 15), insofar as justice names precisely that excess which prevents law from closing in upon itself as a self-sufficient order.

Beyond possible interpretations, what is certain is that to overlook these aporias would be to disregard the entire critical effort aimed at exposing the limits and blind spots of legal orders – an effort that Derrida rightly attributed to "the articulation between literature and philosophy, law and politico-institutional problems" (Derrida 1992, 8). From this perspective, the contribution of authors within the critical tradition has been, and continues to be, indispensable for revealing the constitutive tensions that inhabit modern legal and punitive institutions. Their enduring value lies precisely in the fact that they leave us with "disparities and disputes at least as much as [they leave us] with agreements, with coincidences or consensus" (Derrida 1992, 9).

This also means that a critical dimension is immanent to law itself, by virtue of the very way in which law is constituted: an immanent critique of law thus implies a fundamentally open conception of the juridical order. This does not merely mean that law is historically conditioned, and therefore subject to change, nor simply that it can never be regarded as a definitive system. The issue is not only one of incompleteness, but rather that law's claim to validity as a normative order must always be accompanied by an awareness of the possibility of its own deconstruction, grounded in its constitutive "impurity": this means that its legitimacy can never be taken as fully secured once and for all. It is therefore no coincidence that contemporary critical legal approaches, especially in the field of punishment, have increasingly moved away from a conception of responsibility centred exclusively on the individual offender and have instead sought to take into account the broader structures of domination, inequality, and injustice that legal institutions may reproduce or reinforce. At the same time, they have progressively embraced a more prospective and future-oriented understanding of justice, one that is not exclusively concerned with the retrospective attribution of responsibility, but also with the transformation of the social conditions that give rise to criminal conduct.

The contributions gathered in this issue take shape within such a broader horizon of inquiry, from which arises the awareness that some of the most pressing challenges facing contemporary societies expose the limits of the traditional juridical grammar through which justice has long been conceived. This grammar is structured around a consolidated set of concepts – individual responsibility, fault, imputation, punishment, compensation, and reparation – which play a central role within legal systems. However, when these categories are confronted with phenomena such as structural discrimination, historical injustice, mass violence, and other forms of systemic harm, their insufficiency becomes increasingly evident in capturing the complexity of ongoing processes and the role that law is called upon to play. This tension has generated a growing body of scholarship aimed at rethinking the relationship between responsibility, punishment, and social transformation. Rather than asking exclusively who is to blame for a past wrong, these approaches increasingly inquire into the conditions that enable injustice to persist, the forms of accountability appropriate to structural harms, and the institutional mechanisms capable of fostering social change. The focus shifts, at least in part, from the retrospective logic of desert to prospective concerns with prevention, transformation, repair, and collective responsibility.

One area of inquiry concerns the philosophical foundations of punishment, which have been increasingly revisited in light of the inadequacy of exclusively retributive frameworks for addressing contexts marked by structural violence and enduring inequalities. Indeed, the prison system itself has become the object of increasingly intense debate, driven by longstanding concerns over prisoners' rights and extending the biopolitical critique that highlights the intrinsic link between punishment and power. In light of often alarming prison conditions and the phenomenon of mass incarceration – particularly in the United States – the legitimacy of imprisonment as the default response to crime has been progressively challenged. This crisis concerns not only the effectiveness of punishment, but more fundamentally its normative justification, opening a theoretical space in which abolitionist perspectives (Davis 2003; Coyle & Scott 2021; Ricordeau 2026) and alternative paradigms have emerged. Since the 1990s restorative justice has developed as a response to the limits of retributive criminal law (Zehr 2002; Van Ness & Johnstone 2013). It reorients the response to wrongdoing away from punishment and toward the repair of harm through the active involvement of victims, offenders, and communities, emphasizing dialogue, accountability, and reparation. Its diffusion has been remarkable, both in theoretical scholarship and in practical implementation. A particularly significant example in this regard is Italy, where the Cartabia Reform introduced a comprehensive regulatory framework for restorative justice, marking one of the most ambitious attempts to integrate restorative practices into the criminal justice system (Legislative Decree No. 150 of 10 October 2022, implementing Law No. 134 of 27 September 2021).

Alternative conceptions of responsibility are also explored that move beyond individualistic models and emphasize collective or political forms of accountability, as exemplified by debates in environmental justice and transitional justice, within a perspective according to which legal institutions are capable not only of sanctioning wrongdoing but also of shaping social meanings and promoting cultural change. This broader understanding of responsibility raises further questions concerning the actors and mechanisms through which normative change can be pursued. On the one hand, legal institutions may seek to influence collective understandings and social practices through the expressive dimensions of law. On the other hand, processes of accountability may emerge outside formal institutions, through social practices that enforce shared values and regulate public behaviour within communities. The contributions situated within this broader line of inquiry examine both the transformative ambitions of criminal law and the role of informal social sanctions in shaping norms, expectations, and practices of accountability. Within this framework, transformative justice has emerged as an approach developed primarily within grassroots traditions to address harm without relying on punitive state institutions. Rather than focusing on formal legal procedures, it seeks to respond to violence and harmful conduct through community-based practices oriented toward structural change, with particular attention to the social and cultural conditions that make harm possible and to the long-term transformation of those conditions (Evans 2018; Pelsinger 2025).

Taken together, these debates suggest that the most pressing challenges facing contemporary legal and political thought cannot be reduced to the identification of individual wrongdoers or the retrospective allocation of responsibility. Rather, they call attention to the institutional processes through which harms are generated, sustained, and normalized, inviting a re-examination of the relationship between individual conduct, collective practices, and the broader frameworks within which injustice takes shape. These broader concerns become particularly salient when harms are rooted in enduring relations of domination and affect individuals as members of socially disadvantaged groups. In such cases, the tension between individual accountability and structural injustice emerges with particular clarity, raising questions about whether criminal law can adequately respond to harms whose origins lie in systemic social conditions rather than in isolated acts of wrongdoing. Gender-based violence and femicide provide a paradigmatic illustration of these difficulties.

II. Addressing Structural Injustice: Symbolic Functions and Structural Limits of Criminal Law

Within anti-discrimination literature, a so-called "structural turn" has been gaining ground for quite some time (Bagenstos 2006; Burns 2008). Recently, Shreya Atrey has conceptualized the structural turn as follows:

> (…) in contrast with the liberal view which sees the harm of inequality/ discrimination as something inflicted by and against individuals or collectivities through specific acts or omissions. The structural view places individual victims and perpetrators within the broader dimensions of the social, economic, legal, political, psychic and cultural contexts in which they exist and the power relations within them (Atrey 2025, 154).

This involves a shift in perspective in understanding discriminatory phenomena and in providing an institutional response to counter them. It should be noted that not all structural harms coincide with structural injustices (Young 2011; Wolff 2024). For example, several cases of harmful pollution produce a structural harm that affects the collectivity, without falling into the category of structural injustice. Indeed, structural injustice concerns social structures whose functioning leads to the creation of harms against individuals belonging to social groups, which are deemed an injustice insofar as they are linked to dynamics of domination[2] or generate severe vulnerability (Wolff 2024).

The recognition of certain discriminatory dynamics as systemic has raised demands for such conducts to find a specific criminal typification.[3] The objective is to bring to the fore the disvalue produced by phenomena of structural injustice that hinder substantial equality. In particular, their legal translation can give rise either to autonomous offenses, such as femicide, or to the introduction of specific aggravating circumstances that take into account forms of discrimination based on gender, race, and sexual orientation, among others. Such configurations are traced back and grouped together by some authors through the use of the expression "anti-discriminatory criminal law" (Pelissero 2024a; Perin 2024; Londoño Martínez 2025). In this way, criminal law focuses on those forms of wrongfulness that are linked to membership in a social group. However, if the punitive intervention is isolated from other public policies, the risk is that of slipping into penal populism, exhausting itself in a purely symbolic use of the sanction that leaves the matrix of the problem intact.

This section reflects precisely on the introduction of such offenses into criminal law. In particular, the core of the problem lies in the discrepancy that arises between countering

2 Note that in the definition proposed by Iris Marion Young the element of domination is central. For Young, a structural injustice "exists when social processes put large groups of persons under systematic threat of domination or deprivation of the means to develop and exercise their capacities, at the same time that these processes enable others to dominate or to have a wide range of opportunities for developing and exercising capacities available to them" (Young 2011, 52).

3 On the limits of retributive punishment for structural harms and the proposal of a "transformative sanction" based on Iris Marion Young's Social Connection Model, see (Guljar 2026) in this thematic issue.

a phenomenon of structural discrimination that has deep and widespread roots and an instrument that acts after the harm has been committed by a single individual. In other words, criminal law intervenes on the symptom – the manifestation by an individual of a conduct that falls within a structural injustice – but not on the matrix of the problem – the social structure that allows its maintenance. From this observation, three questions follow: is criminal law suitable for a harm stemming from a structural injustice? Can attention to the structural dimension of the harm be understood and driven forward by institutions? Is the penal instrument sufficient to produce effects that go beyond the single punishable episode? The following analysis addresses these questions, contextualizing them through the case study of the crime of femicide, recently introduced into the Italian legal system by Law 181/2025, which entered into force on December 17, 2025.

From a first perspective, the question of adequacy does not merely revolve around whether such conducts call for punishment. Indeed, the disvalue of those conducts that perpetuate a structural injustice lies in their conflict with the core prerequisites of a democratic and egalitarian society, as well as in the negation of equal human dignity.[4] Such reasons often constitute the foundation of the penal relevance that the legislature recognizes in similar offenses. In this sense, the infliction of punishment finds its foundation also in requirements of negative general prevention (intimidating-deterrent) and positive general prevention (cultural orientation) (Marinucci, Dolcini, & Gatta 2023). In this thematic issue, Jorge Fernández Mejías (Mejías 2026) takes up the promotional function of law as defined by Norberto Bobbio (Bobbio 1969), in order to investigate the conditions that render legitimate the recourse to criminal law as a driver of social change. In this section, instead, I intend to investigate whether punishment can in some way be an adequate tool to respond to a phenomenon of structural injustice. In other words, whether punishment can have some type of impact on the matrix of the structural problem.

Indeed, as previously noted, structural injustices are not mere wrongful actions of an individual agent. In fact, every (discriminatory) action of an individual concurs, together with that of other agents and institutions that are part of a specific social structure, to the maintenance of the objectives and interests defined and accepted within that structure. There is, therefore, a sharing of values and social norms within that community that makes the maintenance of those injustices possible. In other words, injustice also passes through interpretative codes and social meanings that qualify a specific behavior as legitimate or acceptable. It is evident that this is a complex phenomenon, precisely

4 The European Court of Human Rights has also ruled in these terms on a case concerning hate speech. "The present case is characterised, in particular, by the fact that the applicant was punished for statements classified by the domestic courts as 'hate speech'. Having regard to the relevant international instruments (see paragraphs 22-24 above) and to its own case-law, the Court would emphasise, in particular, that tolerance and respect for the equal dignity of all human beings constitute the foundations of a democratic, pluralistic society. That being so, as a matter of principle it may be considered necessary in certain democratic societies to sanction or even prevent all forms of expression which spread, incite, promote or justify hatred based on intolerance (including religious intolerance), provided that any 'formalities', 'conditions', 'restrictions' or 'penalties' imposed are proportionate to the legitimate aim pursued" (European Court of Human Rights 2003, *Gündüz v. Turkey*, n. 35071/97, 4 December).

because it is widespread at a social level: not only due to the subjects who take part in it, but also due to the distinct manifest or implicit ways in which one contributes to the maintenance of a specific belief system. If the objective is to eliminate these phenomena at their roots, it is necessary to diversify counter-strategies, precisely to impact the matrix of the problem. In this sense, criminal law can be a locus of safeguard and an instrument of social influence, but it can in no way be considered a self-sufficient means to trigger social transformation. In this respect, providing a tout-court answer to the question of the adequacy of criminal law in tackling structural injustice is simplistic and partial. The variables that affect this answer necessarily have to do, among others, with the social and legal context, and the presence (or absence) of a solid and structured intervention program for an effective social transformation. Added to this is the fact that even the repeal of legislative provisions linked to forms of structural discrimination – demanded by those who support an extreme liberal position – is anything but free from criticism: both because of the obligations assumed at the international level, and because of the message it conveys regarding the disvalue of certain discriminatory conducts (Pelissero 2024b).

If the strictly repressive and individualizing efficacy of the penal sanction is unable to dismantle structures of oppression, a part of the response to structural injustices nevertheless lies in the expressive-epistemic function of law, that is, in the capacity of legal language to name the structural harm. In this sense, Catharine A. MacKinnon observed that "you can't change a reality you can't name" (MacKinnon 2007, 89). The use of certain terms within the language of the law contributes to the epistemic recognition of phenomena of structural discrimination. This aspect was highlighted by those who supported the legislative reform that led to the integration of the crime of femicide into the Italian legal system (Di Nicola Travaglini 2025; Boiano 2025). In this regard, Ilaria Boiano observes:

> Refusing to legally recognise violence for what it is – an exercise of patriarchal power, fuelled by sexist hatred – means leaving the legal sphere open as a context for legitimising it, whilst overlooking the fact that the attempt to address it in legal terms has made it possible to reshuffle the cards on the table of social relations and legal practice (Boiano 2025, 16, our trans.).

This leads directly into the domain of the symbolic functions that this type of legislative intervention can have. Indeed, the introduction of an autonomous crime of femicide performs a signalling function, in which the legislature gives attention and prominence to a serious and widespread phenomenon, also in relation to the situation of social alarm linked to it (Ferraro 2026). However, the core issue remains the real capacity for social influence, in order to attack the very root of the phenomenon.

Offenses linked to structural injustices, in fact, present extra-legal concepts that render them indeterminate. This determines a problem in terms of institutional translation, raising the doubt as to whether adjudicating bodies are effectively capable of understanding the normative text and applying it to concrete cases. In this perspective, however, one could hypothesize that even symbolic criminal law might fall within an interactive legislation process – a theoretical approach used especially in health law and bio-law. This model, indeed, "builds on broad processes of interaction among a great number of societal actors at all stages of the norm-making and implementing processes" (Van der Burg 2015, 7). Within the paradigm of interaction falls the communicative approach to legislation, in which several further functions are recognized in addition to the traditional ones (protective and instrumental). Specifically, Francesco Ferraro (Ferraro 2026) identifies the following functions: cognitive,[5] rhetorical,[6] constitutive,[7] expressive[8] and reactive.[9] Through these functions, the introduction of extra-legal concepts into the language of the law ceases to be merely a factor of indeterminacy and allows for the identification of a common vocabulary among the subjects belonging to an interpretive community, thus delineating the limits of the interpretative framework in the legal field. In this sense, the impact on the matrix of the structural injustice problem would be achieved through the influence of social meanings:

> [Law] can alter the social norms that provide interpretative codes (that is, determine which meanings to attribute) and, through these changes, it can alter the social norms that govern behaviour. This chain of symbolic changes can lead, and often does lead, to social and behavioural changes: by changing the way people think – their mindset – one also changes the way they act (Ferraro 2026, 213, our trans.).

This potential cultural change that can be conveyed through criminal law must, however, be measured against current social understanding and the modalities through which it is possible to intervene. In this sense, the use of aggravating circumstances can be a more effective and balanced tool to generate an understanding of certain extra-legal concepts by the community of experts. This is due to the fact that the aggravating circumstance rests on an already consolidated base offense, reducing the impact of interpretative uncertainties on the level of the sanction. On the contrary, the provision, instead, of autonomous crimes, such as the case of femicide, can increase applicative complexity, thus undermining its effectiveness. Although it is still too early to evaluate it, this problem is already emerging in some recent measures in Italy (Gatta 2026). The

5 "Communicative legislation provides a vocabulary – a set of terms – that shapes how interpreters – legal and political actors – perceive social reality" (Ferraro 2026, 104, our trans.).

6 "Communicative legislation provides rhetorical arguments and *topoi* that are subsequently used in public debates (legal, political and ethical)" (Ferraro 2026, 104, our trans.).

7 "Communicative legislation constitutes an 'interpretive community', that is to say, it fosters the formation of a group of official and unofficial interpreters: legal and political actors, as well as ordinary citizens, who are involved in matters relating to the application of the law" (Ferraro 2026, 104, our trans.).

8 "It is one of the most significant functions of communicative legislation, consisting of 'giving voice' to the values that society considers most important" (Ferraro 2026, 104, our trans.).

9 "Thanks to its 'open' nature, communicative legislation creates forums for discussion and new decision-making structures, where the general public and officials interact to adapt the law to social needs" (Ferraro 2026, 105, our trans.).

arguments adduced to deny the crime of femicide reproduce discriminatory beliefs that highlight a lack of understanding of the phenomenon of femicide as a structural injustice. In this way, the structural dimension of the harm remains foreign to the legal culture of the courts: the lack of understanding of the phenomenon blocks its practical application, making it a merely formal intervention that provides only an apparent protection. After all, the proposal of the Italian legislature is an example of symbolic criminal law, not oriented towards achieving the declared purpose of countering gender-based violence. It is an "alibi" legislation (Ferraro 2026, 33) that manipulates public perception because the legislature, although apparently providing a response to a structural injustice through the punitive instrument *par excellence*, in fact does not invest sufficient resources in prevention and cultural education. A similar problem had occurred in the past also in the area of hate speech. Barbara Giovanna Bello (Bello 2019), in fact, highlighted several reasons that make protection through the criminal instrument complex, identifying them in the difficulty of socially understanding the disvalue of certain verbal aggressions, in their negative impact on the health of the recipients, and in the epistemic uncertainty linked to their identification. In other words, there can be no structural impact if legislation is not outcome-oriented with respect to the modification of the belief system. This endeavor requires, in fact, a diversified, comprehensive response supported by adequate economic resources.

In conclusion, it has emerged that the penal instrument cannot be the only tool providing a response to a structural injustice; otherwise, it functions as a mere legislative alibi aimed at promoting a distorted perception regarding the commitments undertaken to counter a phenomenon that has a structural matrix. The use of this type of symbolic legislation is inevitably linked to purposes of social consensus. However, this type of response, although it still performs a signaling function, is insufficient to respond to a structural injustice. Instead, the use of the language of criminal law, together with other strategies that act more effectively on the matrix of social injustices – the discriminatory social culture and all those mechanisms that allow its maintenance – can be useful precisely due to the functions performed by a communicative legislation that impacts, among other things, the creation of the interpretative framework within which the recognition of that phenomenon is situated.

III. Overview of the Contributions

Despite their different theoretical premises and objects of inquiry, the seven contributions collected in this issue share a common concern: they interrogate the meaning, justification, and functions of punishment in a context marked by growing dissatisfaction with traditional penal paradigms. Retribution, individual blame, proportionality, and imprisonment increasingly appear inadequate for addressing structural injustice, environmental harm, social conflict, and emerging forms of responsibility. The contributions therefore question the limits of punitivism and explore alternative or complementary responses to wrongdoing and harm, including restorative, transformative, preventive, and prospective forms of justice, as well as new conceptions of responsibility extending beyond individual fault. The first four contributions develop conceptual frameworks for rethinking punishment, responsibility, and justice. The final three examine how such questions arise in concrete legal and social contexts, including environmental harm, social punishment, and affirmative consent legislation. In what follows, the main lines of argument and the principal contributions of each article will be briefly presented.

Marina Beraha's contribution *Utilitarianism and Prison Abolition: A Normative Perspective* mobilizes utilitarianism in an original way by drawing on Bentham's penal thought not to justify imprisonment, as utilitarian theory has traditionally done, but to support its abolition. If deterrence is regarded as the principal justification of punishment and harsher sanctions are assumed to produce greater preventive effects, this logic conflicts with the utilitarian principle of frugality, which requires minimizing suffering. Through a novel reading of Bentham, the paper shows that his primary concern is crime prevention through indirect legislation and the removal of the social causes of offending. Integrating this account with contemporary debates on punishment, incarceration and restorative justice, it argues that a genuinely utilitarian approach should prioritize harm reduction through social restructuring and restorative responses rather than deterrence-centred imprisonment.

María Fernanda Miranda González's article *Beyond Retribution: Walter Benjamin and the Case for Prospective Justice* extends the critique of traditional punitive paradigms by challenging an even more fundamental assumption of modern penal thought, namely that justice is necessarily grounded in retribution and proportional punishment. The author argues that this paradigm, grounded in a logic of debt and repayment, continually postpones justice to the future and is therefore unable to realize it in the present. As an alternative, she proposes a notion of "prospective justice" inspired by Walter Benjamin's Jewish-messianic conception of time. Rather than focusing on punishment, prospective justice seeks to address the social and institutional structures that produce injustice, emphasizing human dignity, collective agency, and the possibility of realizing justice in the present through the interruption of existing relations of domination.

Idoia Quintana Domínguez's article *Punishing and Forgiving: Contributions of Deconstruction to Contemporary Debates on Justice* pursues a related attempt to rethink the meaning of justice beyond inherited punitive categories, approaching this task through Derrida's reflections on forgiveness and its complex entanglement with punishment. Derrida's account of forgiving the unforgivable locates forgiveness beyond logics of proportionality, reciprocity and calculation, while maintaining it as an "impossible" experience that resists institutionalisation and measurable outcomes. Yet the boundary between punishment and forgiveness is neither external nor rigid. This makes forgiveness a critical lens through which to interrogate punitive, restorative and transitional justice, as both a possible opening beyond punishment and a means for managing harm and normalising conflict.

Arif Guljar's article *Reimagining Justice Beyond Blame: Defending Iris Marion Young's Social Connection Model* defends Iris Marion Young's Social Connection Model (SCM) as a framework for addressing responsibility for structural injustice and resolving the "punishment dilemma": the liability model is conceptually ill-suited to systemic harms, yet impunity is morally unacceptable. Reconstructing the SCM's features and parameters, the article argues that responsibility for structural injustice derives from social position and participation in interconnected processes rather than from blame or direct causation. Drawing on Virginia Mantouvalou, Serena Parekh and Young's critique of ressentiment, it develops the idea of transformative sanction, understood as punitive or quasi-punitive measures aimed at forward-looking structural disruption rather than backward-looking desert. The article concludes that transformative sanction offers a third path between retribution and impunity by orienting accountability toward institutional and structural transformation.

Giulia Battistoni's article *Preventing, Repairing, Punishing: The EU Environmental Crime Directive as A Hybrid Response to Environmental Harm* brings these questions into the environmental domain, where diffuse and cumulative forms of harm challenge traditional models of responsibility and punishment. The temporal and diffuse character of environmental harm often obscures causation, the attribution of responsibility, the justification of punishment and even the definition of harm itself. These challenges reveal the need for clearer normative frameworks. Against this background, the article examines contemporary debates on environmental responsibility and punishment and analyses Directive (EU) 2024/1203. It argues that the Directive introduces a hybrid response to environmental harm, combining individual, corporate, and institutional responsibility, retributive and preventive rationales of punishment, and retributive, restorative, and transformative approaches to justice.

Natalia Witosza's article *From Celebrities to Ordinary Citizens: Limits of the Legitimacy of Social Punishment* examines whether social punishment practices should apply differently to celebrities and non-celebrities. Adopting the method of reflective equilibrium, it argues that the legitimacy of social punishment depends on the wrongdoer's social status and, more specifically, on the degree of epistemic power they possess. The paper distinguishes between celebrities, quasi-celebrities and non-celebrities, maintaining that social punishment is generally justified only for celebrities and, in qualified cases, for quasi-celebrities. Because it can reduce epistemic power and mitigate its harmful effects, social punishment is often more appropriate for public figures. By contrast, non-celebrities and most quasi-celebrities should be subject only to private punishment, which poses fewer risks of disproportionate harm and better supports social reintegration.

Jorge Fernández Mejías' article *Can Criminal Law Drive Social Change? The Promotional Function and the Affirmative Consent Debate* brings the discussion to one of the most contested functions of criminal law: its capacity to contribute to social and cultural transformation. Using the Spanish affirmative consent model as a case study, it distinguishes between protecting negative sexual autonomy – the right not to be subjected to non-consensual sexual conduct – and promoting positive autonomy. Drawing on Bobbio's distinction between the conservative and promotional functions of law, the article analyses affirmative consent from three perspectives: declaratory, conditioned promotional and normative framing. Through these lenses, the article demonstrates that the promotional justification of affirmative consent fails to meet their respective requirements, meaning it cannot legitimately be deployed to entrench a specific communicative model. The article concludes that criminal law should play only a subsidiary, consolidating role in cultural transformation, leaving broader promotional aims to social and educational policy.

Taken together, the contributions collected in this issue provide an original and theoretically diverse exploration of how punishment and justice can be reimagined today, challenging entrenched dichotomies between punishment and impunity, individual and collective responsibility, and retribution and social transformation.

References

Atrey S. 2025. "Equality Law: A Structural Turn," *German Law Journal* 26(2):153–169. https://doi.org/10.1017/glj.2025.1

Bagenstos S. R. 2006. *The Structural Turn and the Limits of Antidiscrimination Law* 94(1):1–47. https://doi.org/10.15779/Z38HB00

Bello B. G. 2019. "Riflessioni critiche sulla legislazione penale italiana contro il discorso d'odio nel sistema multilivello: la prospettiva del paradigma del Social Working of Law," *Ragion Pratica* 2. https://doi.org/10.1415/95164

Benjamin W. 2021. *Toward the Critique of Violence. A Critical Edition*, Stanford: Stanford University Press.

Bobbio N. 1969. "The Promotion of Action in the Modern State," in G. Hughes (Ed.), *Law, Reason and Justice. Essays in Legal Philosophy* (pp. 189–215). New York: New York University Press.

Boiano I. 2025. "Nominare la violenza maschile contro le donne: diritto penale e giustizia tra conflitto simbolico e responsabilità politica", *Giustizia Insieme*, April 4:1–18.

Burns T. R. 2008. “Towards a Theory of Structural Discrimination: Cultural, Institutional and Interactional Mechanisms of the ‘European Dilemma’,” in G. Delanty, R. Wodak & P. Jones (Ed.), *Identity, Belonging and Migration* (pp. 152–172). Liverpool: Liverpool University Press. https://doi.org/10.5949/liverpool/9781846311185.003.0009

Caruso D. G. 2021. *Rejecting Retributivism: Free Will, Punishment, and Criminal Justice*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108689304

Coyle M. J. & Scott D. 2021. *The Routledge International Handbook of Penal Abolition*. New York – London: Routledge. https://doi.org/10.4324/9780429425035

Davis A. Y. 2003. *Are Prisons Obsolete?* New York: Seven Stories Press.

Derrida J. 1992. “Force of Law: The ‘Mystical Foundation of Authority’,” in D. Corner, M. Rosenfeld, & D. G. Carlson (Eds.), *Deconstruction and the Possibility of Justice* (pp. 3–67). New York-London: Routledge. https://doi.org/10.4324/9781315539744.

Di Nicola Travaglini P. 2025. “Risposte alle principali obiezioni sull’introduzione del delitto di femminicidio (e delle aggravanti per i delitti di violenza contro le donne)”, Roma, October 21.

European Court of Human Rights (ECtHR) 2003. *Gündüz v. Turkey*, n. 35071/97, 4 December. URL: https://hudoc.echr.coe.int/eng?i=001-61522.

Evans M. 2018. *Transformative Justice. Remedying Human Rights Violations Beyond Transition*. New York – London: Routledge. https://doi.org/10.4324/9781351239462

Fernández Mejías J. 2026. “Can Criminal Law Drive Social Change? The Promotional Function and the Affirmative Consent Debate,” *Ethics in Progress* 17(1).

Ferraro F. 2026. *Le funzioni simboliche del diritto. Lineamenti di una teoria analitica*. Torino: G. Giappichelli Editore.

Gatta G. L. 2026. “Femminicidio o omicidio se il marito tenta di uccidere la moglie dopo aver scoperto che si prostituisce? Note a margine del caso di ventimiglia”, *Sistema Penale* 1:195–199.

Guljar A. 2026. “Reimagining Justice Beyond Blame: Defending Iris Marion Young’s Social Connection Model,” *Ethics in Progress* 17(1).

Hegel G. W. F. 2008. *Outlines of the Philosophy of Right*. Oxford-New York: Oxford University Press.

Kant I. 1991. *The Metaphysics of Moral*. Cambridge: Cambridge University Press.

Kelly E. 2018. *The Limits of Blame. Rethinking Punishment and Responsibility*. Cambridge, London: Harvard University Press. https://doi.org/10.4159/9780674989436

Kervégan J.-F. 2025. *De la violence à l’institution. Philosophie juridique de Hegel*. Paris: Vrin.

Londoño Martínez F. 2025. "¿Derecho Penal Anti Discriminación? Una crítica, a propósito de 'motivi aggravanti e circostanze discriminatorie' (Andrea Perin, Torino, 2024). De Paso, Notas sobre la Atenuante Pasional y el Principio de Subsidiariedad", *Política Criminal* 20(39):438–454. https://doi.org/10.4067/s0718-33992025000100438

MacKinnon C. A. 2007. *Women's Lives, Men's Laws*. Cambridge, Massachusetts: Harvard University Press.

Marinucci G., E. Dolcini, & G. L. Gatta. 2023. *Manuale di diritto penale: Parte generale*. Milano: Giuffrè Francis Lefebvre.

Murphy J. G. 2014. *Punishment and the Moral Emotions: Essays in Law, Morality, and Religion*. Oxford – New York: Oxford University Press. https://doi.org/10.1093/acprof:osobl/9780199764396.001.0001.

Nietzsche F. 2014. "On the Genealogy of Morality. A Polemic," in F. Nietzsche, *Beyond Good and Evil/On the Genealogy of Morality* (pp. 207–349). Stanford: Stanford University Press.

Pelissero M. (Ed.) 2024a. "Focus 'Diritto Penale Antidiscriminatorio. Una indagine, partendo dalla giurisprudenza'", *GENIUS*, October 4. URL: https://www.geniusreview.eu/2024/focus-diritto-penale-antidiscriminatorio/

Pelissero M. 2024b. "Importanza, ragionevolezza e limiti del diritto penale antidiscriminatorio tra presente o (possibile) futuro", *GENIUS*, October 4:1–4.

Pelsinger C. 2025. "Beyond Punishment: Towards a Framework of Transformative Accountability," *Punishment & Society* 27(4):785–804. https://doi.org/10.1177/14624745251319345

Perin A. 2024. *Motivi aggravanti e circostanze discriminatorie. Legittimità e limiti della sanzione penale dell'offesa alla pari dignità*. Torino: G. Giappichelli Editore.

Ricordeau G. 2026. *Tant qu'il y aura des prisons*. Lyon: Le passager clandestin.

Van der Burg W. 2015. "The Emerging Interactionist Paradigm and the Ideals of Democracy and Rule of Law," SSRN Scholarly Paper No. 2589808. Social Science Research Network, April 4. https://doi.org/10.2139/ssrn.2589808

Van Ness D. & Johnstone G. 2013. *Handbook of Restorative Justice*. London – Oxford – New York: Taylor & Francis. https://doi.org/10.4324/9781843926191

Wolff J. 2024. "Structural Harm, Structural Injustice, Structural Repair," in J. Browne & M. McKeown (Eds.), *What Is Structural Injustice?* (pp. 12–30). Oxford – New York: Oxford University Press. https://doi.org/10.1093/oso/9780198892878.003.0002

Young I. M. 2011. *Responsibility for Justice*. Oxford – New York: Oxford University Press.

Zehr H. 2002. *The Little Book of Restorative Justice*. Intercourse: Good Books.

Ethics in Progress (ISSN 2084-9257). Vol. 17 (2026). No. 1,
Art. #2, pp. 20-34. DOI:10.14746/eip.2026.1.2

Utilitarianism and Prison Abolition: A Normative Perspective

Marina Beraha

(Independent Researcher, Milan, Italy; marina.beraha@hotmail.it)
ORCID: 0000-0002-3516-1619

Abstract: This paper develops a utilitarian case for prison abolition grounded on utilitarianism's commitment to the reduction of pain, challenging the traditional forward-looking utilitarian justifications of punishment, whose empirical foundations are increasingly contested. To construct the utilitarian abolitionist argument, traces of an abolitionist theory are identified in Bentham's penal thought, especially in his neglected analysis of indirect legislation as a means of crime prevention. His theoretical analysis is integrated with contemporary debates on punishment, incarceration and restorative justice. The paper argues that a consistently utilitarian approach should prioritize harm reduction through social restructuring, transformative approaches, and restorative responses to wrongdoing, rather than relying on incarceration and deterrence-centred punishment.

Keywords: Abolition; Utilitarianism; Bentham; prison; punishment; justice; prevention.

I. Introduction

Utilitarianism has been linked to the emergence, affirmation, and justification of prison as the sole form of institutionalised punishment.[1] In this paper, I aim to demonstrate that a deeper understanding of utilitarianism and a correct reading of Bentham's works would instead support an abolitionist stance aimed at replacing punishment-centred systems with social, economic, and relational structures that prevent harm and address it without incarceration. My normative framework will be that of hedonistic act utilitarianism in the tradition of Jeremy Bentham, which interprets the principle of "the greatest happiness of the greatest number" in terms of maximising pleasure and minimising pain, evaluating actions on a case-by-case basis by their consequences for all affected.

Bentham's *Introduction to the Principles of Morals and Legislation* is often considered the foundational text of utilitarianism, presenting the first clear-cut formulation of the principle of utility. The book was intended to be an introduction to a penal code that Bentham eventually never wrote. Utilitarianism is so rooted in the 18th-century penal reform that the

1 An earlier exposition of some of the arguments presented here can be found in Chapter 3.3 of my book (Beraha 2025), of which this paper constitutes a revised, updated version.

principle of utility is presented in the *IPML* as a guide for the legislator, specifically with respect to punishment. Punishment, in fact, is conceived as the key instrument for the creation of an artificial harmony between the individual and the general interest, making socially harmful but personally advantageous actions unfavourable for the agent as well, since the prospect of the disadvantage inflicted by punishment would offset the expected personal advantage.

Moreover, Bentham himself designed the prison model that inspired many of the first prisons built after the penal reform, especially in the United States: the (in)famous panopticon (Bentham 1995), which is now the symbol of omnipresent oppressive power, of a quest for efficacy that leaves no room for humanity.

However, it should not be forgotten that he was coming from a system of punishment that relied on physical violence and was part of the movement that aimed at changing it. These efforts now seem utterly insufficient (and the panopticon a dystopian *dispositif*), but they were revolutionary at a time when torture, even during interrogation to ascertain culpability, and the death penalty were the norm: in this context, the panopticon could effectively diminish the violence perpetrated within the penal system. *Now* reminiscent of totalitarian techniques and practices, it was part of the process of softening penalties that had started with Beccaria and the other reformers.

Given this context, the idea of a utilitarian theory of prison abolition might seem less paradoxical: although utilitarianism is rooted in the penal reform that generated the expansion of the prison system and made punishment one of the first and central subjects of its reflection, it really aimed at making it less cruel, minimising the suffering involved. Utilitarianism would, ideally, aim at overcoming punishing practices.

II. Utilitarian Justifications of Punishment and Their Crisis

Nonetheless, utilitarianism has indeed shaped the contemporary debate on punishment in many ways, most notably by introducing its forward-looking justifications. In the *IPML*, Bentham mentioned the three contemporary consequentialist justifications for punishment: deterrence, incapacitation, and rehabilitation. Incapacitation and rehabilitation, however, are addressed as “properties of inferior importance” (Bentham 2015, 180) compared to the main aim of punishment, that is, general deterrence. Concerning incapacitation, Bentham speaks of the possible “disabling efficacy” of penalties, of which he claims:

> The inconvenience is, that this property is apt, in general, to run counter to that of frugality: there being, in most cases, no certain way of disabling a man from doing mischief, without, at the same time, disabling him, in a great measure, from doing good, either to himself or others (Bentham 2015, 181).

It is, therefore, discarded on utilitarian grounds.

Rehabilitation is addressed only in the form of the “reforming tendency” of punishment, which consists of the generation of aversion to the crime. Hence, it does not refer to the reintegration of the offender into the social web in a meaningful way, thereby preventing them from reoffending not out of fear of punishment but, like the great majority of their fellow citizens, because they recognise the legitimacy of the law (Tyler 2006). Bentham's reforming tendency corresponds instead to specific deterrence: the greater the punishment the greater the reforming effect. Again, obviously, this goes against the principle of frugality, which is a crucial constraint on punishment in the utilitarian quest for minimising suffering. Being reformation, just like incapacitation, a "minor property," it cannot outweigh frugality: it is, therefore, not legitimate to inflict disproportionate punishments on the basis of their supposed reformative effect (which, by the way, as Beccaria believed, has been disproven).

For Bentham, the central justification for punishment is general deterrence: punishment has beneficial consequences not primarily for those on whom it falls but for "such persons as are under temptation to commit it [the criminal act]: in which case, in as far as it tends to restrain them from committing such acts, it is of a beneficial nature" (Bentham 2015, 147).

The justification of general deterrence still plays a central role in the contemporary conception of punishment; however, it is commonly believed that utilitarian justifications, based primarily on deterrence, bear an intrinsic moral risk, namely the manipulation of the individual, and therefore must be coupled with the retributive justification in order to avoid it.[2] It is the age-old criticism of utilitarianism, according to which the principle of utility, if not complemented with other moral principles, would lead to the punishment of the innocent if the overall consequences were believed to be positive. This accusation is ironic because Bentham's aim in the *IPML* was precisely to carry forward the Enlightenment quest to reduce arbitrariness in the penal system by classifying crimes and punishments clearly, establishing proportional punishments, and instituting laws from which no *ad hoc* derogation could be made (Burns & Hart 2015, lxv).

Although he does not address the problem of the punishment of the innocent explicitly, it is clear that Bentham takes for granted that only the guilty should be punished and from his insistence on the importance of transparency, education, and shared knowledge, as well as from his anti-tyrannical stance, it is evident that he thinks that passing innocent people off as guilty to have someone to punish in virtue of the supposed deterrent effects of punishment would not be a wise strategy in the long run in terms of utility. This implies a far-reaching utilitarian view that aims to take into account indirect and more distant consequences rather than narrowly focusing on the direct and nearest

2 This, for example, is done by Hart when he distinguishes the General Justifying Aim of punishment from its distributive principle: we may have a utilitarian General Justifying Aim, but it must be coupled with a retributive distributive principle that only allows for the punishment of the guilty. Even when this principle of distribution is breached and the innocent is punished, the fact that the punishment of the innocent is perceived as worse than the punishment of the guilty cannot be explained in utilitarian terms, which means that we intuitively adopt a retributive distributive principle even when generally justifying punishment in utilitarian terms (Hart 2008, 8–13).

outcome: a practice that does, say, reduce crime through deterrence or incapacitation but that substantially contributes to the evolution of society in an autocratic and repressive direction cannot be justified on utilitarian grounds.

The same answer given by Bentham is offered by many present-day utilitarians in response to the same criticism, although without much success in dispelling the objection. Others, like J. J. C. Smart, decided to accept, albeit reluctantly, that the punishment of the innocent is an inescapable consequence of the application of the utilitarian principle. Of course, extreme examples in which the punishment of the innocent would eventually be justified can always be formulated, and I challenge any non-utilitarian thinker to say that the punishment of an innocent would not be acceptable if it could prevent, say, the explosion of an atomic bomb – only Kant might allow the bomb to go off in order to defend his categorical imperative. This demonstrates the hollowness of such criticism, for there is always a quantitative threshold people are not prepared to cross in order to uphold other principles: the fact that people who identify as utilitarians may present lower thresholds might just be the result of greater awareness of this fact.

This does not mean, as McCloskey claimed, that utilitarians are "happy" to punish the innocent (McCloskey 1963). Smart neatly replied to him:

> Even in my most utilitarian moods I am not *happy* about this consequence of utilitarianism. Nevertheless, however unhappy about it he may be, the utilitarian must admit that he draws the consequence that he might find himself in circumstances where he ought to be unjust.[3] Let us hope that this is a logical possibility and not a factual one. In hoping thus I am not being inconsistent with utilitarianism, since any injustice causes misery and so can be justified only as the lesser of two evils. The fewer the situations in which the utilitarian is forced to choose the lesser of two evils, the better he will be pleased (Smart 1973, 71).

However, it is true that an ethical theory based only on the principle of utility, one that does not admit retribution as a value, does not recognise an intrinsic difference between the punishment of the guilty and that of the innocent, except on an emotional level that moral, rational reflection should override. The unease generated by the possibility of justifying the punishment of the innocent should not be overlooked, but it should not be misunderstood either. It constitutes a warning sign of the problematic nature of the infliction of suffering on others: we should not be happy, as McCloskey suggested, about punishing the innocent – but we could be equally scandalised by the punishment of the guilty. Instead of accepting the innocent's punishment with revulsion, as Smart does, this aversion should push us to look for alternative approaches to the problem of crime management.

In addition, it must be noted that if Bentham is to be accused of any deceptive strategy in his approach to punishment, it would rather be in the opposite direction: not to make the public believe that an innocent was guilty in order to punish them, but that

3 The fact that taking the course of action that maximises utility is deemed by a fervent utilitarian like Smart as "unjust" is surprising. "Unfair" may be more accurate.

someone had been punished while, in truth, they had not been harmed – or had been harmed much less than the public was led to suppose. As the principal aim of punishment is general and not specific deterrence, punishment only needs to be *apparent*, a principle established by Bentham and reflected in Foucault's rule of "sufficient ideality" (Foucault 1995, 94). In Bentham's words:

> It is the idea only of the punishment (or, in other words, the *apparent* punishment) that really acts upon the mind; the punishment itself (the real punishment) acts not any farther than as giving rise to that idea. It is the apparent punishment, therefore, that does all the service, I mean in the way of example, which is the principal object (Bentham 2015, 179–180).

Despite all this, the accusation of justifying the innocent's punishment remains a prominent topic in discussions of utilitarian justifications for punishment (see, for example: Sandel 2009, 37–40; Ryberg 2011, 89).

To this theoretical criticism, practical problems have been adding up, as decades of research on the general and specific deterrent effects of incarceration, incapacitation and rehabilitation have yielded very discouraging results (Mosconi 2025, 39–51). Moreover, to justify a practice on utilitarian grounds[4] it is not sufficient to show that it has some positive impact: first, its positive effects need to be greater than the harms the practice itself may cause – harms that are not negligible when we consider the carceral system, especially in this era of mass incarceration. Suffice it to say that, as of 2023, almost 2 million people were incarcerated in the U.S. alone, a number that grows to 5.5 million if those under probation and parole are added to the incarcerated population (Wang 2023); it must also not be forgotten that the negative consequences of incarceration affect not only the individual, but also their families. The relatives of incarcerated people, in fact, don't only suffer from the pain of affective deprivation and the disappearance of a practical support element in their family life. They often have to face harsh economic consequences: the loss of a salary, frequently the main, if not the only, source of income for the family; the legal costs; the burden of time spent on bureaucratic requirements in criminal proceedings. Also, people are not rarely incarcerated far from their homes, forcing their relatives to undertake costly journeys to visit them – travels that can be incompatible with the family's economic situation and work schedule. When mothers are incarcerated, in most Western countries their infants and toddlers are sent to prison with them. Finally, social stigma is no less significant and is often underestimated: as an example, in January 2026 the parents of a man accused of killing his wife committed suicide as a consequence of the pressure of public scorn.[5]

4 As Herbert Hart and James Burns remark in their introduction to the *IPML*, Bentham, like Hume, may have believed that "there might not be a general justification of practices apart from existing conventions", which means that he was not aiming for a utilitarian general justification for punishment, a practice that was not put into question. More limitedly, in the *IPML* he is justifying the reform (Burns & Hart 2015, lxix).

5 See: https://roma.corriere.it/notizie/cronaca/26_gennaio_25/suicidi-i-genitori-di-carlomagno-in-auto-da-roma-per-morire-ad-anguillara-la-lettera-all-altro-figlio-e-il-nuovo-interrogatorio-

Additionally, a utilitarian justification must be shown to be the most economical in terms of suffering: again, it is not enough that more suffering is avoided than is created; the suffering that is generated must be as little as possible: if similar results in terms of crime prevention can be achieved with the infliction of less suffering, this alternative path must be taken. That this is the case for the carceral system is far from being demonstrated.

The disheartening data on deterrence, incapacitation and rehabilitation have led jurists to the common idea that "nothing works" (Duff 2011, 64; Martinson 1974; Ryberg 2011, 86) and that punishment cannot be justified on a utilitarian basis, for it does not have the positive consequences it was hoped it would have. Surprisingly, however, this has not led to questioning the practice of punishment itself. Rather, it has brought retributive justifications back, and they have been, if not dominant, at least central in legal philosophy since the Seventies (Ryberg 2011, 86). Philosophers have started defending them as more *humane* than the instrumentalising utilitarian justifications to the point that Herbert Morris, in his 1968 paper *Persons and Punishment*, advanced the Kantian argument that to punish an offender for retributive reasons is a form of respect towards them, that they even have a "right" to being punished (Morris 1968, 476). These scholarly positions fit well with the popular sentiment that demands that offenders receive "their just deserts."

One question, at this point, arises: if punishment was practised, both individually and institutionally, long before its utilitarian justifications, and it was justified in a retributive manner, and if the utilitarian justifications have been proved ineffective while the practice of punishment in itself has not been questioned, does this not mean that utilitarianism has served as a sort of cover, a satisfactory rationalisation used to hide the real, not-so-moral and not-so-rational reasons why we punish at all? Of the vengeful impulses that could not be accepted in contemporary bureaucratic states aimed at the rationalisation of public life? Utilitarians did provide some rational justifications for a pre-existing practice based on irrational impulses, but did not question the practice.

III. Utilitarian Abolition

III.1. Bentham's Abolitionist Stance

It would be unfair to say the same about Bentham himself. Like all utilitarians, he is unequivocal in underlining that the infliction of suffering is, in itself, always evil; he defines punishment as "mischief" – the same word that he uses when referring to offences (Bentham 2015, 180, 176) – and even refers to penal justice as "a train of evils":

> (…) evils in the threats and constraint of the law – evils in the pursuit of the accused, before the innocent can be distinguished from the guilty – evils in the infliction of judicial sentences – evils in the inevitable consequences which

1b33ad11-15ba-4b2a-94c7-aaafbd616xlk.shtml (accessed: February 27, 2026).

reverberate upon the innocent (Bentham 1962, 534).

Contrary to most utilitarians, however, Bentham dedicated lengthy analysis to the topic of crime prevention, not simply passively acknowledging that punishment, as the lesser of two evils, must be accepted and executed, but actively working to make it unnecessary, thus challenging the idea that the only alternative to it is the greater evil of leaving crime unpunished. He implicitly calls for shifting away from the binary logic that underlies the "lesser evil" kind of reasoning and taking into account the bigger picture. He does so in the *Principles of Penal Law*,[6] a text much less well known than the *IPML*, whose last edition dates back to John Bowring's 19th century collected works of Bentham.

Bentham analyses the topic in the third and last part of the *Principles*, titled *Of Indirect Means of Preventing Crimes* (Bentham 1962, 533). Whereas "direct legislation" concerns punishment and has been addressed in the previous part of the *Principles*, as well as in the *IPML*, indirect legislation covers the actions that can be taken to prevent crimes before any punishment is administered by acting on the factors that make people want to commit crime in the first place (Bentham 1962, 533). Once again, Bentham enters slippery ground, for sly critics might interpret his aim as one of manipulation, of totalitarian thought control, but before jumping to such conclusions, let us first analyse his proposal.

First of all, his recognition of the criminogenic influence of poverty, which is well substantiated by data on the correlation between poverty and offence rates, motivates the many pages Bentham dedicates to the so-called "Poor Laws,"[7] in which he advocates both welfare-state-type aid and employment support (Bentham 1962, 543–544).[8] Poverty aside, in the *Principles,* Bentham focuses primarily on education and knowledge sharing (Bentham 1962, 536–540, 563, 575–578), seeming to espouse a moderate ethical intellectualism: the positive consequences of ignorance, deception, and manipulation are only illusory in that they are limited – they may seem to provide the best outcome in a specific situation, but are negative contributions to the overall picture.

After Bentham, the utilitarian discussion on punishment has been chiefly

6 From now on: *Principles*. A hint to the importance of crime prevention as opposed to repression, however, is present in the *IPML*, too, when Bentham discusses "needless" punishments, which take place "where the purpose of putting an end to the practice may be attained as effectually at a cheaper rate: *by instruction*, for instance, as well as by terror: by informing the understanding, as well as by exercising an immediate influence on the will" (Bentham 2015, 164, my emphasis).

7 Other than the few pages dedicated to the topic in the *Principles*, Bentham's lengthy *Writings on the Poor Laws* are gathered in two volumes (2001 and 2010) of *The Collected Works of Jeremy Bentham*, the outcome of the Bentham Project's (UCL) ongoing effort to publish all of Bentham's writings.

8 Their proposed implementation sounds nowadays anachronistic and paternalistic at best, Bentham being a supporter of the workhouses. This reminds us that Bentham, despite being very progressive on many subjects, was still a man of his time and that *his* utilitarianism bears the marks of the Georgian era. This fact, however, is not a reason to reject utilitarianism, but to distinguish the principles of utilitarianism, which Bentham set out in his work in an exemplary manner, and the practical and political proposals that cannot always be transposed unmodified (and sometimes cannot be transposed at all) to the present day.

focused on deterrence, following solely Bentham's *IPML*. It looks like utilitarians have jumped on the bandwagon of punishment upon shared assumptions about efficacy, but if utilitarianism aims at maximising social utility, and specifically pleasure in its hedonistic variant, shouldn't the intentional infliction of suffering, i.e., punishment, be avoided as much as possible? It is sometimes said that punishment should be considered a last resort, *extrema ratio*, when everything else has proven ineffective. However, its effectiveness compared to the other ineffective measures must also be assessed, whereas *extrema ratio* arguments tend to take it for granted. Utilitarians, as Bentham has shown, should be the first to consider alternatives to punishment, but the path he had traced has remained unexplored in the utilitarian field. Bentham's position is not only consistent with present-day prison abolitionists' demands; it anticipates them. Abolitionists, on the other hand, have neglected engaging with a utilitarian moral framework. It is therefore time to examine the abolitionist stance and assess its compatibility with hedonistic act utilitarianism.

III.2. Prison Abolition: Roots and Limitations

The prison abolition movement emerged in recent decades, proposing a different paradigm, aiming at abolishing the institution of prison and developing alternative responses to harm. Incarceration is not equivalent to punishment; however, the carceral institution has occupied the entire space of institutionalised punishment in the Western world, so much so that we do not even consider that State-administered punishment might take other forms on the grounds that physical punishments (except for death sentences, notably and shamefully in some Western countries) are considered retrograde and barbaric. From this, it follows that the prison abolition movement corresponds to the punishment abolition movement: no prison abolitionist advocates for other kinds of institutionalised punishment.

Prison abolition was initially theorised by anarchist thinkers in the 19th century (for a comprehensive overview of anarchist thinking on prison abolition, see: Nocella, Seis & Shantz 2020). Anarchists, of course, based their abolitionist stances on their rejection of the State and hierarchical systems of power in general: the scope of their project was, as is clear, very broad. Present-day abolition thinking, which developed mainly in the U.S.,[9] has lost that wide breadth. As many abolitionist scholars aim at changing practices of justice, it is vital that their research tackle specific contexts and their peculiarities (e.g., McLeod 2019), and research with a wider approach cannot substitute for the valuable work they are doing. However, the two approaches should complement each other, and more abstract and theoretical work may ground hands-on research and proposals on a firmer foundation, also helping context-sensitive results expand to other settings.

9 A notable exception is European abolitionist sociology, embodied by Norwegian Nils Christie (e.g., Christie 1977) and Dutch Louk Hulsman (e.g., Hulsman 1991), whereas the subject has been completely neglected by philosophers.

More importantly, contemporary abolition thinking is largely intertwined with the history of discrimination in the U.S., specifically with respect to slavery and racism. The American prison system has, in fact, developed dramatically in the aftermath of the American Civil War, which resulted in the presumptive abolition of slavery, as the triad of criminalisation, incarceration, and forced labour became effectively a surrogate for slavery in many of the Southern States after its formal abolition, and conditions of black incarcerated people who were victims of the convict-leasing system were often far worse than they had been when slavery was in force (Davis 2003, 32). Angela Davis wrote:

> If we are already persuaded that racism should not be allowed to define the planet's future and if we can successfully argue that prisons are racist institutions, this may lead us to take seriously the prospect of declaring prisons obsolete (Davis 2003, 25).[10]

The work on the connection between the American carceral system and racism is essential, and the de-racialisation of crime is even more important. However, I believe that the broader prison abolition movement should primarily stand on different grounds. It is true that, almost everywhere, minorities are targets of police violence and face higher incarceration rates; however, the basic problem with incarceration does not lie in racism. Morally, we would need prison abolition even if incarceration rates perfectly mirrored the ethnic, religious, and financial distribution of the general population: the fundamental problem with the punitive system is not discrimination; it is punishment itself. To limit abolitionist thinking to the critique of the racist aspects of the American system leaves intact the punitive structure in its most basic feature, that of the belief that some good is achieved through the voluntary infliction of suffering, whether the backwards-looking good of retribution or the forward-looking good of utility. Present-day abolition thinking lacks universal breadth.

III.3. Framing Abolition: Utilitarianism as a Driver of Moral Progress

It should be clear by now that I think hedonistic act utilitarianism offers here the perfect solution. Utilitarianism has been so closely linked with the justifications for punishment for accidental reasons rather than because of any intrinsic punitive character of the normative framework: the first consequentialist justifications for punishment were elaborated in a specific context, that of the reform that led to the overcoming of the torture system and the softening of penalties. Moreover, Bentham's justifications for punishment were accompanied by a fervent critique of the practice of punishment itself, the recognition that punishment is the mark of a system's failure and that action should primarily be taken through indirect legislation so that people would not commit crimes in the first place.

This latter aspect of Bentham's thought has been remarkably neglected in the

10 Similar positions can be found in many abolitionist thinkers (e.g., Bell 2021, 45, 61; McLeod 2019, 1617; Law 2021, 49–55).

subsequent utilitarian discussion on crime and punishment. This could be partly explained by the *Principles*' publishing vicissitudes, whereas the *IPML*, which treats direct legislation and punishment justifications at length, has established the traditional bond between utilitarian thought and punishment. However, Bentham's justifications have probably been so successful for three other less accidental reasons: first, because they could fit into a system, the retributive system of punishment, that was already in place; second, because, as Bentham himself remarked, punishment is a simple, intuitive solution (Guyau even called it a "popular remed[y], like boiling oil into which (…) the limbs of the wounded were plunged"; [Guyau 1891, 172]) – though the problem it addresses is most complex, hence the utter failure of simple solutions; and, finally, because it complies with deeply-rooted emotional impulses, validating our feelings more than our rational analysis.

So, utilitarianism has been a driver of sentence softening, and many of Bentham's arguments are plainly compatible with – if not entirely in support of – prison abolition. Most importantly, however, utilitarianism offers precisely what abolition thinking currently lacks: universal breadth. The utility principle is inherently universal in scope and has already served as a powerful catalyst for moral inclusivity and the expansion of the moral circle. This was palpable in Bentham's writings, in his famous stances against the criminalisation of homosexuality (Bentham & Crompton 1978), opposing slavery, racism, and even the abuses perpetrated against non-human animals, making him, to all intents and purposes, the father of contemporary antispeciesism (Bentham 2015, 283, n. b). Utilitarian ethics is based on sentience, a characteristic that we share among varying ethnicities and species, as well as across different life choices and pathways. The suffering of people who committed crimes has, from this perspective, the same value as that of the innocent.

This universalism allows for a reformulation of abolition in broader terms. Moreover, hedonistic utilitarianism, which grounds morality in the promotion of pleasure and the prevention of suffering, represents the logical normative theory within which to frame abolition. Utilitarianism, as we have seen, has been focused on crime prevention through deterrence, conceiving punishment as a necessary lesser evil. On the other hand, abolition thinking questions the idea that this evil really is necessary at all and aims at harm prevention through a restructuring of society that reduces the conditions that lead to violence, just like Bentham himself.

This revolutionary role is one that utilitarians have been afraid to take on. Mostly, utilitarians have emphasised the continuity between utilitarianism and common-sense moral rules, showing that adopting the utilitarian perspective would not lead to an upheaval of ordinary morality. This is in part due to the fact that an implicit consequentialist element that weighs the benefits and burdens of a choice is already present, in most cases, in common sense ethics. This is not to say that people do not tend to think about their moral duties in deontological terms, in part because it is easier to commit to definite and supposedly unquestionable rules rather than engage in difficult and always uncertain calculations that need to adapt to specific situations. However, when faced with examples where the positive consequence is evident, as in the aforementioned case of killing an innocent person to prevent the explosion of a nuclear bomb, common sense morality becomes consequentialist.

The utilitarian *felicific calculus* is undoubtedly a critical element for utilitarianism, for calculating precise utilities in concrete situations whose outcomes are ultimately unpredictable is impossible. This has led some authors to endorse rule utilitarianism, which involves the formulation of moral rules according to the utilitarian principle and then following them even when a bad outcome is foreseen for a specific action. However, not allowing exceptions – as in the nuclear bomb example – would entail a form of rule worship that is fundamentally incompatible with utilitarianism; on the other hand, if exceptions are allowed, then rule utilitarianism is reduced to act utilitarianism, for the rule is only followed as long as following it is expected to generate the best consequences.

Common sense morality may, hence, be a kind of lazy utilitarianism, consisting in following rules that are expected to have positive overall consequences, albeit on the basis of intuitions and custom rather than the meticulous analysis required by act utilitarianism, and allowing exceptions in cases where there is clear-cut evidence that waiving the rule will produce positive outcomes. This reading of common-sense morality explains why, in most cases, utilitarians can claim that utilitarianism would mostly respect the moral stances of common sense.

Utilitarians have also been concerned with emphasising the continuity between their normative framework and common sense because they had to counter the dystopian accusation of their critics. The utilitarian ideal world, in fact, is often depicted as a world of "secrecy, deception and manipulation" (Primoratz 1991, 390), as Primoratz put it. These accusations are not groundless. Famously, indeed, one of the founding fathers of utilitarianism, Henry Sidgwick, did defend what is now called "esoteric morality", a morality that allows and even encourages duplicity in some cases (Sidgwick 1907, 489). This view has also been defended by contemporary utilitarians such as Peter Singer and Katarzyna de Lazari-Radek, who maintained that, when shown to have better consequences than honesty, duplicity is morally required (de Lazari-Radek & Singer 2010, 47). However, again, let us think of the nuclear bomb case: if dishonesty was needed in order to avert the explosion, wouldn't it be widely considered morally acceptable to lie and deceive? We face another case where a utilitarian approach might well support a general rule – that of avoiding duplicity and secrecy and, as Bentham recommended, promoting education and knowledge sharing – that might nonetheless be broken in (rare and extreme) clear-cut cases.

A final critical point that helps explain utilitarians' tendency to conservativism – and the general population's wariness towards utilitarianism – is the fact that, in order to promote the utility of the greatest number, one needs to believe that they know where that utility may lie. This assumption generates some legitimate vigilance, for it quickly evokes ethical states that intrude into their citizens' lives to impose moral values and impede self-determination. Again, this accusation is paradoxical since, contrary to deontological morality, which establishes moral rules on intuitive grounds and raises them to the status of objectivity, therefore justifying their imposition on those who have different intuitions, utilitarianism is based on respect for individual preferences, on Bentham's idea that every person is the best judge of their own interests. If the "central point of utilitarianism", as Jan Narveson put it, is "not to project our values indiscriminately on other people" (Narveson 2019, 159), then to think of it as a moral framework that promotes imposition and manipulation means completely misunderstanding it.

Nevertheless, the need to know the other's utility is a problem for utilitarians and has, once again, pushed them to defend a conservative approach: just like it is better, in general, to follow a commonsensical rule instead of calculating the precise utility of a specific action, for the risks of error are too high, so we'd better support traditional liberal stances rather than allow intrusion into people's choices on the assumption that policymakers know their preferences and utility.

Utilitarian policymaking seems to be the most problematic: if one is guided by the aim of maximising the utility of the greatest number, some understanding of the others' utility must be presumed. In this case, negative hedonistic utilitarianism – aiming at reducing pain rather than maximising pleasure – may constitute the preferable framework, being less intrusive into people's lives, hence less dangerous from the liberal point of view, without losing too much of utilitarianism's aspiration of rationality and universality.

The famous critique levelled against negative utilitarianism by Roderick N. Smart, the author who coined the very term "negative utilitarianism" in 1958, is that if one does not give weight to positive experience, then the preferred utilitarian outcome would be extinction (Smart 1958). However, I only defend negative utilitarianism as the safest normative framework for policymaking, as another rule of thumb that may be waived but serves as the most efficient guiding principle in a specific context, avoiding the slippery slope of ethical states. If the State's objective is not to positively maximise utility but to negatively minimise disutility, the risk of intrusion into private life and personal choices is reduced, as the State's action tends rather to alleviate hardships – which are more easily identifiable than utilities – thus supporting the welfare state and social assistance. And if the State's aim is to minimise suffering, it is evidently paradoxical that it should voluntarily inflict suffering on its citizens with a punitive system.

Caution in all these matters must not lead to underestimating the revolutionary power of a normative theory based on a principle that can be tested on empirical data. A moral framework of this kind calls to question institutions and practices on the basis of an analysis of reality and has an intrinsic inclination to change, for the desirable configuration does not depend on respect of fixed rules but on the concrete experience of the actors involved. The fact that a weak form of rule utilitarianism that allows for exceptions and does not end up in rule worship is compatible with the workings of common-sense morality, and that the content of these rules may frequently remain unchanged under a utilitarian perspective, does not mean that a rigorous analysis of real-life cases would not lead to the formulation of different rules, which is precisely what I have aimed at doing in this work. Utilitarians have been focused on rejecting the dystopian accusations that we have seen but have thus depowered utilitarianism itself, transforming it into a justification, rather than a challenge, to the status quo.

With respect to incarceration and punishment, the basic intuition underlying them is that justice is achieved through the imposition of suffering on those who harm us. Utilitarians have elaborated alternative justifications to support this intuition, but utilitarian ethics requires the exploration of alternative systems – an exploration that can be conducted on the basis of Bentham's own suggestions. Here, the alliance with the prison abolition movement comes into play. The question that looms over abolition, once indirect legislation to reduce crimes is put in place, of course is: what do we do with offenders? However successful indirect legislation might be in preventing offences, the prospect of eradicating all forms and occasions of violence is clearly unrealistic. The issue is broad and complex; I have explored it in another publication (Beraha 2025) and will therefore confine my concluding remarks to outlining the main direction for addressing it.

IV. Conclusions and Future Directions: The Restorative Paradigm

Whereas the carceral solution to the problem of violence, which is a major form of social bond breakdown, further accentuates the disconnection of the offender from society (in which they are supposed to be, eventually, reintegrated), an abolitionist proposal should aim at social connection. Restorative justice has emerged, since the Truth and Reconciliation Commission in post-apartheid South Africa, as the principal tool for responding to harm, respecting the survivor's need for truth, understanding, and reparation (which, by the way, are not respected at all in traditional criminal prosecution) while also not persecuting the offender.

Restorative processes are based on the acceptance that the people involved in the harmful situation hold different and legitimate narrative truths about the event, which forms the basis for investigating the conditions that gave rise to it, all within a dialogical process that, by its very nature, creates new forms of social bonds (see, e.g., Mazzucato 2015; Ceretti 2015; Mosconi 2025, 181–210). Restorative justice maintains the symbolic and expressive aspect of punishment, constituting a form of institutionalised response to harm that signals censure of the act and attention to the offended persons' needs, while rejecting the idea that these aims can only be achieved by intentionally inflicting pain on the offender, therefore radicalising the utilitarian principle of frugality: no punishment is needed for general deterrence. Its details, its supposed deterrent effect, its transformative potential and its overall positive consequences, however, must be left to further investigation.

References

Bell M. 2021. "Abolition: A New Paradigm for Reform," *Law & Social Inquiry* 46(1):32–68. https://doi.org/10.1017/lsi.2020.21

Bentham J. 1962. "Principles of Penal Law," in J. Bowring (Ed.), *The Works of Jeremy Bentham* (pp. 365–358). New York: Russell & Russell.

Bentham J. 1995. *The Panopticon Writings*. Edited by Miran Bozovic. London and New York: Verso.

Bentham J. 2015. *An Introduction to the Principles of Morals and Legislation*. Edited by J. H. Burns and H. L. A. Hart. Oxford: Oxford University Press.

Bentham J. & Crompton L. 1978. "Offences Against One's Self," *Journal of Homosexuality* 3(4):389–406. https://doi.org/10.1300/J082v03n04_07

Beraha M. 2025. *Ripensare la giustizia. Oltre il modello retributivo*. Roma: Carocci.

Burns James H. & Hart H. L. A. 2015. "Introduction," in H. L. A. Hart & J. H. Burns (Eds.), *An Introduction to the Principles of Morals and Legislation* (pp. xxxi–lxix). Oxford: Oxford University Press.

Ceretti A. 2015. "Per una convergenza di sguardi. I nostri tragitti e quelli della commissione per la verità e la riconciliazione sudafricana", in G. Bertagna, A. Ceretti, & C. Mazzucato (Eds.), *Il libro dell'incontro. Vittime e responsabili della lotta armata a confronto* (pp. 219–250). Milano: Il Saggiatore.

Christie N. 1977. "Conflicts as Property," *The British Journal of Criminology* 17(1):1–15. https://doi.org/10.1093/oxfordjournals.bjc.a046783

Davis A. Y. 2003. *Are Prisons Obsolete?* New York: Seven Stories Press.

Duff R. A. 2011. "Responsibility, Restoration, and Retribution," in M. Tonry (Ed.), *Retributivism Has a Past: Has it a Future?* (pp. 64–85). Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199798278.003.0004

Foucault M. 1995. *Discipline and Punish. The Birth of the Prison*. New York: Vintage Books.

Guyau J.-M. 1891. *A Sketch of Morality Independent of Obligation and Sanction*. London: Watts & Co.

Hart H. L. A. 2008. *Punishment and Responsibility. Essays in the Philosophy of Law*. New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199534777.001.0001

Hulsman L. 1991. "The Abolitionist Case: Alternative Crime Policies," *Israel Law Review* 25(3–4):681–709. https://doi.org/10.1017/S0021223700010694

Law V. 2021. *"Prisons Make Us Safer" And 20 Other Myths About Mass Incarceration*. Boston: Beacon Press.

Lazari-Radek K. de & Singer P. 2010. "Secrecy in Consequentialism: A Defence of Esoteric Morality," *Ratio* 23(1):34–58. https://doi.org/10.1111/j.1467-9329.2009.00449.x

Martinson R. 1974. "What Works? Questions and Answers about Prison Reform," *The Public Interest* 35:22–54. URL: https://www.nationalaffairs.com/public_interest/detail/what-works-questions-and-answers-about-prison-reform (accessed on February 27, 2026).

Mazzucato C. 2015. "La giustizia dell'incontro. Il contributo della giustizia riparativa al dialogo tra responsabili e vittime della lotta armata", in G. Bertagna, A. Ceretti & C. Mazzucato (Eds.), *Il libro dell'incontro. Vittime e responsabili della lotta armata a confronto* (pp. 251–304). Milano: Il Saggiatore.

McCloskey H. J. 1963. "A Note on Utilitarian Punishment," *Mind* 72(288):599. https://doi.org/10.1093/mind/LXXII.288.599

McLeod A. M. 2019. "Envisioning Abolition Democracy," *Harvard Law Review* 132(6):1613–1649.

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

Mosconi G. 2025. *Decostruire la pena. Per una proposta abolizionista*. Milano: Meltemi.

Narveson J. 2019. *Morality and Utility*. Baltimore: Johns Hopkins University Press.

Nocella A. J. II, Seis M., & Shantz J. 2020. *Classic Writings in Anarchist Criminology: A Historical Dismantling of Punishment and Domination*. Chico: AK Press.

Primoratz I. 1991. "Utilitarianism, Justice and Punishment: Comments on Smart and Flew," *Israel Law Review* 25(3–4):388–397. https://doi.org/10.1017/S0021223700010499

Ryberg J. 2011. "Punishment and Desert-Adjusted Utilitarianism," in M. Tonry (Ed.), *Retributivism Has a Past: Has It a Future?* (pp. 86–100). Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199798278.003.0005

Sandel M. 2009. *Justice. What's the Right Thing to Do?* London: Penguin Books.

Schofield P. 2019. "«The First Steps Rightly Directed in the Track of Legislation»: Jeremy Bentham on Cesare Beccaria's Essay on Crimes and Punishments," *Diciottesimo Secolo* 4:65–74.

Sidgwick H. 1907. *The Methods of Ethics*. London: MacMillan.

Smart J. J. C. 1973. "An Outline of a System of Utilitarian Ethics," in J. J. C. Smart & B. Williams, *Utilitarianism. For and Against* (pp. 1–74). Cambridge – New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511840852.001

Smart R. N. 1958. "Negative Utilitarianism," *Mind* 67(268):542–543. https://doi.org/10.1093/mind/LXVII.268.542

Tyler T. 2006. *Why People Obey the Law*. Princeton: Princeton University Press.

Wang L. 2023. "Punishment Beyond Prisons: Incarceration and Supervision by State," *Prison Policy Initiative*, May 2023. URL: https://www.prisonpolicy.org/reports/correctionalcontrol2023.html (accessed on February 27, 2026).

Ethics in Progress (ISSN 2084-9257). Vol. 17 (2026). No. 1,
Art. #3, pp. 35-50. DOI:10.14746/eip.2026.1.3

Beyond Retribution: Walter Benjamin and the Case for Prospective Justice

María Fernanda Miranda González

(Universitat Autònoma de Barcelona, Barcelona, Spain; MariaFernanda.Miranda@autonomia.cat)
ORCID: 0009-0001-0318-5122

Abstract: The dominant paradigm of justice in Western thought is retributive, framed through an economic logic of debt, guilt, and proportional punishment. Retributive justice promises balance but perpetually postpones its delivery. This article argues that this deferral is not a failure but a structural feature: justice conceived as debt can never be settled in the present. This holds an analogy with Christian eschatological temporality, reproduced in secularized juridical forms. The deferral is actively maintained by what Walter Benjamin identified as "mythical violence," that is, the foundational and preservative force of law that reduces life to administrable "mere life" and functions as a freezer of political time. Through a Benjaminian lens, the article critiques this apparatus and posits an alternative: prospective justice. Reconceived not as a future ideal but as a messianic intervention in the "now-time" [Jetztzeit], prospective justice involves the collective redistribution of agency to repair harm, the restoration of inviolable human dignity, and the use of a "divine violence" directed not at individuals but at the systemic structures of legal domination. The conclusion asserts that true justice requires a shift from inhabiting the frozen time of debt to seizing the charged possibility of the Jetztzeit.

Keywords: Justice; Benjamin; violence; law; collective agency.

I. Introduction

The ideological foundation of the classical notion of justice consists of a principle of retribution. There is a conceptual structure that sustains justice, and that is deeply rooted in our moral sensibility and our legal imaginary, which is expressed in the language of debt, guilt, and compensation. Many modern juridical intuitions point out that to do justice, where harm has been committed, a kind of payment needs to be demanded from a wrongdoer. Justice consists, it seems, in the administration of positive and negative retributive experiences, following a principle of proportionality. Under this "fair calculation" harmful events become seemingly controllable, as imputable acts tied to culpable subjects. Thus, individuals function as adequate causers that are capable of generating moral debts with each other through their actions, and have the right to either demand what they are owed, or are obligated to pay their debts.

Within the retributive frame, punishment appears as an essential component for achieving justice, because it is considered a legitimate currency to obtain a cleared balance. If we examine theories that attempt to justify punishment within legal governance systems, this becomes more evident. Even where some approaches have adopted elements of other theories that assign punishment just an instrumental value, subordinated to prevention or reintegration ends, the ultimate moral assessment remains retributive. Citizens' sense of justice is satisfied when they perceive that criminals "pay" for their crime. Any measure considered lenient is attacked for violating this principle. Even the very language of criminal law is retributive: it speaks of "paying a debt to society", of the punishment being "proportional" to the crime, of "guilt" and "merit" of the punishment.

This economic-juridical grammar of guilt and settlement invites a genealogical examination, yet this is not my purpose. My approach is rather a contrastive conceptual analysis that places two paradigms of justice in tension: the retributive logic embedded in modern legal thought and theological temporality on one hand, and the Benjaminian categories of *Jetztzeit*, and divine violence on the other. Walter Benjamin serves as the gravitational center of this article because his thought, nourished by Jewish messianism and a materialist critique of legal ideology, allows us to break with the very assumptions that sustain retributive justice.

This paper contends that the retributive model is constitutively unable to deliver justice in the present. Its promise is perpetually delayed because it is sustained by what Benjamin termed "mythical violence"; a self-perpetuating legal apparatus that stops time, reduces existence to its manageable elements, and prioritizes systemic preservation over concrete reparation. Against this impasse, I develop a Benjaminian notion of prospective justice. This is not a justice for a distant future, but a radically present outbreak: a community-based practice that interrupts the cyclical violence of the law by redistributing agency and following the imperative principle of human dignity.

Although the paper suggests a parallelism between Christian eschatology and modern juridical forms, this does not imply a claim about a direct or linear historical derivation from one to another. The argument rather concerns the persistence of homologous structures of intelligibility: configurations of temporality, debt, guilt, and deferred redemption that continue to shape juridical imagination in secularized forms. In this sense, the article does not reconstruct a genealogy of influences, but identifies paradigmatic constellations whose logic survives across heterogeneous historical formations. Benjamin's own notion of "mythical violence" should be understood in this way. His reference to Greek myth does not suggest that modern law historically originates in archaic religion, but that both share a structural logic of law-founding and law-preserving violence, of guilt administration, border sacralization, and order reproduction. What persists is not a continuous historical lineage, but a recurring configuration through which violence legitimizes itself as juridical order. Such an approach follows Benjamin's method of historical constellations and non-linear correspondences rather than a simple causal history of ideas.

To make this case, I first (II) dissect the retributive logic and its theological-temporal underpinnings. I then (III) analyze Benjamin's concept of mythical violence as the hidden engine of legal order. Finally (IV), I articulate the contours of a prospective justice, linked to Benjamin's "divine violence", as a form of collective healing that operates in the *Jetztzeit* to redeem past and future in a transformative present.

II. Retribution and Time

II.1. Balance and Individual Punishment

To understand the temporal structure of retributive justice, we must begin where it reaches one of its most rigorous formulations: the work of Immanuel Kant. Kant is not merely a theorist of punishment, he is the philosophical voice of legal modernity itself, as he articulates a vision of justice as rational calculation, individual accountability, and submission to universal law. Kant is the philosopher who purges punishment of all utilitarian function – rehabilitation, deterrence, social utility – and reduces it to a pure principle of desert. This drastic purity makes Kant the unavoidable interlocutor for any grounded critique of retributivism, because it is in his system that the logic of debt, individual imputation, and deferred balance appears most clearly, and thus most vulnerably, exposed. Benjamin, that critiques modern legal reasoning disguised as progress, finds in Kant's categorical framework his necessary bull's-eye.

Kant has a retributive sense of justice at heart when he says that individual punishment of criminals must not be used merely as a means to other ends, neither for the wrongdoer's nor for society's greater good. Justice is understood by him as a principle of talionic equality that can be only achieved by penal legislation. An undeserved harm, Kant thinks, is like an imbalance of the justice scale that needs to be straightened back by a tribunal. But there is something important to notice in this line of Kantian thinking: the State must ensure proportional retributive harm as a response to crime not just to provide a social sensation of fairness, but most importantly because, as he writes, "only by the law itself can dishonor (the stain of punishment) be wiped out" (Kant 1996, 142). That Kantian justice is a matter of retaliation is especially evident in the following passage:

> Even if a civil society were to dissolve itself by common agreement (for example, if the people inhabiting an island decided to separate and disperse throughout the world), the last murderer remaining in prison would first have to be executed, so that each person receives what his deeds deserve and blood guilt does not cling to the people for not having insisted on this punishment (Kant 1996, 142).

This explains why one of the conditions of possibility for the juridical doing of justice is being able to identify a proper causer for unlawful events. As evident as this premise may sound, we must stress since this point onwards that the notion of retributive justice functions on the individual imputation basis. This individualist lens not only shapes the response to harm but also its very perception: it directs our gaze towards the singular wrongdoer and away from the general, social net that plays a great part in the equation of the wrongdoing. We have become so familiar with this practice that even if we are inclined to say that some occurrences in our experiences are undeservedly harmful (e.g., the destruction that a hurricane produces), we cannot demand *justice* from a situation that has not an adequate and punishable culpable. The harm needs to be imputed to an agent that, by action or omission, would have generated some liability; someone must pay, and her unpleasantness is somehow a reassuring payment.

This model of retributive justice seems to be apt for guaranteeing the fair collection of debt of wrongdoers, and therefore a sense of order and equilibrium in our societies, but a question about temporality remains. When can this ideal of justice be achieved by our current juridical mechanisms? Experience tells us that our legal systems have not yet attained a balanced and expurgated state of affairs, that justice has not been properly done. This is not only a question of bureaucracy; the issue goes beyond the concrete protocols and institutions. Why does it seem that justice never arrived for many victims in the past, and continues to not be arriving for the present generations? The premise that I would develop next is that this model of retributive justice is not constructed to deliver justice in the present. In no one's present.

II.2. Christian Temporality

We must bring to the surface the theological dimension of this subject, as it will show the tight bonds that our sense of justice maintains with a specific understanding of temporality. From the Christian teleology point of view, the present moment is not occurring in the habitual chronologic sense of time. Temporality would have suffered a major break in its lineal course at the time that Christ incarnated in Earth. This entails a partial stop in the sequence of time. It is like if that moment would have striated itself to an elongated epilogue before the second coming of Christ. This introduces a juridical structure into temporality, as this awaited event is also the Final Judgment of every human being. What follows after is the cessation of linear time and the beginning of eternal punishment or reward.

History, from a Christian point of view, has a clear teleology, within which the stretched present time is the last chance for humanity to follow God's commandments and earn a place in the Kingdom of Heavens. From this perspective justice is but a promise, unachievable here and now, and strongly centered on individual accountability. Although the destiny of humanity is at stake, each one must look after their own salvation. The ultimate righteous balance of accounts can then only be integrally achieved at the end of times and will be somewhat imperfect in the prelude-world.

This Christian-juridical structure of time is not, of course, Benjamin's own position. It is rather the unacknowledged theological blueprint of the retributive logic that his work seeks to dismantle. Against this economy of deferred debt, Benjamin mobilizes a different inheritance: that of Jewish messianism. Where Christian eschatology postpones redemption to an end of time, the messianic tradition insists that every instant may be the one in which justice irrupts. The "now-time" that Benjamin retrieves from this tradition is not a waiting room but a field of interruption, a moment in which the legal continuum can be shattered without founding a new one. This is not to claim that Benjamin himself would have framed his project as a direct response to the Christian paradigm I have sketched. It is rather to argue that his notion of divine violence, and the prospective justice it enables, becomes intelligible when set against the temporality of deferral that he implicitly rejects. His proposal is a different temporality of justice altogether.

We can, nevertheless, object that our modern notion of retributive justice does not share this Christian background. But the secularization of the issue rearticulates basic Christian assumptions. When we speak about injustice in general terms, about an event, a gesture, an undeserved occurrence, we tacitly allude to a sort of hypothetical transcendental record – or scale, if you wish – that registers offenses and damages *that must be cleared*. This is not a mere reverie, but the fundamental assumption that moral actions are being always assorted through a quasi-providential invisible system. Because of this, we do not need to always share a concrete normative frame or scale to qualify something as unfair. The notion of justice that articulates these assumptions suggests that there will be a future compensation or loss, since everything is being balanced.[1] This is the principle on which we base the order and regulation we assume our moral and juridical systems warrant.

Thus, when we assert a value judgment about something being unjust, we are not just saying that the particular issue is wrong, but that if it does not generate charges, it is unacceptable, that it should not exist as such. Therefore, the violence that a "just" penal collection entails is considered perfectly adequate, but not the unsanctioned type of violence that threatens the very functioning of the systemic record. However, the sole penal calculation is in itself an unfair procedure. In fact, for unlawful events to be registered into such record, a lot of non-quantifiable variables must be compromised in their complexity. Considering such events integrally, with their own weight and value, can prevent offenses being reduced to quantifiable individual faults. Retributive justice functions as an abstraction in that sense; as a mathematical operation that aims to end in zero.

The result has not been the effective reduction of crime, here and now, because

1 For this reason, the notion of *karma* is also part of our common secular vocabulary; it expresses the intuition that misfortunes are not random or contingent events, but are intimately linked to their author and generate credit. Giorgio Agamben, in *Karman* (2018), traces an etymological and conceptual link between the Buddhist notion of karma and the logic of imputation in archaic Roman law. His point is not to establish a direct historical derivation, but to identify a common conditional structure: "if this exists, then that exists" (Agamben 2018, 27). According to Agamben, both karmic causality and juridical guilt presuppose the existence of imputable actions that generate consequences across time, linking conduct, debt, punishment, and redemption within a shared ethical grammar (Agamben 2018, 28). What matters here is not the validity of a unified "Indo-European ethics," but the persistence of a model of justice grounded in calculability, deferred balance, and the administration of guilt.

such abstraction permits at best a calculated justification of concrete penalties. It is not meant to address the broader causes that continue to cause strangling living conditions, but only to condemn errors. Focusing on the guilty individual dehistoricizes harm. It excludes from the realm of justice the systemic structures – racism, capitalism, patriarchy – that produce the conditions for delinquency. The abstraction is also political: it reduces a multiform social problem to an individual moral failing. This sense of justice cannot reach the general social, political, and economic conditions to shape dignified human lives. But to think justice beyond individual prosecution may feel incorrect to our basic intuitions. As if there were a quasi-holy status that the principle of authorized punitive reparation held.

The Christian resonance is noticeable not just by the fact that we think of justice in this transcendental and sacred terms, but also because we assume that proper justice takes time. Retributive traditions tend to postulate the fairness of the final balance, so any sense of urgency is neutralized. This echoes Christian conceptions of time and redemption. But if retributive justice is this postponed, abstract, and individualizing blaming machine, what force actively maintains it in the present, preventing any other justice from emerging? The promise of a cleared score as the regulative ideal of justice is, moreover, sustained by a specific mechanism of law, which is called *mythical violence* by Walter Benjamin. This is the apparatus that establishes and maintains the legal "order" by deferring justice from the present moment.

III. Mythical Violence and the Civilizational Iceberg

That law depends on violence is broadly discussed in philosophical debates. A blueprint for a sovereign power that wields force to guarantee social order is referred by Thomas Hobbes when he writes that covenants "without the sword, are but words, and of no strength to secure a man at all" (Hobbes 1994, 106). Max Weber, on his part, articulates a modern definition of the state, which sees its core in the "monopoly of the legitimate use of physical force" (Weber 1946, 78). However, this view often treats violence as a neutral tool; a regrettable but necessary means to the end of legal order and security. Precisely this hypothesis – that law uses violence in a fair, adequate way – is what permits us to differentiate between just punishment and vengeance. We assume that maltreatment can adopt a civilized form if it is a legitimate authority who exercises it in order to guarantee an orderly state of affairs.

Walter Benjamin's critique begins precisely by challenging this instrumental view. In his *Critique of Violence*, he argues that mythical violence is not secondary but essential to law. And to explain this thesis, Benjamin uses the myth of Niobe, which illustrates the constitutional core of law. The myth tells the story of a proud queen who boasted of her 14 children, claiming she was more blessed than the goddess Leto, mother of Apollo and Artemis. Angered by this hubris, Leto sent her children to slay all of Niobe's progenies with their arrows. Overwhelmed by grief, Niobe fled to Mount Sipylus in Asia Minor, where the gods turned her into a weeping rock.

This mythical violence mirrors, Benjamin thinks, the type of violence that produces and preserves modern law as a governance apparatus. Let us analyze it: Niobe challenges the border, the established order where everything has its place – inferior humans and higher gods –. Niobe's pride did not simply break a commandment but defied the cosmic arrangement of things that law protects. Nevertheless, the violence displayed in her punishment was not meant to neutralize her as a threat to harmony and peace, by vanishing her from earth. The aim was to teach a lesson, not just to her, but to everyone, on the inviolability of the borders. By challenging the lawful order, Niobe challenged law itself. And this is why she remained forever as a weeping statue, a crystallized warning, a monument to the scope of legal force.

If we are talking about vengeful gods or perverse dictators, we may be able to see the analogies, but we believe in the legitimacy and fairness of legal systems that operate in contemporary democracies. That belief, nevertheless, is but the tip of an iceberg, which shows only the manifest institutional role that we assign to such systems. The typical philosophical concern about this issue is the question of how they can function better. But if we delve deeper into the anatomy of the iceberg, we can realize the fitness of Benjamin's analogy. In the middle section of the body of ice, we find mythical violence supporting institutional procedure. This sort of violence creates a circular, self-perpetuating system where law posits itself as both the sole means and the ultimate end; this is how its authority is produced. Civil order is not law's greater end, but its position to govern.

In law's own interest most of its mechanisms are concealed, under water. Every offense to the precepts must be reframed as condemnable insurrection to society structure. Law's governance rationality rooted in violence remains unchallenged as a state apparatus because is well-founded by the belief that norm, in its modern institutionalized form, is essential to civilization. Additionally, since we have the power to elect our rulers, the system is considered fair. But if we go even deeper to explore the base of the iceberg, we will encounter Benjamin's indications once more, this time from his book *On the concept of history* (Benjamin 2003). Benjamin reminds us, in a famous formulation, that "there is no document of civilization which is not at the same time a document of barbarism" (Benjamin 2003, 392). That is, that the foundation of the mythical system as we know it – as the only alternative to chaos – consists of terror and a state of exception; it is not its contradiction but its beginning.

This is what we find at the innermost part of the mass of ice: that mythical power arises from a founding act of violence. Its legitimacy is not natural but historical, an achievement of imposition. We can affirm, in line with Benjamin's materialist conception of history, that this founding violence is not an anomaly but the origin of every legal order, including the democratic state. There is somewhat a passive submission to that power in the sense of recognition by default, by acquiescence, by lack of alternative. But it is not an active and reflexive consent. Once recognized, its violence is legitimate because it is exercised from that recognition. If the imposition that commences lawful power in our democracies (and not simply a certain administration or political orientation), were to exhibit this raw core its legitimate authority would weaken, it would be naked frontal aggression, yearning for domination.

One of the fundamental traits of mythical violence is, thus, its hidden imposition as self-legitimation: by creating and conserving jurisdiction through violence, it articulates itself as a normalized governance system. This is violence's potential to create law that Benjamin specifies: "Lawmaking is power making, assumption of power, and to that extent, an immediate manifestation of violence" (Benjamin 1996, 248). The "power making" is the hidden base of the iceberg; the law is its visible, legitimized tip. And this explains why, if justice is the end that law claims to pursue, it cannot be but a retributive penal kind that is not meant to fix and effectively repair here and now, but to perpetuate the mythical procedure that maintains law's authority.

Law is not immediately perceived as triumphant institutionalized violence, also because it recognizes the right of the opposite party. Indeed, to sustain the fiction of neutrality we are constantly reminded that law is applied to everyone, and that each of us has the same rights. But law can be so surreptitious, Benjamin reminds us, quoting Anatole France, that it will forbid spending the night under the bridge to both the poor and the rich (Benjamin 1996, 249). This irony echoes the saying that we can find in Plato's Republic, by Thrasymachus, about justice being the privilege of the powerful (Plato 1992, I, 338c).

Within these coordinates, individual punishment is deliberately enhanced to protect the position of the system. By fixating on the misuse of causal capacities of agents, it downplays the role of structural factors – whether prescriptive or contextual – in producing harmful or unlawful events. And without attending to structural violence, no true present justice can be done; theft will be the norm where food is scarce. Thus, individual prosecution captures and condemns fallible insubordinate agency to reinforce the belief that the offense – the violation of the untouchable order – generates a sort of credit that needs to be scrapped to repair the harm. But since every individual act from inequal social coordinates, guilt is just a mechanism to perpetuate those very inequalities.

And this also explains, according to Benjamin, why law has produced a *mere life* [*das bloße Leben*] (Benjamin 1996, 250), an individual that has been stripped out of her dignity, and has been reduced to a quantifiable culpable, manageable through mythical violence. But this production does not operate through the overt, literal domination of chattel slavery. Instead, it captures fundamental human capacities and relationships, transforming them into entitlements that it dispenses. For instance, in our neoliberal societies, we became accustomed to speaking the language of "rights": we do not refer to our capacity for movement but to the right to circulate; our creative and cognitive potential is funneled into a right to an education; our ability to shape a habitable world is reconfigured as a right to work. While we may object that this means that law has the duty to provide the conditions for the rights that it grants, the other side of the coin is the sovereign privilege that it exercises in deciding and administrating the type of education, health, food and labor that people ultimately will receive.

The underlying message is that life, under this penal-governance system, lacks the authority to exist outside its regulation; the primary horizon in which human existence unfolds appears to be preceptive. And this strategy of reducing a person into a vulnerable, dispensable organic life that can be used and sacrificed with impunity is what allows the system to function as a freezer. Freezer of the present time – like the Christian epilogue – and of meaningful transformative political action. The circle of self-perpetuating lawful violence spins to make its mythical time to last endlessly. If violence can at all represent the vitalizing power Nietzsche defends, here it is alienated by legal capture. What remains for the governed is submission. This is a political and symbolic gesture.

III.1. The Chronic Injustice

Modern juridical institutions preserve themselves not only through the monopoly of legitimate violence, but through the administration of political temporality itself. The argument developed here is not primarily a critique of the modern state as such, but of a juridical rationality that structures justice through debt, punishment, and temporal adjournment. The modern state appears insofar as it historically administers and preserves this rationality. This is particularly evident if we reflect on bureaucracy as the quotidian administrative correlate of mythical violence. Its endless stream of protocols, procedural delays, and classification grids do not administer justice, but rather actively postpone it, transforming living demands for redress into perpetually pending legal cases. In this sense, bureaucracy should not be understood merely as administrative inefficiency. Within the modern state, it functions as a political technology that translates concrete social antagonisms into manageable procedural categories. Thus, it depoliticizes structural conflict through its dominance protocols disguised as neutrals.

We can better picture this if we imagine society as a body that has a chronic disease: As a governance mechanism, modern law identifies in such body a symptom. And this symptom, that it names crime, is treated individually with a punitive medicine. Without the penal drugs, the system argues, the body will be in an emergency state of sickness and chaos. And so, crime can be “managed” without addressing the root structural causes, if not healing but stable illness is what is sought. And this is precisely what mythical violence seeks. The cure must extend the sickness; the sanction is a palliative that perpetuates the disease. Only if the illness persists, the existence of the practitioner is justified. Ultimately, this kind of retributive treatment does not aim at the termination of all monopolized use of violence. It does not aim at a real state of balance where there will be no more need for a penal apparatus. Its objective is to perpetuate indefinitely the legitimate administration of violence.

In a formulation that anticipates this diagnosis, Marx described the state as an "appalling parasitic body, which enmeshes the body of French society like a net and chokes all its pores" (Marx 1979, 185). Benjamin's critique of violence can thus be read as a radicalization of this insight, carried from the political economy of class to the metaphysics of law itself.

If the freezer of mythical violence sustains the chronic illness, the project of a prospective justice must begin by imagining what a true healing, a melting of the frozen present, would entail, and what kind of "violence" (or non-violent force) could break this circularity without founding a new myth. Therefore, prospective justice is the attempt to think about this healing protocol, which necessarily involves dismantling the legal fiction that confuses the stability of the disease with health. If we follow, yet again, the ideas of Benjamin, we may be able to outline a particular ensemble that combines prospective justice and what he calls "divine violence."

IV. Justice in the *Jetztzeit*

"Prospective justice" is a fairly recent notion. The term has been used in political philosophy and constitutional law to discuss institutional designs that prevent future injustices. In ethics and Justice Theory, is posited as a counterpoint to "retrospective justice", which focuses on correcting past wrongs. There are some thinkers, such as Hans Jonas (Jonas 1984), that have referred to a notion of justice that considers obligations to future generations. And philosophers like Iris Marion Young (Young 1990) have developed a critical concept of non-oppressive justice, which intends to dismantle systemic structures of subjugation, such as domination, exploitation and marginalization, in order to forward-thinking justice beyond retribution.

While these approaches inform the prospective horizon, the Benjaminian inflection I propose is more radical. It is a messianic cessation of the mythical continuum that allows for the redemption of both past and future in the now. Even though the very name may indicate an ideal of justice that belongs to the forthcoming time, the Benjaminian prospective justice that I will develop next denotes instead the present effective rendering of justice.

The temporality proper to historicism is, for Benjamin, a progressive and accumulative conception of history in which events are arranged as moments within a continuous causal sequence. Such temporality is structurally compatible with retributive justice because both depend on deferral. As we have considered, justice is always projected toward a future equilibrium that never fully arrives in the present. The *Jetztzeit*, by contrast, designates an interruption of this continuum. It is not another moment within chronological succession, but a condensation of historical time capable of "blasting open the continuum of history" (Benjamin 2003, 395). Through this interruption, the past appears not as concluded but as something that addresses the present with a "weak messianic power," (Benjamin 2003, 390) demanding redemption here and now.

This redemption, however, is not an individual act but a collective one. Prospective justice enhances a strong bond between generations and places collective agency at its center: it conceives political community as a fabric that connects past, present, and future generations, enabling justice to reach the deeper structural layers that have produced harm and continue to do so. Justice thus becomes conceivable not as postponed compensation, but as an interruption of the temporal structure of penitence. This interruption secures access to truth, reparation, and non-repetition not as legal guarantees but as collective achievements. In this sense, prospective justice extends its concern beyond the living to include those whose suffering was never juridically repaired. We are dealing with a sense of justice that inhabits beyond law's mechanisms and rationality.

Christian temporality that relies on "homogeneous, empty time" (Benjamin 2003, 395) cannot sustain such sense of justice. Benjamin's distinction between mythical and divine violence becomes decisive at this point. This is the "bloody power over mere life for its own sake" (Benjamin 1996, 250) that characterizes the juridical order, a power that captures life within cycles of culpability to maintain its authority. Divine violence, by contrast, does not administer guilt but expiates it. It is "bloodless" insofar as its force is not directed toward the production of culpable subjects or the preservation of legal order.

Therefore, prospective justice fundamentally reorients the question. It stops asking "How can we dispense better institutionalized violence?" and instead asks, "How can we redistribute the capacity to act – the agency – so that communities can repair harm with autonomy?". This is a question that seeks its answer outside retributive logic, and that conceives faults in its wider contextual frame. It is not about creating a new administration or a kinder, gentler legal code. It is a meta-juridical or a-juridical practice that seeks to unleash the community's own creative and restorative force. Thus, prospective justice functions without legal mediation of the individual and collective powers. It is not based on granted "rights", on authorized capacities. To return to our medical analogy, it is about giving the social body immunological autonomy. The goal is no longer to treat the symptom by punishing the offender, but to strengthen the community's health so it can generate its own antibodies, its own forms of repair, accountability, and reconciliation.

This shifts the entire paradigm. Justice is no longer an abstract calculation of debt, dependent on finding legitimate means and just ends. It becomes a singular, contingent event: the community's collective capacity to address a specific harm. It is a practice, not a protocol. By focusing on all the aspects involved in each conflict, especially the structural ones, prospective justice moves these tasks from the legal sphere back to where they belong: in the diverse, autonomous space of collective praxis. Its guiding principle is not to preserve power, but to articulate better conditions for the multiple ways of living. It would therefore be a post-mythical exercise who wants to repair with the living Niobe, not turn her into a stone monument to be a warning to others.

Far from proposing a pacifist ideal about moral and political relationships, Benjamin will say that to achieve this other type of justice, it is necessary to use divine violence. He writes that:

> Since, however, any conceivable solution to human conflicts, not to speak of deliverance from the confines of all the world-historical conditions of existence obtaining hitherto, remains impossible if violence is totally excluded in principle, the question necessarily arises as to what kinds of violences exist other than all those envisaged by legal theory (Benjamin 1996, 247).

It is, then, a matter of thinking a type of violence that is not juridically captured. For instance, something other than the violence that law comprise in its "right to make a strike"; that is, the permission to display violence in some non-threatening way to the system. Benjamin proposes a type of violence that is "lethal without spilling blood" (Benjamin 1996, 249), in the sense that it aims to destroy the very system, and not to punish individual bodies. In his words, "if mythic violence is lawmaking, divine violence is law-destroying" (Benjamin 1996, 249). The latter does not preserve the system by managing symptoms, but shatters the icy continuum itself to restore a flow of unmediated life. In a sense, it is perhaps more violent than the mythical kind, because it will disrupt as well most classic intuitions about order, civilization, and restorative balance. But it will leave untouched what should be sacred: human dignity.

Divine justice redeems the unjust production of mere life in the name of all the living beings. Because despite that we possess, as humans, the biological body that can be managed and controlled, we can't ever be reduced to that sort of abstract existence; the value of our life cannot be lessened to be administrable. Divine justice, in this sense, restores the status of life that was always beyond the legal categories of guilt and innocence; "for the sake of the living" (Benjamin 1996, 250), he writes. Certainly, unless it is reduced to a life that could bleed, labor, sicken, be confined, or killed, law cannot capture human beings. At their core they are invulnerable. As Benjamin puts it, "man cannot, at any price, be said to coincide with the mere life in him" (Benjamin 1996, 251). Their states, its physical life, this can be the profane. It is individuals reduced to their contingent attributes: their biological body, their health, their property, their legal status. This is indeed vulnerable. But divine violence aims to restore the inexpugnable sacred quality of humankind by destroying first and foremost a system that has instructed individuals out of their worthiness. Only mere life can be punished, but an integral dignified being has the capacity to dialogue, learn and repair.

Benjamin still talks about divine punishment. "God's judgement strikes privileged Levites" (Benjamin 1996, 250), he says, but as power positions that are to be demolished. Divine punishment does not castigate an internal transgression of transcendental law, but rather the law's very claim to be the ultimate horizon of life. It is an annihilation of the system of privilege as such. A dissolution of the very social relationship that produces the "criminal" and the "Levit". The Benjaminian thesis reminds us that there is a horizon of justice beyond the law, and that horizon has a name: the inviolable dignity of the living.

Every practice of justice must ultimately be seized by this higher principle: Does it expands the sacred power of the living or does it subject it to the profane logic of domination? This ethical maxim functions as the guideline to ensure true reparation here and now. Justice can be, in this sense, a matter of the present only if it reaches the core of the system of mythical violence, that has structurally produced unjust bloodshed and it will continue to do so, until the present becomes the messianic moment that Benjamin proposes.

In this sense, justice can be understood as the vindicative incoherence that interrupts the chain of progress of the system. Because progress itself, as Benjamin shows, is nothing but a narrative that accumulates ruin after ruin in the name of an expected greater natural end that could ultimately justify devastation. Prospective justice would, in contrast, aim for a present moment that is not run over by the unstoppable course of history. This sense of justice is the only way, Benjamin argues, to stop the catastrophe now, to think of an otherwise unachievable historical justice to past bloodshed. To illustrate this intuition, he notoriously describes the Angelus Novus painting:

> A Klee painting named "Angelus Novus" shows an angel looking as though he is about to move away from something he is fixedly contemplating. His eyes are staring, his mouth is open, his wings are spread. This is how one pictures the angel of history. His face is turned toward the past. Where we perceive a chain of events, he sees one single catastrophe which keeps piling wreckage upon wreckage and hurls it in front of his feet. The angel would like to stay, awaken the dead, and make whole what has been smashed. But a storm is blowing from Paradise; it has got caught in his wings with such violence that the angel can no longer close them. This storm irresistibly propels him into the future to which his back is turned, while the pile of debris before him grows skyward. This storm is what we call progress (Benjamin 2003, 392).

Only outside the fiction of progress, a moment for the past to be present becomes conceivable. Or as Benjamin put it "only a redeemed mankind receives the fullness of its past" (Benjamin 2003, 390). Not as a debt, that would be in any case beyond price, but as something that is still present, and it urges us to create the conditions so that prior oppression is not perpetuated. The present must be, thus, the radical insubordination to structural lawful violence, the redemption of the possibilities foreclosed by past oppression. This means a clean cut with the accumulation of debt and a deactivation of the record that vertically distributes just harm. What must be suspended is the time of normative alienation of existence.

V. Objections to Justice Without Law

Any attempt to conceive justice beyond the law necessarily provokes a set of concerns rooted in the dominant juridical imagination. The a-juridical account of justice developed here, following Walter Benjamin, is no exception. The objections it faces are not merely pragmatic concerns, but symptoms of a legal paradigm that equates justice with normativity, procedure, and institutional control. Rather than dismissing these concerns, this section addresses four central objections: (i) objection of arbitrariness / anomie; (ii) objection of the lack of protection for vulnerable people; (iii) objection of apology of violence; and (iv) objection of utopianism or impracticality.

The following objections should not be understood as entirely discrete categories. They frequently overlap and reinforce one another because they emerge from a shared juridical horizon of meaning. Distinguishing them analytically is useful only insofar as it clarifies different dimensions of the same underlying issue: the fear that justice without legal administration collapses into indeterminacy.

i. The first objection can be formulated as follows: If justice is removed from the law, institutional norms, and stable procedures, what prevents it from becoming an arbitrary practice, dominated by local power dynamics, covert revenge, or oppressive forms of consensus? This concern, however, rests on the assumption that normativity is the sole source of constraint. A Benjaminian perspective reverses this assumption. It is precisely the normative fixation of justice – its procedural stabilization – that enables the perpetuation of harm. Prospective justice is constrained not by procedure but by the inviolability of human dignity and by the collective responsibility to repair concrete damage. Justice, in this sense, is not decided from a position of sovereignty, but enacted by a dialogical praxis that emerges from contingent conflict.

ii. A second objection holds that, historically, the most vulnerable have relied on law for protection. Does this critique not imply a withdrawal of such protections, leaving the most exposed at the mercy of unaccountable powers? To this I will respond that said objection presupposes that vulnerability is an external condition requiring juridical management. In fact, it is precisely legal reduction of life to administrable categories that actively produces vulnerability. Ultimately, law does not grant capacities but manages them. Juridical protection is thus conditional and revocable. Law monopolizes protection to preserve its power, not to eliminate vulnerability. To this extent, prospective justice does not withdraw protection; it seeks to return protective capacities to the collective itself.

iii. A third objection concerns the notion of *divine violence*. Even if it is said to be bloodless, does it open the door to justifying destruction or even terror in the name of a higher justice? In fact, Benjamin himself refers to this reply, when he writes:

> The premise of such an extension of pure or divine power is sure to provoke, particularly today, the most violent reactions, and to be countered by the argument that, if taken to its logical conclusion, it confers on men even lethal power against one another. This, however, cannot be conceded. For the question "May I kill?" meets its irreducible answer in the commandment "Thou shalt not kill" (…) [*the commandment*] exists not as a criterion of judgment, but as a guideline for the actions of persons or communities who have to wrestle with it in solitude and, in exceptional cases to take on themselves the responsibility of ignoring it (Benjamin 1996, 250).

Benjamin's own words show that divine justice is not directed at individuals. As we considered, its targets are the structures. Since it neither administrates punishment, nor founds a new legal order, nor functions as an instrument of power, it cannot be compared with terrorism or revolution in the classic sense. Benjamin's notion of divine violence should not be confused with political violence aimed at seizing power; its destructive force is directed exclusively at juridical relations of domination.

iv. Finally, although prospective justice may sound attractive in theory, will it ultimately be an ethical ideal without application criteria or possibility of institutionalization? But this very demand belongs to the mythical narrative of progress. To require such guarantees before acting is to inhabit frozen temporality. Prospective justice is not a program but an event. This is not a rejection of practicality, but of the demand for institutionalizability as such. One of the defining features of prospective justice is its contingent, complex nature; it is not meant to be a general model, but a mode of intervention that becomes possible only when the mythical continuum of law is interrupted.

VI. Conclusions

Prospective justice of the *Jetztzeit* would require to return to the collective the possibilities and capacities that were captured by the mythical violence of law. And therefore, it operates beyond the false opposition between law/chaos. From this elemental demolition, organization can be reclaimed as an autonomous, effective force. Benjamin reminds us that justice is a category of temporality. Although it is assumed to occur in a homogenous time that "progresses," it is necessary to irrupt this dominant sense of linearity of history to show the structural continuity between a past that we do not want to repeat and our present; between democratic "normality" and open fascism. In this sense Benjamin qualifies as "not philosophical" (Benjamin 2003, 392) the amazement that some expressed about dated events still occurring in the 20th century. Time is the first thing to reclaim, in order to do justice.

This is precisely what he recognizes when he writes that the critique of violence is the philosophy of its history. For only the idea of its development, rather than a gaze directed at immediate information, makes possible a "critical, discriminating and decisive approach to its temporal data" (Benjamin 1996, 251). Prospective justice is such an approach: it refuses the myopia of retributive logic and seeks to break the cycle of lawmaking and law-preserving violence. Benjamin is clear about the destination of this rupture. The breaking of the mythical cycle requires nothing less than "the abolition of state power," (Benjamin 1996, 252) for only then can a new historical epoch be founded.

In the end, the choice Benjamin leaves us with is not between systems. It is between two ways of inhabiting time. Do we choose to inhabit the frozen time of debt, where a glacial present can only punish a petrified past? The time of alienated progress? Or do we dare to inhabit the "now-time", the instant charged with possibility, where a community can interrupt the cycle and begin to repair the unfrozen present? Prospective justice is not a future ideal. It is the courage to redeem even a past that insists on remaining present.

References

Agamben G. 2018. *Karman*. Santford: Stanford University Press.

Benjamin W. 1996. "Critique of Violence," in W. Benjamin, *Selected Writings*. Vol. 1. Cambridge: The Belknap Press of Harvard University Press.

Benjamin W. 2003. "On the Concept of History," in W. Benjamin, *Selected Writings*. Vol. 4. Cambridge: The Belknap Press of Harvard University Press.

Hobbes T. 1994. *Leviathan*. Indianapolis: Hackett Publishing.

Jonas H. 1984. *The Imperative of Responsibility*. Chicago: University of Chicago Press.

Kant I. 1996. *The Metaphysics of Morals*. Cambridge: Cambridge University Press.

Marx K. 1979. "The Eighteenth Brumaire of Louis Bonaparte," in *Marx/Engels Collected Works*. Vol. 11. London: Lawrence & Wishart.

Plato. 1992. *Republic*. Indianapolis: Hackett Publishing.

Young I. M. 1990. *Justice and the Politics of Difference*. Princeton: Princeton University Press.

Weber M. 1946. "Politics as a Vocation," in H. H. Gerth & C. Wright Mills (Eds.), *From Max Weber: Essays in Sociology* (pp. 77–128). New York: Oxford University Press.

Ethics in Progress (ISSN 2084-9257). Vol. 17 (2026). No. 1,
Art. #4, pp. 51-64. DOI:10.14746/eip.2026.1.4

Punishing and Forgiving: Contributions of Deconstruction to Contemporary Debates on Justice

Idoia Quintana Domínguez

(Centre for Applied Ethics, Deusto University, Bilbao, Spain; idoia.quintana@deusto.es)
ORCID: 0000-0002-9087-5366

Abstract: This article challenges the widespread tendency to frame forgiveness either as a moral alternative to punishment or as a practice set in opposition to and disconnected from criminal law. It instead focuses on the tension between punishment, forgiveness and justice by examining the work of Jacques Derrida. The paper argues that the prevailing readings of Derridean forgiveness as unconditional and alien to any legal logic overlook its place within the broader deconstructive task, particularly in the constitutive tension between law and justice. In light of this reconsideration, the article explores the critical scope of deconstruction to interrogate the role of forgiveness, both within punitive logics and within frameworks that advocate responses to harm beyond punishment, including transitional justice and restorative justice.

Keywords: Forgiveness; pardon; punishment; Derrida; justice; law; punitivism; transitional justice; restorative justice.

I. Introduction.
The Emergence of Forgiveness in Contemporary Debates on Justice

Research on forgiveness has generated an increasingly extensive body of scholarship that now extends beyond its traditional disciplinary boundaries.[1] This concept was historically addressed within moral philosophy and theology, where attention centred on its subjective and spiritual dimensions. However, forgiveness later transcended this framework and moved into political, legal and institutional debates on responsibility and reparation. This shift is inextricably linked to changing conceptions of punishment and criminal justice, as well as to growing critiques of punitivism as the dominant framework for managing harm and its effects on the social fabric.

A significant example of this shift is the growing interest in what have come to be termed *political apologies*, public requests for forgiveness made by political leaders or institutions in response to human rights violations committed in the recent or distant past. In this context, forgiveness is debated as an instrument for acknowledging harm, driving reparations and

1 Funding: This research has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie grant agreement No. 894400.

promoting reconciliation within and between States (Schaafsma & Zoodsma 2021). These practices highlight the prominent place that forgiveness has in contemporary political discourse and reveal the ambiguity of its effects when incorporated into institutional frameworks which are oriented towards specific outcomes. In what has been described as the "age of apology" (Brooks 1999; Lind 2008), public performances of repentance and forgiveness have become increasingly frequent. This development raises both the risk of banalisation and the need to think forms of justice shaped by imperatives other than traditional ones (Derrida 2019).

Forgiveness occupies a controversial position in processes of political and social reconstruction following episodes of mass violence within the realm of peace and conflict studies, particularly in transitional justice. In these contexts, it has become a practice traversed by persistent tensions between the demands for truth and justice and the aims of reconciliation. The specialist literature has emphasised its potential to foster recognition of harm and restore social bonds; the risks associated with its political instrumentalization; its use as a vehicle for impunity; and its imposition upon victims (Helmick & Petersen 2001; Bentley 2016). A paradigmatic example was the Truth and Reconciliation Commission in South Africa, which explicitly linked political forgiveness to national reconciliation. This process institutionalised a model that exposed the friction between so-called politics of forgiveness – often oriented towards exemption from criminal liability through amnesties – and interpersonal forgiveness. It also brought into view the complex interweaving of legal, moral and religious registers within a transition that sought to inaugurate a new horizon of coexistence (Cassin 2004; Lefranc 2002).

Within legal and criminological scholarship, several critical approaches have questioned the centrality of retributive criminal justice and explored alternative or complementary models, most notably restorative justice, which reconsiders the function of punishment, the role of victims and the forms of reparation. By emphasising the relational dimension between victim and offender, restorative justice may appear conducive to forgiveness. However, its leading theorists maintain that forgiveness should be understood neither as a guiding principle nor as an explicit objective. Rather, it is to be treated as a contingent element within a process focused on victim reparation, acknowledgement of responsibility and the reintegration of individuals (Zehr 2002; de Greiff 2008; Bessone 2025).

At the same time, forgiveness has begun to occupy complex spaces within classical punitive justice. Its use in legal systems suggests that the boundary between punishing and forgiving can be porous, which raises questions that challenge traditional categories: Can punishment and forgiveness coexist within the same institutional framework, for example by having criminal and restorative pathways work together? Alternatively, should punishment be understood as a requirement for forgiveness to create a new relational horizon? Moreover, some criminal codes have afforded forgiveness a specific place. This raises a further question: if forgiveness is required of offenders, whether as a mitigating factor or as an explicit objective, does it risk being absorbed into punitive logic and reduced to a mere instrument of the judicial system? (Smith 2014; Quintana 2026; Nayfeld et al. 2025). These brief remarks are not intended to be exhaustive. They instead account for the diverse landscape of contemporary invocations of forgiveness, the functions of which remain uncertain and vary according to their assumptions and aspirations. In this context, recourse to forgiveness introduces a decisive ambiguity. While it may open a space for thinking about justice beyond punishment, it may also operate as yet another technology for managing harm and normalising conflict. How is it possible to conceive of a form of forgiveness that is neither reduced to a function of the punitive or disciplinary order, nor neutralised by the legal and political frameworks that seek to administer it?

This article explores forgiveness as a response to harm that is distinct from punishment and from logics of proportionality, at least in principle or in substance. To do this, it provides a reading of Jacques Derrida's contributions, as they are among the most rigorous theoretical frameworks for engaging with the complexity of forgiveness and its diverse manifestations. Derrida's approach has often been interpreted as a defence of unconditional forgiveness located in an abstract sphere. Such readings tend to minimise its critical force in relation to existing punitive practices and legal frameworks. Against this view, the article argues that, in Derrida's thought, unconditionality does not pretend to represent a pure or transcendent form; rather, it seeks to interrogate how it is articulated within the normative assumptions that regulate harm, responsibility and their possible responses. From this perspective, I intend to move beyond the familiar claim that forgiveness must forgive the unforgivable, with a view to reconsider the issue in terms of what forgiveness does – and undoes – within the political, legal and relational contexts in which it operates and is regulated. These processes are shaped by logics that range from reconciliation, reintegration and healing, to forms of inner transformation. I will focus on the aporetic and phenomenological dimensions of Derrida's proposal, highlighting the potential of his analysis to reactivate the paradoxical character of forgiveness and to illuminate the critical force that these tensions exert in contemporary debates on justice and punitivism.

II. The End of Forgiveness: The Unforgivable

One of Derrida's most suggestive and complex definitions of forgiveness is the one that condenses his approach into a paradox: "forgiveness forgives only the unforgivable" (Derrida 2001, 32).[2] This statement was formulated largely in opposition to post-Holocaust

2 In the late 1990s, Derrida gave a two-year seminar entitled *Le parjure et le pardon* (*Perjury and Pardon*) (1997-1999) as part of a series of seminars under the title *Questions of Responsibility* (1991-2003). The seminar was held at the École des hautes études en sciences sociales in Paris in 1997-1998 and 1998-1999 at the University of California, Irvine in 1998 and 1999, and later at New York University in 2001. Some of the sessions were reconstructed and published later

reflections on the limits of forgiveness and the notion of the unforgivable (Banki 2018; De Warren 2014). In this context, it was argued that there are some crimes that exceed the capacity to forgive or that are inexpiable, insofar as no criminal penalty could ever be commensurate to them. Derrida took up and problematised this idea of the limit or the end of forgiveness. He highlighted that in justifying this, a close correlation is established between the power to forgive and the power to punish: where one proves impossible, the other is deemed impossible as well.

For Hannah Arendt, the power to forgive is rendered powerless in the face of "radical evil".[3] In cases where the harm inflicted is so extreme that it overwhelms human capability to respond, neither forgiveness nor punishment can fulfil their function of bringing closure. Both fail in their attempt to put an end to something that, without intervention, could be indefinitely perpetuated. Arendt underlined the structural relationship between punishment and forgiveness, which she understood not as opposites, but as alternative responses to harmful action. Both share a defining feature: they are capable of breaking the cycle of violence that sustains the logic of revenge.[4] In *The Human Condition* she argued:

> The alternative to forgiveness, but by no means its opposite, is punishment, and both have in common that they attempt to put an end to something that without

in special issues or in other works by Derrida (see "Editors' Note", in Derrida 2022, XVII–XVIII, n. 13; "Editors' Note", in Derrida 2023, XXIII, n. 26). This seminar has been published in two volumes (2019 and 2020) and translated into English under the title *Perjury and Pardon* (2022 and 2023). Some sessions had circulated earlier, but, especially in the English-speaking sphere, Derrida's analysis of forgiveness was disseminated in the early 2000s through two key texts. The first of these was *On Forgiveness* (Derrida 2001). This was based on an interview in which Derrida discussed forgiveness in specific historical contexts, including apartheid in South Africa and the Vichy regime (Derrida 1999). The second was *To Forgive: The Unforgivable and the Imprescriptible* (2001), the first session of the seminar, where he set out the main problems developed throughout the course. Read and discussed primarily within specialist circles, these texts appeared at a time when forgiveness was gaining particularly problematic prominence in political and media debates. In 1998, for example, the report of the South African Truth and Reconciliation Commission was published, the Good Friday Agreement was signed in Northern Ireland, and the President of Japan made a public apology for the suffering caused during Japan's colonial rule over Korea. A few years later, reconciliation processes unfolded in Rwanda in the aftermath of the genocide. These are just some of the most significant events that have placed forgiveness at the centre of discussions on accountability, reconciliation and justice since the late 1990s.

3 "Radical evil" is an expression borrowed from Kant and refers to a deviation from the moral law, the radical nature of which is linked to a propensity inherent in human nature. Arendt gave a different meaning to it, transferring its moral sense to the historical and political sphere (Arendt 1976, 459).

4 It is worth noting the extent to which Arendt distanced herself from the view, developed in a qualified sense by Paul Ricœur, that criminal punishment may retain the structure of an institutionalised form of revenge. Although Ricœur, drawing explicitly on Arendt in the epilogue of *La mémoire, l'histoire, l'oubli* (Ricœur 2000), presents criminal law as one of the major achievements of rationality insofar as it substitutes discourse for violence and interrupts the cycle of vengeance, in *Avant la justice non violente, la justice violente* (Ricœur 2004) he also argues that modern penal justice has not entirely overcome the logic of revenge. He describes it as a form of "violent justice" because, under the concept of retribution, there persists a rationalised form of vengeance – a repressed remainder [*refoulé*] within juridical culture. Penal justice thus succeeds in halting private vengeance, but it does so through institutionalised violence (punishment) whose rational justification remains incomplete, placing it at a stage prior to what he calls "non-violent" or restorative justice. Arendt, by contrast, conceived punishment less as a continuation of vengeance than as a dischargeable debt capable of bringing to an end a self-perpetuating cycle of violence.

> interference could go on endlessly. It is therefore quite significant, a structural element in the realm of human affairs, that men are unable to forgive what they cannot punish and that they are unable to punish what has turned out to be unforgivable. This is the true hallmark of those offenses which, since Kant, we call "radical evil" (…) we can neither punish nor forgive such offenses and that they therefore transcend the realm of human affairs and the potentialities of human power (Arendt 1958, 241).

In a similar line of thought, Vladimir Jankélévitch introduced a significant twist with respect to the analysis he had developed in his essay *Le pardon* (Jankélévitch 1967), in which he defended a form of unlimited forgiveness, unconstrained by reason or calculation. In a later text (Jankélévitch 1986) which was part of the debates on the imprescriptibility of crimes against humanity in France in the 1970s, he argued that there is no punishment capable of atoning for the magnitude of certain crimes and that, in this context, forgiveness is morally unjustifiable ("Forgiveness died in the death camp"). Jankélévitch argued that, in the face of crimes that go beyond all human measure, punishment loses its compensatory and corrective function, just as forgiveness loses its ethical legitimacy: "Properly speaking, this grandiose massacre is not a crime on a human scale. (…) One cannot punish the criminal with a punishment proportional to his crime (…) strictly speaking, what happened is inexpiable" (Jankélévitch 1986, 29).

In Arendt's and Jankélévitch's reflections, punishment and forgiveness operate according to a shared metric grounded in correlation and proportionality. There is a symmetry between a crime and its punishment, as well as between the ability to punish and the ability to forgive. Situating forgiveness within a horizon of measurability suggests that the wrongdoing is still actionable through an economy of exchange, whether as punishment or reparation. The incommensurability of a crime means, on the contrary, the impossibility of resorting to punishment or forgiveness as adequate responses to this harm and as a way of putting an end to it. For Derrida, however, the unforgivable does not mark the limit of forgiveness but its point of departure. It is precisely the extreme nature of the crime, where no exchange between wrongdoing and any possible reparation can be conceived, that summons forgiveness and simultaneously defines its double status. Forgiveness thus appears, on the one hand, as a condition of possibility, since only the unforgivable truly calls for it. On the other hand, it appears as a condition of impossibility, since forgiving the unforgivable names an act that cannot, in fact, be realised.

III. Impossible Forgiveness: Double Bind between Conditionality and Unconditionality

To prevent forgiveness from being reduced to a mere exchange, Derrida pointed out that forgiving the unforgivable must be an experience that transcends the transactional logic that links harm to reparation, or wrongdoing to penalty: "If there is forgiveness, it must forgive the unforgivable – such is the logical aporia (...). If one had to forgive only what is forgivable (...) then one would not forgive. One would excuse, forgive, erase, one would not be granting forgiveness" (Derrida 1997, 158). He also detached it from any exercise of sovereignty – be it individual, legal or state-based – and moved it towards an experience that exceeds the will. In doing so, Derrida notably took a radical distance from classical definitions that conceive forgiveness as a faculty or moral power of the individual – a position shared, among others, by Arendt (Arendt 1958), Murphy and Hampton (Murphy & Hampton 1988), Verdeja (Verdeja 2004), Griswold (Griswold 2007) and Konstan (Konstan 2010). He also questioned whether forgiveness can emanate from a subject capable of administering it as a resource at their disposal. In the face of the social logics of reciprocity or proportionality, forgiveness should be that exceptional instance that interrupts them; a pole of heterogeneity that defies the frameworks of symmetry and transaction, standing out not for its usefulness, but for its extraordinary character.

> I shall risk this proposition: each time forgiveness is at the service of a finality, be it noble and spiritual (atonement or redemption, reconciliation, salvation), each time that it aims to re-establish a normality (social, national, political, psychological) by a work of mourning, by some therapy or ecology of memory, then the "forgiveness" is not pure – nor is its concept. Forgiveness is not, it should not be, normal, normative, normalising. It should remain exceptional and extraordinary, in the face of the impossible: as if it interrupted the ordinary course of historical temporality (Derrida 2001a, 32).

However, pure forgiveness, or forgiveness "worthy of the name" – terms used by Derrida to underscore its excessive degree in relation to the order of the possible – is characterised by features that render it irreducible to the order of the possible. These same features have led forgiveness to appear abstract and unrealisable. It does not constitute a transcendent ideal or an abstraction alien to history; rather, it operates as a disruptive force. While it is defined as "impossible," its nature as an event lies in its capacity to interrupt the course of the possible. This unconditioned dimension coexists in a conflictual tension with the conditioned forms of forgiveness.[5] For that reason, the purity of the concept is not exhausted by mere transcendence. Instead, it operates as a demand that fractures and unsettles the economy of regulated exchange (confession, repentance and the promise of non-repetition). Although the aims of reconciliation and healing claim for themselves the name of forgiveness, Derridean unconditionality interrogates these

5 Derrida argued that, while the Abrahamic legacy largely inscribes forgiveness within an economy – or even a transeconomy – aimed at reconciliation, redemption or salvation through confession, repentance, sacrifice or expiation ("the greatest religious and spiritual traditions of forgiveness never remove forgiveness from the horizon of reconciliation, hope for redemption and salvation, through avowal, remorse or repentance, sacrifice and expiation" [Derrida 2022, 9]), it also opens onto the possibility of unconditional forgiveness in a tense and conflictual manner. Unconditional forgiveness would demand that forgiveness be granted without being requested, without requiring any transformation on the part of the offender and without relying on a historical horizon in which the past can be relieved or overcome. Derrida does not, however, locate the origin of forgiveness in the Judeo-Christian tradition. Instead, he problematises the widely accepted view that forgiveness was absent in Greco-Roman antiquity. According to this interpretation that is widely predominant in the general literature, only a form of forgiveness understood as clemency existed in antiquity, while forgiveness as grace emerged only within Judeo-Christian thought (see especially the third, sixth and seventh sessions of the first volume of *Perjury and Pardon* (forgiveness), in Derrida 2022).

structures. It shows that forgiveness, if it occurs at all, does so only insofar as it moves away with the logic of conditionality and resists reduction to a compensatory transaction.

The double bind between the unconditioned and the conditioned thus constitutes the axis of Derrida's deconstructive strategy and is articulated under the same logic as his analyses of justice (see especially Derrida 1992) and hospitality (see especially Derrida & Dufourmantelle 2000) and the recently published seminars on this subject). In these cases, an indissoluble yet necessary antinomy comes into view, one that describes two heterogeneous but inseparable orders. Just as justice is not law but the incalculable demand that interrogates it, and pure hospitality is that which opens to the stranger before implementing any legal identification or regulation, pure forgiveness appears as a force that destabilises the conditions of possibility and transcends all transactional normativity. On the one hand, the unconditional opens up the possible beyond itself; on the other hand, specific forms of forgiveness, regulated hospitality and positive law necessarily operate under specific conditionalities and purposes.

This structure implies that pure forgiveness does not stand in opposition to conditioned forms as though they were merely degraded versions of an ideal, but confronts the very conditionalities that organise experience. The incalculable, whether justice, pure hospitality, or unconditional forgiveness, does not occupy an external or transcendent plane. Rather, it operates as that which persistently fractures, displaces and reconfigures the field of the possible. Derrida therefore does not conceive these categories as moral ends that rely on a sovereign faculty. He understands them as expressions of an original mismatch that compels one to "calculate with the incalculable" (Derrida 1992, 16).

Understanding the unforgivable through this bind and through the intervention of the impossible allows us to situate Derrida's approach within contemporary debates on justice. My intention is to show that Derridean notion of forgiveness cannot be reduced to a technology of political stabilisation or grounding, nor can it be transformed into an instrument for securing closure for conflicts. On the contrary, its unconditionality constitutes a form of resistance to any use of forgiveness that seeks to bring about closure or to certify a final word. Precisely because of this, it keeps open the very possibility of forgiveness in an uncertain but indispensable way, while rendering suspect any form of forgiveness that is fully recognisable, verifiable or normatively defined within institutional frameworks.

IV. Pure Forgiveness and Impossibility of Closure: Normativity, Excess and Perjury

For Derrida, to speak of pure forgiveness is not to propose a normative ideal capable of institutional implementation, but rather to insist on the irreducibility of forgiveness in the face of political and juridical uses that tend to turn it into an end-oriented instrument, such as social pacification or national reconciliation. Forgiveness as defined by its unconditional, incalculable and unpredictable character cannot be directly translated into public policy and regulated mediation practices. However, this position has raised important questions.

Among others, Ernesto Verdeja (Verdeja 2004) has questioned the practical and moral validity of this model (Evans 2013). According to Verdeja, the Derridean demand for non-conditionality, whereby forgiveness is a free gift that does not expect the perpetrator's repentance, carries significant risks. First, he argues that forgiveness which demands nothing in return may be perceived by the offender as a strategy of moral superiority on the part of the victim, thereby generating a "debt of *resentment*" that hinders true transformation. Secondly, Verdeja identifies a dangerous paradox in Derrida's work. To prevent forgiveness from becoming an economic transaction or a burden of gratitude, the act should be accompanied by an erasure of memory. From this perspective, unconditional forgiveness would run the risk of becoming a second crime against victims by dismantling efforts towards truth, accountability and reparation. Against what he regards as the "estrangement" of the private encounter between victim and perpetrator, Verdeja defends a firm, conditional model of forgiveness in which the acknowledgement of responsibility and the preservation of memory are prerequisites for any social closure.

Nevertheless, this critique presupposes that Derridean forgiveness is intended to replace or deactivate these devices of justice and memory. Derrida does not propose an alternative politics of forgiveness, nor does he deny the need for truth and reparation processes aimed at enabling coexistence:

> Of course, no one would decently dare to object to the imperative of reconciliation. It would be better to put an end to the crimes and discords. Once again, however, I believe it necessary to distinguish between forgiveness and this process of reconciliation, this reconstitution of a health or a "normality", as necessary and desirable as it would appear through amnesties, the "work of mourning", etc. (Derrida 2001a, 50).

He introduces an aporetic displacement: pure forgiveness does not function as an alternative institutional model, but as the critical instance that prevents the full identification of forgiveness with closure: "A 'finalised' forgiveness is not forgiveness; it is only a political strategy or a psycho-therapeutic economy" (Derrida 2001a, 50).

Under this reading, Derrida's proposal does not deactivate criminal, transitional or restorative justice policies, but rather destabilises them by denouncing any pretence that they can exhaust the meaning of forgiveness. What is at stake is not an "erasure of memory" – which for Derrida would be another form of normalisation – but resistance to forgiveness being assimilated as a compensation mechanism that is automatically activated after repentance. By maintaining forgiveness as something "impossible," Derrida ensures that it is not absorbed into a logic of calculation that could, paradoxically, ultimately absolve the perpetrator of any responsibility.

This displacement also affects the way in which the relationship between punishment and forgiveness is conceptualised. The two are usually presented as ontologically heterogeneous orders. Punishment belongs to the domain of calculation, proportionality, equality before the law and normative repeatability; forgiveness, by contrast, is an exceptional gesture that exceeds legality, disrupts equivalence and suspends calculation. However, the boundary between the two is neither external nor rigid. As Derrida pointed out:

> The difference between the unconditional and conditional, or here between the juridical and nonjuridical, between the imprescriptible and unforgivable, isn't simply a distinction between two external terms. However heterogeneous those two values be, it is possible that one is inscribed within the other, that one comes to interrupt the other. In other words, that in law there is a moment of the interruption of the law, or conversely (…) that in forgiveness, that in the nonjuridical domain of forgiveness, there is a hidden juridical instance, and reference to a penalty, a potential punishment even where that isn't provided for by juridical legislation (Derrida 2022, 8, n. 19).

The heterogeneity between the two orders therefore does not entail total mutual impenetrability. Punishment is neither pure calculation as payment of a debt, nor is forgiveness a wholly disinterested gesture entirely beyond the reach of normativity. The order of punishment, albeit structured by generality criteria, is not reduced to mere automatic calculation. The notion of imprescriptibility, the appeal to single responsibility or the expectation of transformation of the subject introduce elements that go beyond the arithmetic of equivalence. Even where proportionality is sought, the penalty refers to a dimension that cannot be measured exhaustively in terms of symmetry between the harm and the sanction. Conversely, forgiveness is not purely external to normativity. In order to be intelligible, forgiveness adopts cultural forms, gestures and grammars that inscribe it in the public sphere. It is structured by expectations – repentance, recognition of harm, an explicit request – that shape its phenomenology without being ontological conditions in Derrida's framework. Precisely because of this, forgiveness can be feigned, instrumentalised or falsified. The very possibility of simulation reveals that there is an underlying normativity that makes it identifiable. If forgiveness can be a representation, it is because it is somewhat encoded in social and legal language.

This intertwining leads to a second decisive displacement: fully-verifiable forgiveness does not exist. The phenomenal dimension of forgiveness brings to light the impossibility of certifying its occurrence. Whether someone utters "I forgive you" or "I ask for your forgiveness," in neither case is there a criterion for determining conclusively whether the act has genuinely taken place. This fragility is not contingent, but structural, as forgiveness is constitutively exposed to the risk of perjury and simulation. As Derrida warns:

> forgery would be immediately and fatally contemporaneous with the presentation, the putting into presence of forgiveness. A forgiveness that presents itself (forgiveness requested or granted) and says "here I am" would already be a forgiveness contaminated by the said forgery (Derrida 2022, 12).

Under this premise, the possibility of feigning forgiveness is not an accidental deviation, but an internal condition of its functioning. The moment forgiveness is performed (when it appears through word or rhythm) it is fatally contaminated by the possibility of simulacra. There is, therefore, no access to a pure interiority that guarantees the authenticity of the gesture. This fragility also affects the institutional mechanisms that seek to inscribe forgiveness in formal reconciliation processes. This means that no commission, court or programme can certify that forgiveness has actually taken place or ensure its genuineness.

Far from invalidating justice policies, this impossibility forces them to recognise their own limits. Processes aimed at truth, reparation, public memory and guarantees of non-repetition can – and must – establish verifiable conditions, responsibilities and mechanisms. But they cannot produce or demand forgiveness as a result. When they turn it into an indicator of success or an implicit requirement for reconciliation, they run the risk of instrumentalising it. In this sense, Verdeja's critique can be reformulated. Pure forgiveness does not deactivate justice, but prevents justice from being fully identified with a logic of equivalence and closure. It does not replace memory or reparation, but neither does it allow them to be exhausted in a transactional scheme (Nussbaum 2016). Where punishment and compensation tend to be structured as proportional exchanges, forgiveness introduces an openness that does not cancel responsibility, but neither does it reduce it to calculation.

There is, therefore, neither a pure order of punishment nor a form of forgiveness wholly external to norm. There are heterogeneity and mutual inscription. And there is no fully certifiable genuine forgiveness. Its possibility is traversed by exposure to perjury, simulation and indeterminacy. Precisely for this reason, forgiveness cannot be a normative pillar of criminal, transitional or restorative justice. Its place, if it occupies one at all, is that of a critical "non-place" that is unverifiable and inoperative, although not thereby rendered non-existent or irrelevant.

V. Conclusions

Contemporary discussions of forgiveness have tended to situate the practice within a normatively defined framework. The aim is to specify when it is possible, how far it can go and what ends it can legitimately achieve. From this perspective, forgiveness appears either as a valuable – even desirable – component of processes of reparation and reconstruction of the social fabric, or as a problematic element that may undermine the demands for justice, truth, memory and guarantees of non-repetition now regarded as indispensable pillars of collective justice. With this framework, forgiveness is also assessed in relation to the rationality of punishment, as that which may temper it, interrupt it or call its legitimacy into question. Forgiveness therefore oscillates between functional inclusion in reparative frameworks and preventive exclusion. The latter arises from concerns that it may trivialise harm, enable impunity or prematurely bring closure to conflict. Jacques Derrida's proposal makes a decisive contribution precisely at the point where this polarity emerges, between treating forgiveness as essential and rejecting it as an obstacle. He makes a decisive contribution that extends beyond a simple denunciation of its instrumental use.

Derrida invites us to think of an impossible forgiveness the impossibility of which is not merely logical or formal, but phenomenal in nature. Forgiveness can never present itself in a fully identifiable and verifiable form as such. It cannot offer itself as a certain or fully performative act, in the sovereign sense of the saying that acts as a doing, as if its enunciation alone were sufficient to transform reality and guarantee its effects. The characteristics that Derrida associates with forgiveness (a face-to-face encounter, unconditionality, its unique and unrepeatable character, and its excessive degree in relation to any political, legal or historical calculation or purpose) cannot occur in a fully present, transparent or immediate form. It is always traversed by conditions, mediations, interests, institutions or expectations, without these being able to exhaust it or contain its tension (demand for pure, unconditional, impossible forgiveness). It is precisely this overflow that shows how unconditional forgiveness resists, questions and reconfigures the limits, logics and tensions of these structures.

The impossibility of forgiveness, as Derrida describes it, does not render it irrelevant. On the contrary, it defines its mode of operation. Forgiveness does not present itself as a verifiable fact or a guarantee that can be secured. It appears instead as a demand that exceeds any concrete realisation. It does not replace that logic, but unsettles it by preventing forgiveness from being definitively stabilised as a closed equivalence between wrongdoing and compensation, between offence and punishment. From this perspective, Derrida problematises a recurrent assumption in contemporary debates on justice, whether punitive, transitional or restorative. According to this assumption, forgiveness could or should function as definitive closure or as a secure point of departure for a new temporal horizon. The force of his thesis lies precisely here. To transform forgiveness into an instrument of closure or a component in symbolic exchange is to eliminate the uncertainty, unverifiability and incalculability that define it, which is what it introduces. This does not mean that forgiveness should be excluded from practices of justice; rather, it involves that its relationship to them must be understood differently.

In processes of truth, reparation and the reconstruction of public memory, forgiveness, understood through the phenomenological instability proposed by Derrida, subjected to an inherent undecidability that marks its authenticity, cannot be established as either a normative requirement or an indicator of success. It is instead assigned a kind of "non-place" that reveals its critical force as resistance to logics of closure. Because it cannot be institutionalised or translated into measurable outcomes, forgiveness challenges the transactional dynamics of reparation, whether articulated through punishment or compensation. These dynamics tend to be structured as forms of equivalence: harm for punishment, offence for compensation, recognition for reintegration. Forgiveness, by contrast, introduces an asymmetry that exceeds calculation. It does not eliminate the need for justice or substitutes reparation obligations. However, it prevents them from being reduced to a closed framework. From this perspective, forgiveness does not seek to complete a process or to bring conflict definitively to an end. It resists closure. It prevents justice from being reduced to equivalence, a reduction that Derrida repeatedly associates with the mismatch between justice and law. It also resists the temptation to treat reconciliation as a guaranteed resolution. This is the gesture of forgiveness: it opens a space while withdrawing any assurance that closure has been achieved.

References

Arendt H. 1958. *The Human Condition*. Chicago: University of Chicago Press.

Arendt H. 1976. *The Origins of Totalitarianism*. New York: Harcourt Brace.

Banki P. 2018. *The Forgiveness to Come: The Holocaust and the Hyper-Ethical*. New York: Fordham University Press. https://doi.org/10.2307/j.ctt1xhr62h

Bentley T. 2016. *Empires of Remorse: Narrative, Postcolonialism, and Apologies for Colonial Atrocities*. Abingdon: Routledge.

Bernstein R. J. 2002. *Radical Evil: A Philosophical Interrogation*. Malden, MA: Polity.

Bessone M. 2025. "Le pardon, effet essentiellement secondaire de la justice réparatrice", in N. Nayfeld, L. Jaffro, S. Lemaire, & M. Bessone (Eds.), *La justice pénale aux frontières du pardon* (pp. 197–216). Paris: Garnier.

Brooks R. L. (Ed.). 1999. *When Sorry Isn't Enough: The Controversy Over Apologies and Reparations for Human Injustice*. New York: NYU Press.

Cassin B. 2004. "Amnistie et pardon: pour une ligne de partage entre éthique et politique", *Le Genre humain* 43(2):35–58. https://doi.org/10.3917/lgh.043.0035

Cornell D., Rosenfeld M., & Carlson D. G. (Eds.). 1992. *Deconstruction and the Possibility of Justice*. New York: Routledge.

De Greiff P. 2008. *The Handbook of Reparations*. Oxford: Oxford University Press.

De Warren N. 2014. "L'impardonnable chez Jankélévitch", *Archives de Philosophie* 77(3):421–433. https://doi.org/10.3917/aphi.773.0421

Derrida J. 1992. "Force of Law: The 'Mystical Foundation of Authority'," in D. Cornell, M. Rosenfeld, & D. Gray Carlson (Eds.), *Deconstruction and the Possibility of Justice* (pp. 3–67). New York: Routledge. URL: https://fswg.wordpress.com/wp-content/uploads/2018/02/derrida-force-of-law.pdf

Derrida J. 1997. "D'abord, je ne savais pas….", *Les Cahiers du GRIF* 3:131–166.

Derrida J. 2000. "On Forgiveness," *Studies in Practical Philosophy* 2(2):81–102. https://doi.org/10.5840/studpracphil2000229

Derrida J. & Dufourmantelle A. 2000. *Of Hospitality*. Trans. By R. Bowlby. Stanford, California: Stanford University Press.

Derrida J. 2001a. *On Forgiveness. On Cosmopolitanism and Forgiveness*. New York: Routledge.

Derrida J. 2001b. "To Forgive: The Unforgivable and the Imprescriptible," in J. D. Caputo, M. Scanlon, & M. Dolley (Eds.), *Questioning God* (pp. 21–51). Bloomington – Indianapolis: Indiana University Press.

Derrida J. 2001c. "On Forgiveness: A Roundtable Discussion with Jacques Derrida," in J. D. Caputo, M. Scanlon, & M. Dolley (Eds.), *Questioning God* (pp. 52–72). Bloomington – Indianapolis: Indiana University Press.

Derrida J. 2019. *Le parjure et le pardon. Volume I. Séminaire (1997–1998)*. Paris: Seuil.

Derrida J. 2020. *Le parjure et le pardon. Volume II. Séminaire (1998–1999)*. Paris: Seuil.

Derrida J. 2022. *Perjury and Pardon. Volume I*. Chicago: University of Chicago Press.

Derrida J. 2023. *Perjury and Pardon. Volume II*. Chicago: University of Chicago Press.

Evans M. 2013. "Derrida and Forgiveness," in E. J. Alam (Ed.), *Compassion and Forgiveness* (pp. 17–32). Notre Dame: University of Notre Dame Press.

Griswold C. 2007. *Forgiveness: A Philosophical Exploration*. Cambridge: Cambridge University Press.

Helmick R. G. & Petersen R. (Eds.) 2001. *Forgiveness & Reconciliation: Public Policy & Conflict Transformation*. Philadelphia: Templeton Press.

Jankélévitch V. 1967. *Le Pardon*. Paris: Aubier-Montaigne.

Jankélévitch V. 1986. *Imprescriptible. Pardonner? Dans l'honneur et la dignité*. Paris: Seuil.

Kant I. 1960. *Religion Within the Limits of Reason Alone*. New York: Torchbooks.

Konstan D. 2010. *Before Forgiveness: The Origins of a Moral Idea*. Cambridge: Cambridge University Press.

Lefranc S. 2002. *Politiques du pardon*. Paris: Presses Universitaires de France.

Lind J. 2008. *Sorry States: Apologies in International Politics*. New York: Cornell University Press.

Murphy J. G. & Hampton J. 1988. *Forgiveness and Mercy*. Cambridge: Cambridge University Press.

Nayfeld N., Jaffro L., Lemaire S., & Bessone M. (Eds.) 2025. *La justice pénale aux frontières du pardon*. Paris: Classiques Garnier (Bibliothèque de la pensée juridique).

Lacey N. & Pickard H. 2015. "To Blame or to Forgive? Reconciling Punishment and Forgiveness in Criminal Justice," *Oxford Journal of Legal Studies* 35(4):665–696. https://doi.org/10.1093/ojls/gqv012

Nussbaum M. 2016. *Anger and Forgiveness. Resentment, Generosity, Justice*. Oxford: Oxford University Press.

Quintana D. I. 2023. “Entre una ética hiperbólica y un principio político: la incondicionalidad del perdón en Jacques Derrida”, *Pensamiento. Revista De Investigación E Información Filosófica* 79(303):293–311. https://doi.org/10.14422/pen.v79.i303.y2023.001

Quintana D. I. 2026. “Forgiveness as a Legal Condition: A Study of the Uses of Forgiveness in the Basque Conflict,” *Punishment & Society* 0(0):1–19. https://doi.org/10.1177/14624745251410458

Ricœur P. 2004. “Avant la justice non violente, la justice violente”, *Le Genre humain* 43(2):157–172. https://doi.org/10.3917/lgh.043.0157

Ricœur P. 2000. *La Mémoire, l’Histoire, l’Oubli*. Paris: Le Seuil. https://doi.org/10.3917/ls.ricoe.2014.05

Schaafsma J. & Zoodsma M. 2021. *The Political Apology Database*. Retrieved from the Political Apologies across Cultures website: http://www.politicalapologies.com/

Smith N. 2014. *Justice Through Apologies: Remorse, Reform, and Punishment*. Cambridge: Cambridge University Press.

Verdeja E. 2004. “Derrida and the Impossibility of Forgiveness,” *Contemporary Political Theory* 3:23–47. https://doi.org/10.1057/palgrave.cpt.9300103

Zehr H. 2002. *The Little Book of Restorative Justice*. Intercourse: Good Books.

Ethics in Progress (ISSN 2084-9257). Vol. 17 (2026). No. 1,
Art. #5, pp. 65-104. DOI:10.14746/eip.2026.1.5

Reimagining Justice Beyond Blame: Defending Iris Marion Young's Social Connection Model

Arif Guljar

(Indian Institute of Technology Bombay, Mumbai, India; arifguljar125@gmail.com)
ORCID: 0009-0003-5624-0281

Abstract: This paper defends Iris Marion Young's Social Connection Model (SCM) as a superior framework for understanding responsibility for structural injustice and proposes a conceptual intervention into what I term the punishment dilemma. The dilemma is this: the liability model's backward-looking focus on blame is conceptually mismatched to structural phenomena (since no identifiable perpetrator exists), yet impunity is morally intolerable. The liability model – built on identifiable wrongdoing, direct causation, and culpable intent – fails because structural harms are processual, systemic, and normalized. Young's SCM performs a foundational shift from fault to social position: responsibility derives not from culpable action but from constitutive participation in interconnected social processes that produce unjust outcomes. I reconstruct the SCM's philosophical architecture, showing how its five features and four parameters provide a coherent alternative. Drawing on Virginia Mantouvalou's analysis of state-mediated vulnerability and Serena Parekh's work on gender violence and refugee policy, I demonstrate that structural injustice generates a demand for accountability that retributive punishment cannot satisfy. I then develop transformative sanction: punitive or quasi-punitive measures justified not by backward-looking desert but by forward-looking structural disruption. Young's four parameters – power, privilege, interest, and collective ability – provide criteria for operationalizing this framework. I address the indeterminacy critique (Christian Barry and Luara Ferracioli) and the motivational challenge (Martha C. Nussbaum), showing that the SCM survives both. While further specification of procedural safeguards is needed, the SCM stands as a philosophically coherent and politically actionable framework for transforming structural injustice.

Keywords: Iris Marion Young; structural injustice; social connection model; liability model; transformative sanction; ressentiment; state-mediated injustice.

I. Introduction

How should we hold agents accountable for harms that no single person intends and no single act causes? This question lies at the heart of contemporary debates about structural injustice. Consider a low-wage single mother (Sandy) facing eviction not because any landlord or policymaker has acted maliciously, but because the combined effects of urban redevelopment, unaffordable housing, and inadequate transportation systems produce predictable patterns of displacement (Young 2011, 43–48). Consider communities living near industrial facilities who suffer from pollution not because any single executive ordered illegal dumping, but because regulatory frameworks, economic incentives, and routine business practices systematically externalize environmental costs onto marginalized populations. In such cases, harm emerges not from isolated wrongdoing but from the normal operation of social structures (Young 2011, 43–64). This paper addresses what I term the punishment dilemma for structural justice, and it offers a way forward through Young's Social Connection Model (SCM). The paper does not claim to resolve the punishment dilemma definitively. Rather, it offers a conceptual intervention within an ongoing theoretical debate, proposing transformative sanction as one possible way forward rather than the definitive solution. The aim is to suggest that Young's SCM, when extended through the concept of transformative sanction, provides a coherent and actionable framework for accountability in structural contexts – without claiming to be the only or final word on the matter.

The punishment dilemma consists of three claims that cannot be simultaneously maintained. First, the prevailing liability model of responsibility – with its juridical focus on blame, voluntary action, and direct causation – is constitutively mismatched to structural injustice. Young demonstrates this through the paradigmatic case of "Sandy," a low-wage single mother facing eviction through no individual's malicious intent (Young 2011, 43–48). The harm Sandy experiences cannot be traced to any single act of voluntary wrongdoing; it is the emergent effect of multiple interdependent processes operating according to their normal logic. The liability model cannot locate a culpable perpetrator because none exists. Second, simply abandoning accountability is unacceptable. Impunity for systemic harms – environmental destruction, labour exploitation, institutional violence – compounds injustice with indifference. Moreover, as Parekh (Parekh 2011, 684–686) argues, victims of structural injustice often demand accountability as a form of recognition. To dismiss these demands as mere *ressentiment*[1] would be philosophically arrogant and politically disabling. Third, the institutions through which we might pursue accountability are themselves embedded in the structures that produce injustice. As Mantouvalou (2023, 11–25) demonstrates, the state often does not merely fail to protect vulnerable populations but actively constructs vulnerability through law. The legal system that would punish is the same system that, through its ordinary operations, legitimizes exploitation and shields powerful actors from consequences. We are trapped: neither retributive punishment nor impunity offers an acceptable path forward.

This paper defends Young's Social Connection Model as the framework that can navigate this dilemma. I make three interconnected contributions. First, I reconstruct

1 In this paper, ressentiment (a term from Nietzsche, adapted by Young) refers to a reactive, blame-seeking moral psychology that demands a guilty individual to punish. As Young (Young 2011, 115) puts it, ressentiment "allows those who lay blame to wallow in the past." Applied to structural injustice – which has no single perpetrator – ressentiment traps us in backward-looking blame rather than forward-looking transformation. The problem is not that victims' demands are illegitimate; the problem is that retributive punishment channels those demands into a search for a culprit that structural injustice lacks.

Young's SCM in its full philosophical architecture (Young 2003a, 2003b, 2004, 2006, 2011). I show how its foundational shift from fault to social position, its five defining features (non-isolating, judging background conditions, forward-looking, shared, and requiring collective action), and its four parameters (power, privilege, interest, and collective ability) provide a coherent alternative to liability thinking. Responsibility, on this view, is not about identifying culpable wrongdoers but about understanding how our social positions implicate us in ongoing structural processes. Second, I demonstrate through contemporary applications that structural injustice generates a demand for accountability which retributive punishment cannot satisfy. Drawing on Mantouvalou's (2023) analysis of state-mediated vulnerability and Parekh's (2011, 2020) work on gender violence and refugee policy, I show how the punishment dilemma emerges in concrete contexts – from precarious workers trapped by employment law to women subjected to violence rooted in cultural norms to asylum seekers punished for irregular entry. Third, I develop the concept of transformative sanction: punitive or quasi-punitive measures whose justification is not backward-looking desert but forward-looking structural disruption. I show how Young's four parameters provide criteria for determining when such sanctions are appropriate and how they should be structured. I operationalize this framework through concrete institutional mechanisms and a practical decision-making matrix, demonstrating that transformative sanction is not merely a theoretical proposal but a usable analytical tool.

The argument proceeds in seven stages. Section II demonstrates why the liability model fails, engaging with extensions by Christopher Kutz, David Miller, and Robert E. Goodin. Section III reconstructs Young's SCM, explaining its five features and four parameters. Section IV draws on Mantouvalou and Parekh to show how structural injustice generates the punishment dilemma. Section V deepens this diagnosis through Young's Nietzsche critique of ressentiment. Section VI develops transformative sanction as a constructive alternative, distinguishes it from restorative and transformative justice, operationalizes it through institutional mechanisms and a decision-making matrix, and defends it against co-optation. Section VII addresses the two strongest objections to Young's framework: Barry and Ferracioli's indeterminacy critique and Nussbaum's motivational challenge. Section VIII concludes by synthesizing the paper's central claims and reorienting the philosophy of punishment toward structural transformation.

Before proceeding, a brief note on methodology is in order. This paper is a work of analytic political philosophy. It proceeds by conceptual analysis (clarifying the concepts of liability, responsibility, and structural injustice), normative argument (defending the SCM as a superior framework), and constructive theorizing (developing the concept of transformative sanction). It draws on the history of philosophy (Arendt, Nietzsche, Jonas, Kant, Hegel) to ground its arguments and on contemporary political theory (Mantouvalou, Parekh, Barry, Ferracioli, Nussbaum) to engage with current debates. It does not aim to provide empirical evidence for its claims, though it draws on empirical examples to illustrate its arguments. Its primary contribution is philosophical: to clarify the logic of accountability for structural injustice and to propose a new framework for thinking about punishment beyond retribution.

This paper demonstrates why liability thinking fails for structural injustice, shows how Young's SCM offers a superior alternative, develops transformative sanction as a way to hold agents accountable without falling back into retributive logic, and establishes how Young's four parameters can guide practical judgment in structural contexts. While further work on institutional design and procedural safeguards is needed, this paper provides a philosophically coherent framework for reimagining justice beyond blame.

II. The Liability Model and Its Limits

Before we can appreciate the necessity of Young's alternative – and, consequently, before we can understand why punishment must be reconceived for structural contexts – we must understand with precision why the established paradigm fails. This failure is not incidental but stems from the liability model's core philosophical commitments, which render it conceptually unequipped to comprehend, let alone address, harms of a structural nature.

II.1. The Philosophical Architecture of Liability

Young defines structural injustice as occurring when "social processes put large groups of persons under systematic threat of domination or deprivation of the means to develop and exercise their capacities, at the same time that these processes enable others to dominate or to have a wide range of opportunities" (Young 2011, 52). This definition highlights three essential characteristics that liability thinking cannot adequately address: the processual nature of harm (it emerges from ongoing social processes rather than discrete events), its systemic character (it affects groups through patterned social positions rather than individuals through unique misfortune), and its normalized operation (it functions through ordinary, often rule-governed conduct rather than exceptional wrongdoing).

The liability model, in contrast, is built upon what we might call the juridical triad[2]: identifiable agents, voluntary wrongdoing, and direct causation. Young makes this structure explicit when she writes that the liability model "seek[s] to identify liable parties for the purposes of sanctioning, punishing, or exacting compensation or redress" (Ibid., 98). Its natural home is the courtroom or the moral assessment of discrete interpersonal wrongs like assault, fraud, or breach of contract. When we apply this triad to structural cases, a profound conceptual mismatch emerges.

Consider the paradigmatic case Young develops of "Sandy," a low-wage single mother

2 This triadic formulation is my synthesis of Young's analysis across Young (2011, 95–122).

facing eviction due to urban redevelopment, unaffordable housing, distant employment, and inadequate transportation (Ibid., 43–48). As Young demonstrates, no individual agent – not the developer, the employer, the landlord, nor the policymaker – acts illegally or with malicious intent. Each operates within the accepted rules of the market and policy, pursuing legitimate goals. The harm Sandy experiences cannot be traced to any single act of voluntary wrongdoing or to a direct causal chain; it is the emergent effect of multiple interdependent processes operating according to their normal logic. This is the core of the mismatch: liability models require linear, traceable causation, but structural injustice operates through diffuse, emergent, non-linear processes. As Young puts it, the actions of particular persons: "do not contribute to injustice for other persons directly, moreover, but rather indirectly, collectively, and cumulatively through the production of structural constraints on the actions of many and privileged opportunities for some" (Ibid., 96). The liability model, in seeking to isolate a culpable perpetrator, fundamentally misidentifies the ontological character of the harm. It attempts to impose an agent-centric, event-based ontology onto what is fundamentally a systemic, processual phenomenon.

This mismatch has direct political consequences. As Young observes, the application of liability thinking to structural contexts tends to produce "blame-switching" (Ibid., 117), a dynamic where agents whose actions contribute to structural processes point the finger at others, and in the ensuing game of moral musical chairs, responsibility evaporates. This political fragmentation flows directly from the philosophical premises of liability: if responsibility requires finding the culpable party, then exonerating oneself becomes the rational response. The model thus systematically undermines the very collective agency needed to address structural problems.

II.2. Extended Liability and Its Failures

Before turning to Young's alternative, we must ask a basic question: can liability thinking be stretched to cover structural harms? Perhaps if we relax its assumptions – about finding a specific wrongdoer, proving they acted voluntarily, or tracing a direct causal chain – liability might still work. Kutz, Miller, and Goodin each offer theories of responsibility for collective harms that might seem to do this. But if we examine their theories as potential extensions of liability, each fails. Why does this matter? Because their failures reveal something important: structural injustice violates the very assumptions that make liability thinking work at all. The conclusion is clear: we do not need a weaker version of liability. We need a completely different model of responsibility.

Kutz on participatory intention. Kutz (Kutz 2000, 74–112) develops the idea of "participatory intention." His argument is simple: when you intentionally join a group project that causes harm, you share responsibility for that harm – even if your own action seems small. This works well for organized groups: a military unit, a corporation, a conspiracy. Everyone involved shares a common plan. But if we try to use Kutz's model as an extension of liability to structural harms, we run into trouble. Consider harms such as climate change and housing shortages. Here, no one planned the harm. No group project exists. Kutz calls these "unstructured collective harms"[3] because they come from millions of uncoordinated, everyday actions. To make his model work for such cases, Kutz invents a "quasi-participatory" relationship based on shared habits (Kutz 2000, 188). But this is a stretch. Structural harms emerge from routine actions – buying goods, renting apartments, driving cars – not from any shared intention. Without a shared plan, there is no "participatory intention." As an extension of liability, Kutz's model fails because it cannot explain why anyone bears responsibility for structural harms (Kutz 2000, 188; Young 2011, 102–103). What this teaches us is that structural injustice does not require shared intentions. Any adequate model must assign responsibility without requiring a shared plan.

Miller on outcome-responsibility. Miller (Miller 2007, 87–90) develops the idea of "outcome-responsibility." Imagine a careful driver who hits a pedestrian. The driver did nothing wrong. They took all precautions. But they still caused harm. Miller says they are "outcome-responsible" for the injury – not because they are to blame, but because their action produced the outcome. This seems promising for structural injustice. Perhaps we can hold people responsible for harmful outcomes without proving they intended harm or acted badly. But if we try to use Miller's model as an extension of liability, we encounter a problem. Miller's model looks backward: it asks who caused a harm that has already happened. It also looks at individuals one by one: it tries to trace each person's contribution to the outcome. This works when one driver hits one pedestrian. But structural harms come from thousands or millions of people acting independently. No single person's action caused any particular victim's eviction or illness. As Young argues, Miller's model "presupposes that we can trace individual contributions to outcomes" – but structural emergence makes this impossible (Young 2011, 104–105). As an extension of liability, Miller's model fails because it cannot handle emergent, forward-looking harms. What this teaches us is that structural injustice is forward-looking (it will continue unless we change it) and emergent (no single cause can be found). Any adequate model must look forward and focus on systems, not isolated individuals.

Goodin on vulnerability. Goodin (Goodin 1985, 109–144) develops a "vulnerability model"[4] of moral duties. His argument is that when one person's interests are vulnerable to another's actions, the powerful person has a special duty to protect the vulnerable one. This captures something true about structural injustice: it does create deep inequalities of power and vulnerability. But if we try to use Goodin's model as an extension of liability to structural injustice, we find a different problem. Goodin's model turns collective

3 Kutz defines unstructured collective harms as those "produced by the aggregate effects of large numbers of individuals acting independently, without any coordinating intention or agreement" (Kutz 2000, 188).

4 Goodin defines vulnerability as "a condition in which one person's interests are subject to damage or diminution by the actions of another" (Goodin 1985, 110). His model argues that vulnerability creates special moral duties of protection.

political problems into private, one-on-one relationships. It asks: who is vulnerable? Who has power over them? What duty does that powerful person have? This shifts attention away from changing unjust systems and toward individual acts of charity or protection. As Young puts it, Goodin's framework tends to "individualize and privatize" collective political problems (Young 2011, 166–168). The state becomes the default protector, which can actually reduce ordinary citizens' sense of political responsibility rather than increasing it. As an extension of liability, Goodin's model fails because it cannot capture the collective, political nature of structural injustice. What this teaches us is that structural injustice is collective and political, not just a collection of individual vulnerabilities. Any adequate model must be collective and political, not private and dyadic.

What all three share. Notice the pattern. Each philosopher offers a theory of responsibility for collective harms. But if we try to use these theories as extensions of liability – as ways to make liability work for structural injustice – each fails. Kutz needs a shared intention that does not exist. Miller needs traceable individual causes that cannot be found. Goodin needs one-on-one relationships where only systems exist. Structural injustice violates all three assumptions that liability requires: identifiable agents, voluntary wrongdoing, and direct causation.

Young draws the right conclusion: "responsibility in relation to structural injustice should not be thought of as an attenuated form of responsibility as complicity" (Young 2011, 103–104). We do not need a weaker version of liability. We need "a different conception of responsibility altogether" (Ibid., 104). That is what the Social Connection Model provides. I now turn to explaining Young's alternative.

III. The Social Connection Model

Young's Social Connection Model (SCM) represents a genuine alternative to liability thinking – not merely a modification but a distinct paradigm built upon different ontological and normative premises. In this section, I reconstruct the model's core architecture, demonstrating its internal coherence and its superiority for conceptualizing responsibility in structural contexts.

III.1. The Foundational Shift: From Fault to Position

The foundational move of the SCM is to decouple responsibility from fault and recouple it with social position. Young articulates this pivot with crystalline clarity:

> Individuals bear responsibility for structural injustice because they contribute by their actions to the processes that produce unjust outcomes. Our responsibility derives from belonging together with others in a system of interdependent processes of cooperation and competition through which we seek benefits and aim to realize projects (Young 2011, 105).

This formulation accomplishes several crucial philosophical moves simultaneously.

First, it establishes that the basis of responsibility is not culpable action but participation in structural processes. One is responsible not because one has done something wrong, but because one is constitutively situated within webs of social interdependence that produce unjust outcomes. This represents a shift from a morality of intention to a politics of position. Second, it universalizes the scope of responsibility beyond specific relationships to encompass anyone implicated in the relevant structural processes. Third, it reorients the temporal direction of responsibility: the obligation is not to account for the past but to transform the future.

This reorientation resonates with Hans Jonas's argument in *The Imperative of Responsibility*, where the cumulative, long-term effects of collective technological actions call for a new ethics suited to technological civilization (Jonas 1984, Ch. 1–2). Young draws on this forward-looking orientation to clarify that political responsibility is concerned with preventing future injustice rather than assigning liability for past wrongdoing (Young 2011, 108). However, Young advances beyond Jonas by providing a social theoretical framework that specifies how we are connected. Where Jonas issues a general imperative of responsibility for the future of humanity, Young offers a structural map of specific social connections – through labour markets, trade regimes, legal systems – that allows us to determine who bears responsibility for what. This move from ontological urgency to structural specificity is what grants the SCM its political traction.

III.2. The Five Features: A Systematic Alternative

Young elaborates the SCM through five interrelated features that systematically counter the limitations of liability thinking (Young 2011, 105–117).

First, the SCM is **non-isolating**. The liability model operates through a logic of separation: it seeks to identify and isolate specific wrongdoers, thereby distinguishing them from – and implicitly exonerating – all others. As Young observes, drawing on Hannah Arendt,[5] "the purpose of finding guilt is to focus on what particular agents have done in their singularity" (Ibid., 105). In direct contrast, the SCM implicates "all the persons who participate by their actions in the ongoing schemes of cooperation that constitute these structures" (Ibid., 106). They are implicated, as Young stresses, "without being specifically at fault" (Ibid., 106). Young rejects the isolating logic because it is practically incoherent in structural contexts – where none of us can trace our individual actions to specific aspects of the injustice – and because it is politically corrosive, allowing other implicated parties to be absolved while triggering a paralyzing blame-switching dynamic (Ibid., 117). The SCM, therefore, replaces the search for the guilty with an acknowledgment of shared implication.

Second, the SCM requires that we **judge background conditions**. Liability

5 Arendt distinguishes between guilt and political responsibility. Guilt is personal, backward-looking, and requires an identifiable perpetrator. Political responsibility, by contrast, is collective, forward-looking, and derives from membership in a political community. Young adapts this distinction in developing the Social Connection Model (see Arendt 1963; Arendt 1998).

assesses actions against background rules it assumes as normatively neutral. A wrong is conceived as “a specific deviation from an acceptable baseline” (Ibid., 107). The SCM, by contrast, demands we subject the “normal and accepted background conditions of action” themselves to critical scrutiny (Ibid., 107). When we judge that structural injustice exists, we are saying that at least some of the normal background conditions are not morally acceptable. Most of us contribute to structural injustice precisely because we follow accepted rules and conventions, often habitually, without explicit reflection on their wider implications (Ibid., 107). This represents a fundamental reversal: where liability takes structure as given, political responsibility interrogates the structure’s justice.

Third, the SCM is **forward-looking**. This feature marks the decisive temporal reorientation from the liability model. While liability practices may sometimes look to deter future acts, their “primary purpose (...) is backward-looking” (Ibid., 108). The SCM, by contrast, assigns responsibility for structural injustice that “has existed recently, is ongoing, and is likely to persist unless social processes change” (Ibid., 109). The point is not to compensate for the past, but for all who contribute to processes that produce unjust outcomes to work to transform those processes. This forward-looking imperative means responsibility “cannot be discharged by isolated individuals but requires collective action” aimed at altering future trajectories (Ibid., 109).

Fourth, the SCM is **shared**. From the premise that the SCM does not isolate liable parties in ways that implicitly absolve others, it follows that all who contribute to structural processes share responsibility for those harms. As Young clarifies:

> As I understand it, shared responsibility is a responsibility I personally bear, but I do not bear it alone. I bear it in the awareness that others bear it with me; acknowledgment of my responsibility is also acknowledgment of the inchoate collective of which I am a part, which together produces injustice (Young 2011, 110).

Young’s point is this: responsibility is personal (I bear it) but not solitary (I do not bear it alone). When I acknowledge my responsibility, I necessarily acknowledge that I belong to an ‘inchoate collective’ – a group of people who, through their interconnected social positions, together produce structural injustice. This collective is ‘inchoate’ (meaning ‘just beginning’ or ‘not yet fully formed’) because it is not yet organized as a political actor, but it exists as a matter of social fact. For example, all consumers who participate in an exploitative supply chain form an inchoate collective: they are not organized, but they share structural responsibility.

Fifth, this responsibility **requires collective action**. This feature follows from the essentially shared nature of the responsibility. Young argues that forward-looking responsibility “can be discharged only through collective action” aimed at transforming structures (Ibid., 111). Where thousands or millions contribute to structural processes, no individual can act alone to change those processes. As Young observes, a homeowner concerned about housing injustice cannot alone alter property values or incentive structures; such processes “can be altered only if many actors from diverse positions within the social structures work together to intervene in them” (Ibid., 111). This reorients responsibility from the private realm of individual moral conscience to the public realm of political engagement.

An important corollary follows: victims of structural injustice also share responsibility for its transformation. Young argues that those in less advantaged positions “perhaps should take the lead in organizing and proposing remedies for injustice, because their interests are most acutely at stake” and because their social position provides “a unique understanding of the nature of the problems” (Ibid., 113). This is not to blame victims but to recognize that structural transformation requires the epistemic and political agency of those most affected.

These five features – non-isolating, judging background conditions, forward-looking, shared, and requiring collective action – define the conceptual architecture of the Social Connection Model. However, they do not yet tell us how to apply the model in practice. To determine who bears how much responsibility and what specific actions they should take, Young introduces four parameters that weigh different agents’ responsibilities. I now turn to these parameters.

III.3. The Four Parameters: From Philosophical Framework to Practical Guidance

A persistent criticism of forward-looking, shared responsibility models is that they lack determinacy. Young anticipates this objection with four parameters that gauge the relative weight of an agent’s forward-looking responsibility (Young 2011, 144–148; 2003b, 16–19).

Power refers to an agent’s relative capacity to influence the structural processes in question. Young writes: “[Power] refers to the degree to which an agent is in a position to influence the processes that produce the outcomes” (Ibid., 144). A multinational corporation’s power to alter its global supply chain practices is categorically different from an individual consumer’s power to choose among products.

Privilege denotes the benefits an agent accrues from the structural injustice. Young observes that “persons and institutions that are relatively privileged within structural processes have greater responsibilities than others to take actions to undermine injustice” (Ibid., 145). The question is not whether one chose to benefit, but whether one’s position within the structure confers unearned advantages that create a special responsibility to transform the conditions that produce them.

Interest refers to the stake an agent has in transforming the structures. Young writes: “Victims of structural injustice in particular have unique interests in undermining injustice, and they ought to take responsibility for doing so” (Ibid., 145). Their interest carries epistemic and political priority in defining the injustice and shaping solutions.

Collective ability concerns the access an agent has to organized means of collective action. Young defines this as “the availability of collective forms of action that can effectively aim to affect the processes" (Ibid., 146). Unions, advocacy networks, and professional associations can amplify individual efficacy. These parameters do not function as a mechanical formula but as a framework for practical-political judgment. They interact dynamically. Young explains: "Such parameters ought to respond to the intuition that different agents properly have different kinds and degrees of forward-looking responsibility for justice" (Ibid., 144). A European state, for instance, bears heavy responsibility for refugee justice due to its power (geopolitical influence, resources), its privilege (benefits from the global economic order that produces displacement), and its collective ability (membership in the EU, treaty-making capacity).[6] An individual citizen of that state bears responsibility through different parameters: privilege (enjoying the security of citizenship), some power (as voter, consumer), and collective ability (through supporting refugee rights organizations). This graduated, contextual approach sustains the model's universal implication while providing the differentiation needed to guide action.

Having established the philosophical architecture of the SCM in Section III, I now turn to its application. Section IV demonstrates how the SCM illuminates concrete structural injustices and, crucially, reveals why criminal punishment cannot serve as the primary mechanism of accountability in structural contexts.

IV. Contemporary Applications: From Structural Diagnosis to the Punishment Dilemma

The SCM's philosophical architecture provides a powerful framework for analysing structural injustice. Yet its true test lies in its capacity to illuminate concrete problems and, crucially for our purposes, to reveal why punishment cannot be understood within the liability model's terms. This section examines two significant contemporary applications of the SCM: Mantouvalou's (2023) analysis of the state as a constitutive agent of injustice and Parekh's (2011, 2020) work on gender violence and refugee policy. I argue that each application, when pressed, reveals a central tension: structural injustice demands accountability, yet the mechanisms of criminal punishment seem ill-suited to address it. This tension is not a failure of the SCM but a feature that points toward the need to reconceptualize punishment itself.

IV.1. Mantouvalou: The State as Architect of Vulnerability

Although Young's SCM powerfully captures the diffuse production of structural injustice, her framework tends to position the state primarily as a coordinator or remedy-

6 For a detailed application of Young's parameters to European state responsibility in the refugee context, see Parekh (Parekh 2020, 165–175), who argues that "wealthy, powerful states that disproportionately benefit from the global political and economic order bear the greatest forward-looking responsibility for providing refuge and leading systemic reform" (see also Young 2011, 144–148 for the general formulation of the parameters).

provider (Young 2011, 166–168).[7] Mantouvalou's work, particularly in *Structural Injustice and Workers' Rights*, forces a decisive correction to this view. She reconstructs the SCM to demonstrate that the state is frequently not a passive responder to injustice but its active architect through law (Mantouvalou 2023, 11–25). Central to this intervention is Mantouvalou's concept of "state-mediated structures of injustice" (Ibid., 21; 2024, 250).[8] She defines these as exploitative patterns that arise when the state enacts legal rules possessing an "appearance of legitimacy" which nevertheless "increase workers' vulnerability that is then systematically exploited, primarily by private actors" (Ibid., 21). This represents a crucial theoretical refinement: law is reconceived from a neutral backdrop into the primary mechanism that produces the terrain of vulnerability. The state, in this view, does not merely "hold the ring" for market actors; it designs the ring itself, determining its shape and the vulnerabilities inherent within it.

Mantouvalou's analysis of the "Marcell" case – a lawful agency worker in the UK trapped in precarity by the combined force of employment law and welfare conditionality – vividly illustrates this constitutive role (Ibid., 18–21). In this scenario, no individual employer commits a legal wrong; the exploitation is enabled by the legal framework itself. Mantouvalou holds that state authorities "know or ought to know of the vulnerability that they create, perpetuate, and increase" (Ibid., 21).[9] This formulation is pivotal. It adapts a foreseeability criterion to the state's constitutive role, establishing a ground for political responsibility for the foreseeable injustices baked into legal design.

This analysis generates a profound tension for punishment. If the state constructs vulnerability through law, can the state's courts legitimately punish individuals who act within the structures the state itself has created? Consider a precarious worker who, facing eviction due to legal regimes that enable exploitative housing practices, commits a crime of survival. The state is simultaneously the architect of the conditions that produced the act and the institution that punishes the actor. More fundamentally, if structural injustice is produced through the ordinary operation of law – law that appears legitimate, that follows proper procedures, that is enacted by democratically accountable institutions

7 Young discusses the state's role in coordination and remediation, arguing that "political responsibility aims to bring about structural change," often through state action, but does not centrally theorize the state as a primary structural cause (Young 2011, 166–168).

8 Mantouvalou's concept of "state-mediated structures of injustice" captures a specific form of structural injustice: the state creates legal rules that appear legitimate but systematically increase workers' vulnerability to exploitation. Private employers then exploit this vulnerability, often acting within the law. The state does not directly cause the exploitation, but it is responsible for constructing the legal framework that enables it. Unlike isolated cases of employer misconduct, these structures are widespread, routine, and entrenched. Because the state has the power to change the law, it bears a special responsibility for remedying such injustice (Mantouvalou 2023, 21–24).

9 The phrase strategically adapts a foreseeability standard from tort law – where agents are held responsible for harms they could reasonably anticipate – to hold the state politically accountable for the predictable effects of its legislative choices. This move bridges the liability model's retrospective logic (foreseeability) with the SCM's forward-looking political responsibility, creating a hybrid accountability framework that Mantouvalou argues is necessary for state-mediated structural injustice.

– then the criminal law becomes not a neutral arbiter of wrongdoing but a mechanism that reinforces the very structures it purports to adjudicate. The punishment dilemma emerges with clarity: if we punish, we risk legitimizing structures of vulnerability; if we do not punish, we risk impunity for harms that demand accountability.

IV.2. Parekh: Structural Violence and the Limits of Criminal Justice

Parekh's work offers a parallel intervention that sharpens this dilemma. In her philosophical analysis of gender violence, Parekh challenges the dominant liberal-legal framing that views such violence as a series of discrete, private crimes (Parekh 2011, 673). Drawing explicitly on Young, she reconceives it as a structural injustice rooted in "unquestioned norms, habits, and symbols" that systematically subordinate women (Parekh 2011, 676; Young 1990, 41). This structural diagnosis directly confronts theorists like Thomas Pogge, who restrict human rights violations to "official" state acts, thereby rendering pervasive private violence politically unintelligible within a human rights framework (Parekh 2011, 674; Pogge 2001, 192).[10]

Parekh's central argument is that Young's notion of forward-looking political responsibility provides the philosophical foundation for holding states accountable for sustaining the background conditions that make gender violence a predictable social outcome, even when the immediate perpetrators are private actors (Parekh 2011, 677–679). She demonstrates that this reconception of responsibility is not merely theoretical but is embodied in international law, specifically Article 5 of the Convention on the Elimination of All Forms of Discrimination Against Women (CEDAW), which obligates states to "modify the social and cultural patterns of conduct" that perpetuate gender hierarchy (Parekh 2011, 675).[11]

Yet this analysis, like Mantouvalou's (2023), generates a tension for criminal punishment. When violence against women is structural – rooted in norms and symbols that the state has a duty to transform – criminal prosecution of individual perpetrators, while necessary, cannot alone address the injustice. Moreover, a criminal justice system that prosecutes individual acts of violence while leaving intact the norms and institutions that produce systematic vulnerability may actually reinforce structural injustice. It individualizes responsibility, locating the harm in the aberrant perpetrator rather than in the background conditions that make such violence predictable. The system punishes the individual while exonerating the structure – including the state's own failure to discharge its forward-looking responsibility to transform the norms that generate violence.

10 Parekh argues that the legalistic approach "fails to capture the systemic nature of the problem" (Parekh 2011, 673). Her critique of Pogge is that his institutional model lacks the conceptual tools to assign responsibility for harms produced through diffuse social structures and private conduct, thereby making them "invisible" as human rights violations (Ibid., 674).

11 Parekh cites CEDAW Article 5(a) as a legal instantiation of the duty to address structural injustice by modifying the underlying social and cultural patterns (Parekh 2011, 675). For the full text, see *Convention on the Elimination of All Forms of Discrimination Against Women*, opened for signature Dec. 18, 1979, 1249 U.N.T.S. 13 (entered into force Sept. 3, 1981).

In *No Refuge*, Parekh applies the same structural lens to the global refugee regime, demonstrating how refugee harms result from "the cumulative outcome of many uncoordinated norms and decisions by different states and the UNHCR" (Parekh 2020, 165). She employs Young's parameters of power, privilege, and interest to argue that wealthy, powerful states that disproportionately benefit from the global political and economic order bear the greatest forward-looking responsibility for providing refuge and leading systemic reform (Parekh 2020, 171).[12] Yet here too, the punishment dilemma emerges: if states are collectively responsible for producing displacement through their policies, can individual asylum seekers be punished for irregular entry? If the structure itself is unjust, what does it mean to enforce its boundaries through criminal sanctions? In both cases – labour exploitation and refugee displacement – the same pattern emerges: the state is implicated in producing the conditions that make harm predictable, yet the criminal law is deployed to punish individuals who act within those conditions. The dilemma is not a flaw in the SCM but a structural feature of accountability under conditions of injustice.

The global dimension of structural injustice deserves more attention than this paper can provide. Young (2006; 2011) herself argued that the SCM applies globally – that wealthy nations bear responsibility for global structural injustice through their participation in trade regimes, financial institutions, and migration policies. Parekh (2020) develops this argument in the refugee context. Mantouvalou (2023) focuses on domestic labour exploitation but acknowledges its global dimensions. The punishment dilemma is even more acute globally: there are no global courts with the authority to impose transformative sanctions on powerful states or corporations, and the institutions that might pursue accountability are themselves embedded in the structures that produce global injustice. The framework developed in this paper applies globally, but its institutional implementation requires further work.

IV.3. The Dilemma Articulated

Together, the interventions of Mantouvalou and Parekh accomplish two crucial theoretical tasks. First, they demonstrate the SCM's robust analytical power across disparate domains – from the legal architecture of labour markets to the cultural norms of gender and the international system of sovereignty. Second, and more importantly for our purposes, they generate the punishment dilemma with philosophical precision. Let me articulate this dilemma explicitly.

The SCM teaches us that structural injustice is produced through the ordinary, normalized operations of social structures, including legal systems (Young 2011, 43–52). When we respond to structural injustice by punishing individuals, we risk:

12 Parekh explicitly uses Young's parameters, arguing that "wealthy and powerful states (…) have a greater responsibility because they have greater power to effect change and because they benefit from the global order that produces refugees" (Parekh 2020, 171).

Risk	Explanation
(a) Individualizing responsibility	Punishment focuses on discrete actors, obscuring the collective, structural nature of the harm (Young 2011, 105–106)
(b) Legitimizing unjust structures	Treating structural operations as normatively neutral while punishing only “deviant” actors (Young 2011, 107–108)
(c) Exonerating the state	Focusing on private actors while ignoring the state’s constitutive role in constructing vulnerability (Mantouvalou 2023, 18–25)
(d) Perpetuating impunity	Powerful actors whose actions, though lawful, reproduce injustice face no consequences (Young 2011, 148–149; Mantouvalou 2023, 157–160)

Table 1: Risks of responding to structural injustice through individual punishment, as derived from Young’s Social Connection Model (SCM).

Yet the alternative – abandoning punishment entirely – is equally untenable. Impunity for systemic harms – environmental destruction, labour exploitation, institutional violence, refugee displacement – communicates that certain forms of suffering do not matter, that certain perpetrators are beyond reach, that certain victims are not entitled to redress (Arendt 1994, 126–128). As Young argues, following Arendt, the failure to hold accountable is itself a form of world-building, a way of constituting what matters and who counts (Young 2011, 91–94; Arendt 1994, 126–128). Moreover, victims of structural injustice often demand accountability as a form of recognition, and to dismiss these demands is to ignore the epistemic and political priority of those most affected (Parekh 2011, 684–686; Mantouvalou 2023, 179–182). To dismiss such demands for accountability would not only be philosophically arrogant but also politically disabling.

The punishment dilemma, then, is this: How can we hold agents accountable for structural injustice without falling back into the liability model’s individualistic, backward-looking logic, yet without abandoning accountability altogether? The SCM’s resources – its forward-looking orientation, its parameters for differentiated responsibility, its emphasis on collective action – seem to rule out retributive punishment. Yet the demand for accountability persists.

Before we can determine how to navigate this dilemma, we must first understand with precision why one of its horns – retributive punishment – is so deeply mismatched to structural injustice. This is not merely a practical problem of application; it is a conceptual problem rooted in the very logic of retributive justice. Section V diagnoses this conceptual mismatch. Section VI then builds on this diagnosis to develop transformative sanction as the constructive alternative.

IV.4. The SCM in Practice: A Case Illustration

To see why the punishment dilemma is not merely theoretical, consider the case of Marcell, the agency worker Mantouvalou (2023, 18–21) analyses. Marcell is a lawful worker in the UK, yet he is trapped in precarity by the combined force of employment law and welfare conditionality. No individual employer acts illegally; no single policy is obviously unjust. Yet the structure produces predictable outcomes: workers like Marcell face housing insecurity, food poverty, and chronic stress.

The SCM illuminates this case with precision:

• **Power:** The state holds the power to redesign employment law and welfare conditionality. Employers hold power over working conditions.

• **Privilege:** Employers and shareholders benefit from the flexibility and low costs enabled by precarious labour. The state benefits from reduced welfare expenditure.

• **Interest:** Marcell and workers like him have the clearest interest in structural transformation.

• **Collective ability:** Unions and advocacy organizations can amplify workers' voices and demand reform.

Yet here the punishment dilemma emerges: if no individual employer has acted illegally, retributive punishment cannot be applied. But impunity is unacceptable – the structure produces predictable suffering. Transformative sanction, as developed below, offers a third path: structural injunctions requiring changes to employment practices, reparative obligations requiring employers to fund worker protections, and public reporting requirements that enable ongoing collective oversight.

This case illustrates that the punishment dilemma is not a philosopher's puzzle but a lived reality for millions of workers. It also demonstrates that the SCM provides the analytical tools to diagnose structural injustice and identify appropriate accountability mechanisms.

V. Why Retributive Punishment Fails Structural Injustice

The applications examined in Section IV reveal a persistent tension: victims and communities demand accountability, yet the dominant framework for accountability in modern legal systems – retributive punishment – seems ill-suited to structural harm. In this section, I deepen this diagnosis by examining why retributive punishment is constitutively mismatched to structural injustice. I argue that the mismatch is not merely practical (retributive punishment is difficult to apply) but conceptual (retributive punishment presupposes conditions that structural injustice systematically violates).

V.1. The Logic of Retributive Punishment

Retributive punishment, as developed in the philosophical tradition, is built upon three interconnected assumptions that I have called the juridical triad: (1) identifiable agents, (2) voluntary wrongdoing, and (3) direct causation. Each of these assumptions is essential for retributive justice to function coherently.

First, **identifiable agents**. Retributive punishment requires a specific person or persons whose actions can be isolated and evaluated. As Kant (Kant 1797/1996, 6:331) argues, punishment can only be imposed on someone who "has committed a crime." Without an identifiable perpetrator, there is no one to punish. This is why criminal law operates through individual defendants, not abstract entities. Second, **voluntary wrongdoing**. Retributive punishment presupposes that the agent acted voluntarily, with some degree of culpable intent (*mens rea*). Hegel (Hegel 1821/1991, 126) grounds punishment in the need to "annul the crime," implying that the crime is a voluntary act that negates the right. Even Hart (Hart 1968, 22), who offers a more complex pluralist account, acknowledges that "the principle that punishment should be restricted to those who have voluntarily broken the law" is a constraint any adequate theory must respect. Third, **direct causation**. Retributive punishment assumes a traceable causal chain from the agent's action to the harmful outcome. The punishment is meant to be proportional to the harm caused. If the causal link is attenuated or indirect, the justification for punishment weakens. These three assumptions are perfectly appropriate for discrete interpersonal wrongs: assault, theft, fraud, breach of contract. In such cases, there is a clear perpetrator, a voluntary act, and a direct causal link. Retributive punishment answers the question "what does this agent deserve?" with coherence. The problem is that structural injustice violates all three assumptions simultaneously.

Consider again Young's paradigmatic case of "Sandy," a low-wage single mother facing eviction due to the combined effects of urban redevelopment, unaffordable housing, distant employment, and inadequate transportation (Young 2011, 43–48). No individual agent – not the developer, the employer, the landlord, nor the policymaker – acts illegally or with malicious intent. Each operates within accepted rules pursuing legitimate goals. The harm Sandy experiences cannot be traced to any single voluntary act or direct causal chain. It is the emergent effect of multiple interdependent processes operating according to their normal logic.

To ask 'what does any particular agent deserve?' in this context is to ask a question that has no coherent answer. There is no identifiable perpetrator whose voluntary act directly caused Sandy's eviction. The liability model's juridical triad does not merely face practical difficulties here; it is conceptually mismatched to the phenomenon it purports to address.

V.2. Ressentiment and the Politics of Blame

One might respond: 'Even if retributive punishment is conceptually mismatched, victims still demand punishment. Should we not honor those demands?' This objection has force, but it requires careful examination of what victims actually demand and why.

Young's engagement with Nietzsche's (Nietzsche 1887/1967) *On the Genealogy of Morals* illuminates why the demand for punishment can be politically problematic when applied to structural harms. Nietzsche diagnoses ressentiment as a characteristic pathology in moral responses to suffering: the reactive demand for a powerful agent upon whom to vent affects, the insistence on equivalence between harm and punishment. As Young summarizes, the ressentiment-driven response seeks "some living thing upon which he can (...) vent his affects" (Young 2011, 114; Nietzsche 1967, 127). The problem with ressentiment is not merely that it misidentifies causes – though it often does – but that it traps us in a debilitating temporal orientation. It "allows those who lay blame to wallow in the past" (Young 2011, 115). Those caught in this dynamic, in Nietzsche's haunting phrase, "tear open their oldest wounds, they bleed from long-healed scars" (Nietzsche 1967, 127–128). Ressentiment looks backward, seeks a perpetrator, and demands equivalence. It is reactive rather than creative, vengeful rather than transformative.

A more expansive spirit, Young argues, "will affirm that equivalence cannot be found for every harm and that going forward into a transformed future sustains greater power than reinscribing the relationships of the past" (Young 2011, 115). This is not to dismiss victims' demands for accountability. On the contrary, it is to take those demands seriously enough to ask: What form of accountability actually serves victims' interests? Does backward-looking punishment help Sandy find housing? Does imprisoning a single officer prevent future police misconduct? Does retribution address the structural conditions that produced the harm?

When applied to structural harms, retributive punishment risks becoming an expression of ressentiment – a demand for a perpetrator who can be made to suffer, a backward-looking fixation that forecloses the forward-looking orientation that structural transformation requires. The demand for equivalence cannot be satisfied because the harm is not equivalent to any individual act. The search for a perpetrator cannot succeed because the harm is not traceable to any individual agent.

However, Young carefully notes that blame and fault-finding "arguably are even appropriate in some political contexts – where a public official has made a disastrous decision without proper information and deliberation, for example, or when actions producing a crime come through a bureaucratic chain of command" (Ibid., 115–116). The problem is not blame as such. The problem is the misapplication of a blame-oriented framework to harms that are "a repeated product of structural social processes" (Ibid., 116). For such harms, we need a different logic of accountability – one that is forward-looking, structural, and collective rather than backward-looking, individual, and retributive.

V.3. The Punishment Dilemma Sharpened

We are now in a position to articulate the punishment dilemma with philosophical precision. The dilemma consists of three irreducible claims that cannot be simultaneously maintained:

First, retributive punishment is conceptually mismatched to structural injustice. Because structural injustice lacks identifiable agents, voluntary wrongdoing, and direct causation, the retributive question “what does this agent deserve?” has no coherent answer. Punishing individuals for structural harms either scapegoats low-level actors (while leaving power holders untouched) or misidentifies the nature of the harm altogether. As Young (2011, 95) argues, structural injustice “cannot be reduced to discrete wrongs committed by identifiable agents.”

Second, abandoning accountability altogether is unacceptable. Impunity for systemic harms – environmental destruction, labour exploitation, institutional violence, refugee displacement – compounds injustice with indifference. Moreover, as Parekh (2011, 684–686) argues, victims of structural injustice often demand accountability as a form of recognition. To dismiss these demands as mere ressentiment would be philosophically arrogant and politically disabling, as argued in Section I. Victims are not simply bitter; they are demanding that their suffering be acknowledged and that those who benefit from unjust structures be held to account.

Third, the institutions through which we might pursue accountability are themselves embedded in the structures that produce injustice. The criminal law that would punish is the same system that, through its ordinary operations, legitimizes exploitation and shields powerful actors from consequences. As Mantouvalou (2023, 18–25) demonstrates, the state often constructs vulnerability through law itself. To expect the criminal justice system to remedy structural injustice is to expect the source of the problem to be the solution.

The dilemma, then, is this: How can we hold agents accountable for structural injustice without falling back into the liability model’s individualistic, backward-looking logic, yet without abandoning accountability altogether?

If retributive punishment is conceptually mismatched, and impunity is morally intolerable, then we require a third path – a form of accountability that is neither retributive (backward-looking, individual, blame-oriented) nor impunity (no accountability at all). The following section develops transformative sanction as precisely such a third path: punitive or quasi-punitive measures whose justification is not backward-looking desert but forward-looking structural disruption.

VI. Transformative Sanction: An SCM Approach to Accountability

I have argued that retributive punishment is constitutively mismatched to structural injustice, yet impunity is morally intolerable. In this section, I develop a constructive alternative: transformative sanction.

VI.1. Defining Transformative Sanction

I define transformative sanction as punitive or quasi-punitive measures whose primary justification is not backward-looking desert but forward-looking structural disruption.

To understand what this means, compare transformative sanction with retributive punishment across three dimensions:

Dimension	**Retributive Punishment**	**Transformative Sanction**
Justificatory structure	What does this agent deserve?	What intervention will disrupt unjust structures?
Temporal orientation	Backward-looking (annul past wrong)	Forward-looking (alter future trajectories)
Institutional logic	Logic of equivalence (proportionate suffering)	Logic of structural disruption (shift power, alter incentives, enable collective action)

Table 2: Retributive punishment and transformative sanction compared across three dimensions.

The key distinction is this: retributive punishment asks 'Who is to blame?' and seeks to impose suffering proportional to the wrongdoing. Transformative sanction asks 'What is needed to change the structure?' and seeks to intervene at the level of systems, not merely individuals.

Consider an example. In 2013, a federal court found that the New York Police Department's stop-and-frisk policy violated the constitutional rights of hundreds of thousands of New Yorkers.[13] The court could have responded by imprisoning individual officers who conducted the stops. That would be retributive punishment: it asks what each officer deserves and imposes individual suffering. Instead, the court ordered structural reforms: it required the NYPD to change its training, supervision, and monitoring practices. That is a transformative sanction: it targets the structure that produced the harm, not the individuals who operated within it. This example clarifies three features of transformative sanction. First, it targets power: the court's order applied to the institution (the NYPD) and its practices, not low-level officers. Second, it disrupts privilege: the order dismantled a policy that disproportionately affected marginalized communities while leaving those with positional authority accountable for reform. Third, it enables collective ability: the court's monitoring requirements created public documentation that community organizations could use to demand ongoing accountability.

Transformative sanction differs from consequentialist justifications of punishment. Consequentialism asks whether punishing an individual will deter future crimes or reform the offender's character. Transformative sanction does not focus on the individual criminal at all. Its aim is not to make the sanctioned agent a better person but to disrupt the structural conditions that produce and reproduce injustice. This orientation follows directly from Young's Social Connection Model. As Young (2011, 95) argues, structural

13 *Floyd v. City of New York*, 959 F. Supp. 2d 540 (S.D.N.Y. 2013). In this case, the court held that the NYPD's stop-and-frisk policy violated the Fourth and Fourteenth Amendments and mandated structural reforms including the appointment of an independent monitor, retraining of officers, policy changes, and a joint remedial process. Individual officers were not criminally punished.

injustice "cannot be reduced to discrete wrongs committed by identifiable agents" and therefore requires intervention at the level of structures, not merely individual behaviour.

With this definition in hand, I now distinguish transformative sanction from related frameworks – restorative and transformative justice – before operationalizing it through Young's four parameters.

VI.2. Distinguishing Transformative Sanction from Restorative and Transformative Justice

To avoid conceptual confusion, it is necessary to distinguish transformative sanction from two related but distinct frameworks: restorative justice and transformative justice.

Restorative justice, as developed by Braithwaite (Braithwaite 2002) and Zehr (Zehr 2015), focuses on repairing harm through inclusive processes that bring together offenders, victims, and communities. Its central mechanism is dialogue aimed at acknowledgment, apology, and restitution. While restorative justice shares transformative sanction's rejection of purely retributive logic, it differs in three crucial respects. First, restorative justice typically presupposes identifiable offenders who accept responsibility for discrete harms. Transformative sanction addresses structural contexts where no identifiable offender exists. Second, restorative justice emphasizes voluntary participation and consensus, while transformative sanction may involve coercive measures imposed on powerful actors who refuse to engage. Third, restorative justice's primary aim is repairing relationships between specific parties, whereas transformative sanction aims to disrupt structural conditions that produce systematic harm.

Transformative justice, emerging from grassroots responses to interpersonal violence within marginalized communities, emphasizes community accountability, survivor safety, and addressing root causes of harm (Kaba 2021; Kim 2018). It shares transformative sanction's structural orientation and rejection of state-centric punishment. However, transformative justice typically operates outside state institutions and focuses on interpersonal rather than systemic harm. Transformative sanction, while also critical of state carcerality, does not rule out state-mediated accountability measures provided they satisfy the four parameters. Moreover, transformative sanction specifically targets the structural position of powerful agents, whereas transformative justice often emphasizes horizontal accountability among community members. Table 3 summarizes these distinctions.

Dimension	Retributive Punishment	Restorative Justice	Transformative Justice	Transformative Sanction
Primary aim	Desert/harm annulment	Repair harm	Address root causes	Disrupt unjust structures
Target	Individual offender	Offender-victim dyad	Community relationships	Structural position of power

Dimension	Retributive Punishment	Restorative Justice	Transformative Justice	Transformative Sanction
Mechanism	State-imposed suffering	Facilitated dialogue	Community accountability	Variable (may include state)
Identifiable wrongdoer required	Yes	Yes	Sometimes	No
Coercion	Yes	No (voluntary)	Variable	Yes (when necessary)
Temporal orientation	Backward-looking	Past/present	Present/future	Forward-looking

Table 3: Comparative framework of four accountability paradigms: retributive punishment, restorative justice, transformative justice, and transformative sanction.

To make the distinctiveness of transformative sanction explicit, three contrasts are worth emphasizing. First, unlike retributive punishment, transformative sanction does not ask 'what does this agent deserve?' but 'what intervention will disrupt unjust structures?' Retributive punishment is backward-looking and individualistic. Transformative sanction is forward-looking and structural. Second, unlike restorative justice, transformative sanction does not presuppose identifiable offenders willing to accept responsibility for discrete harms. Restorative justice operates through dialogue between specific victims and specific offenders. Transformative sanction operates in structural contexts where no single perpetrator exists and where harm emerges from diffuse, normalized social processes. Third, unlike transformative justice, transformative sanction does not rule out state-mediated accountability measures, provided such measures satisfy Young's four parameters. Transformative justice often rejects state institutions entirely and seeks community-based alternatives. Transformative sanction is more pragmatic: it accepts state-mediated measures (such as structural injunctions, consent decrees, or disqualification) when they target power, disrupt privilege, center affected communities, and enable collective mobilization.

This combination of features – forward-looking orientation, structural focus, openness to coercive measures when necessary, and explicit grounding in Young's parameters – positions transformative sanction as a genuinely novel framework. It is not a variant of retributive punishment, a subset of restorative justice, or an instance of transformative justice. It occupies its own conceptual space, answering a question that none of the other frameworks adequately addresses: how to hold agents accountable for structural injustice without falling back into individualistic blame or abandoning accountability altogether.

Transformative sanction thus occupies a distinctive conceptual space: it shares both restorative and transformative justice's rejection of retributivism but differs in its focus on structural disruption, its willingness to employ coercion against powerful actors, and its explicit grounding in Young's parameters for evaluating accountability mechanisms.

VI.3. Young's Four Parameters as Criteria

Young's four parameters provide a framework for determining when transformative sanction is appropriate and how it should be structured (Young 2011, 144–148).

Power: The first question is whether sanctioning a particular agent can disrupt unjust structures. Transformative sanction should target agents whose position gives them the capacity to maintain or change the structure (Ibid., 144). When corporate executives shape environmental practices across an industry, sanctioning them can alter corporate behaviour. By contrast, sanctioning low-level actors may do little to disrupt structures; it may even reinforce injustice by providing scapegoats while leaving power holders untouched.

Privilege: The second question is whether impunity protects those who benefit from injustice. Young (Ibid., 145) defines privilege as "the fact that many people have a relative benefit in the structure." Those who benefit from unjust structures are not necessarily wrongdoers, but their benefit creates heightened responsibility. Transformative sanction asks whether impunity allows beneficiaries to continue enjoying unearned advantage without accountability. When corporations benefit from exploitative labor practices or wealthy nations benefit from a global order that produces displacement, impunity protects structural privilege.

Interest: The third question is whether affected communities seek accountability. Young (Ibid., 145–146) argues that those most subjected to injustice have epistemic and political priority in defining what accountability means. The question for transformative sanction is whether accountability mechanisms respect the perspectives and priorities of those most affected. In contexts of environmental harm, affected communities often prioritize prevention, restoration, and institutional reform over punishment of individual wrongdoers. Transformative sanction must honor such priorities.

Collective ability: The fourth question is whether sanction can catalyse broader collective action. Individual sanctions, however justified, are politically insignificant without collective organization (Ibid., 146). Transformative sanction should be evaluated not only by its direct effects on sanctioned agents but by its capacity to enable collective mobilization. Public accountability processes that generate testimony, create records, and enable victims to speak may mobilize communities and build solidarity across different groups.

The discussion below operationalizes these four parameters through concrete institutional mechanisms and a decision-making matrix, using a case of corporate environmental harm as an illustration.

VI.4. Operationalizing Transformative Sanction: Institutional Mechanisms and Practical Decision-Making

A theory of accountability remains incomplete unless it demonstrates how accountability operates in practice. Transformative sanction risks remaining a general normative orientation unless it can be shown to function through concrete institutional forms. This subsection addresses that concern by presenting five existing institutional mechanisms that instantiate transformative sanction, a decision-making matrix that operationalizes Young's four parameters, and a case illustration that demonstrates the matrix's application.

The argument proceeds in three steps. First, I identify five existing institutional mechanisms that illustrate how transformative sanction can be implemented. Each mechanism is grounded in one or more of Young's four parameters. Second, I provide a Decision-Making Matrix (Table 5) that translates Young's parameters into four practical questions. Third, I illustrate the matrix's application through a concrete case of corporate environmental harm.

Not every accountability mechanism qualifies as transformative sanction. A mechanism qualifies only if it meets three conditions. First, it must target structures rather than individuals. Second, it must look forward rather than backward. Third, it must be evaluable using Young's four parameters. The five mechanisms described below satisfy these conditions. They are drawn from existing legal and political practices: courts, corporate regulation, financial settlements, truth commissions, and consent decrees. Each is already in use. My contribution is to show that these mechanisms can be understood as instances of transformative sanction and evaluated using Young's parameters.

First, structural injunctions. A structural injunction is a court order that requires institutional reform rather than punishing past conduct. In *Brown v. Board of Education* (1954), the Supreme Court ruled that segregated schools violated the Equal Protection Clause. The Court did not punish individual school board members. Instead, it ordered the desegregation of public schools. (The following year's remedial order in *Brown II* required desegregation to proceed "with all deliberate speed," 349 U.S. 294 [1955]). This is a transformative sanction because no single school board member caused segregation. Punishing individuals would miss the structural nature of the harm. As Young (2011, 95) argues, structural injustice "cannot be reduced to discrete wrongs committed by identifiable agents." The injunction transformed the institutional arrangement itself. Structural injunctions target power because they apply to institutions and their practices, not to low-level actors. They also enable collective ability by creating public documentation that community organizations can use to demand ongoing accountability. Contemporary examples include court-ordered police reform following patterns of misconduct (*Davis v. City of New York*, 2011).

Second, disqualification and decertification. Disqualification removes agents from positions where their power enables structural harm. This practice includes barring corporate executives from serving as officers of public companies following systematic fraud, decertifying police departments with patterns of unconstitutional conduct, and disqualifying public officials who design policies producing predictable vulnerability.[14] This is transformative sanction because it does not punish the individual for past wrongs. Instead, it removes the structural lever through which harm was produced. Young (2011, 145) argues that privilege creates heightened responsibility: "persons and institutions that are relatively privileged within structural processes have greater responsibilities than others to take actions to undermine injustice." When those with positional privilege fail to act, removing them from that position disrupts the structural advantage that allowed harm to persist. This mechanism specifically targets privilege.

Third, reparative obligations tied to structural position. Reparative obligations require powerful agents to fund institutional reforms. Following the 2008 financial crisis, banks responsible for predatory lending practices were required to fund housing counselling and legal aid programs. This was not punishment for past wrongs. It was a forward-looking measure to address structural vulnerability.[15] This is transformative sanction because it operationalizes the beneficiary-pays principle: those who have benefited from unjust structures bear reparative obligations to fund their transformation. As Young (Ibid., 145) notes, "where there are structural injustices, these usually produce not only victims of injustice, but also persons with relative privilege in relation to the structures." The relevant question is not whether beneficiaries acted culpably. The relevant question is whether their position within the structure confers unearned advantages that create a special responsibility to fund reform. This mechanism targets privilege and interest – the stake that affected communities have in transformation. Similarly, the Global Fund for AIDS, Tuberculosis and Malaria requires wealthy nations to contribute based on their position within global economic structures.[16]

Fourth, public naming and shaming through investigative mechanisms. Public naming and shaming operates through truth commissions and public inquiries that document structural harms. The South African Truth and Reconciliation Commission's mandate was not punishment. It was public acknowledgment of institutional patterns of violence (Tutu 1999). This is transformative sanction because it provides recognition

14 On director disqualification as a public enforcement mechanism, see Loughrey (Loughrey 2024). On police decertification, see Goldman and Puro (Goldman & Puro 1987) and Smith and Alpert (Smith & Alpert 1993). On disqualification of public officials, see Ginsburg, Huq and Landau (Ginsburg, Huq, & Landau 2023).

15 In October 2008, Bank of America (which acquired Countrywide Financial) agreed to a multi-state settlement to resolve claims of unfair lending practices. Washington State received $1.8 million for foreclosure prevention programs, including nearly $600,000 for housing counselling to assist up to 2,300 homeowners and $320,000 for pro bono legal services for families facing foreclosure (Washington State Office of the Attorney General 2010).

16 The Global Fund to Fight AIDS, Tuberculosis and Malaria operates on a replenishment model in which donor countries pledge funds every three years based on their economic capacity. The Fund allocates funding to countries using a formula based predominantly on each country's disease burden and economic capacity, refined by qualitative adjustments for contextual factors (see The Global Fund 2025).

without retribution. Jean Améry, a survivor of torture and Auschwitz, argued that a victim's demand for punishment is not a wish to see the perpetrator suffer. It is a demand for recognition – the restoration of the victim's status as a moral agent whose suffering matters (Améry 1966/1980, 68–80). For Améry, punishment serves as a "moral turning-back of the clock." It forces the wrongdoer to confront what they have done and, in doing so, acknowledges that the victim was wronged. Améry also defended the victim's right to resentment, arguing that resentment toward un-repaired injustice is a rational and dignified response, not a petty emotion. Impunity destroys the possibility of such recognition. However, Améry's analysis assumes an identifiable perpetrator – the torturer, the interrogator, the camp guard. Structural injustice presents a different problem: there is no single perpetrator whose "turning-back of the clock" can restore the victim's dignity. This is where transformative sanction departs from Améry while retaining his insight about recognition. Truth commissions, public inquiries, and investigative mechanisms offer recognition without retribution. They document harms, name perpetrators, and create public records that acknowledge victims' suffering – even when no single perpetrator can be punished. They preserve Améry's demand for recognition while freeing it from the retributive form that structural injustice cannot accommodate. This mechanism targets interest by centering affected communities and enabling them to demand accountability.

Fifth, collective sanctions against corporate and institutional actors. Collective sanctions target organizations through fines, consent decrees, and monitorships. These mechanisms target organizational structures rather than individual employees. When the United States Department of Justice enters a consent decree with a police department requiring systemic reforms, it imposes a forward-looking sanction aimed at structural transformation rather than backward-looking punishment of individual officers.[17] This is transformative sanction because it recognizes that individual actions are politically insignificant without collective organization. Young (2011, 146) argues that individual sanctions "are politically insignificant without collective organization." Structural transformation requires what she calls "collective ability" – the capacity of organized groups to exercise power. Consent decrees target the institution, not individuals, and require ongoing monitoring. This monitoring empowers community organizations to demand compliance. The monitor's role is not to punish but to ensure that structural changes are implemented and sustained. This mechanism targets collective ability.

The five mechanisms follow a logical progression. The first mechanism changes the rules and practices of institutions. The second mechanism removes bad actors from positions of power. The third mechanism makes beneficiaries pay for reform. The fourth mechanism provides recognition to victims. The fifth mechanism creates ongoing oversight. Together, these mechanisms cover the full range of what transformative

17 On consent decrees as a mechanism for police reform, see 34 U.S.C. § 12601 (*Violent Crime Control and Law Enforcement Act of 1994*). On court-appointed monitors in institutional reform litigation, see the monitor's role in *Floyd v. City of New York*, 959 F. Supp. 2d 540 (S.D.N.Y. 2013).

sanction can accomplish: changing structures, removing power, redistributing resources, acknowledging harm, and sustaining accountability over time.

VI.4.1. Selecting Among Mechanisms: A Decision Framework

The five mechanisms identified above – structural injunctions, disqualification and decertification, reparative obligations, public naming and shaming, and collective sanctions – are not mutually exclusive. In many cases, a combination of mechanisms will be appropriate. However, the choice of mechanism should be guided by two considerations:

First, the nature of the structural injustice. Some structures are maintained primarily through institutional rules (e.g., employment law, zoning regulations). For these, structural injunctions are most appropriate. Other structures are maintained through the concentration of power in particular agents (e.g., corporate executives, public officials). For these, disqualification and decertification may be more effective. Still others are maintained through the absence of accountability and transparency. For these, public naming and shaming mechanisms may be most appropriate.

Second, the capacity of affected communities. The SCM emphasizes that those most affected by structural injustice should lead in defining remedies (Young 2011, 145–146). Therefore, the selection of mechanisms should be guided by the priorities of affected communities. If communities prioritize prevention and institutional reform, structural injunctions are appropriate. If they prioritize removing bad actors from positions of power, disqualification is appropriate. If they prioritize recognition and acknowledgment, truth commissions are appropriate.

Third, the risk of co-optation. As will be discussed in Section VI.5, accountability mechanisms can be co-opted by powerful actors. To resist co-optation, mechanisms should include procedural safeguards: public justification, affected community consent, and mechanisms for appeal. They should also create public documentation that enables ongoing collective oversight. The following table summarizes the decision framework:

Structural Feature	Primary Mechanism	Secondary Mechanisms
Institutional rules produce harm	Structural injunctions	Reparative obligations
Power concentrated in specific agents	Disqualification/ decertification	Public naming and shaming
Absence of accountability/ transparency	Public naming and shaming	Collective sanctions
Economic exploitation through market mechanisms	Reparative obligations	Collective sanctions
Systemic patterns across multiple institutions	Collective sanctions	All of the above

Table 4: Decision Framework for Selecting Transformative Sanction Mechanisms.

This framework is not algorithmic – it requires practical judgment. But it provides structured criteria for choosing among mechanisms based on the nature of the injustice, the priorities of affected communities, and the risk of co-optation.

VI.4.2. Procedural Safeguards: Preventing Abuse and Co-optation

Any system of accountability that employs coercive measures must include procedural safeguards to prevent abuse. Transformative sanction is no exception. The following safeguards are essential:

First, public justification. Any transformative sanction must be accompanied by a public justification explaining: (a) the nature of the structural injustice; (b) why the agent being sanctioned bears responsibility; (c) why the particular mechanism was chosen; and (d) how the mechanism will advance structural transformation. This justification should be accessible to affected communities, not merely to legal or technical experts.

Second, affected community consent. Because affected communities have epistemic and political priority in defining what accountability means (Young 2011, 145–146), transformative sanctions should be designed with the meaningful participation of those communities. This requires not merely consultation but genuine consent – affected communities should have the power to veto mechanisms that would harm their interests.

Third, mechanisms for appeal. Agents subject to transformative sanctions should have the right to appeal on grounds that: (a) the structural injustice does not exist; (b) the agent does not bear responsibility; (c) the mechanism is disproportionate or ineffective; or (d) procedural requirements were violated. Appeals should be heard by independent bodies that include representatives of affected communities.

Fourth, periodic review. Transformative sanctions should be subject to periodic review to assess whether they are achieving their structural aims. If they are not, they should be modified or replaced. If they are achieving their aims, they should be phased out to avoid becoming permanent instruments of control.

Fifth, transparency and documentation. All transformative sanctions should create public documentation that enables ongoing collective oversight. This includes publishing data on outcomes, providing regular reports to affected communities, and ensuring that monitoring is independent and transparent.

These safeguards are not merely procedural niceties. They are essential to ensuring that transformative sanction serves structural transformation rather than becoming a new form of oppression. As Young (Ibid., 148–149) warned, powerful actors often respond to claims of structural injustice by demanding "moderation" and "compromise" – and when victims refuse to back down, they are frequently repressed by violent means. Procedural safeguards are the bulwark against this repression.

To guide application, I now provide a practical matrix that operationalizes Young's four parameters. The matrix translates abstract philosophical criteria into four practical questions, as shown in **Table 5**.

Parameter	Questions for Assessment	Indicators of Appropriateness
Power	Does this agent have capacity to alter structural processes?	Agent controls resources, sets policies, or holds positional authority over relevant systems
Privilege	Does impunity protect unearned advantage?	Agent benefits from the very structure they could help transform; accountability would redistribute advantage
Interest	Do affected communities seek this mechanism?	Affected communities have organized, articulated demands, and participate in decision-making
Collective ability	Can this mechanism enable broader mobilization?	Mechanism creates documentation, public record, precedents, or resources that strengthen collective action

Table 5: Decision-Making Matrix for Transformative Sanction.

A transformative sanction is warranted when all four questions receive affirmative answers. If power is absent, the mechanism targets the wrong agent. If privilege is absent, the mechanism addresses no structural benefit. If interest is absent, the mechanism imposes external solutions. If collective ability is absent, the mechanism substitutes individual accountability for collective action.

Case Illustration: Return to the environmental pollution case from Section VI.3. A transformative sanction approach would proceed through the matrix as follows. Corporate executives and board members control production decisions, while low-level employees lack this power; therefore, power is present for targeting executives but absent for targeting employees. Shareholders and executives benefit from externalized costs through profits and compensation, while local residents do not share this benefit; therefore, privilege is present. Affected communities have organized, demanding remediation and healthcare monitoring through documented community meetings and legal representation; therefore, interest is present. Public documentation of the corporation's practices strengthens community organizing, and a structural injunction creates accountability that supports rather than substitutes for collective action; therefore, collective ability is present. Since all four conditions are satisfied for targeting executives and shareholders, transformative sanction is appropriate. The mechanism might combine a structural injunction requiring production changes, disqualification of executives who knowingly maintained harmful practices, reparative obligations funding community health monitoring, and public reporting requirements that enable ongoing collective oversight.

This operationalization does not eliminate judgment. No matrix can replace the need for practical wisdom. But it provides structured criteria that transform transformative sanction from a general orientation into a usable analytical tool. The reader now knows what mechanisms qualify as transformative sanction, how each mechanism connects to Young's parameters, how to decide which mechanism to use, and how to apply the matrix to a concrete case.

VI.5. Philosophical Resources for Defense Against Co-optation

The philosophical literature on punishment, collective responsibility, and structural injustice provides resources for ensuring that accountability mechanisms serve structural transformation rather than merely simulating it. Four insights are particularly important.

First, **Arendt (Arendt 1963) on the limits of criminal law**. Arendt argues that criminal law is designed for individual actors who possess mens rea – the intention to do wrong. In *Eichmann in Jerusalem*, she shows that this framework breaks down when confronted with what she calls the "banality of evil": harms produced through routine institutional operations, where no single actor intends the final outcome. This directly parallels Young's claim that structural injustice "cannot be reduced to discrete wrongs committed by identifiable agents" (Young 2011, 95). The failure of criminal justice in structural contexts is not merely practical; it is conceptual. Therefore, transformative sanction is not an ad hoc solution. It is a principled response to a conceptual mismatch built into criminal law itself.

Second, **Nietzsche (1887/1967) and Améry (1966/1980) on moving beyond retribution**. Nietzsche argues that punishment has no single essence. Throughout history, it has served multiple functions: compensation, deterrence, ritual purification, and the expression of collective anger. The modern equation of punishment with retribution is a contingent historical development, not a metaphysical necessity. Améry adds a crucial dimension. Writing as a torture survivor, he argues that the victim's demand for punishment is not a wish to see the perpetrator suffer. It is a demand for recognition – the restoration of one's status as a moral agent whose suffering matters (Améry 1966/1980, 68–80). For Améry, punishment serves as a "moral turning-back of the clock." It forces the wrongdoer to confront what they have done and, in doing so, acknowledges that the victim was wronged. Améry also defended the victim's right to resentment, arguing that resentment toward un-repaired injustice is a rational and dignified response, not a petty emotion. Impunity destroys the possibility of such recognition. However, Améry's analysis assumes an identifiable perpetrator – the torturer, the interrogator, the camp guard. Structural injustice presents a different problem: there is no single perpetrator whose "turning-back of the clock" can restore the victim's dignity. This is where transformative sanction departs from Améry while retaining his insight about recognition. Truth commissions, public inquiries, and investigative mechanisms offer recognition without retribution. They document harms, name perpetrators, and create public records that acknowledge victims' suffering – even when no single perpetrator can be punished. They preserve Améry's demand for recognition while freeing it from the retributive form that structural injustice cannot accommodate.

Third, **Fraser (Fraser 2000) and Ullrich (Ullrich 2024) on the risk of co-optation**. Fraser distinguishes between two types of remedies. Affirmative remedies redistribute resources or recognize harms within existing structures. Transformative remedies restructure the generative framework that produces injustice (Fraser 2000, 113–114). Fraser warns that many recognition struggles that claim to be transformative are actually affirmative: they address symptoms while leaving the underlying structure intact. Ullrich (2024, 229–231) documents precisely this risk in her analysis of the International Criminal Court's reparations mandate. The promise of "transformative justice" shifts responsibility onto individual victims while leaving institutions untouched. Accountability mechanisms become "mechanisms that help reproduce the status quo rather than disrupting it" (Ullrich 2024, 189).

Young herself warned about this same risk. She writes that powerful agents often respond to claims of structural injustice by demanding "moderation" and "compromise" (Young 2011, 149). She also notes that victims who refuse to back down from confrontation are frequently repressed by violent means (Young 2011, 148–149). This is where my argument begins. Young identified the problem of co-optation. But she did not develop a systematic framework to resist it. That is what I aim to provide. The four parameters I have articulated – power, privilege, interest, and collective ability – are not merely analytical tools. They are also weapons in political struggle. They help victims and their allies answer four concrete questions: Who has the power to change the structure? Who enjoys unearned privilege from it? Who has a stake in transforming it? Who can mobilize collective action? Transformative sanction, grounded in these parameters, is designed precisely to resist the co-optation that Young feared.

Fourth, Young's four parameters directly address the risks each theorist identifies. The table below summarizes this connection:

Risk	**Theorist**	**Young's Parameter That Addresses It**
Criminal law cannot handle systemic harm	Arendt (1963)	Power – target power, not low-level actors
Punishment protects privilege	Nietzsche (1887/1967)	Privilege – disrupt unearned advantage
Recognition must be defined by victims	Améry (1966/1980)	Interest – center affected communities
Individual accountability substitutes for collective action	Fraser (2000); Ullrich (2024)	Collective ability – enable mobilization

Table 6: Philosophical Resources for Defense Against Co-optation.

Transformative sanction, grounded in these four parameters, offers a principled framework for distinguishing mechanisms that serve structural transformation from those that merely simulate it. It retains accountability as a demand for recognition

(Améry) while rejecting retribution as its necessary form (Nietzsche). It targets power rather than low-level actors (Arendt) and enables collective mobilization rather than substituting individual accountability for collective action (Fraser, Ullrich). This is what makes transformative sanction a genuine alternative to the liability model: it does not abandon accountability but reorients it toward the goal that matters – transforming the structures that produce injustice.

VII. Objections and Replies

I now address the most significant philosophical objections to Young's framework. Each objection is stated explicitly, followed by a rigorous reply.

VII.1. The Indeterminacy Objection (Barry & Ferracioli)

Barry and Ferracioli (2013, 251) contend that Young's Social Connection Model, even with its four parameters, remains "intolerably vague" and fails to provide individuals with sufficient guidance, resulting in a responsibility "of everyone in general and no one in particular." Their critique rests on the assumption that a viable model of responsibility must generate determinate, action-guiding duties for each individual agent. This objection, however, rests on a misunderstanding of the kind of guidance appropriate for structural contexts. Young's SCM is not designed to function as a mechanical algorithm that outputs specific duties for every individual. To expect algorithmic certainty is to impose the liability model's juridical logic onto a political problem that is inherently relational, processual, and context-dependent (Young 2011, 103–104; 144–148).

Responsibility for structural injustice requires what Aristotle (1999, Book VI) termed phronesis (practical wisdom) understood as the capacity for situated judgment about appropriate action within particular social relations and constraints. Young's four parameters – power, privilege, interest, and collective ability – provide the essential framework for exercising such practical wisdom (Young 2011, 144–148). These parameters do not eliminate the need for judgment; rather, they supply critical criteria for practical deliberation.

Consider how this reasoning proceeds in practice. An individual concerned about labour exploitation might deliberate as follows. First, assessing power: 'My individual power as a consumer is limited'. Second, evaluating collective ability: 'However, I possess collective ability through membership in a labor union'. Third, recognizing privilege: 'My privilege, evident in the low prices enabled by exploitative labor, creates a responsibility to support worker-led movements demanding fair wages'. Fourth, centering interest: 'The interests most at stake are those of the workers themselves; therefore, I ought to follow their leadership rather than imposing my own solutions'. This form of guidance is not algorithmic – it does not specify 'perform action X at time T' – but it is nonetheless genuine guidance. It directs the agent toward relevant questions, relevant factors, and relevant voices. The liability model, by contrast, would simply declare the individual not guilty of any specific wrongdoing and offer no positive direction whatsoever (Ibid., 103–104).

Barry and Ferracioli's critique mistakes the necessary absence of an algorithmic shortcut for a fatal lack of guidance. In the domain of structural injustice, a framework for practical judgment is not a defect; it is the only philosophically coherent form of guidance possible. As Young (Ibid., 95–99; 144–148) argues, the very features that make structural injustice resistant to liability thinking – diffuseness, emergence, normalization – also make algorithmic precision impossible.

VII.2. The Motivational Challenge (Nussbaum)

Nussbaum's foreword to *Responsibility for Justice* presents a more profound challenge. She argues that Young's attempt to sever political responsibility from notions of blame or guilt risks rendering the former motivationally inert. Nussbaum contends that if an agent possesses a forward-looking responsibility and fails to discharge it, then "when the relevant time passes, she is guilty of not having shouldered her responsibility" (Nussbaum 2011, xxi). Without the prospect of blame or guilt, Nussbaum suggests, agents lack sufficient motivation to act.

I offer three responses to this challenge.

First, Nussbaum's objection mischaracterizes the motivational architecture of the SCM. The model does not eliminate accountability; it reorients it. The SCM shifts the agent's focus from retrospective culpability ('Am I to blame for past harms?') to prospective agency ('What is my role in transforming ongoing structures?'). This reorientation can be illuminated by Arendt's (Arendt 1998, 177–178) account of natality – the human capacity to begin something new through public engagement. The motivational question is not 'Will I be punished if I fail?' but rather 'What kind of world do I want to help build?' Moreover, empirical research on moral psychology suggests that guilt and shame often induce defensive withdrawal – denial, avoidance, blame-shifting – rather than constructive engagement (Tangney & Dearing 2002). The SCM's approach – grounding responsibility in social position rather than individual fault – removes this defensive dynamic and opens space for positive, forward-oriented action.

It is worth noting that Nussbaum herself has argued that emotions such as compassion and political love can motivate action more effectively than guilt and shame (Nussbaum 2001, 2006). Her objection to Young, then, is somewhat puzzling: if guilt and shame are less reliable motivators than positive emotions, why insist on retaining guilt-based motivation? The SCM's reorientation from retrospective culpability to prospective agency aligns more closely with Nussbaum's own account of political motivation than she acknowledges.

Second, the SCM provides mechanisms for accountability that do not depend on internal guilt. Young's four parameters create accountability through publicity, deliberation, and collective action (Young 2011, 144–148). When agents are publicly identified as occupying positions of power over unjust structures, they are held accountable not by internal guilt but by external political pressure. The anti-sweatshop movement illustrates this dynamic: it succeeded not by inducing universal consumer guilt but by strategically organizing pressure against high-power corporate actors, exposing their role in structural injustice, and mobilizing collective action to demand change (Young 2004, 386–387).

Third, the free-rider problem that Nussbaum anticipates is addressed by the SCM's emphasis on collective ability. The SCM does not rely on voluntary individual compliance; it recognizes that structural transformation requires collective organization – unions, advocacy networks, coalitions – capable of holding even unwilling agents accountable (Young 2011, 111; 146). The question is not whether each individual will voluntarily discharge their responsibilities, but whether those most affected by injustice can build the collective power to hold all implicated agents accountable.

The concept of transformative sanction, developed in Section VI, directly addresses Nussbaum's concern that agents might receive "a free pass." When international tribunals prosecute architects of mass violence, when truth commissions publicly name perpetrators, when corporate executives face disqualification from positions of power – these are not escapes from responsibility but robust expressions of forward-looking accountability.

VII.3. The Category Error Underlying Both Objections

Both objections share a common philosophical flaw: they evaluate the Social Connection Model by criteria drawn from the liability model.

Barry and Ferracioli (2013) demand the algorithmic determinacy that liability thinking promises. They expect the SCM to specify, for each individual agent, exactly what they must do – analogous to how criminal law specifies penalties for specific offenses. But the SCM is not a criminal code. It is a framework for political responsibility in contexts where no single perpetrator exists and where harms emerge from diffuse, normalized social processes (Young 2011, 95–99). Nussbaum (Nussbaum 2011) demands the guilt-based motivation that liability thinking relies upon. She expects the SCM to say "you are guilty" to motivate action. But the SCM is not a system of blame assignment. It is a framework for collective transformation grounded in social position rather than individual fault (Young 2011, 103–104; 143–148).

Both objections thus misjudge the SCM by applying the wrong evaluative standards. The SCM is not a failed liability model; it is a different kind of framework altogether, designed for a different domain. This is not to say that the liability model has no place. On the contrary, the liability model remains appropriate for discrete wrongs involving identifiable agents, voluntary wrongdoing, and direct causation – paradigm cases of assault, theft, fraud, or breach of contract (Young 2011, 98–99; 186). The SCM does not deny this. What it denies is that liability thinking, even in extended forms, is adequate for structural injustice. As Young herself puts it: "The liability model is the appropriate model for assigning responsibility for specific harms that are traceable to the actions of particular agents. The social connection model is the appropriate model for assigning responsibility for structural injustice" (Young 2011, 186).

The complementarity of the two models is not merely a matter of different domains. It is also a matter of different *functions*. The liability model serves the function of determining guilt and imposing punishment for discrete wrongs. The SCM serves the function of assigning political responsibility and guiding collective action for structural transformation. These functions are not in competition; they serve different purposes in a just society. A just society needs both: it needs to hold individuals accountable for discrete wrongs, and it needs to transform structures that produce systematic harm. The error of liability theorists is to assume that the liability model can serve both functions. The error of some SCM proponents is to assume that the SCM can replace the liability model entirely. Both errors are mistakes. The two models are complementary, and a complete theory of justice must include both.

The two models are complementary, not competitive. Each has its proper domain. The error of Barry and Ferracioli and Nussbaum is to demand that the SCM perform functions that belong properly to the liability model – and then to declare the SCM deficient when it cannot perform those functions.

VIII. Concluding Remarks

This paper has defended Young's Social Connection Model (SCM) against the objection that it leaves no legitimate role for accountability in response to structural injustice. Rather than claiming to resolve the punishment dilemma definitively, I have proposed a conceptual intervention: transformative sanction – punitive or quasi-punitive measures whose justification is forward-looking structural disruption rather than backward-looking desert.

Before summarizing the argument, it is necessary to restate what makes transformative sanction distinct from related frameworks. Unlike retributive punishment, transformative sanction does not ask 'what does this agent deserve?' but 'what intervention will disrupt unjust structures?' Unlike restorative justice, transformative sanction does not presuppose identifiable offenders willing to accept responsibility for discrete harms. Unlike transformative justice, transformative sanction does not rule out state-mediated accountability measures, provided such measures satisfy Young's four parameters. Table 3 (in Section VI.2) summarizes these distinctions across six dimensions: primary aim, target, mechanism, identifiable wrongdoer required, coercion, and temporal orientation.

The argument has established three interconnected claims. First, the liability model's juridical triad – identifiable agents, voluntary wrongdoing, and direct causation – is conceptually mismatched to structural harms. Structural harms are processual, systemic, and normalized. They emerge from diffuse social processes, not from discrete culpable acts. As demonstrated in Section II, even sophisticated extensions of liability by Kutz, Miller, and Goodin remain captive to an individualistic, retrospective frame that cannot grasp structural emergence. Second, Young's SCM provides a genuine alternative to liability thinking. As reconstructed in Section III, the SCM has five defining features: it is non-isolating, it requires judgment of background conditions, it is forward-looking, it is shared, and it requires collective action. Its four parameters – power, privilege, interest, and collective ability – provide a framework for determining the relative weight of different agents' forward-looking responsibilities. Together, these features and parameters transform responsibility from a backward-looking search for the guilty into a forward-looking project of structural transformation. Third, transformative sanction operationalizes this framework. Section VI.4 identifies five institutional mechanisms that illustrate its practical application: structural injunctions, disqualification and decertification, reparative obligations, public naming and shaming through investigative mechanisms, and collective sanctions against corporate and institutional actors.

The Decision-Making Matrix presented in Table 5 provides a practical tool for determining when transformative sanction is warranted. A transformative sanction is warranted when the agent has the power to alter structural processes, when impunity would protect unearned privilege, when affected communities seek accountability, and when the mechanism enables broader collective mobilization. The four parameters are further operationalized through Table 4 (Decision Framework for Selecting Transformative Sanction Mechanisms) and Table 6 (Philosophical Resources for Defense Against Co-optation), which together provide a comprehensive analytical toolkit for distinguishing mechanisms that serve structural transformation from those that merely simulate it.

The practical significance of these claims lies in reorienting the philosophy of punishment away from the false binary of retribution or impunity. For structural injustices – environmental devastation, labour exploitation, institutional violence, refugee displacement – the liability model offers only two unacceptable options. The first option is to punish an individual who is not the cause of the harm, thereby scapegoating low-level actors while leaving power holders untouched. The second option is to do nothing, thereby compounding injustice with indifference. Transformative sanction offers a third path. It holds agents accountable without scapegoating. It intervenes in structures without ignoring agency. It demands change without wallowing in the past.

This account remains vulnerable to several limitations. Whether transformative sanction requires new institutional forms or can be integrated into existing legal systems is an open question that requires further research. Procedural safeguards against abuse – including public justification, affected community consent, and mechanisms for appeal – remain to be designed. The boundary between structural and individual wrongdoing, and consequently the boundary between cases requiring transformative sanction and those properly addressed by liability frameworks, demands further specification. These are not failures of the framework but invitations for future research.

What ultimately emerges from this paper is not a finished theory but a direction for future work – a reorientation of the philosophy of punishment toward the demands of structural justice. The persistence of the punishment dilemma, I have suggested, reflects not a failure of existing frameworks but a deeper tension within modern moral and political life: the simultaneous demand to condemn wrongdoing, to preserve the moral standing of persons, to respond seriously to harm, and to transform the conditions that make systematic suffering predictable. Transformative sanction does not dissolve this tension. Rather, it clarifies the conditions under which coercive measures may claim legitimacy in structural contexts – namely, when they target power, disrupt privilege, center affected communities, and enable collective mobilization. In doing so, it reimagines accountability not as a backward-looking calculus of desert but as a forward-looking project of structural transformation. The question for theorists and practitioners alike is no longer ‘who is to blame?’ but ‘what shall we build together?’ It is to the practical and institutional elaboration of this question – the design of legal frameworks, the creation of monitoring bodies, and the empowerment of affected communities – that future work should turn. The architecture of transformative justice is not yet built, but this paper has laid its foundations.

References

Améry J. 1980 (orig. 1966). *At the Mind’s Limits: Contemplations by a Survivor on Auschwitz and Its Realities*, trans. S. Rosenfeld and S. P. Rosenfeld. Bloomington: Indiana University Press.

Arendt H. 1963. *Eichmann in Jerusalem: A Report on the Banality of Evil*. New York: Viking Press.

Arendt H. 1994. “Organized Guilt and Universal Responsibility,” in J. Kohn (Ed.), *Essays in Understanding, 1930-1954* (pp. 121–132). New York: Harcourt-Brace.

Arendt H. 1998. “Action,” in *The Human Condition*, 2nd Ed. (pp. 175–247). Chicago: University of Chicago Press. https://doi.org/10.7208/chicago/9780226924571.001.0001

Aristotle. 1999. *Nicomachean Ethics*, trans. T. Irwin. 2nd Ed. Indianapolis: Hackett Publishing.

Barry C. & Ferracioli L. 2013. “Young on Responsibility and Structural Injustice,” *Criminal Justice Ethics* 32(3):247–257. https://doi.org/10.1080/0731129X.2013.861981

Braithwaite J. 2002. *Restorative Justice and Responsive Regulation*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780195136395.001.0001

Brown v. Board of Education, 347 U.S. 483 (1954).

Convention on the Elimination of All Forms of Discrimination Against Women, opened for signature Dec. 18, 1979, 1249 U.N.T.S. 13 (entered into force Sept. 3, 1981).

Davis v. City of New York, 959 F. Supp. 2d 324 (S.D.N.Y. 2011).

Floyd v. City of New York, 959 F. Supp. 2d 540 (S.D.N.Y. 2013).

Fraser N. 2000. "Rethinking Recognition," *New Left Review* 3:107–120. https://doi.org/10.64590/pup

Ginsburg T., Huq A. Z., & Landau D. 2023. "Democracy's Other Boundary Problem: The Law of Disqualification," *California Law Review* 111(6):1633–1704. https://doi.org/10.15779/Z388P5VB2H

The Global Fund to Fight AIDS, Tuberculosis and Malaria 2025. "Sources of Funding." URL: https://resources.theglobalfund.org/en/grant-life-cycle/sources/#si (accessed on 15 March 2026).

Goldman R. & Puro S. 1987. "Decertification of Police: An Alternative to Traditional Remedies for Police Misconduct," *Hastings Constitutional Law Quarterly* 15(1):45–80.

Goodin R. E. 1985. "Vulnerabilities and Responsibilities," in *Protecting the Vulnerable: A Reanalysis of Our Social Responsibilities* (pp. 109–144). Chicago: University of Chicago Press. https://doi.org/10.1017/CBO9780511625053.003

Hart H. L. A. 1968. *Punishment and Responsibility: Essays in the Philosophy of Law*. Oxford: Clarendon Press.

Hegel G. W. F. 1991 (orig. 1821). *Elements of the Philosophy of Right*, trans. H. B. Nisbet. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511808012

Jonas H. 1984. "The Altered Nature of Human Action," in *The Imperative of Responsibility: In Search of an Ethics for the Technological Age* (pp. 1–24). Chicago: University of Chicago Press.

Kaba M. 2021. *We Do This 'Til We Free Us: Abolitionist Organizing and Transforming Justice*. Chicago: Haymarket Books.

Kant I. 1996 (orig. 1797). *The Metaphysics of Morals*, trans. M. Gregor. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511809644

Kim M. E. 2018. "From Carceral Feminism to Transformative Justice: Women-of-Color Feminism and Alternatives to Incarceration," *Journal of Ethnic & Cultural Diversity in Social Work* 27(3):219–233. https://doi.org/10.1080/15313204.2018.1474827

Kutz C. 2000. *Complicity: Ethics and Law for a Collective Age*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511663758

Loughrey J. 2024. "Individual Accountability for Corporate Harms: Disqualification and the Role of Public Enforcement," *Griffith Law Review 33(4):*534–555. https://doi.org/10.1080/10383441.2025.2502196

Mantouvalou V. 2023. *Structural Injustice and Workers' Rights*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780192857156.001.0001

Mantouvalou V. 2024. "Structures of Injustice, the Law, and Exploitative Work," in J. Browne & M. McKeown (Eds.), *What Is Structural Injustice?* (pp. 241–259). Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198892878.003.0014

Miller D. 2007. "Two Concepts of Responsibility," in *National Responsibility and Global Justice* (pp. 81–109). Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199235056.001.0001

Nietzsche F. 1967 (orig. 1887). *On the Genealogy of Morals*, trans. W. Kaufmann. New York: Vintage Books.

Nussbaum M. C. 2001. *Upheavals of Thought: The Intelligence of Emotions*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511840715

Nussbaum M. C. 2006. *Hiding from Humanity: Disgust, Shame, and the Law*. Princeton: Princeton University Press.

Nussbaum M. C. 2011. "Foreword," in I. M. Young, *Responsibility for Justice* (pp. ix–xxv). Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195392388.001.0001

Parekh S. 2011. "Getting to the Root of Gender Inequality: Structural Injustice and Political Responsibility," *Hypatia* 26(4):672–689. https://doi.org/10.1111/j.1527-2001.2010.01159.x

Parekh S. 2020. "Structural Injustice," in *No Refuge: Ethics and the Global Refugee Crisis* (pp. 151–176). Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780197507995.001.0001

Pogge T. 2001. "How Should Human Rights Be Conceived?" in P. Hayden (Ed.), *The Philosophy of Human Rights* (pp. 187–210). St. Paul, MN: Paragon House.

Smith W. C. & Alpert G. P. 1993. "Law Enforcement: Policing the Defective Centurion – Decertification and Beyond," *Criminal Law Bulletin* 29(2):147–157.

Tangney J. P. & Dearing R. L. 2002. *Shame and Guilt*. New York: Guilford Press.

Tutu D. 1999. *No Future Without Forgiveness*. New York: Doubleday.

Ullrich L. 2024. "Reparations, Abolitionist Imaginaries, and Self-Transforming Victims: Transformative Justice at the ICC," in *Victims and the Labour of Justice at the International Criminal Court: The Blame Cascade* (pp. 188–231). Oxford: Oxford University Press. https://doi.org/10.1093/9780191915048.003.0005

Violent Crime Control and Law Enforcement Act of 1994, 34 U.S.C. § 12601.

Washington State Office of the Attorney General 2010. "Countrywide Settlement Cash Will Refinance Washington Foreclosure Prevention Efforts," February 10, 2010. URL: https://www.atg.wa.gov/news/news-releases/countrywide-settlement-cash-will-refinance-washington-foreclosure-prevention. Accessed 10 January 2026

Young I. M. 1990. "Five Faces of Oppression," in *Justice and the Politics of Difference* (pp. 39–65). Princeton: Princeton University Press.

Young I. M. 2003a. "From Guilt to Solidarity: Sweatshops and Political Responsibility," *Dissent* 50(2):39–45. https://search.opinionarchives.com/Summary/Dissent/V50I2P39-1.htm

Young I. M. 2003b. *Political Responsibility and Structural Injustice*. University of Kansas. URL: https://hdl.handle.net/1808/12416

Young I. M. 2004. "Responsibility and Global Labor Justice," *The Journal of Political Philosophy* 12(4):365–388. https://doi.org/10.1111/j.1467-9760.2004.00205.x

Young I. M. 2006. "Responsibility and Global Justice: A Social Connection Model," *Social Philosophy and Policy* 23(1):102–130. https://doi.org/10.1017/S0265052506060043

Young I. M. 2011. *Responsibility for Justice*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195392388.001.0001

Zehr H. 2015. *The Little Book of Restorative Justice* (Rev. Ed.). New York: Good Books.

Ethics in Progress (ISSN 2084-9257). Vol. 17 (2026). No. 1,
Art. #6, pp. 105-120. DOI:10.14746/eip.2026.1.6

Preventing, Repairing, Punishing: The EU Environmental Crime Directive as a Hybrid Response to Environmental Harm

Giulia Battistoni

(Universität Münster, Münster, Germany/Università degli Studi di Verona, Verona, Italy; giulia.battistoni.90@gmail.com)
ORCID: 0000-0002-5391-8590

Abstract: Starting from the principle of criminal legality – *nullum crimen, nulla poena sine lege* – and from selected environmental cases, this paper examines the difficulties of applying traditional criminal law categories to environmental offences. In environmental harm, causation, *mens rea*, individual responsibility, and even the identification of harm itself are often opaque. Given this complexity, the paper then addresses contemporary debates, focusing in particular on alternative ways of conceptualizing responsibility and punishment in cases of environmental offences. Finally, it analyses Directive (EU) 2024/1203, to be implemented by Member States by May 2027, arguing that it offers a hybrid model of responsibility, punishment and justice through a combination of individual, corporate and institutional responsibility, of retributive and consequentialist theories of punishment, as well as retributive, restorative, and transformative approaches to justice.

Keywords: Environmental harm; responsibility; punishment; Directive (EU) 2024/1203; retributive justice; restorative justice; transformative justice.

Conceptual Premises

In the context of increasingly widespread debates concerning the need to reconsider both the function of punishment within contemporary society and the category of responsibility itself (see Battistoni & Magni 2025), as well as the necessity of rethinking traditional legal categories in light of contemporary challenges, environmental offences represent a particularly urgent and complex case.[1] The EU Directive 2024/1203 "on the protection of the environment through criminal law" attempts to address this challenge from the perspective of criminal law. In doing so, it responds to longstanding debates concerning responsibility, punishment, and environmental harm. It is therefore worth examining the

1 This essay is part of the project *Collective Responsibility towards Nature and Future Generations (ReNa)*, that has received funding from the European Union's Horizon Europe Research and Innovation Programme under the Marie Skłodowska-Curie; grant agreement No. 101064728; funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

issues at stake more closely.

First of all, the generally accepted principle of legality holds that *nullum crimen, nulla poena sine lege*, meaning that both the act constituting an offence and the penalty attached to it must be expressly provided by law.[2] Consequently, this principle implies, first, the existence of a law defining a particular wrongdoing as a crime; second, the clear determination of the corresponding penalty; and third, the possibility of identifying concrete cases falling under the relevant legal provisions.[3]

Moreover, criminal offences are traditionally understood as involving both an objective element (*actus reus*) and a subjective element (*mens rea*), although these categories are conceptualized differently across legal traditions. The *actus reus*[4] constitutes the external and objective element of the offence, while the *mens rea* refers to the psychological or subjective element of the crime, linked to the agent's intentions, knowledge, beliefs, or state of mind.[5] Together, the *actus reus* and *mens rea* reflect a particular, predominantly individual and retrospective conception of responsibility upon which traditional theories of punishment have largely rested. Depending on the legal framework, responsibility may be conceived either as subjective responsibility, which in criminal law takes into account the intentional and voluntary character of the action, or as objective responsibility, based primarily on the wrongfulness of the act itself, as is often the case in civil law contexts.

2 The principle is commonly traced back to the German jurist Paul Johann Anselm von Feuerbach, who, in his *Lehrbuch des gemeinen in Deutschland gültigen peinlichen Rechts*, wrote in § 20: "*Jede Zufügung einer Strafe setzt ein Strafgesetz voraus*. (Nulla poena sine lege)", namely that every infliction of punishment presupposes a penal law. He further stated that "*Die Zufügung einer Strafe ist bedingt durch das Daseyn der bedrohten Handlung*. (*Nulla poena sine crimine*)", that is, the infliction of punishment is conditioned by the existence of a prohibited act. Finally, he added that "Die gesetzlich bedrohte That (...) ist bedingt durch die gesetzliche Strafe. (Nullum crimen sine poena legali)", namely that the legally prohibited act is conditioned by the legal punishment itself, insofar as it is through the law that a violation of right is linked to its necessary legal consequence (Feuerbach 1832[11], § 20, 19).

3 The principle *nullum crimen sine lege* is also known through the broader notion of the "rule of law" and is generally regarded as one of the fundamental principles of criminal law in Europe (Ashworth & Horder 2009[6], 57). The formulation presented above can also be derived from Article 7 of the *European Convention on Human Rights*, adopted in 1950, which states in paragraph 1 that: "No one shall be held guilty of any criminal offence on account of any act or omission which did not constitute a criminal offence under national or international law at the time when it was committed. Nor shall a heavier penalty be imposed than the one that was applicable at the time the criminal offence was committed." The first part of the provision clearly recalls the principle *nullum crimen sine lege*. However, Article 7 itself is entitled "No punishment without law," corresponding to *nulla poena sine lege*, thus suggesting that the two dimensions are strictly interconnected (Ashworth & Horder 2013[7], 63).

4 "The *actus reus* consists of the prohibited behaviour or conduct, including any specified consequences arising therefrom" (Ashworth & Horder 2013[7], 83).

5 "The *mens rea* is usually described as the mental element – the intention, knowledge, or recklessness of the defendant in relation to the proscribed conduct" (Ashworth & Horder 2013[7], 83). Aristotle's discussion of voluntary and involuntary action in Book III of the *Nicomachean Ethics* may be regarded as an important philosophical antecedent of later conceptions of subjective responsibility and *mens rea*. As Aristotle observed, when agents act under duress or in ignorance of the relevant circumstances, and therefore lack voluntariness, they cannot properly be blamed or praised for their actions, nor fully held accountable for them. Consequently, punishment itself appears to lose its justification (Aristotle 2012, Book 3).

From a temporal perspective, responsibility can be understood either retrospectively or prospectively. Retrospective responsibility is linked to the concept of imputation and is therefore backward-looking: something that an agent did in the past can be imputed, namely attributed to them as their responsibility, insofar as they caused it, at least indirectly, voluntarily, and with awareness of the relevant circumstances. If the act is legally recognized as wrongful, the agent may consequently be punished for it. Aristotle already articulated criteria underlying retrospective responsibility that still largely apply today (Aristotle 2012, Book 3; Werner 2016, 134). Prospective responsibility, by contrast, is future-oriented and arises when agents set normative expectations and obligations for themselves or for other actors (Werner 2016).

Against this background, grounding punishment and responsibility for environmental offences means addressing the nature of environmental harm itself, the kind of responsibility that may be attributed to the agents involved, and the legal frameworks defining both offences and penalties.[6] This paper seeks to engage with these questions by first highlighting the difficulties that environmental damage poses for traditional conceptions of responsibility and punishment (I.), and then by advancing possible solutions, at least at a theoretical level, through engagement with some of the most relevant positions in the contemporary debate (II.). Finally, the paper will examine Directive (EU) 2024/1203 on the protection of the environment through criminal law (III.). The hypothesis to be tested through references to the Directive itself is that it addresses several of the problematic issues previously identified (I.), incorporates some of the theoretical insights emerging from contemporary debates on environmental responsibility and punishment (II.), and advances a hybrid model both of punishment – combining retributive and consequentialist elements – and of responsibility and justice, integrating retributive, restorative, and transformative dimensions – that is, punishment proportionate to wrongdoing, the repair of harm and relationships, and the transformation of the social and institutional conditions underlying harm. In this sense, the Directive may represent a promising attempt to cope more effectively with environmental offences in the future.

I. Punishment and Responsibility for Environmental Harm: Difficulties and Challenges

First of all, in the environmental context, the definition of the *actus reus* becomes particularly problematic, since environmental damage may materialise long after the original act or omission. Here, human actions produce effects that unfold across complex

6 All these aspects fall within the broader field of environmental law. In this context, the importance of clearly defining legal prohibitions and establishing shared international rules has increasingly emerged as a necessary condition for grounding environmental responsibility. Indeed, it becomes difficult to determine whether a state, corporation, or individual has violated environmental obligations in the absence of sufficiently clear normative standards to which responsibility may be referred (see Louka 2006).

and often indeterminate temporal horizons, making both the anticipation and the precise delineation of environmental harm inherently difficult. Moreover, most polluters do not intentionally seek to cause irreversible environmental harm, but rather pursue economic or personal interests while underestimating or ignoring the possible consequences of their actions. As a consequence, the identification of the *mens rea* may also prove difficult, insofar as environmental harm is frequently not the direct object of intention. This makes it increasingly difficult to define environmental crime itself and to establish its relationship to individual agents. Precisely because of this temporal and causal complexity, human responsibility can no longer be limited to immediate causality, but must also take into account effects that are remote in time and structurally interconnected, thereby assuming a necessarily collective dimension.[7]

To illustrate these challenges and understand them more concretely, let us consider a couple of examples among the many cases raising questions of environmental justice and ethics. First, we shall refer to a case decided by the Supreme Court of Sweden.[8] The case concerned the sale of twelve coyote skins marketed as "wolfskins" during a medieval-themed market. Although wolves were protected under the CITES Regulation[9] while coyotes were not, the seller argued that he had used the label "wolfskin" merely for historical and thematic reasons, *without intending* to market specimens belonging to a protected species. The legal problem arose from the wording of the European regulation itself. In the English version, the definition of "specimen" also included goods that "appear" to belong to protected species, whereas the Swedish version lacked this expression. As a consequence, it was not sufficiently clear, in the Swedish wording of the provision, that the conduct in question constituted a criminal offence.

The Swedish Supreme Court therefore acquitted the seller, holding that a conviction would have violated the principle of legality, since the provision did not define the punishable conduct with sufficient clarity. The case illustrates how environmental criminal law may become problematic when the legal provision does not clearly define the environmental offence and when the agent lacks the direct intention to harm a protected environmental interest. In this sense, it raises difficulties concerning both the *actus reus* – the precise identification of the prohibited conduct – and the *mens rea* – the subjective connection between the agent and the environmental wrong.

A further example of the difficulties surrounding responsibility and punishment in the environmental context can be found in the *Sharma v. Minister for the Environment* case in Australia.[10] In 2020, eight teenagers, represented by a nun, filed a class action

7 These issues have been addressed by Serena Vantin through the notion of the necessary "metamorphoses of responsibility" (Vantin 2024).

8 This case is discussed by Olsen Lundh 2024, 3–4 and was published in the legal periodical *Nytt Juridiskt Arkiv (NJA)* 2016, 680 (Olsen Lundh 2024, n. 33).

9 *Convention on International Trade in Endangered Species of Wild Fauna and Flora* 1973, available at: https://cites.org/eng/disc/text.php.

10 At the following link it is possible to access all the documents relating to the Sharma v. Minister

lawsuit before the Australian Federal Court seeking to prevent the approval of a coal mine extension project, arguing that the Minister for the Environment owed a duty of care to young people. According to the plaintiffs, approving the project would have aggravated climate change and caused serious harm to future generations. Under the *Environment Protection and Biodiversity Conservation Act*,[11] they therefore sought an injunction preventing the Minister from authorizing the project.

Although, since 2021, a duty of care aimed at preventing personal harm to children had been recognized under the *Negligence Act*, the Federal Court ultimately refused to grant the injunction. After several judicial developments, the project was approved and no effective duty of care towards the youth was recognized on the part of the Minister. The reasons provided were the following: first, the absence of sufficient proximity between the Minister and the children – a point that can more generally be understood as the *lack of proximity* between present generations and future ones; second, the *indeterminacy of the duty of care* itself; and third, the *unpredictability* of the consequences deriving from the approval of the project, namely the lack of certainty that the coal mine extension would actually cause personal harm to children. These arguments reveal the inadequacy of traditional conceptions of causal, retrospective responsibility and punishment when confronted with the long-term and indirect consequences of human actions on the environment and future generations. More broadly, the case highlights a normative, legal, and socio-political vacuum in environmental matters, linked to the lack of adequate conceptual and legal frameworks capable of grounding justice in this field.

While the Swedish case reveals the tensions between environmental law and the principle of legality – particularly when the environmental offence is not clearly defined and difficulties arise concerning both *actus reus* and *mens rea* – the Sharma v. Minister for the Environment case exposes a different but complementary set of problems concerning proximity, foreseeability, and responsibility for long-term environmental harm affecting future generations.[12]

---

for the Environment case: Federal Court Online File. I first addressed this case in Battistoni 2023.

11 *Environment Protection and Biodiversity Conservation Act* 1999 (Cth), available at: Federal Register of Legislation.

12 For the first time, human beings are called to answer for their actions not only before other human beings, but also before nature and future generations, which are themselves unable to directly assert rights against present agents. The question of responsibility towards nature has been widely discussed within environmental ethics (see at least Passmore 1974). Even if nature and non-human animals may not be considered "moral subjects" in a Kantian sense, this does not exclude them from becoming objects of ethical inquiry and, as such, at least "moral objects" toward which human agents may bear duties of care and responsibility. This position has been developed by some revisionist approaches to discourse ethics, which attempt to extend ethical consideration beyond the community of rational and linguistically competent subjects. The idea is that nature and future generations are not excluded from becoming objects of ethical reflection, nor are their interests prevented from being represented within the community of presently existing subjects engaged in discourse (on this debate, see at least Eckersley 2007, 24–49 and Morar 2008, 129–158). The reference to future generations (Nolt 2015) also points to the broader debate on intergenerational justice, which in recent years has become the object of an increasing number of studies, volumes, and book symposia. See at least the entry "Intergenerational Justice" in the *Stanford Encyclopedia of Philosophy*, which addresses, among other issues, Derek Parfit's non-

Against this background, several attempts have been made to rethink responsibility and punishment in the field of environmental offences. Without any claim to completeness, the following discussion will touch upon some of these approaches, which will later prove useful for the analysis of the European Directive.

II. Rethinking Punishment and Responsibility for Environmental Offences

In light of the difficulties outlined above, it is useful to consider some of the theoretical responses emerging from contemporary debates on responsibility, punishment, and environmental justice. The aim is not to offer definitive solutions, but rather to explore conceptual resources capable of addressing the challenges that environmental offences pose to conventional legal and moral categories.

The cases examined above highlight three interrelated challenges: 1. the indeterminacy of environmental harm, as evidenced by the Swedish case; 2. the inadequacy of existing legal and normative structures, exemplified by the contested notion of *duty of care* in Sharma; and 3. the resulting uncertainty surrounding responsibility, including the attribution of conduct and the justification of punishment. These challenges prompt the following reflections.

(1) First of all, the need to define environmental harm as clearly and precisely as possible emerges with particular force in order to satisfy the principle of legality embodied in the maxim *nullum crimen sine lege.*[13] Accordingly, the *actus reus* must be clearly identifiable and legally defined as constituting a criminal offence.

(2) Second, institutions bear what has been described as a responsibility to take responsibility (Shockley 2020), namely the duty to establish adequate legal frameworks, mechanisms of accountability, and systems of punishment capable of preventing environmental offences and responding effectively to environmental harm (Werner

identity problem and the debate concerning the rights of future generations (see also Parfit 1984). On the question of intergenerational justice, see also the contributions collected in Gosseries & Meyer 2009; Cordonier Segger, Szabó, & Harrington 2021. As is well known, Hans Jonas was among the first thinkers to recognize with remarkable foresight the necessity of rethinking responsibility in light of nature, technological power, and future generations (Jonas 1984).

13 This requirement has led, for example, to increasingly precise legal definitions of environmental damage. At the Italian national level, for example, environmental damage is defined by Legislative Decree No. 152/2006 as "any significant and measurable direct or indirect deterioration of a natural resource or of the utility ensured by it." The definition refers, among other things, to damages affecting protected species and natural habitats, inland and marine waters, coastal areas, land contamination creating significant risks for human health, as well as forms of air pollution regulated under national legislation. Traditionally, environmental offences in many legal systems mainly concerned unauthorized emissions or violations of administrative licenses and therefore largely belonged to the sphere of administrative law (Faure & Visser 1995). However, the growing recognition of environmental harm as a matter requiring criminal protection led, also under the influence of Directive 2008/99/EC, to the progressive introduction of environmental crimes into national criminal systems. In Italy, Directive 2008/99/EC was first implemented through Legislative Decree No. 121/2011, while Law No. 68/2015 subsequently introduced a specific title devoted to environmental crimes into the Italian Criminal Code.

2016). The Sharma case illustrates the importance of this meta-responsibility: where the relevant normative framework remains uncertain or underdeveloped, the attribution of responsibility and the justification of punishment become equally problematic. Meta-responsibility therefore concerns the capacity of institutions to create the normative conditions under which responsibility can be recognised and enforced in practice. In this sense, it points towards a transformative conception of justice, understood as an approach that seeks not merely to sanction past wrongdoing but to reshape legal and institutional arrangements so as to prevent future environmental harm.

(3) Third, these considerations call for a rethinking of both the concept of responsibility (3.1.) and its normative foundations, as well as of the foundations and functions of punishment in the context of environmental harm (3.2).

(3.1.) With regard to the concept of responsibility, environmental harm makes it necessary to move beyond purely individual and retrospective models towards a broader, collective, and proactive conception, capable of grounding preventive legal frameworks and adequate forms of punishment for both individuals and corporations involved in environmental offences.[14] To this end, environmental theorists have proposed contributory, collective, and differentiated accounts of responsibility for environmental harm. According to this perspective, every human being bears at least a responsibility not to contribute to environmental damage and to help reduce the conditions that generate it (Shockley 2017). As is evident, this is no longer a purely retrospective and guilt-based form of responsibility grounded exclusively on past actions and their imputability, but rather a proactive and prospective conception of responsibility oriented toward the future. Individual responsibility still plays an important role insofar as each person may contribute, directly or indirectly, either to the mitigation of harmful environmental conditions or to their worsening. In this sense, each individual may be regarded at least as an *indirect cause* in the development of better or worse environmental conditions. Environmental responsibility is also *differentiated*, insofar as it recognizes that some states, communities, corporations, or social actors bear a greater degree of responsibility than others for causing environmental damage. In this sense, environmental responsibility has been said to be both *common but differentiated* (Rodeiro 2022): common, because all human beings participate, at least indirectly, in environmental processes affecting the planet; differentiated, because responsibility exists in different degrees, depending on the concrete contribution to environmental harm.[15]

14 The urgency of strengthening national and international legislation in this direction has therefore increasingly emerged, possibly also through more uniform forms of regulation, for instance within the European context, where significant differences between legal systems still persist (Olsen Lundh 2024).

15 In this sense, reference has also been made to a "historical responsibility" approach, according to which responsibility for environmental harm should be assessed not only on the basis of present conduct, but also in light of the cumulative and long-term contribution that particular states, corporations, or communities have made to environmental degradation over time. This also constitutes the basis for forms of compensation and reparation for environmental harm through the payment of an "ecological debt", thereby entering into a broader logic of restorative justice (see

(3.2.) The way in which the legal framework and the concept of responsibility are conceived has significant implications for the justification, functions, and forms of punishment. For example, the "polluter pays principle" has been recognized as an important starting point for grounding environmental accountability. At the same time, however, its application requires significant modifications and adaptations in light of concrete circumstances and presents several practical difficulties closely related to the attribution of responsibility. In many cases, pollution results from the cumulative contribution of multiple actors, making it difficult to identify who exactly the "polluter" is. Moreover, questions arise as to when, according to which criteria, and to whom compensation should be paid, as well as what should count as compensable damage. Consequently, the "polluter pays principle" cannot operate in isolation but must be complemented by the principle of common but differentiated responsibilities (see Bugge 2009).

Moreover, given that environmental offences are often not directly intended by the agents involved – as illustrated by the Swedish case discussed above – establishing *mens rea* may become particularly difficult. This raises the question of the proper basis of punishment. The main positions oscillate between strict liability, which dispenses with proof of a culpable mental state and attaches responsibility to the occurrence of harm itself, and negligence-based models centred on breaches of duties of care and due diligence.

Strict liability does not require proof of guilt, intention, or negligence, but grounds responsibility primarily on the commission of the prohibited act itself. From this perspective, what matters is that a prohibited harmful activity has been carried out in violation of legal provisions, permits, or licensing requirements. Several studies have emphasized the preventive and deterrent function of this model, arguing that punishment based on strict liability may encourage greater attention and caution with regard to environmentally risky activities. In this sense, it has been proposed that environmental punishment should partially shift from a purely guilt-based model toward a more preventive logic linked to the broader need for environmental protection (see Ali, Permana et al. 2022).[16]

By contrast, another line of thought rejects the exclusive reliance on strict liability and instead grounds environmental responsibility and the justification of punishment in duties of care and standards of due diligence, thus framing environmental offences primarily in terms of negligence. Under this approach, liability arises not merely from the occurrence of harm but from the failure to exercise the level of care reasonably required to prevent environmental damage. In this regard, one particularly interesting aspect of the Sharma case is precisely the role attributed to the *duty of care*. The duty of care generally

Schlosberg & Collins 2014).

16 These would therefore be cases involving offences grounded in the violation of permits, authorizations, or regulatory requirements, namely prohibited forms of conduct resulting from the absence or breach of administrative norms.

requires that, whenever a person knows – or could reasonably be expected to know – that their actions may harm the environment, they must take all the measures that can reasonably be demanded in order to *prevent* the danger or limit its consequences. Failure to comply with such duties may give rise to forms of liability grounded in negligence. It is precisely within the law of negligence that the notion of duty of care has its roots in the common law tradition. Indeed, liability for negligence presupposes, first of all, the existence of a duty of care owed by the agent:

> Once a duty is established, the two additional elements to be satisfied before establishing liability in negligence are (1) that there was a failure to observe the standard of care required, and (2) that that failure has caused harm which was reasonably foreseeable. Simply put, the three essential elements of negligence are: (1) a duty to take care; (2) a failure to take care; and (3) harm resulting from that failure to take care (Greiner 2013, 529–530).[17]

By further developing legal standards concerning negligence, states may strengthen the recognition of duties of care owed to the environment. The advantage of this approach lies in the introduction of standards of due diligence that individuals, corporations, and institutions are required to follow, together with the possibility of applying forms of liability in cases of negligent failure to comply with them. From this perspective, liability may arise whenever the agent was aware – or reasonably should have been aware – of a significant risk that the unlawful conduct or harmful consequence could occur, and nevertheless failed to exercise the required standard of care to prevent it (see Cohen 1068). Also, this approach would produce a shift from a retrospective and largely reactive approach to a proactive and preventative one involving a positive environmental duty (Gunningham 2017, 202), thereby taking the survival of both future generations and the environment into account in a preventive way. In this way, responsibility would shift from a model grounded primarily on causal connection to one increasingly based on normative relations and expectations that the agent must be able to understand and comply with (see Werner 2016).[18]

Consequently, two broad approaches have emerged in contemporary debates on environmental responsibility and punishment. On the one hand, some scholars have defended forms of strict liability grounded primarily in the commission of the prohibited act itself. On the other hand, others have sought to ground environmental responsibility and punishment in duties of care and standards of due diligence, thereby linking liability to negligence rather than intentional wrongdoing (see Egteren & Smith 2002). This latter approach has increasingly supported a shift from predominantly administrative responses towards the genuine criminalization of environmental harm and more preventive forms of

17 As also stated in *Newhall Land & Farming Co. v. Superior Court*, actionable negligence presupposes "a legal duty to use due care, a breach of such legal duty, and the breach as the proximate or legal cause of the resulting injury" (as quoted in Kay 2006, 170).

18 The bridge connecting prospective and retrospective responsibility is therefore constituted by the normative expectations underlying both forms of responsibility (Werner 2016, 132).

responsibility. It is precisely this shift that Directive (EU) 2024/1203 appears to embrace. The Directive will now be examined in order to assess how it addresses the three issues identified above and whether its response can ultimately be regarded as convincing.

III. Directive (EU) 2024/1203 as a Hybrid Model of Justice, Responsibility and Punishment

Following an evaluation of the 2008 European Directive on the protection of the environment through criminal law, the European Commission concluded that that Directive had produced only limited practical effects and that too few environmental cases had been successfully investigated and prosecuted (see Lundh 2024). This led to its revision and replacement by Directive (EU) 2024/1203, which introduces several new elements and features deserving closer attention as possible responses to the problems previously discussed concerning environmental harm.

First of all, as a response to the need to define environmental harms as clearly and precisely as possible in order to ensure the applicability of the principle of legality expressed by the maxim *nullum crimen sine lege* (1.), Article 3 of Directive (EU) 2024/1203 provides a detailed list of twenty *criminal offences* – compared to the nine *offences* contained in the previous 2008 Directive. While this one mainly focused on more traditional forms of environmental harm – such as pollution, waste management, protected species, and habitats – the new Directive adopts a broader and more systematic understanding of environmental harm by extending and further detailing environmental protection in relation to a wider range of ecological risks and regulated activities, including water extraction, ship recycling, invasive alien species, greenhouse gases, and deforestation-related products. The offences are now defined through more precise references to specific European regulations and legal obligations, thereby strengthening legal certainty and reinforcing the applicability of the principle *nullum crimen sine lege*. At the same time, the Directive clearly reflects a broader shift from a predominantly administrative approach toward the genuine criminalization of environmental harm, while also moving from a merely retrospective model of punishment toward a more preventive and proactive conception grounded in due diligence, duties of care, and environmental protection.

This preventive logic emerges particularly clearly in provisions such as Article 3(e), which criminalizes the carrying out of projects without the required environmental authorization, thereby linking criminal liability to the violation of preventive procedures designed to assess and avoid environmental risks in advance. A similarly preventive rationale emerges from Article 3(r), concerning invasive alien species, where criminal liability arises in connection with activities capable of generating systemic ecological harm even before such harm has fully materialized. The provision therefore reflects an anticipatory form of environmental protection aimed at preventing systemic harm to ecosystems and biodiversity.

As regards the conception of responsibility underpinning environmental punishment (3.), the new Directive exhibits both important elements of continuity and significant developments in comparison with its predecessor. On the one hand, Article 6 on the liability of legal persons remains substantially unchanged. The Directive thus confirms a conception of environmental responsibility that extends beyond individual conduct. On the other hand, it continues to attach considerable importance to the personal culpability of natural persons, thereby preserving a central role for individual fault within the attribution of environmental criminal responsibility.

Article 3 provides, as a general rule, that the listed conducts constitute criminal offences when they are unlawful and committed *intentionally*. For several particularly serious environmental offences, however, criminal liability may also arise where the conduct is carried out with at least *serious negligence*. Unlawful conduct may therefore constitute a criminal offence even in the absence of direct intent. The Directive does not, however, adopt a model of pure strict liability. Rather, it progressively broadens the scope of environmental responsibility beyond intentional wrongdoing to include forms of negligence, duties of due diligence, and preventive obligations.

Beyond the corporate dimension of environmental liability, the new Directive further develops institutional and shared forms of environmental responsibility that were less explicit in the previous Directive. This emerges, first of all, from Recital 12, according to which Member States remain free to adopt more stringent rules, including provisions concerning the criminal liability of public institutions. Responsibility for environmental harm is thus no longer conceived exclusively at the level of private actors, but may also extend to public and institutional structures. A similar logic underlies Article 21, absent from the 2008 Directive, which requires Member States to establish national strategies for combating environmental crime by 21 May 2027. These strategies must define the roles and responsibilities of all competent authorities involved in combating environmental offences, including forms of coordination and cooperation between national authorities, Union bodies, and European networks dealing with environmental crimes, particularly in cross-border cases. Environmental responsibility therefore increasingly acquires a coordinated, transnational, and shared dimension.

But one of the most significant innovations introduced by Directive (EU) 2024/1203 concerns precisely the definition of more precise and articulated sanctions, something that was largely absent from the 2008 Directive. This absence probably contributed to the limited practical effectiveness of the previous framework: as already observed, the principle *nulla poena sine lege* requires not only the clear definition of criminal offences, but also sufficiently determinate sanctions. In this regard, Article 5 now provides that *penalties for natural persons* must be effective, proportionate, and dissuasive, thereby explicitly emphasizing the deterrent function of punishment. At the same time, the Directive introduces different levels of imprisonment depending on the gravity of the offence and the consequences produced. In this respect, the Directive clearly preserves an important retributive and proportional dimension of punishment, since sanctions are calibrated according to the seriousness of the offence and the magnitude of the resulting harm.

At the same time, however, the Directive does not rely exclusively on retributive punishment, but also introduces restorative measures, that is, measures primarily concerned with repairing environmental harm and restoring damaged ecosystems and affected communities rather than merely imposing sanctions. For example, Articles 5 and 7, concerning sanctions applicable respectively to natural and legal persons, introduce a wide range of accessory criminal and non-criminal penalties that reveal a restorative understanding of environmental justice. Particularly significant are the obligations to restore the environment where the damage is reversible, or to compensate for environmental damage where restoration is impossible. These measures move beyond a purely punitive logic and clearly open toward forms of restorative justice centered on reparation, remediation, and environmental recovery. Other sanctions – such as the withdrawal of permits and authorizations, exclusion from public funding, judicial supervision, closure of establishments, and publication of judicial decisions – may in some cases exert an even more effective deterrent and preventive impact than imprisonment itself.

Especially relevant, in light of the previous theoretical discussion, is Article 7(i), which allows legal persons to be required to establish due diligence systems aimed at strengthening compliance with environmental law. In this respect, the Directive appears to take a clear position in favor of a model grounded not on pure strict liability, but rather on due diligence, preventive responsibility, and the management of environmental risks. Punishment is therefore no longer conceived solely as retrospective retribution for past harm, but increasingly also as a means of preventing future environmental damage and transforming organizational practices and institutional behavior.

This preventive and transformative orientation emerges even more clearly in Article 16, explicitly entitled "Prevention," which requires Member States to adopt appropriate preventive measures, including information and awareness-raising campaigns directed at both public and private stakeholders, as well as research and educational programs aimed at reducing environmental offences and the risk of environmental crime.

Other highly significant innovations suggest that the Directive also moves in the direction of a transformative understanding of environmental justice, by introducing elements aimed at modifying the structural conditions underlying environmental harm. Article 14, for instance, requires Member States to adopt measures protecting persons who report environmental offences, thereby recognizing the importance of participation, transparency, and access to justice in environmental matters. Equally significant are the provisions concerning information and education. Article 15 refers to the publication of information in the public interest, while Article 18 explicitly requires specialized training for judges, prosecutors, police officers, judicial staff, and competent authorities involved in environmental proceedings. The Directive thus shows a clear awareness that knowledge and information are essential instruments for combating environmental crime: the public has a right to be informed, while judicial and administrative authorities have a corresponding duty to acquire the expertise necessary to act effectively.[19] Similarly, Article 19 further emphasizes the importance of coordination among competent national authorities, while Article 22 requires Member States to collect registration and statistical data concerning environmental offences in order to monitor the phenomenon more effectively. Taken together, these provisions suggest a conception of environmental justice that is not limited to punishment after the fact, but seeks to transform the social, institutional, cultural, and informational conditions that allow environmental harm to emerge and persist. At the same time, they reflect an awareness of the meta-responsibility of institutions to establish clear normative frameworks and conditions of accountability within the environmental field (2.).

A final noteworthy aspect is that the Directive explicitly refers to ecocide, thereby showing sensitivity to the ongoing debate concerning the possible inclusion of ecocide within the Rome Statute of the International Criminal Court (see Nowak 2022). In particular, the Directive refers to conduct "comparable to ecocide," thus acknowledging the growing international discussion surrounding the criminalization of severe and widespread environmental destruction.[20]

IV. Concluding Remarks

This article has shown how the context of the contemporary environmental crisis profoundly challenges several traditional categories of legal and moral philosophy, especially those concerning responsibility and punishment. Classical guilt-based, retrospective, and individual models of responsibility appear increasingly insufficient when confronted with environmental harms that are often collective, systemic, transnational, and projected toward the future. Environmental harm forces legal and philosophical reflection to rethink responsibility beyond narrow causal and individualistic paradigms, while at the same time preserving the relevance of personal culpability, due process, and the principle of legality. Likewise, punishment itself can no longer be understood exclusively either in retributive terms or according to a purely utilitarian and deterrent rationale.

The article has therefore reconstructed some of the main theoretical difficulties and proposals emerging within contemporary debates, including discussions surrounding

19 For an interesting reflection on the necessary cooperation between science and legal judgment – in a way that still allows judges to preserve their critical autonomy while addressing contemporary challenges, including environmental ones – see Farano (Farano 2024).

20 On the connection between ecocide and structural injustice, and on the argument that it is the behaviours producing ecocide that must be changed, together with the need for coordinated public action aimed at prevention, see Rodeiro (Rodeiro 2022).

strict liability, negligence, duties of care, collective and corporate responsibility, restorative justice, and transformative approaches to environmental harm. Against this background, the analysis of Directive (EU) 2024/1203 has shown how the new European framework appears to incorporate several of these developments and tensions within a single legal instrument that Member States will be required to implement in the coming years.

The Directive emerges, first of all, as the expression of a mixed conception of punishment. On the one hand, it preserves important retributive and proportional elements, particularly through the differentiation of sanctions according to the gravity of offences and the harm caused. On the other hand, it explicitly emphasizes deterrence and prevention by requiring sanctions to be dissuasive. At the same time, the Directive adopts a similarly hybrid conception of responsibility: while continuing to recognize the personal culpability of natural persons, it also extends responsibility to legal persons and increasingly to institutional and collective dimensions of environmental governance. Moreover, through the precise definition of environmental criminal offences and sanctions, the Directive strengthens the applicability of the principles *nullum crimen sine lege* and *nulla poena sine lege*, thereby addressing some of the major weaknesses of the previous framework.

References

Ali M., Permana W. P. N., Nurhidayat S., Syafi'le M., Triwantoro A. N., & Hakim A. L. 2022. "Punishment without Culpability in Environmental Offences," *Cogent Social Sciences* 8:1–15. https://doi.org/10.1080/23311886.2022.2120475

Almer C. & Goeschl T. 2010. "Environmental Crime and Punishment: Empirical Evidence from the German Penal Code," *Land Economics* 86(4):707–726. https://doi.org/10.3368/le.86.4.707

Aristotle 2012. *Nicomachean Ethics*. Translated, with an interpretative essay, notes, and glossary by R. C. Bartlett and S. D. Collins. Chicago & London: The University of Chicago Press.

Ashworth A. & Horder J. 2013[7]. *Principles of Criminal Law*. Oxford: Oxford University Press.

Battistoni G. & Magni C. 2025. "The Right to Punish Today: Navigating Scientific and Cultural Shifts," *Ordines. Per un sapere interdisciplinare sulle istituzioni europee* 1:262–279. URL: https://www.ordines.it/punishment-between-narrative-dimensions-and-todays-challenges/

Bugge H. C. 2009. "The Polluter Pays Principle: Dilemmas of Justice in National and International Contexts," in J. Ebbesson & P. Okowa (Eds.), *Environmental Law and Justice in Context* (pp. 411–428). Cambridge: Cambridge University Press.

Cohen A. M. 1992. "Environmental Crime and Punishment: Legal/Economic Theory and Empirical Evidence on Enforcement of Federal Environmental Statutes," *The Journal of Criminal Law and Criminology (1973-)* 82(4):1054–1108. https://doi.org/10.2307/1143716

Cordonier Segger M.-C., Szabó M. & R. Harrington A. 2021. *Intergenerational Justice in Sustainable Development Treaty Implementation: Advancing future generations rights through national institutions*. Cambridge: Cambridge University Press.

Eckersley R. 1999. "The Discourse Ethic and the Problem of Representing Nature," *Environmental Politics* 8(2):24–49. https://doi.org/10.1080/09644019908414460

Egteren H. va. & Smith T. 2002. "Environmental Regulations Under Simple Negligence or Strict Liability," *Environmental and Resource Economics* 21:367–394. https://doi.org/10.1023/A:1015144713068

Environment Protection and Biodiversity Conservation Act. Available online: https://www.legislation.gov.au/C2004A00485/latest/versions.

European Convention on Human Rights 1950. Available online: https://www.echr.coe.int/documents/d/echr/convention_ENG

Farano A. 2024. *Il giudice e la buona scienza. L'istituzione del fatto scientifico in giudizio*. Torino: Giappichelli.

Faure M. & Visser M. 1995. "How To Punish Environmental Pollution? Some Reflecions on Various Models of Criminalization of Environmental Harm," *European Journal of Crime, Criminal Law and Criminal Justice* 4:316–368. URL: https://brill.com/view/journals/eccl/3/4/article-p316_1.xml?srsltid=AfmBOoq85invhyccOo9RMJQhkf1nU-i3DzIzUpim710H-xFDxWb9uEHl

Feuerbach P. J. A. 1832[11]. *Lehrbuch des gemeinen in Deutschland gültigen peinlichen Rechts*. Giessen: Druck und Verlag von Georg Friedrich Heyer, Vater.

Gosseries A. & Meyer H. L. 2009. *Intergenerational Justice*. New York: Oxford University Press.

Greiner R. 2014. "Environmental Duty of Care: From Ethical Principle Towards a Code of Practice for the Grazing Industry in Queensland (Australia)," *Journal of Agricultural and Environmental Ethics* 27:527–547. https://doi.org/10.1007/s10806-013-9475-6

Gunningham N. 2017. "Should a General 'Duty of Care' for the Environment Become a Centerpiece of a 'Next Generation' Environment Protection Statute?," *Environmental and Planning Law Journal* 34(3):198–208. Link

Jonas H. 1984. *The Imperative of Responsibility: In Search of an Ethics for the Technological Age*. Chicago: University of Chicago Press.

Kay R. M. 2006. "Environmental Negligence: A Proposal for a New Cause of Action for the Forgotten Innocent Owners of Contaminated Land," *Colorado Law Scholarly Commons* 94(1):149–174. https://doi.org/10.2307/20439029

Louka E. 2006. *International Environmental Law. Fairness, Effectiveness, and World Order*. Cambridge, New York: Cambridge University Press.

Morar N. 2008. "The Limits of Discourse Ethics Concerning the Responsibility toward Nature, Nonhuman Animals, and Future Generations," in B. Olaru (Ed.), *Autonomy, Responsibility, and Health Care. Critical Reflections* (pp. 129–158). Bucharest, Romania: Zeta Books.

Nolt J. 2015. "Future Generations in Environmental Ethics," in S. M. Gardiner & A. Thompson (Eds.), *The Oxford Handbook of Environmental Ethics* (pp. 344–354). Oxford: Oxford University Press.

Nowak E. 2022. "From Genocide to Ecocide. Essentials of a New Category of International Crime against Humanity," *Undecidabilities and Law. The Coimbra Journal for Legal Studies* 2:75–98. https://doi.org/10.14195/2184-9781_2_4

Parfit D. 1984. *Reasons and Persons*. Oxford: Clarendon Press.

Passmore J. 1974. *Man's Responsibility for Nature: Ecological Problems and Western Traditions*. London: Duckworth.

Rodeiro M. 2022. "Common but Differentiated: A Theory of Responsibility for Environmental Harm," *Ethics & the Environment* 27(1):79–100. https://doi.org/10.2979/ethicsenviro.27.1.05

Schlosberg D. & Collins B. L. 2014. "From Environmental to Climate Justice: Climate Change and the Discourse of Environmental Justice," *WIREs Climate Change* 5:359–374. https://doi.org/10.1002/wcc.275

Shockley K. 2020. "Enabling Collective Responsibility of Environmental Justice," in S. Bazargan-Forward & D. Tollefsen (Eds.), *The Routledge Handbook of Collective Responsibility* (pp. 486–500). New York: Routledge.

Shockley K. 2017. "Individual and Contributory Responsibility for Environmental Harm," in S. M. Gardiner & A. Thompson (Eds.), *The Oxford Handbook of Environmental Ethics* (pp. 265–275). Oxford: Oxford University Press.

Vantin S. 2024. *Le metamorfosi della responsabilità. Tecnica, diritto, bioetica*. Torino: G. Giappichelli Editore.

Wang Y., Li Y., Ma Z., & Song J. 2019. "The Deterrence Effect of a Penalty for Environmental Violation," *Sustainability* 11(4226):1–19. https://doi.org/10.3390/su11154226

Werner M. H. 2016. "Verantwortung," in K. Ott, J. Dierks, & L. Voget-Klescin (Eds.), *Handbuch Umweltethik* (pp. 132–135). Stuttgart: J.B. Metzler.

Ethics in Progress (ISSN 2084-9257). Vol. 17 (2026). No. 1,
Art. #7, pp. 121-137. DOI:10.14746/eip.2026.1.7

From Celebrities to Ordinary Citizen. Limits of the Legitimacy of Social Punishment

Natalia Witosza

(Jagiellonian University, Doctoral School in the Social Sciences, Faculty of Law and Administration, Cracow, Poland; natalia.witosza@doctoral.uj.edu.pl)
ORCID: 0009-0002-7863-353X

Abstract: Public discourse and academic debate on the permissibility of social punishment range from strong endorsement to categorical rejection. This paper defends a moderate view: social punishment practices such as shaming, canceling, or deplatforming may be legitimate, but only under certain conditions. It focuses on one key criterion: the social status of the wrongdoer and asks whether celebrities and non-celebrities should be treated differently. I argue that the epistemic power of celebrities justifies greater exposure to social punishment, whereas non-celebrities should generally be subject only to private punishment. The paper outlines the normative implications of this distinction. Due to the subject of this paper and given that it is a work in political philosophy, the method used here is reflective equilibrium.

Keywords: Social punishment; epistemic power; celebrities; public shaming; deplatforming.

I. Introduction

Let me begin by considering two cases. The first one relates to the ban on Donald Trump's social media accounts.[1] In January 2021, United States Capitol in Washington, D.C., was attacked by a group of supporters of President Donald Trump. This happened because Trump and his supporters made false claims of election fraud in the 2020 US presidential election. The attackers believed that Trump was the winner of the election. In a series of posts on various social media platforms, Trump encouraged his followers not to accept the result of the election, which his opponent, Joe Biden, had won. In consequence, the main social media platforms, including Twitter, YouTube, Facebook, Reddit, Twitch, decided to suspend his accounts (Sitaraman 2023, 500).

The second case that I would like to consider concerns the reaction of Twitter users to a series of racist posts published by Justine Sacco, who had previously been an unknown employee of an American corporation. In December 2013, during a plane trip to Cape Town,

1 The research for this paper has been supported by the National Science Centre, Poland ("Shaming, Canceling, No-platforming: Boundaries of Social Punishment Practices"; grant no. 2024/53/B/HS5/04312; PI: Dr hab. Wojciech Ciszewski).

Sacco posted on her Twitter account 'Going to Africa. Hope I don't get AIDS. Just Kidding. I'm white!'. Prior to this incident, Sacco's account was followed by 170 people, but after this message was posted, she became the number one worldwide trend on Twitter. The response of Twitter users was the hashtag #HasJustineLandedYet and thousands of angry tweets posted in response to her joke. As a result, her employer decided to terminate her employment, and she reportedly suffered long-term psychological harm (Ronson 2015, 85–102).

The two presented cases are examples of two different types of social punishment practices, meaning non-legal forms of response to violations of moral norms that happen between people outside of their institutional roles (Radzik et al. 2020, 11). In the case of Trump, deplatforming was applied (Sitaraman 2023, 500) and in the case of Sacco, public shaming was applied (Ronson 2015, 85–102). These two cases serve as paradigmatic cases of non-legal forms of punishing individuals of two different social standings – one is a well-known politician, and the other is an ordinary citizen.

This paper seeks to address a key question: what criteria should guide our assessment of the legitimacy of social punishment practices? It focuses on one particular criterion: the social status of the individual who committed the wrongdoing. More specifically, in this paper I will ask whether social punishment practices should be applied differently to celebrities compared to non-celebrities, and how this difference should be understood. In other words, how should we assess the punishments described above of Justine Sacco and Donald Trump?

I begin by introducing the concept of social punishment. In the next section, I present how the role of celebrities, especially their duties, is understood. I then introduce the concept of epistemic power and argue that celebrities possess it to a considerable degree. I examine the legitimacy of social punishment with regard to the social status of the person being sanctioned, providing a classification of such persons into (a) celebrities, (b) quasi-celebrities, and (c) non-celebrities. I argue that only celebrities and quasi-celebrities in qualified cases can be subjected to social punishment practices. In the following section, I propose that outside of these cases private punishment should be applied. I conclude with final remarks. Due to the subject of this paper and given that it is a work in political philosophy, the method used here is reflective equilibrium. In brief, reflective equilibrium is the mutual adjustment of principles and judgments in light of relevant arguments and theories (Knight 2025).

II. Deplatforming, Public Shaming, and other Social Punishment Practices

Typically, when discussing the issue of punishment, the first thing that comes to mind is legal punishment imposed by state authorities in a formal procedure. However, social punishment should be distinguished from these more familiar and more extensively researched forms of punishment. The most comprehensive account of social punishment was proposed by Linda Radzik, particularly in her book *The Ethics of Social Punishment: The Enforcement of Morality in Everyday Life* (Radzik et al. 2020). This paper relies on Radzik's account of social punishment. She defines social punishment as "non-legal forms of authorized, intentional, reprobative, reactive harming between people who are not acting within hierarchically-structured institutional roles" (Radzik et al. 2020, 11).

Radzik claims that people of equal social status (e.g., friends, family members, colleagues, or people from the same community) are entitled to engage in certain punitive practices. Radzik bases her definition of social punishment on the definition of punishment provided by David Boonin (Boonin 2008), and she also draws on classical notions of informal punishment by society developed by John Locke and John Stuart Mill. She gives five conditions for social punishment to occur, which are: (a) the harm condition, (b) the reaction condition, (c) the reprobation condition, (d) the intention condition, and (e) the authorization condition. According to Radzik's account, the authorization condition (e) means that social punishment occurs in everyday relations among social equals, where authority is symmetrical (e.g., between friends), rather than in institutional roles, as in the case of legal punishment. The intention condition (d) requires that the punishment be intentional, rather than a mere side effect of some other action. The reprobation condition (c) holds that, for an action to count as a form of social punishment, it must express moral disapproval. The reaction condition (b) requires that social punishment be a response to a perceived wrongdoing and that it be directed at the individual who is the wrongdoer. Finally, with the harm condition (a), the relevant harm need not be limited to the deprivation or restriction of a right. It may also consist in hard treatment, exposure to the critical gaze of others, or damage to one's reputation (Radzik et al. 2020, 11–20).

Social punishment consists of various practices, the most frequently discussed in the literature being deplatforming, canceling, and public shaming. At the beginning of this paper, I introduced two cases. I identified the suspension of Donald Trump's social media accounts in 2021 as an instance of deplatforming. By contrast, I argued that the condemnation of Justine Sacco for racism constituted a case of public shaming. Canceling is a form of boycotting that consists in the withdrawal of any kind of support (viewership, social media follows, purchases of products associated with the person, etc.) from individuals who have said or done something regarded as morally reprehensible or highly problematic (Ng 2020, 623). For this reason, by definition, canceling applies only to widely recognized figures, since it presupposes the withdrawal of support. Consequently, it cannot apply to private individuals and will therefore not be considered in this article.

Deplatforming is the practice of excluding a particular individual from access to a platform through which they could express their views (Veber 2024, 971). The reason for such exclusion is the moral condemnation of that person's actions or beliefs. A platform may include not only social media (X, formerly Twitter; YouTube; Facebook; Reddit; Twitch), but also any forum in which invited speakers express their views, such as spaces associated with a particular university (e.g., a public lecture at this university) or particular traditional media (radio broadcasts, television programs, or newspaper articles) (Veber 2024, 971). The described sanction is imposed and enforced by the person or body that controls access to the given platform (Sitaraman 2023, 532–533). This may include a social media platform administrator, a festival organizer, an editor-in-chief of a journal, or a university rector.[2]

The literature contains three slightly different definitions of the practice of public shaming. The first holds that it is an action that involves expressing moral disapproval toward a person who has violated moral norms (Billingham & Parr 2020a, 727), for instance by posting a racist comment on their social media. The second account maintains that shaming must be a stigmatizing action directed at a particular individual who has violated a moral norm. The person in question must be portrayed as inferior to a "normal" member of the community (Adkins 2019, 2). The third position, in turn, equates shaming with especially malicious forms of public criticism, such as highly personalized, abusive, and vitriolic attacks that humiliate their targets (Billingham & Parr 2020a, 727). Despite these differences in how shaming is conceptualized, one feature remains constant: it must contain an element of publicity. If it does not occur before an audience, either directly or online, there must at least be a sense that it could be made public (Adkins 2019, 2). Moreover, an important element of shaming is the action through which it is expressed, rather than mere inaction. In this sense, shaming has a distinctly active character and cannot be expressed simply by refraining from acting.

Public discourse, as well as academic debate on the permissibility of social punishment, encompasses a very broad spectrum of positions: from views that affirm social sanctioning practices as potential instruments of justice (Janssens & Spreeuwenberg 2022, 89–90) to views that categorically reject them as incompatible with fundamental values such as freedom and equality (Frye 2022, 19). In this context, a more moderate position appears to be the most reasonable and plausible: namely, the view that social punishment practices may be legitimate, but only under certain conditions. Framing the issue in this way leads naturally to a crucial question: what guidelines should be used to assess the legitimacy of social sanctions in a particular case? In contrast to legal punishment, social punishment lacks clear guidelines as to what counts as a punishable act (Janssens & Spreeuwenberg 2022, 95), although the literature does propose several possible criteria that seem relevant here. Examples include the relationship between the punisher and the punished, the nature and severity of the wrongdoing, and the legitimate

2 It might be difficult to find as social equals, for example, a social media platform administrator and an individual user, or a university rector and an academic teacher. There is a significant asymmetry in their resources and power. However, I think that in this context, "social equals" should be understood broadly as people who do not act in state-related institutional roles that can impose legal punishment. By using the term "social equals" in this broader sense, I do not claim that these actors are equal in wealth, influence, or power. I rather emphasize that they interact outside the system of state-backed legal punishment and within privately or institutionally defined frameworks. Moreover, in the case of deplatforming, exclusion from a given platform occurs as a result of social responses such as criticism for a norm violation.

interests of victims and communities (Radzik et al. 2020, 47). Billingham and Parr introduce other criteria for the moral justification of punishment for moral wrongdoing: proportionality, necessity, respect for privacy, non-abusiveness, and reintegration (Billingham & Parr 2020b, 5–10). Of course, there are other possible criteria that come to mind, such as whether the person has already faced any form of penalty (for example, a legal one) or the person's awareness of the facts relevant to the case.

However, this paper focuses on one particular criterion that appears to be relevant in this context: the social status of the individual who committed the wrongdoing. Whether the wrongdoer is a celebrity attracting significant public and media attention, or a non-celebrity with no real public presence, may affect how the harmfulness of the situation is assessed. This paper seeks to explore and develop that intuition.

III. Celebrities and Their Duties

To clarify my idea of the criterion for assessing the legitimacy of social punishment – namely, the social status of the individual who committed the wrongdoing – and to explain why celebrities should be treated differently from non-celebrities, it is necessary to define these terms. Various accounts of celebrity have been proposed in the literature, ranging from relatively narrow to broad definitions. For example, Daniel Boorstin defines a celebrity as "a person well-known for their well-knownness" (Boorstin 1962, 57). Alternatively, celebrity is "a quality or status characterized by a capacity to attract attention, generating some 'surplus value' or benefit derived from the fact of being well known (highly visible) in itself in at least one public arena" (Van Krieken 2019, 10).

It seems that a common feature of these definitions is that a celebrity is someone who is known beyond the circle of those with whom they have direct contact (Archer, Alfano, & Dennis 2024, 760). This intuition is well captured by Antoine Lilti. According to him, the "celebrated individual is not known simply to his family, his colleagues, his neighbors, his peers or his customers, but to a vast group of people with whom he has no direct contact, who have never met him and will never meet him, but who frequently encounter his public image" (Lilti 2017, 6). Following the account of Alfred Archer and colleagues (Archer et al. 2020, 28), I understand celebrity as an individual whose life is subject to high levels of public attention in ways that go beyond their specific talents, expertise, or professional role. I also maintain that celebrities are known to a vast group of people with whom they have no direct contact, as Lilti (2017, 6) argues. On the basis of this definition, a celebrity is not limited to people who are "well-known for their well-knownness" (Boorstin 1962, 57) but also includes individuals whose fame derives primarily from media visibility rather than from recognized achievements or expertise, such as members of the Kardashian family, Paris Hilton, or Addison Rae. A celebrity may also be a well-known politician who holds public office, such as Emmanuel Macron or Ursula von der Leyen, as well as prominent figures in the cultural sphere, including actors, film directors, and writers such as Angelina Jolie, Quentin Tarantino, or J. K. Rowling.

It is relatively uncontroversial that celebrities exert significant influence over large audiences (Wellman 2003, 335). Some argue that their influence gives them special moral duties towards society (Archer, Alfano, & Dennis 2024, 762; Croce, Liberti, & Vaccarezza 2025, 135–162). A common argument for attributing special moral duties to celebrities is that they bear the duties of role models. Such duties consist in serving as positive moral examples for others (Osman & Harrison 2025). Although, to some extent, all individuals may have a responsibility to set a good example, the influence of celebrities provides strong grounds for holding them to high standards in this respect. Given their public visibility and impact on audiences, celebrities can be seen as having a particular responsibility to function as moral exemplars and to encourage virtuous rather than harmful attitudes and behavior (Archer, Alfano, & Dennis 2024, 762; Wellman 2003, 335).

Another argument for the special obligations of celebrities can be called the salience magnets argument. On this view, celebrities are highly visible figures who attract and concentrate collective attention. Because many social problems are coordination problems, societies benefit from central points that help individuals align their behavior with that of others (Archer, Alfano, & Dennis 2024, 763). In large modern societies, people often lack direct information about what others are doing or what norms are being followed. Celebrities, due to their exceptional visibility, provide publicly observable signals that many people might notice at the same time. When a celebrity behaves a certain way, audiences are aware not only of the behavior itself but also that countless others are observing it as well. This shared awareness makes certain norms more salient and easier to coordinate around (Archer, Alfano, & Dennis 2024, 763).

However, the strongest argument for the claim that celebrities are subject to special obligations is that they possess a high level of epistemic power, that is, the power to influence what others believe and know (Archer, Alfano, & Dennis 2024, 762). In this paper, I employ this argument to justify my proposed account of the legitimacy of social punishment, which differentiates between wrongdoers based on their social status. The relevant criterion for this distinction is the degree of epistemic power possessed by the individual.

IV. What is Epistemic Power?

Some people’s voices are heard more than others, especially when they belong to socially dominant groups and are highly visible in the media. Some individuals are treated as authorities not only on their own experiences but also on the experiences of others who are different from them. Others are regarded as unqualified to describe even their own experiences. These are manifestations of epistemic injustice, understood as a form of injustice done to someone in their capacity as a knower (Fricker 2007).

Closely related to the notion of epistemic injustice is the concept of epistemic power, developed within social epistemology. Epistemic power can constitute a form of epistemic injustice. Archer and colleagues (Archer et al. 2020) define epistemic power as follows: "A person has epistemic power to the extent she is able to influence what people think, believe, and know, and to the extent she is able to enable and disable others from exerting epistemic influence" (Archer et al. 2020, 29). It is difficult to identify another group that is as visible in the media and attracts as much attention as celebrities, whose lives are subject to high levels of public interest that extend beyond their specific talents, expertise, or professional roles. Celebrities are situated within structured testimonial networks that allow them to reach and influence far larger audiences than the credibility of their claims would typically warrant.

Epistemic power refers to a type of power possessed by an epistemic agent. Almost everyone has some epistemic power, but some have much more than others. A teacher possesses epistemic power over her students, a mother has epistemic power over her child, and a famous influencer possesses epistemic power over his followers. Epistemic power is described in two aspects, related to the abilities possessed by the subject in question. It involves, first, the capacity to shape what other people think, believe, or know. Second, it involves the capacity to either grant or restrict other people's ability to exercise such epistemic influence themselves (Archer et al. 2020, 28–29). An example of exercising these two abilities is the relation between a teacher and her students. This teacher can influence what students believe, think, or know. She can also increase student's epistemic power by praising her student's mathematical skills, leading others to believe that student when they provide answers, even if they are occasionally wrong (Archer et al. 2020). Epistemic power can be possessed not only by individuals but also by institutions, such as universities (Lechner et al. 2026).

There are various sources of epistemic power, and some of them are entirely legitimate (Archer et al. 2020, 29). For example, a professor who specializes in the topic of gender and law possesses detailed knowledge in this area and is likely able to explain it in a clear and systematic way. Moreover, as a university employee, she holds a distinctive status as an expert in her field. However, if she were to publicly express her views, for instance on the topic of speedway racing, while relying on her academic title and the platform provided by the university, she would be exercising her epistemic power beyond the scope of her competence. The following risk arises: individuals may occupy institutional positions that grant epistemic power despite lacking the relevant expertise (Archer et al. 2020, 29).

Archer and colleagues (2020) identify two primary sources of epistemic power: perceived credibility and attention. The literature suggests that celebrities are often regarded as more credible than warranted, even on topics outside their areas of expertise. For example, after Angelina Jolie published an editorial about her preventive mastectomy, there was a significant increase in the number of women seeking genetic testing for breast cancer risk (Desai & Jena 2016). Studies also indicate that some parents consider celebrities to be credible sources on issues such as vaccine safety, even though medical professionals remain the most trusted source of information (Freed et al. 2011). Moreover, Oprah Winfrey's public support substantially increased votes for Barack Obama in the 2008 presidential primary, which also shows that celebrity endorsements can shape political behavior (Garthwaite & Moore 2012). Aside from more typical celebrity cases, there are also academics who enjoy a high degree of perceived credibility and are well known outside the academic community because of their media engagement. Academic titles carry a degree of implicit credibility even beyond one's specific domain of competence, so their perceived credibility can be even higher than that of celebrities. The degree to which a person is perceived as credible affects the extent of their epistemic power. The more credible someone is perceived to be, the greater their power to influence what people think, believe, and know, as well as whose voices can exert epistemic influence.

When it comes to attention as a source of epistemic power, it is relatively uncontroversial to claim that celebrities are often the focus of public attention. By attracting large audiences, they gain a platform that allows their testimony to be widely heard, enables them to direct public attention toward issues or voices, and even shape the political agenda. Although some politicians and experts also receive attention, their epistemic power is typically grounded in decision-making authority or relevant expertise. In contrast, celebrities often receive significant attention regardless of expertise, which gives them epistemic influence without a corresponding epistemic justification (Archer et al. 2020, 31). In 2020, during the COVID-19 pandemic, Arnold Schwarzenegger recorded a video of himself smoking one of his trademark cigars while explaining the collective benefits of individual self-isolation (Archer, Alfano, & Dennis 2024, 761). By presenting an attractive vision of what quarantine could look like, he used attention as a source of epistemic power, hoping that his followers would emulate his behavior. Schwarzenegger's self-isolation clip was viewed several million times due to the attention he attracts (Archer, Alfano, & Dennis 2024, 761).

V. Epistemic Power as a Criterion for Differentiating Social Responsibility

The most important justification for treating social punishment applied to celebrities differently from that applied to non-celebrities is that, due to their epistemic power, celebrities have a greater potential to cause harm through morally repugnant behavior. A similar thought appears in recent legal discussions, where it is argued that those who abuse their epistemic power to spread false information should face criminal, and not merely social, punishment (Mamak 2026, 6–11). We can imagine a case in which a well-known person publicly makes a harmful statement: for example, by expressing hostility toward a vulnerable group or rejecting a medical procedure. If such a statement were made by an unknown individual, it would likely have little impact. However, because of their epistemic power, a celebrity can influence a large audience, whose members may take their claims seriously and act on them. Celebrities can influence what others believe, set agendas, and help establish a shared common ground. Due to their visibility and the attention they attract, they are often vocal about issues in which they lack expertise, which can be potentially harmful (Archer & Sie 2023, 340–342). The level of attention celebrities receive can also contribute to epistemic injustice and to the harms associated with it (Smith & Archer 2020). Based on the idea of epistemic power, I treat this as both the main justification for my division and as its normative criterion, since I classify individuals into these categories according to the degree of epistemic power they possess. This does not mean that one could not argue that there are other reasons to impose certain bans on members of particular institutions, for example, a university to impose a ban on a student, in light of those institutions' special obligations. However, I maintain that such additional considerations fall outside the scope of this paper, which focuses exclusively on epistemic power as the relevant basis for differentiating the permissibility of social punishment.

Due to celebrities' significant visibility in public debate, they express their views in various situations, for example by giving interviews, publishing books, hosting live sessions, and doing streams, and gossip portals track their activities. Whereas social punishment imposed on ordinary people may be based on only fragmentary information, such as a single comment, the views and actions of celebrities are relatively easy to access (Archer & Sie 2023, 343). The status of celebrities requires them to consciously manage their image, including dealing with the negative consequences of such exposure. Through their visibility in public debate, celebrities may, at least to some extent, relinquish aspects of their privacy. Rebecca Wallbank (Wallbank 2025) claims that celebrities are unreliable narrators, because social media allows them to control the narrative and their image, so the messages they send are often misleading or even deceptive. People's trust in celebrities is precisely a result of "epistemic recklessness." Users are aware that celebrities' content can be unreliable, yet despite this knowledge, many of them still base their beliefs on such content (Wallbank 2025).

Epistemic power therefore provides both the justification for my distinction and its normative criterion. I classify individuals into three categories according to the degree of epistemic power they possess:

(a) celebrities;

(b) quasi-celebrities;

(c) non-celebrities.

In the first group (a), I include individuals whose status as celebrities is, in general, unquestionable. They are the ones known for their media statements, the ones newspapers write about, and whose actions are analyzed on social media. It does not matter whether this concerns a politician's statement about the new state budget or a celebrity's new plastic surgery. These are individuals whose public visibility and media presence give them significant epistemic power.

The crucial factor in being a celebrity is a certain form of epistemic power and the source from which it derives. The source of this epistemic power is not entirely legitimate: for example, it is not grounded solely in expertise in a particular field, since such expertise would grant epistemic authority only within that field. For the purposes of this paper, the sources of celebrities' epistemic power are attention and perceived credibility. They may also possess epistemic power from legitimate sources: for instance, a speedway driver may be an expert in speedway racing. However, if that person is famous and attracts considerable attention, they also possess a degree of epistemic power that may create a risk of harm. This epistemic power is not entirely legitimate because it derives from attention, perceived credibility, or both. Thus, even if a celebrity is an expert in a particular field, the high level of attention they receive, extending beyond their specific talents, expertise, or professional role. Their recognition by large audiences with whom they have no direct contact, means that they could be subject to social punishment practices if they commit a moral wrong, provided that the other criteria mentioned at the beginning of this paper are also met.

It is, of course, fairly uncontroversial that celebrities have a sphere of private life and may, in certain contexts, function as non-celebrities, for example in matters concerning their family life. Nevertheless, I argue that this does not undermine the legitimacy of socially punishing them for morally reprehensible behavior. For instance, if it were revealed that a famous actress had used violence against her child, the fact that this matter concerns her private life would not provide a good reason against social punishment. According to the framework I propose, celebrities do not possess any sphere that is categorically exempt from the application of social punishment.

Some individuals become famous by accident and gain attention only temporarily, especially due to a negative event, for example, by being the victim of a crime, a tragic accident, or public shaming, as in the case of Justine Sacco. Such individuals may indeed be known to large groups of people with whom they have no direct contact, and they may attract attention beyond their expertise. However, if their fame is merely temporary, they do not possess epistemic power. If, after the tragic event, they attempt to return to a normal life and do not engage in public events, issue public statements, or try to set public agendas, they remain in category (c): non-celebrities.

The second group, category (b), includes individuals who are not celebrities, as a rule, but who function as such in certain contexts. By "context" I mean a specific social group, for example a professional community. Given that epistemic resources are also distributed within social groups, the issue of epistemic power is definitely applicable here. Within its hierarchy, some individuals play more important roles than others and are subject to high levels of attention within that group, as well as being known to large numbers of people with whom they have no direct contact. They possess a higher degree of epistemic power in this particular community: they attract attention and are perceived as credible. Examples include a university professor, as opposed to a lower-ranking administrative employee or a student, or the head of a hospital department, as opposed to a staff nurse.

According to this view, a given individual functions as a celebrity only within a particular context, while outside it they remain a non-celebrity. In these specific contexts of particular communities, the person is known by a large group of people, even though they may be completely unknown outside them. They occupy institutional roles and positions that grant them a significant degree of epistemic power. I propose that, within these contexts, they should be treated as celebrities for the purposes of applying social punishment (i) by an authorized group or (ii) in connection with the function they perform.

For example, one can imagine that the aforementioned professor tells a racist joke during their class in connection with the performance of their function, in which case members of the university community, as the authorized group, may apply certain punishment practices toward them. Due to the special role they play in this particular community, I argue that they should be treated as celebrities only within it when assessing the legitimacy of social punishment practices. Consequently, social punishment should not be applied to them for wrongdoing related to their private life, except by the authorized group. For instance, the aforementioned professor may be subjected to shaming for a sexual scandal not taking place within the university context only by the entitled group, namely the academic community, rather than publicly on social media. However, if such a sexual scandal were to occur in the course of performing their professional function, for example, with a student, it could be publicly condemned.

In the third group (c), I include individuals whose status as non-celebrities is, in principle, unquestioned. The key characteristic of this group is that they possess very little or no epistemic power. Therefore, they do not participate in the public sphere, for instance, in discussions or on social media, from a position of authority. They are generally known mainly within their close social circles. It is to these individuals that the protection of privacy, in the context of the processing and distribution of personal data, applies most strongly, in contrast to persons performing important public roles (Sartor 2014, 5–6). Quite often, when non-celebrities are accidentally and unintentionally thrust into the public eye, they must face the negative consequences of this situation.

The category of non-celebrities includes most ordinary citizens. A widely discussed example of the negative consequences of a non-celebrity's accidental appearance online is the so-called "dog poop girl" incident (Billingham & Parr 2020b, 371–390). In 2005, a young woman in South Korea refused to clean up after her dog, which had defecated in a subway train. Another passenger photographed the woman and the incident, and the images were posted on a popular Korean website and widely disseminated. She was later identified, and her personal data were published online, as a result of which she was publicly shamed (Knobel & Lankshear 2007, 216).

These individuals possess epistemic power only within their close circles of friends and family. They have no lasting influence on public opinion and no recognized authority in public discourse or within their communities, in contrast to quasi-celebrities. They often face disproportionate consequences after being publicized (Billingham & Parr 2020b, 371–390).

The cases of Donald Trump and Justine Sacco clearly illustrate the difference in epistemic power between celebrities and non-celebrities. Trump possesses enormous epistemic power: he is a globally recognized political figure whose speeches, interviews, and online statements reach millions of people and actively shape public debates. Even when they are controversial, his words influence what people think and believe, and their impact is further amplified by extensive media coverage and political commentary. Justine Sacco, by contrast, was a private individual who possessed little epistemic power. Her tweet was not supported by any recognized status or reputation; it reached a wide audience only because it went "viral" due to the outrage it provoked. She had no lasting influence on public opinion and no recognized authority in public discourse (Ronson 2015, 85–102). The university professor, who illustrates the category of quasi-celebrities, also possesses significant epistemic power, but primarily within the academic environment. Through lectures, publications, and academic presentations, a professor can shape which ideas are regarded as credible and valuable within a given discipline. As a result, even if a professor is not widely known outside the university, within his community he possesses considerable epistemic power. It is precisely this that justifies treating him as a quasi-celebrity: his words and actions in the academic context have the potential to produce real epistemic consequences for others.

Returning to the question posed at the beginning of this paper – namely, whether there is a fundamental difference between the legitimization of the social punishment of Donald Trump and Justine Sacco – my answer is affirmative. Accordingly, considering their epistemic power, I argue that (a) celebrities who have committed a morally reprehensible act may legitimately be subject to social punishment, provided that other relevant criteria, which I do not discuss in this paper, are also satisfied. Regarding category (b), quasi-celebrities, I propose that social punishment may be applied to them in the same way as to celebrities only if at least one of the following two conditions is met:

(i) the punishment is imposed by an authorized group, or

(ii) the punishment is imposed in connection with the function they perform.

By contrast, social punishment should not be applied to (c) non-celebrities. This does not mean that they cannot be punished by members of society outside the legal system, a point that is discussed in more detail in the next section of this paper.

VI. Private Punishment as a Response

The question that arises is whether any form of non-legal punishment of non-celebrities should be prohibited. What if a private person commits a moral wrong for which we would like to punish them? An answer to these questions may be provided by so-called private punishment. This is an idea introduced by Linda Radzik, which she distinguishes both from the legal punishment and social punishment. The most important difference between social punishment and private punishment concerns the role of the audience. Social punishment is imposed publicly or in such a way that information about it can easily reach a wide audience. By contrast, private punishment is imposed in a private setting or even secretly (Radzik et al. 2020, 66–68). An example of such punishment is gossip or private shaming.

Radzik defines gossip as private, informal, evaluative communication about persons who are absent (Radzik 2016, 187). This understanding excludes discussions about celebrities in published forms, since they are neither private nor informal and are moreover accessible to the celebrities themselves. According to this view, gossip may be justified, but probably not as a tool for administering justice. When criticism is expressed in a private manner, as in the case of gossip, the wrongdoer may preserve their public face. As a social practice, gossip may also have positive functions, such as emotional release, the strengthening of social bonds, and the encouragement of moral reflection (Radzik 2016, 197–202).

Returning to the example of Justine Sacco, one can imagine a reaction that would take the form of private shaming in her case. For instance, her coworkers or family members could have criticized her for the racist comment on Twitter. She might have received a phone call from a disgusted cousin who condemned her behavior, or messages from a colleague saying that they did not expect such behavior from her and were deeply disappointed.

Radzik prefers the use of private punishment to social punishment in situations where both private and social punishment could be applied and argues that there are good reasons to adopt a principle that gives priority to private punishment (Pettigrove 2020, 119). This follows from her opposition to shaming (especially on social media) as a first response to a person's wrongdoing (Radzik et al. 2020, 66–68). She identifies several advantages of private punishment. First, it seems to involve a lower risk of unjust punishment, since the accused can more easily present their perspective on the matter. Second, private sanctions avoid many of the mechanisms that can lead to escalation in the form of the issue spreading "virally." Third, they may also be more effective in motivating reparation, because punishment takes place within a close social circle, whose pressure is often particularly strongly felt by the person at whom it is directed. Finally, private sanctions create fewer obstacles to the wrongdoer's reintegration into the community (Radzik et al. 2020, 66–68).

Practices aimed at sanctioning moral wrongdoing may have significant negative consequences for an individual's professional life. Individuals may experience psychological harms such as distress, humiliation, embarrassment, and shame (Billingham & Parr 2020b, 371). Social punishment practices can also inflict reputational harm by exposing a wrongdoer to the shaming gaze of an audience following the violation of social norms (Aitchison et al. 2021, 2–27). I believe that the possibility of harm is a reason to refrain from social punishment practices when there is a more suitable option. In the case of quasi-celebrities outside qualified contexts and in the case of non-celebrities, such an option is private punishment.

Of course, in cases of moral wrongdoing by celebrities, they may also be subjected to private punishment. For example, Donald Trump might receive a phone call from a distant cousin who condemns his behavior. I believe that, in the case of celebrities, this is not an appropriate response for three reasons. First, private punishment may be insufficient, because it does not reduce epistemic power in the way that social punishment does. Second, private punishment may be disproportionately lenient in the case of celebrities. Third, private individuals typically face more disproportionate consequences from social punishment than celebrities do (Ronson 2015, 85–102). I agree with Radzik that private punishment involves a lower risk of unjust punishment, avoids many of the mechanisms that can lead to escalation, and creates fewer obstacles to the wrongdoer's reintegration into the community. Because quasi-celebrities outside qualified cases and non-celebrities possess relatively little epistemic power, the risk of harm to others is lower.

VII. Conclusions

Not all instances of the application of social punishment are justified. Some are wrong, irresponsible, and disproportionately harmful to the person who becomes their target. However, there is no doubt that social punishment is a part of social reality. Today, it has gained additional significance in connection with the development of social media (Billingham & Parr 2020b, 371). In this paper, I argue for the permissibility of applying social punishment in certain cases. However, I maintain that it may be applied only when certain criteria are met, and my proposal concerns one of these criteria.

In this paper I propose a threshold for the legitimacy of applying social punishment practices, given the social status of the wrongdoer and the degree of epistemic power they possess. Because celebrities possess a large amount of epistemic power, social punishment practices are legitimate primarily in their case (provided other criteria not discussed in this paper are met). Therefore, I propose that only category (a), celebrities, should be subjected to social punishment, and category (b), quasi-celebrities, only if at least one of the following two conditions is met: (i) the punishment is imposed by an authorized group, or (ii) the punishment is imposed in connection with the function they perform. In all other cases, that is, (b) quasi-celebrities outside these qualified contexts and (c) non-celebrities in all cases, I propose the use of private punishment.

References

Adkins K. 2019. "When Shaming Is Shameful: Double Standards in Online Shame Backlashes," *Hypatia: A Journal of Feminist Philosophy* X(X):1–22. https://doi.org/10.1111/hypa.12456

Aitchison G. & Meckled-Garcia S. 2021. "Against Online Public Shaming: Ethical Problems with Mass Social Media," *Social Theory and Practice* 47(1):1–31. https://doi.org/10.5840/soctheorpract20201117109

Archer A., Cawston A., Matheson B., & Geuskens M. 2020. "Celebrity, Democracy, and Epistemic Power," *Perspectives on Politics* 18(1):27–42. https://doi.org/10.1017/S1537592719002615

Archer A. & Sie M. 2023. "Using Stars for Moral Navigation: An Ethical Exploration into Celebrity," *Journal of Applied Philosophy* 40(2):340–357. https://doi.org/10.1111/japp.12637

Archer A., Alfano M., & Dennis M. 2024. "On the Uses and Abuses of Celebrity Epistemic Power," *Social Epistemology* 38(6):759–773. https://doi.org/10.1080/02691728.2022.2153351

Billingham P. & Parr T. 2020a. "Online Public Shaming: Virtues and Vices," *Journal of Social Philosophy* 51(3):371–390. https://doi.org/10.1111/josp.12308

Billingham P. & Parr T. 2020b. "Enforcing Social Norms: The Morality of Public Shaming," *European Journal of Philosophy* 28:1–20. https://doi.org/10.1111/ejop.12543

Boonin D. 2008. *The Problem of Punishment*. Cambridge: Cambridge University Press.

Boorstin D. J. 1962. *The Image or What Happened to the American Dream*. Harmondsworth: Penguin.

Croce M., Liberti M., & Vaccarezza M. S. 2025. "The Power of Fame: Celebrities and Influencers as Moral Exemplars," in C. M. Robb, A. Archer, & M. J. Dennis (Eds.), *Philosophy of Fame and Celebrity* (pp. 135–161). London: Bloomsbury Academic.

Desai S. & Jena A. B. 2016. "Do Celebrity Endorsements Matter? Observational Study of BRCA Gene Testing and Mastectomy Rates after Angelina Jolie's New York Times Editorial," *British Medical Journal* 355:1–4. https://doi.org/10.1136/bmj.i6357

Freed G. L., Clark S. J., Butchart A. T., Singer D. C., & Davis M. M. 2011. "Sources and Perceived Credibility of Vaccine-Safety Information for Parents," *Pediatrics* 127(Supplement 1):107–112. https://doi.org/10.1542/peds.2010-1722P

Fricker M. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press.

Frye H. 2022. "The Problem of Public Shaming," *Journal of Political Philosophy* 30(2):188–208. https://doi.org/10.1111/jopp.12252

Garthwaite C. & Moore T. J. 2012. “Can Celebrity Endorsements Affect Political Outcomes? Evidence from the 2008 US Democratic Presidential Primary,” *Journal of Law, Economics, & Organization* 29(2):355–384. https://doi.org/10.2307/23487325

Janssens J. & Spreeuwenberg L. 2022. “The Moral Implications of Cancel Culture,” *Ethical Perspectives* 29(1):89–114. https://doi.org/10.2143/EP.29.1.3290737

Knight C. 2025. “Reflective Equilibrium,” *The Stanford Encyclopedia of Philosophy* (Spring 2025 Edition), E. N. Zalta & U. Nodelman (Eds.). URL: https://plato.stanford.edu/archives/spr2025/entries/reflective-equilibrium/

Knobel M. & Lankshear C. 2007. “Online Memes, Affinities, and Cultural Production,” in M. Knobel & C. Lankshear (Eds.), *A New Literacies Sampler*. New York: Peter Lang.

Lechner I. M., de Ridder J., van Woudenberg R., & Anttila S. 2026. “The University, Epistemic Power, and Institutional Epistemic Responsibilities,” *Social Epistemology* 40(2):214–228. https://doi.org/10.1080/02691728.2025.2553111

Lilti A. 2017. *The Invention of Celebrity*. Cambridge: Polity Press.

Mamak K. 2026. “Disinformation and Superspreaders: With Epistemic Power Should Come Criminal Responsibility,” *Episteme* 1–15. https://doi.org/10.1017/epi.2026.10107

Ng E. 2020. “No Grand Pronouncements Here...: Reflections on Cancel Culture and Digital Media Participation,” *Television & New Media* 21(6):621–627. https://doi.org/10.1177/1527476420918828

Osman Y. & Harrison T. 2025. “Online Celebrities as Moral Role Models: Considerations for Character Educators,” in C. M. Robb, A. Archer, & M. J. Dennis (Eds.), *Philosophy of Fame and Celebrity* (pp. 162–177). London: Bloomsbury Academic.

Pettigrove G. 2020. “Punishment and Protest,” in L. Radzik, C. Bennett, G. Sher, & G. Pettigrove, *The Ethics of Social Punishment: The Enforcement of Morality in Everyday Life* (pp. 113–134). Cambridge: Cambridge University Press.

Radzik L. 2016. “Gossip and Social Punishment,” *Res Philosophica* 93(1):185–204. https://doi.org/10.11612/resphil.2016.93.1.10

Radzik L., Bennett C., Pettigrove G., & Sher G. 2020. T*he Ethics of Social Punishment: The Enforcement of Morality in Everyday Life*. Cambridge: Cambridge University Press.

Ronson J. 2015. *So You’ve Been Publicly Shamed*. New York: Riverhead Books.

Sartor G. 2014. “The Right to be Forgotten: Dynamics of Privacy and Publicity,” in L. Floridi (Ed.), *Protection of Information and the Right to Privacy – A New Equilibrium? Law, Governance and Technology Series* 17. Cham: Springer.

Sitaraman G. 2023. “Deplatforming,” *The Yale Law Journal* 133:497–568.

Smith L. & Archer A. 2020. “Epistemic Injustice and the Attention Economy,” Ethical Theory and Moral Practice 23(5):777–795. https://doi.org/10.1007/s10677-020-10123-x

Van Krieken R. 2019. *Celebrity Society: The Struggle for Attention*. London: Routledge.

Veber M. 2024. “The Ethics of No Platforming: Defending the Defense of Immoral Ideas on University Campuses,” *Journal of Controversial Ideas* 4(2):1-16. https://doi.org/10.35995/jci04020012

Wallbank R. 2025. “Celebrity and Epistemic Influence: Evidence Resistance and Epistemic Recklessness,” in C. M. Robb, A. Archer, & M. J. Dennis (Eds.), *Philosophy of Fame and Celebrity* (pp. 178–200). London: Bloomsbury Academic.

Wellman C. 2003. “Do Celebrated Athletes Have Special Responsibilities to Be Good Role Models? An Imagined Dialog Between Charles Barkley and Karl Malone,” in J. Boxill (Ed.), *Sports Ethics: An Anthology* (pp. 333–336). Oxford: Blackwell Publishing.

Ethics in Progress (ISSN 2084-9257). Vol. 17 (2026). No. 1,
Art. #8, pp. 138-154. DOI:10.14746/eip.2026.1.8

Can Criminal Law Drive Social Change? The Promotional Function and the Affirmative Consent Debate

Jorge Fernández Mejías

(Universidad Católica del Norte, Antofagasta, Chile; jorge.fernandez@ucn.cl)
ORCID: 0009-0005-1604-9065

Abstract: The relationship between criminal law and social change has conventionally been conceived as unidirectional: penal statutes reflect, rather than shape, the normative hierarchy of a given society. This article examines the conditions under which the promotional function of criminal law – its deployment, in Bobbio's sense, as an instrument for installing new social norms rather than merely conserving existing ones – can be considered legitimate. The affirmative consent model in sexual criminal law, as enacted in Spain, serves as the test case against which three analytical perspectives map the boundaries of permissible penal promotion. First, the declaratory perspective holds that, with respect to negative sexual autonomy, the affirmative consent model performs a conservative function, declaring a pre-existing public wrong by reference to shared civic values, while assuming a constitutive function with respect to positive autonomy, thereby exceeding the declaratory warrant of criminal law. Second, the conditioned promotional perspective identifies four cumulative legitimacy criteria and finds the affirmative consent model only partially compliant with them. Third, the normative framing perspective contends that the political rationality structuring affirmative consent reform is oriented toward the legal entrenchment of an interpretative grammar rather than behavioral transformation. The three perspectives converge in finding that the promotional justification of affirmative consent fails to meet the legitimacy requirements applicable to each. The shared conclusion is that whereas criminal law may legitimately perform a conservative function in protecting negative sexual autonomy, it may not legitimately be deployed to entrench a specific communicative model for exercising positive sexual autonomy. The subsidiarity principle is advanced as the criterion that confines criminal law to a consolidating role in cultural transformation, allocating its promotional ambitions to the domain of social and educational policy.

Keywords: Affirmative consent; criminalization; sexual autonomy; social change; criminal law; consent.

Introduction

The relationship between punishment and culture is bidirectional. Garland captures this reciprocity with precision:

> Punishment is shaped by broad cultural patterns which have their origins elsewhere, but it also generates its own local meanings, values, and sensibilities which contribute – in a small but significant way – to the bricolage of the dominant

> culture. Penal institutions are thus 'cause' as well as 'effect' with regard to culture (Garland 1990, 249).

Within criminal law, and particularly in the domain concerned with defining what is "prohibited", the prevailing view has treated it primarily as an "effect" of dominant culture rather than as a potential "cause" of it. Durkheim examined the relationship through the concept of the "collective sentiments." According to the author:

> The collective sentiments to which crime corresponds must… singularize themselves from others by some distinctive property; they must have a certain average intensity. Not only are they engraven in all consciences, but they are strongly engraven. They are not hesitant and superficial desires, but emotions and tendencies which are strongly ingrained in us (Durkheim 1893/1964, 77–78).

For Durkheim, the strong bond – deeply inscribed in every collective consciousness – was discernible in the slowness with which criminal law evolved. Underlying this conception of criminal law was the idea of a discipline that confined its intervention to the protection of values of minimal yet significant content within society, a selection grounded in the breadth of social consensus regarding the need to protect them. Criminal law was thus conceived as an instrument of restraint. Over time, however, it underwent a process of expansion driven by shifting social models and growing demands for protection – a process sustained by confidence in "the 'virtues' of criminal law as an instrument for the protection of citizens" (Silva Sánchez 2001, 24). This expansionary trajectory – running counter to the minimalist tradition grounded in liberal principles – gradually transformed criminal law into a mechanism for managing social problems. As Farmer has argued, criminal law has consistently functioned both as an instrument for constructing civil order and as a vehicle for normative change. Across different historical periods, its purpose has been understood "as articulating norms of a social or collective nature for advancing civil order and civilization" (Farmer 2016, 116). During the welfare state period, criminal law "became a key tool for the legislator in responding to social problems"; and the prevailing conception of the state's capacity "to do good" structurally entailed an expansion in the reach of criminal law (Farmer 2016, 94). Subsequently, in the contemporary neo-classical period, the paradox lay in the fact that, while criminal doctrine sought to contain and constrain criminal law through principled limitations, "the development of these ideas has accompanied a large-scale expansion in the use of the criminal law" (Farmer 2016, 115). Drawing on Norberto Bobbio's distinction between traditional[1] and promotional functions of law, this article explores how that tension persists today, with particular focus on sexual criminal law, and specifically on the positive obligation to criminalize sexual conduct on the basis of consent, as established by the European Court of Human Rights

1 As will be seen in Section II, Bobbio distinguishes two traditional functions of law: the protective function and the repressive function. Both are subsumed here, for the purposes of this article, under the single term "traditional function," which at the level of criminal law corresponds to what Cadoppi terms the "conservative function."

and the Istanbul Convention. The central question is whether the state's deployment of criminal law as a driver of social change can be considered legitimate and, if so, under what conditions.

The argument proceeds in four sections. Section II reconstructs Bobbio's distinction between the traditional and promotional functions of law, tracing its implications for criminal law and locating the legitimacy problem that arises when the promotional function is pursued through penal sanction. Section III describes the affirmative consent model and maps the two distinct levels on which its justification operates: a protective justification grounded in negative sexual autonomy, understood as the right not to be subjected to non-consensual sexual conduct, and a promotional justification grounded in positive sexual autonomy, understood as the prima facie entitlement to engage in sexual conduct (Green 2020, 21), on terms defined by a specific communicative model. The Spanish legislation serves as the primary reference point and the distinction between these two justificatory levels is the conceptual instrument that structures the analysis throughout. Section IV brings three analytical perspectives to bear on the central question: the declaratory perspective, the conditioned promotional perspective, and the normative framing perspective, presenting each perspective's theoretical commitments and applying it to the affirmative consent model. The three perspectives converge in finding that the promotional justification of affirmative consent fails to meet the legitimacy requirements applicable to each. The shared conclusion is that whereas criminal law may legitimately perform a conservative function in protecting negative sexual autonomy, it may not legitimately be deployed to entrench a specific communicative model for exercising positive sexual autonomy. Section V advances the subsidiarity principle as the criterion that confines criminal law to a consolidating role in cultural transformation, allocating its promotional ambitions to the domain of social and educational policy.

II. The Promotional Function in Norberto Bobbio

The concept of law's promotional function is attributed to the legal theorist Norberto Bobbio, who uses it to describe the novel role that legal systems came to perform within the framework of modern welfare states. As Bobbio argues, "after the profound transformation that gave rise everywhere to the welfare state, public bodies began to pursue the new goals proposed for state action through techniques of social control that differed from traditional ones" (Bobbio 1969, 189). Two traditional conceptions of law may be distinguished, each projecting a different image of its function. The first assigns to law an exclusively protective role, aimed at safeguarding lawful conduct and repressing unlawful acts.[2] Instead, the second attributes to law an essentially repressive purpose,

2 According to Bobbio, the clearest representative of this protective perspective is Christianus Thomasius, who "characterized [the law] by the fact that it achieves its goal, which is essentially protective, by issuing negative commands or prohibitions" (Bobbio 1969, 190).

achieved through the organization of negative sanctions (Bobbio 1969, 190).[3] Both images are reductive, insofar as neither suffices to account for law in its transition from the liberal state to the social state. In that transition, the threat of negative sanctions – the defining instrument of the general theory of law – gave way to new forms of social control. These new techniques, which Bobbio theorizes as the promotional function, operate through inducement and encouragement rather than restraint, directing the conduct of individuals and groups towards ends the legal order regards as socially desirable. As Bobbio himself puts it, the promotional function "marks the movement from a passive type of control, which is chiefly concerned with putting marks of disapproval on undesirable behavior, to an active type of control, which is chiefly concerned with stimulating socially desirable behavior" (Bobbio 1969, 200).

In principle, the promotional function of law does not appear particularly troubling. Bobbio derives from it a class of measures of "encouragement," through which a legal order seeks to influence the performance of socially desirable conduct, either by facilitating that conduct or by attaching a reward to its performance. Moreover, this function is broadly consistent with a conception of the state oriented toward "creating and maintaining the conditions for living an autonomous life [by…] protecting and promoting the creation of the environment which makes such a life a possibility" (Raz 1994, 120–121).

Although Bobbio originally theorized the promotional function in direct connection with the concept of reward or incentive law [*diritto premiale*], understood as the subset of legal ordering that attaches positive sanctions to socially meritorious conduct, subsequent scholarship has decoupled these two notions. As Trocino has observed, while reward law necessarily presupposes the conferral of a benefit in response to meritorious conduct already performed, the promotional function extends to any domain of law that deploys techniques of encouragement, including facilitation techniques that render certain conduct particularly attractive without implying the award of any prize (Trocino 2025, 614).

The problem of legitimacy arises, however, when this function is pursued through the threat of criminal punishment. Transposing Bobbio's functional taxonomy to the specific domain of criminal law, Cadoppi distinguishes between a conservative and a promotional function of criminal law, and assigns to the former the protection of legal goods, conceived as interests or values that the legislator finds already established in society (Cadoppi 1988, 399). On this account, criminal law assumes a function proper to social control. The promotional function, by contrast, does not confine itself to control, it operates as an instrument of social change, fostering values that have not yet achieved sufficiently broad consensus within society (Cadoppi 1988, 401).

3 Bobbio identifies this repressive perspective across the legal positivist tradition from Austin to Ihering and Kelsen, each of whom understood law essentially as a coercive apparatus (Bobbio 1969, 190).

III. Affirmative Sexual Consent and the Promotional Function

Contemporary criminalization of sexual violence rests on the protection of sexual autonomy, articulated through the concept of consent.[4] This was not always the case. Historically, the Western legal tradition configured the normative structure of sexual offenses by conceiving them as a disturbance of the political and familial order. From a feminist perspective, women were regarded as objects of patrimonial interests: rape injured the authority of the *pater familias*, first that of the father, and subsequently that of the husband (Canan & Levand 2019, 4), while female sexuality remained subordinated to legitimate reproduction within marriage. Concern for the protection of women's individual autonomy was, accordingly, a legally secondary interest (Caletti 2023, 420). Rape was further defined by the relationship between physical force and resistance, the latter being required as an evidentiary condition for the victim's access to criminal protection (Canan & Levand 2019, 4). This regulatory framework remained largely unchanged through much of the twentieth century.

The conceptual reorientation of rape toward personal autonomy proved decisive in expanding the punitive response to sexual violence. The European Court of Human Rights' (ECtHR) contribution was decisive in two aspects. First, it established autonomy and consent as the two core normative standards by which the harm of sexual violence is defined (Pitea 2005, 448). Second, it imposed positive obligations on states directed at the criminalization of sexual offenses, which were to be formulated based on those same two elements (ECtHR 2003, 166). Particularly significant was the Court's recognition that restricting the prosecution of sexual offenses to conduct involving physical violence "risks leaving certain types of rape unpunished and thus jeopardizing the effective protection of the individual's sexual autonomy" (ECtHR 2003, 166).

The Council of Europe Convention on Preventing and Combating Violence against Women and Domestic Violence (Istanbul Convention 2011) constitutes a landmark in the configuration of international obligations with respect to sexual and gender-based violence. The Convention imposes on member states the obligation to criminalize non-consensual sexual conduct, stipulating that "[c]onsent must be given voluntarily as the result of the person's free will assessed in the context of the surrounding circumstances" (Art. 36.2). States nonetheless retain a margin of appreciation in determining the precise formulation of their domestic legislation and the factors that will preclude the possibility of freely given consent (Council of Europe 2011, 193).

The analysis centers on the affirmative consent model. Following Gruber, "affirmative consent laws direct decision-makers to focus on the external manifestations themselves and decide whether they are sufficient expressions of consent" (Gruber 2022,

4 Sexual violence is employed throughout as the broadest category, encompassing both behaviors involving attempted or completed bodily contact – such as sexual assault – and behaviors that may not include bodily contact, such as sexual harassment; rape denotes the most specific term within this spectrum (Canan & Levand 2019, 3).

64). Conceptions within this family differ in stringency. In its most demanding versions, affirmative consent has been assimilated to a contract, to the manifestation of an "enthusiastic yes," or to a verbal "yes," and has even been cast as a duty on the parties to secure an external expression of agreement through a "stop-and-ask" ritual (Gruber 2022, 65–68). Less demanding is the formulation that "requires an agreement communicated by words or positive actions – for example nodding or tearing off one's clothes in response to a sexual proposal" (Schulhofer 2023, 62). The Spanish legislation supplies the reference point for the critique. Article 178.1 of the Criminal Code (Ley Orgánica 10/1995) provides that "[c]onsent shall be deemed to have been given only where it has been freely expressed through acts which, having regard to the circumstances of the case, clearly manifest the person's will." Consent thus qualifies as valid only where the person authorizes it through signals that express their will in a clear and unequivocal manner (Faraldo 2021, 271).

The review that follows is built around the requirement of an external manifestation of agreement, communicated by words or positive actions sufficient to express consent. On that basis, the analysis sets aside formulations which, though labeled affirmative consent, relocate the inquiry to the circumstances or context surrounding the encounter. The contextual communicative conception illustrates the point, since it allows consent to "be inferred from the totality of a person's behavior in the context of all the circumstances" (Schulhofer 2023, 62). Here, the requirement of an external manifestation dissolves, and the standard ceases to be affirmative in the operative sense. The same occurs, as a matter of positive law, with the Swedish model. Its provision sanctions sexual activity "with a person who is not participating voluntarily" and requires no expression of agreement. The clause adding that "particular consideration is given to whether voluntariness was expressed by word or deed or in some other way" governs proof rather than the substantive definition of the wrong. Classified by the Group of Experts on Action against Violence against Women and Domestic Violence (GREVIO) as affirmative consent (GREVIO 2025, 78), the model in fact turns on the voluntariness of participation assessed through the surrounding circumstances, which is why Hörnle reads it as a "circumstances model" (Hörnle 2024, 57).

In the discussions surrounding these reforms, two distinct strands of justification may be observed, structured here through Green's distinction between negative and positive sexual autonomy (Green 2020, 21), which, within the domain of sexual offences, loosely parallels Berlin's account of negative and positive liberty (Berlin 2002, 166–217). The first is a protective strand grounded in negative sexual autonomy; the second is a promotional strand grounded in positive sexual autonomy. The protective strand operates in the sense given by Bobbio, in that the reform's standard aims to capture more accurately the unlawfulness of non-consensual sexual conduct, by placing the weight of criminal law on the safeguarding of negative sexual autonomy, understood as "the right not to engage in, or be subject to, one or another form of sexual conduct" (Green 2020, 21). On the assumption that the notion of autonomy has achieved sufficiently broad social recognition to constitute a shared normative standard,[5] the punitive expansion entailed by the affirmative consent model is justified on the basis of a conservative function against impunity, operating on two fronts: it reaches conduct that a coercion-based model failed to capture, and it establishes – in contrast to other consent models – that "passivity, silence, lack of protest or lack of resistance cannot be deemed to mean consent" (GREVIO 2025, 73).

In turn, the promotional strand bears on positive sexual autonomy, understood as "the prima facie right to engage in" sexual conduct (Green 2020, 21). On the assumption that consent manifests itself in a wide variety of forms within social interaction, an affirmative model shifts the analytical core toward the absence of a specific form of communicative verification, thereby offering a reductive conception of sexual consent and replacing the object of protection – sexual autonomy – with the procedure designed to verify it (Díaz y García & Trapero 2021, 566). Through this proceduralization, the model loses contact with the social norms that govern sexual interaction in everyday life. As Halley observes, it demands "something more than the relationship and its understandings" (Halley 2016, 270), imposing a formal verification requirement where lived practice operates through context and implicit communication.

This promotional justification for affirmative consent, grounded in the perceived need for social change, is reflected in the discourse of its proponents. Notably, GREVIO maintains in its reports that affirmative consent "is also a powerful way of changing the mindset of law-enforcement and judicial officers and of providing better protection to victims, by putting them at the centre of any intervention" (GREVIO 2025, 77). In the Spanish academic literature, Faraldo assigns to the model a function of social pedagogy, invoking the need to awaken social consciousness and maintaining that the law should function as a manifesto (Faraldo 2021, 277). In the U.S. context, Leary proposes a cultural shift through an "affirmative consent climate" encompassing three dimensions: education, social stigma, and legislative reform (Leary 2016, 34–37). From a critical standpoint, Gruber addresses the aspirational argument underlying affirmative consent, noting that the model, in its pursuit of cultural transformation, exposes a significant portion of society to the risk of criminalization (Gruber 2022, 73–75). This promotional ambition, moreover, is not confined to affirmative consent strictly understood. The Swedish reform is instructive: although the provision finally adopted does not qualify as an affirmative consent model, for the reasons set out above, it was first drafted as one. The 2016 commission proposed that only the outward, communicated signs of voluntariness should count as consent (Lernestedt & Kagrell 2023, 174); the 2018 proposition moderated that design on the operational plane while leaving its ideological premise intact. Wegerstad observes that the resulting law "is woven into a sexual education imperative" (Wegerstad

5 The recent legislative record corroborates this assumption. As Uhnoo et al. have documented, consent-based rape laws spread from seven to twenty European jurisdictions after 2017 (Uhnoo, Erixon, & Bladini 2024, 1), a cross-jurisdictional convergence that would be difficult to explain absent a prior, shared recognition that sexual autonomy constitutes a socially cognizable standard.

2021, 736), while it "sends a clear message about what sex should be – namely, something voluntary – but does not accurately describe the crime and the conduct that deserves criminal censure" (Wegerstad 2021, 737). The discussion of the promotional function attributed to affirmative consent thus converges on a single essential question: can the use of criminal law to achieve such transformation be considered legitimate?

IV. Can Criminal Law Be Legitimately Employed to Transform Social Norms?

As the preceding section has shown, two strands of justification can be found in the debate surrounding affirmative consent reforms: one grounded in negative sexual autonomy, which falls within the conservative function of criminal law, and one grounded in positive sexual autonomy, which conceives criminal law – and affirmative consent specifically – as an instrument of social transformation. It is in this second strand, the promotional rationale attached to certain affirmative consent conceptions, that questions of legitimacy arise. Building on this distinction, this section examines whether criminal law may legitimately function as an instrument of social change, from three analytical perspectives: the declaratory perspective, the conditioned promotional perspective, and the normative framing perspective. The first two assess whether and under what conditions the promotional function may be legitimately pursued through penal sanction. The third is concerned with the political rationality behind the demand for this specific form of regulation, rather than with the normative conditions of its legitimacy.

The declaratory perspective addresses the legitimacy of criminal law's promotional function by asking what criminal law is entitled to declare. It conceives criminal law as a "communicative enterprise" whose substantive dimension "declares certain kinds of conduct to constitute public wrongs that merit a formal public response" (Duff 2018, 109). The notion of declaration is determinative in Duff's definition of criminal law: it is a discipline that neither creates nor constitutes public wrongs but rather identifies and pronounces – in the collective voice of citizens – those forms of conduct already recognized as wrongs by reference to the shared values of the political community. For Duff, criminal law:

> is (or aspires to be) a common law in the sense that it is a law that citizens make for themselves and make (or should make) their own. On this view, the voice of the criminal law is the (collective) voice of the citizens – a voice in which they talk, as it were, to themselves in terms of the shared values that define them as a polity (Duff 2018, 118).

The penal prohibition thus emerges for citizens not as a mandate imposed by an external sovereign, but as a declaration operating on two distinct strata. The first concerns wrongfulness: the wrong is constituted independently of its inclusion in the penal code, deriving instead from the values shared within society. The second concerns the consequence that this collective valuation produces in relation to the wrong itself, conferring upon it its character as a public wrong (Duff 2018, 120). Duff acknowledges, however, that criminal law may function authoritatively even when some citizens dissent, and he draws a distinction between "reasonable" and "unreasonable" dissent. The former constitutes a disagreement about the content of an evolving civic morality, while the latter involves a refusal to recognize a wrong that is identifiable by reference to shared values (Duff 2018, 132). When strong communicative formulations of affirmative consent prescribe a particular model of sexual verification as legally required, the dissent of those who engage in genuinely consensual relations under different models is plausibly reasonable. The authoritative declaration therefore loses its communicative legitimacy and becomes coercive imposition.

The declaratory perspective offers a clear illustration of the distinction between declaring and constituting a public wrong. With respect to negative sexual autonomy, the declaratory constraint is satisfied: the assessment that a non-consensual sexual act harms the person's sexual autonomy is recognizable by reference to values shared and pre-existing within society. Broadly speaking, the consent model has been normatively consolidated from the international to the domestic level through the jurisprudence of the ECtHR and the subsequent ratification by states of the Istanbul Convention. The affirmative consent model, however, stands on the other side of that distinction. The reduction of consent to the verification of a formality may not declare a wrong that the community recognizes as such, it establishes the sole legally valid mode of expressing consent. Those who engage in genuinely consensual – that is, voluntary – relations through a different expressive framework constitute reasonable dissenters in Duff's sense (Duff 2018, 132). What confronts them is not the collective voice of citizens articulating a shared wrong, but the entrenchment of one conception of how sexual interactions ought to be conducted.

The conditioned promotional perspective rests on the premise that criminal law is not a passive receptacle of pre-existing social consensus, but an active agent in its configuration (Corda 2020, 606; Robinson 2017, 1591). Without rejecting the promotional function of criminal law, this perspective makes its legitimacy conditional upon four cumulative requirements: (a) that its transformational use be outcome-oriented rather than purely symbolic; (b) that it not amount to an unwarranted form of social dirigisme, or undue interventionism in the conduct of social life; (c) that it not treat the defendant merely as a means to an end; and (d) that it not impose partisan moral views that run counter to the fundamental rights and perspectives of a pluralistic democratic society.

The first condition requires that penal intervention be directed at tangible effects on social conduct or attitudes, and that it operates as an integral instrument of a broader policy effort that extends beyond penal measures (Corda 2020, 609–612). Where it fails to produce such effects, it amounts, as Robinson observes, to political proclamation clothed in legal dress (Robinson 2017, 1578–1579). The Spanish affirmative consent reform was embedded within a broader array of measures: enacted as a law of "comprehensive protection" of sexual autonomy, it incorporated provisions for prevention, awareness, detection, and education alongside the redefinition of the offences (Cancio Meliá 2023, 221). This accumulation of non-penal measures lends partial support to the condition. This assessment must nonetheless be qualified. Applied to the Spanish case, the condition serves less as an objection than as a demonstration of where the reform's transformative ambition actually lies. As Castellví argues, article 178.1 of the Criminal Code effects no genuine paradigm shift: the prior model never held that silence amounted to consent, and the reformed provision does not hold that silence amounts to its refusal (Castellví 2024, 483). With respect to sexual autonomy, the Spanish reform changes little, since protection against non-consensual conduct was already secured under the conservative strand. What it adds operates instead on positive sexual autonomy: a requirement that consent be manifested through acts clearly expressing the person's will (Castellví 2024, 514). Its aim is behavioral: to shape how sexual interaction is conducted. This duty is the promotional component the first condition tests, and the question is whether it produces the conduct it seeks or exhausts itself in expressive value.

The second condition concerns penal initiative that runs ahead of settled social attitudes. As Corda suggests, elected bodies may act as trailblazers, so that legislation can build consensus rather than merely reflect it and may even depart from prevailing norms (Corda 2020, 618–619). That authority is conditional: its legitimate exercise requires "some prior social/political/non-penal action aimed at setting a certain cultural shift into motion" (Corda 2020, 619). In Robinson's formulation, the law's normative initiative must remain legible to citizens as the articulation of an emerging consensus rather than as an arbitrary imposition of political will (Robinson 2017, 1592–1593). Failing this requirement, penal innovation collapses into what Corda conceives as a form of dirigisme: the attempt to inculcate the legislator's views in "a more or less silent majority," detached from the popular will and free of any significant constraint (Corda 2020, 617). The object of assessment is the affirmative consent model that requires a prior communicative act as a condition of permissible sexual conduct together with the promotional discourse that advances it. So delimited, the model's social anchorage is, at least, doubtful. Even granting that a consent-based understanding of sexual conduct has gained considerable social currency, it does not follow that the affirmative consent model enjoys comparable settlement. A recent briefing for the European Parliament points in the same direction at the level of the evidence base: explicit affirmative standards, however attractive in theory, sit uneasily with the practice of negotiating consent, where parties rely predominantly on non-verbal and behavioral cues (Bavré & Michielsen 2025, 5). A standard that contrasts with the forms in which consent is ordinarily manifested in practice cannot be read as the articulation of an emerging consensus. That being so, what the affirmative consent model does is impose a particular mode of understanding sexual interaction, displacing the plurality of forms through which consent is ordinarily expressed. Imposed in the absence of settled attitudes that would sustain it, that requirement amounts, on Corda's own terms, to *dirigisme* rather than legitimate trailblazing.

The third condition prohibits instrumentalizing the defendant merely as a means to an end. Desert must operate as a constraining factor: punishment imposed in pursuit of ends detached from the offender's blameworthiness clashes with the principles that legitimize the imposition of penal sanctions. Social change cannot in itself justify punishing those who did not act with the degree of culpability that punishment presupposes (Corda 2020, 627–629). This constraint yields a regulative ideal for any prohibition grounded in sexual autonomy: to "adequately criminalize all cases in which an offender subjects a victim to nonconsensual sex but do so without criminalizing any cases involving sex that is consensual" (Green 2020, 76). On the further view that consent is a communicative act, its performance must also be intentional and accompanied by a consenting mental state (Schnüriger 2018, 24). When a criminal provision constructs the wrong of sexual violence solely as a communicative deficit, a constitutive tension follows: the more demanding the prescribed form of manifestation, the wider the gap between the communicative act and the mental state beneath it, shifting the inquiry from what actually occurred to what ought to have been done. The requirement of a clear or unequivocal manifestation of consent thus "risks treating as criminal conduct cases in which genuine consent was present but did not conform to the legal formula" (Díaz y García & Trapero 2021, 567). This is a displacement of the punishable conduct from the *malum in se*, the actual violation of another person's will, toward the *malum prohibitum*, the failure to satisfy a communicative formality (Green 2020, 92). As Green himself concedes, under the affirmative consent standard "it is inevitable that we will capture at least some tokens that do not involve nonconsent" (Green 2020, 90). The doctrinal consequences confirm this diagnosis. According to Castellví, the requirement of manifestation alters liability directly: certain attempts become consummated offences, and certain instances of negligence become intentional ones (Castellví 2024, 495–509). The defect thus operates in two directions. On the one hand, it is overinclusive, criminalizing conduct in which genuine consent was present. On the other hand, it distorts the proportionality of the criminal response. Both are forms of dissociation of censure from desert that the third condition forbids.

The fourth condition bars the imposition of partisan moral values, in the sense that criminal law must not be used to impose the views of specific groups upon the political community as a whole, in contradiction with the foundational principle of pluralism that informs democratic societies (Corda 2020, 631–633). The Spanish debate, however, reveals significant tensions. Cancio observes that the reform was driven primarily by the parliamentary initiative of a single political sector (Unidas Podemos), and that the debate was marked by a "cultural struggle" in which substantive arguments were displaced by identity-based positions on both sides of the spectrum (Cancio Meliá 2023, 226). This is particularly significant if the discussion unfolded in terms of a "collective perspective on sexual identity, moralizing use of criminal law, and an identity-based approach to the subject matter" (Díez Ripollés 2020, 1551–1557).

While the first two perspectives interrogate the legitimacy conditions of criminal law's promotional function, a third shifts attention from those conditions to the political rationality that structures symbolic criminalization claims. According to Wilenmann, symbolic criminalization conflicts concern the framing of meaning rather than the regulation of conduct: what value-oriented groups seek through criminalization is not the reduction of harmful conduct but the legal recognition of their interpretative framework as socially valid (Wilenmann 2019, 10). On this account, criminal law is valuable less for its coercive capacity than for its symbolic power to confer or deny legal validity to conduct whose meaning is contested between social groups. As Wilenmann explains, "value-oriented groups struggle for a definition of the desired content that mimics the meaning which the group attributes to the actions that are to be criminalized or decriminalized, thus seeking to confirm the validity of the framework they use when attributing meaning to those actions" (Wilenmann 2019, 13).

The affirmative consent model furnishes direct evidence of this rationality, since its proponents have made their framing ambitions explicit. On this view, enactment functions for the group as the entrenchment of the affirmative consent grammar as the grammar of law, satisfying its objective whether or not behavioral transformation ultimately occurs. The resort to criminal law for expressive purposes has become an increasingly common feature of criminalizing discourse, which appeals to the need to signal public condemnation and treats punishment as "an instrument of popular pedagogy, in order to 'sensitize' the people. No attention is paid to the crucial question of whether it is appropriate or fair to use punishment for this purpose – the message is all that counts" (Jareborg 2005, 528). In the case of affirmative consent, the message in question is the legal validation of a particular conception of positive sexual autonomy.

V. Conclusion: Sexual Criminal Law as an Instrument of Social Change?

The analysis developed in the preceding sections has examined the legitimacy of employing criminal law as an instrument of social change, taking the affirmative consent model as its test case. The declaratory perspective offers perhaps the most compelling objection: criminal law does not create public wrongs but declares them by reference to the shared values of the political community. The affirmative consent model, by reducing positive sexual autonomy to a process of communicative verification, constitutes rather than declares such wrong. The conditioned promotional perspective does not categorically reject assigning a promotional function to criminal law, but subjects its legitimacy to structural requirements that the affirmative consent model satisfies only partially. The normative framing perspective operates at a different analytical level, shifting attention from the normative conditions of legitimacy to the political rationality that structures symbolic criminalization claims. On this view, what value-oriented groups pursued is the recognition of their reading of sexual interaction as the authoritative one, the establishment of affirmative consent as the normative framework whose validity the law itself confirms.

Subjecting criminal law's transformative claims to critical scrutiny is not equivalent to endorsing legal passivity in the face of practices incompatible with genuine respect for sexual autonomy. Such practices constitute a serious social problem, and the legal and political system has well-founded reasons to address it. The question is not whether criminal law has a role in protecting sexual autonomy against conduct that violates it, but whether it may also legitimately assume the function of actively promoting a specific communicative model for exercising that autonomy. The answer turns on a limit that derives not merely from prudential concerns but from the institutional character of criminal law itself. Educational initiatives, cultural interventions, and social policy may legitimately advocate demanding relational ideals while still proposing rather than compelling: they leave the proposed norm open to argument and refusal, so that it takes hold, if it does, through acceptance over time. The recourse to criminal law, by contrast, carries a cost that these instruments do not. As Van Kempen observes, the criminal justice system:

> 'encompasses the most severe instrument at the state's disposal in times of peace. It is principally coercive in nature (…) The intrinsic violation by the application of criminal law and criminal procedure of an individual's freedom and social sphere is a universal characteristic'. Moreover, involvement in the criminal justice process may have other adverse consequences, including the delinquent's social exclusion and stigmatization… (Van Kempen 2019, 5).

The coercion does not begin with the sentence. It inheres in the investigative and procedural powers that criminalization sets in motion, and it reaches beyond the accused, to third parties such as their family (Van Kempen 2019, 5). The desirability of a social norm does not, therefore, by itself establish the legitimacy of pursuing its consolidation through penal means. In a context of penal expansion and functional distortion, recovering the principle of subsidiarity may still serve a precise purpose. On the one hand, it cautions against "the inadequacy of a policy based exclusively, or almost exclusively, on the criminal justice system" (Londoño Martínez 2025, 12). There is a further risk in this regard: the persistent use of criminal law as a promotional instrument may, over time, generate a displacement effect, eroding the capacity of extra-penal mechanisms to perform that function, not because those mechanisms are ineffective, but because no formal instrument transmits public condemnation with equivalent force (Husak 2005, 538). On the other hand, entrusting structural social change to the domain of social and cultural policy, where strategies for promoting integration and removing the conditions that generate discrimination and violence can be articulated with greater legitimacy and efficacy, reflects a more principled allocation of regulatory functions.

Formulas such as affirmative consent, enthusiastic consent, and negotiated exchange are not without value. Their proper domain, however, is education. As Hörnle observes, "it makes sense to set standards high – they will be diluted in real life. Good advice for adolescents could be: treat the other as if this were a relationship with duties of care and consideration (it might develop this way)" (Hörnle 2018, 239). That register – aspirational, relational, formative – is precisely what criminal law cannot replicate without distortion. Its grammar is prohibition and sanction. To insist that criminal law is the vehicle through which such transformation should be pursued is to ignore the force of Zipf's caution that criminal law is:

> the most inadequate instrument conceivable for the transformation of society. Whoever seeks to reshape society through criminal justice resembles one who attempts to devise a pedagogical program on the basis of [school] punishment (Zipf 1975/1979, 43).

References

Bavré A. & Michielsen K. 2025. *Understanding and Raising Awareness on Sexual Consent: An Exploratory Review.* PE 773.630. Brussels: European Parliament, Policy Department for Citizens, Equality and Culture.

Berlin I. 2002. "Two Concepts of Liberty," in H. Hardy (Ed.), *Liberty: Incorporating Four Essays on Liberty* (pp. 166–217). Oxford: Oxford University Press.

Bobbio N. 1969. "The Promotion of Action in the Modern State," in G. Hughes (Ed.), *Law, Reason and Justice. Essays in Legal Philosophy* (pp. 189–215). New York University Press.

Cadoppi A. 1988. *Il reato omissivo proprio. Vol. I: profili introduttivi e politico-criminali*. Padova: Cedam.

Caletti G. M. 2023. *Dalla violenza al consenso nei delitti sessuali. Profili storici, comparati e di diritto vivente*. Bologna: Bologna University Press.

Canan S. N. & Levand M. A. 2019. "A Feminist Perspective on Sexual Assault," in W. O'Donohue & P. Schewe (Eds.), *Handbook of Sexual Assault and Sexual Assault Prevention* (pp. 3–16). Cham: Springer. https://doi.org/10.1007/978-3-030-23645-8_1

Cancio Meliá M. 2023. "Sexual Assaults under Spanish Law," in T. Hörnle (Ed.), *Sexual Assault*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198863397.003.0011

Castellví C. 2024. "¿Actos que expresen de manera clara la voluntad de la persona? Sobre el significado del 'solo sí es sí' y la definición del consentimiento sexual", *Anuario de Derecho Penal y Ciencias Penales* 77(1):483–514. https://doi.org/10.53054/adpcp.v77i1.10613

Corda A. 2020. “The Transformational Function of the Criminal Law: In Search of Operational Boundaries,” *New Criminal Law Review* 23(4):584–635. https://doi.org/10.1525/nclr.2020.23.4.584

Díaz y García Conlledo M. & Trapero Barreales M. A. 2021. “¿Razones Válidas para la Reforma de los Delitos Sexuales?”, in D.M. Santana Vega, S. Fernández Bautista, S. Cardenal Montraveta, D. Carpio Briz, & C. Castellví Montserrat (Eds.), *Una Perspectiva Global del Derecho Penal: Libro Homenaje al Profesor Dr. Joan J. Queralt Jiménez* (pp. 545–562). Barcelona: Atelier.

Díez Ripollés J. 2020. “Las propuestas de reforma del derecho penal sexual de 2018: su fundamento”, in J. De Vicente, M. Díaz y García, J. Paredes, I. Olaizola, M. Trapero, R. Roso, & J. Lombana (Eds.), *Libro Homenaje al Profesor Diego-Manuel Luzón Peña con Motivo de su 70º Aniversario*, Vol. II (pp. 1551–1563). Madrid: Reus.

Duff R. A. 2018. *The Realm of Criminal Law*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780199570195.001.0001

Durkheim É. 1964 (orig. 1893). *The Division of Labor in Society*. New York: Free Press.

European Court of Human Rights (ECtHR) 2003. *M.C. v. Bulgaria*, Application No. 39272/98, December 4. URL: https://hudoc.echr.coe.int/eng?i=001-61521

Faraldo P. 2021. “‘Solo Sí es Sí’: hacia un modelo comunicativo del consentimiento en el delito de violación”, in M. Alcale, A. Miranda, & A. Nieto (Coords.), *Reformas Penales en la Península Ibérica: A ‘Jangada de Pedra’? Jornadas de las Secciones Española y Portuguesa de la AIDP, Cádiz, 16 y 17 de julio de 2020* (pp. 265–279). Madrid: Agencia Estatal Boletín Oficial del Estado.

Farmer L. 2016. *Making the Modern Criminal Law: Criminalization and Civil Order*. Oxford: Oxford University Press.

Garland D. 1990. *Punishment and Modern Society: A Study in Social Theory*. Chicago: University of Chicago Press.

GREVIO 2025. *Thematic Perspectives on the Implementation of the Istanbul Convention: Compilation of Focus Sections No. 1 to 5 as Presented in GREVIO Activity Reports (June 2015 to December 2023)*. Strasbourg: Council of Europe.

Green S. P. 2020. *Criminalizing Sex: A Unified Liberal Theory*. New York: Oxford University Press.

Gruber A. 2022. “Affirmative Consent,” in E. Hoven & T. Weigend (Eds.), *Consent and Sexual Offenses: Comparative Perspectives* (pp. 57–81). Baden-Baden: Nomos. https://doi.org/10.5771/9783748930242-57

Halley J. 2016. “The Move to Affirmative Consent,” *Signs: Journal of Women in Culture and Society* 42(1):257–279. https://doi.org/10.1086/686904

Hörnle T. 2018. “Rape as Non-consensual Sex,” in A. Müller & P. Schaber (Eds.), *The Routledge Handbook of the Ethics of Consent* (pp. 235–246). London: Routledge.

Hörnle T. 2024. “The Challenges of Designing Sexual Assault Law,” *Current Legal Problems* 77(1):49–80. https://doi.org/10.1093/clp/cuae002

Husak D. 2005. "The Criminal Law as Last Resort," *Ohio State Journal of Criminal Law* 2(2):535–545.

Jareborg N. 2005. "Criminalization as Last Resort (Ultima Ratio)," *Ohio State Journal of Criminal Law* 2(2):521–534.

Leary M. G. 2016. "Affirmatively Replacing Rape Culture with Consent Culture," *Texas Tech Law Review* 49(1):1–56.

Lernestedt C. & Kagrell M. 2023. "The Swedish Move Towards (In)Voluntariness," in T. Hörnle (Ed.), *Sexual Assault: Law Reform in a Comparative Perspective* (pp. 163–184). Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198863397.003.0009

Londoño Martínez F. 2025. "¿Derecho Penal Anti Discriminación? Una crítica, a propósito de *motivi aggravanti e circostanze discriminatorie* (Andrea Perin, Torino, 2024). De Paso, Notas sobre la Atenuante Pasional y el Principio de Subsidiariedad," *Política Criminal* 20(1):1–17. https://doi.org/10.4067/S0718-33992025000100438

Pitea C. 2005. "Rape as a Human Rights Violation and a Criminal Offence: The European Court's Judgment in M.C. v. Bulgaria," *Journal of International Criminal Justice* 3(2):447–462. https://doi.org/10.1093/jicj/mqi037

Raz J. 1994. *Ethics in the Public Domain: Essays in the Morality of Law and Politics*. Oxford: Clarendon Press.

Robinson P. H. 2017. "Democratizing Criminal Law: Feasibility, Utility, and the Challenge of Social Change," *Northwestern University Law Review* 111(6):1565–1595.

Schulhofer S. 2023. "What Does 'Consent' Mean?" in T. Hörnle (Ed.), *Sexual Assault: Law Reform in a Comparative Perspective* (pp. 53–71). Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198863397.003.0004

Schnüriger H. 2018. "What is Consent?" in A. Müller & P. Schaber (Eds.), *The Routledge Handbook of the Ethics of Consent* (pp. 21–31). London: Routledge.

Silva Sánchez J. M. 2001. *La expansión del derecho penal: aspectos de la política criminal en las sociedades postindustriales* (2nd Ed.). Madrid: Civitas.

Trocino S. 2025. "Punire o premiare? Le nuove sfide del sistema sanzionatorio: opportunità e limiti delle sanzioni positive", *Materiali per una storia della cultura giuridica* 55(2):599–623. https://doi.org/10.1436/118869

Uhnoo S., Erixon S., & Bladini M. 2024. "The Wave of Consent-Based Rape Laws in Europe," *International Journal of Law, Crime and Justice* 77:1–16. https://doi.org/10.1016/j.ijlcj.2024.100668

Van Kempen P. H. 2019. "Criminal Justice and the Ultima Ratio Principle: Need for Limitation, Exploration and Consideration," in P. H. Van Kempen & M. Jendly (Eds.), *Overuse in the Criminal Justice System: On Criminalization, Prosecution and Imprisonment* (pp. 3–22). Cambridge: Intersentia. https://doi.org/10.1017/9781780688398.002

Wegerstad L. 2021. "Sex Must Be Voluntary: Sexual Communication and the New Definition of Rape in Sweden," *German Law Journal* 22(5):734–752. https://doi.org/10.1017/glj.2021.32

Wilenmann J. 2019. "Framing Meaning through Criminalization: A Test for the Theory of Criminalization," *New Criminal Law Review* 22(1):3–33. https://doi.org/10.1525/nclr.2019.22.1.3

Zipf H. 1979 (orig. 1975). *Introducción a la Política Criminal*. Trad. M. Izquierdo Macías-Picavea. Madrid: Edersa.