Battaglino G, 'The Kleinberg epistemic paradox within contractual discrimination' (2025) 2 EJPLT.
DOI: https://doi.org/10.57230/EJPLT252GB

EJPLT
ISSN 2704-8012



The Kleinberg epistemic paradox within contractual discrimination

Il paradosso epistemico di Kleinberg e la discriminazione contrattuale

GIOVANNI BATTAGLINO
PhD candidate
University of Camerino
LAST-JD PhD Candidate
Vrije Universiteit Brussel

Abstract

This paper examines contractual discrimination in the age of Artificial Intelligence through the epistemic paradox formulated by Jon Kleinberg. While the principle of non-discrimination is firmly recognised in European private law, its assessment remains epistemically complex, especially where intent is obscured by human bias or algorithmic processes. Two interpretative models are considered: the formalist, which confines protection to statutory indices and categories, and the substantive-functional, which evaluates discrimination retrospectively and contextually. The paper endorses the latter, showing that Kleinberg's paradox offers limited utility in such a framework, yet regains relevance for formalist approaches. Through doctrinal, legislative, and jurisprudential analysis, reinforced by hypothetical examples, the argument highlights how AI systems may serve as "moral mentors," exposing fractures in a positivist system dominated by binarisation, confinement, and market interests. Ultimately, the interpreter, not legislation alone, must ensure that remedies evolve with the complexity of technological and economic development, safeguarding human dignity.

© The author(s) 2025, published by Suor Orsola Benincasa Università Editrice.
This contribution is licensed under a Creative Common Attribution 4.0 International Licence CC-BY-NC-ND, all the details on the license are available at:
https://creativecommons.org/licenses/by-nc-nd/4.0/

Abstract
Il presente contributo esamina la discriminazione contrattuale nell'era dell'Intelligenza Artificiale alla luce del paradosso epistemico formulato da Jon Kleinberg. Sebbene il principio di non discriminazione sia saldamente riconosciuto nel diritto privato europeo, il suo accertamento rimane epistemicamente complesso, specialmente quando l'intento è oscurato da bias umani o da processi algoritmici. Sono presi in considerazione due modelli interpretativi: quello formalista, che circoscrive la tutela agli indici e alle categorie previsti dalla legge, e quello sostanziale-funzionale, che valuta la discriminazione in modo retrospettivo e contestuale. Il contributo aderisce a quest'ultimo, mostrando come il paradosso di Kleinberg offra un'utilità limitata in tale prospettiva, ma riacquisti rilevanza negli approcci formalisti.
Attraverso un'analisi dottrinale, legislativa e giurisprudenziale, corroborata da esempi ipotetici, l'argomentazione evidenzia come i sistemi di Intelligenza Artificiale possano fungere da "mentori morali", mettendo in luce le fratture di un sistema positivistico dominato da binarizzazione, confinamento e interessi di mercato. In ultima analisi, spetta all'interprete e non alla sola legislazione garantire che i rimedi si evolvano in sintonia con la complessità dello sviluppo tecnologico ed economico, salvaguardando la dignità umana.

Keywords: contractual discrimination; artificial intelligence; epistemology; moral mentor; Jon Kleinberg

Summary: 1. Introduction: AI and contractual discrimination within Jon Kleinberg's paradox. – 2. How Discriminatory Treatment Is Evaluated. – 3. *(Follow)* The Reasonable Approach to the Evaluation of Discriminatory Treatment. – 4. Justification of the Alleged Discriminatory Choice and Epistemic Centrality. – 5. Conclusions: The Old Problems of Technique and Liberal Rhetoric in the field of Contractual Liability. The Machine as Moral Mentor and a Revival of Kleinberg's Thesis.

1. AI and contractual discrimination within Jon Kleinberg's paradox.

The use of Artificial Intelligence in contractual relationships raises concerns regarding potential discrimination.[1]

[1] Indeed, the assumption of machine neutrality is now widely contested. AI systems not only replicate but may also amplify the biases embedded in human behavior and social structures. See, for example, part of the legal scholarship specifically addressing algorithmic systems that have demonstrated discriminatory effects: E Falletti, *Discriminazione algoritmica. Una prospettiva* comparata (Giappichelli 2022) 220-229; PM Law and others, 'The Impact of Presentation Style on Human–In–The–Loop Detection of Algorithmic Bias' [2020] GI, arXiv.2004.12388; E Oosterloo, GV Schie, 'The politics and Biases of the "Crime Anticipation System" of the Dutch Police' [2020] Bias Inf. Algorithms Syst.; A Završnik, 'Criminal justice, artificial intelligence systems, and human rights' (2020) 20 (4) Era Forum 567; E Tristram Albert, 'AI in talent acquisition: A review of AI applications used in recruitment and selection' (2019) 18 (5) SHR 215; R Berk and others, 'A convex framework for fair regression' [2017], arXiv:1706.02409.

Non-discrimination is recognised as both a foundational principle and a legal value;[2] nevertheless, its practical assessment remains contested and epistemically complex.[3]

This paper does not address the philosophical or etymological roots of discrimination, nor its conceptual links to human dignity, vulnerability, or equality.[4]

The focus will be on the *ex-post* evaluation of presumed contractual discrimination[5] mediated[6] by a Large Language Models (LLMs).[7]

---

[2] Despite the breadth and fragmented nature of the relevant legal sources, and notwithstanding the existence of general treaties provisions that directly or indirectly address discrimination, the Court of Justice of the European Union has affirmed the principle of non-discrimination as grounded in the principle of equality and as a cornerstone of EU law. On this point, see Court of Justice of the European Communities, Case C-13/94, *P v S and Cornwall County Council*, 30 April 1996 (ECLI:EU:C:1996:170); Court of Justice of European Union, Case C-144/04, *Mangold v Helm*, 22 November 2005 (ECLI:EU:C:2005:709). In other cases, however, the Court of Justice has interpreted the principle of non-discrimination as symbiotic with the principle of human dignity: Court of Justice of the European Communities, Case C-152/82, *Forcheri v Belgian State*, 13 July 1983 (ECLI:EU:C:1983:217); Court of Justice of the European Union, Case C-36/02, *Omega Spielhallen- und Automatenaufstellungs-GmbH v Oberbürgermeisterin der Bundesstadt Bonn*, 14 October 2004 (ECLI:EU:C:2004:614); Case C-144/04 *Werner Mangold v Rüdiger Helm* [2005] ECR I-9981 For scholarly contributions that acknowledge the principle of non-discrimination, see *infra* note 4; for the more specific normative framework concerning contractual matters, see Section 2 below.

[3] On this point, see Sections 2 and following.

[4] For a discussion of the etymology of discrimination, see PS Shin, 'Is There a Unitary Concept of Discrimination?', in D. Hellman, S. Moreau (eds.), *Philosophical foundation of discrimination Law* (Oxford University Press 2013) 161; L Halldenius, 'Dissecting Discrimination' (2005) 14(4) Camb., Q. Healthc. Ethics, 455. For the philosophical foundations of discrimination law, see E Consiglio, *Che cos'è la Discriminazione? Un'introduzione teorica al diritto antidiscriminatorio* (Giappichelli 2020), 131 ff., see also the related bibliography cited therein. In contrast, for a different legal grounding of discrimination in the context of contract law: C Nardocci, 'Artificial Intelligence-based Discrimination: Theoretical and Normative Responses. Perspectives from Europe' (2023) 3 DPCE Online 2385; P Femia, 'Discriminazione (divieto di)', *Enciclopedia del Diritto – I Tematici, Contratto* (Giuffrè Francis Lefebvre 2021) 499 ff.; H Steege, 'Algorithm-Based Discrimination by Using Artificial Intelligence: Comparative Legal Considerations and Relevant Areas of Application' (2021) 1 EJPLT 56; FJ Zuiderveen Borgesius, 'Strengthening Legal Protection Against Discrimination by Algorithms and Artificial Intelligence' (2020) 24(10) Int. J. Hum. Rights. 1572; HH Harvey, 'Justifying the Efficacy of Contract Discrimination' (2019) 2 J. Law Innov., 47; M Ciancimino, 'La discriminazione contrattuale: profili rilevanti per la tutela della persona. Note a margine di un recente dibattito dottrinale' (2018) 2 Dir fam. Pers. 668; G Carapezza Figlia, 'Il divieto di discriminazione quale limite all'autonomia contrattuale' (2015) 6 Riv. dir. civ. 1387; D Maffeis, 'Il diritto contrattuale antidiscriminatorio nelle indagini dottrinali recenti' (2015) 1 N. leg. civ comm. 161; L Gentili, 'Il principio di non discriminazione nei rapporti civili' (2009) 2 Riv. crit. dir. priv. 207; AS Vandenberghe, 'The Economics of the Non-Discrimination Principle in General Contract Law' (2007) 4(3) ERCL 411; J Neuner, 'Protection Against Discrimination in European Contract Law' (2006) 1 ERCL 36.

[5] In addition, the notion of discrimination is consistently construed in a strictly negative and legally adverse sense: FJ Zuiderveen Borgesius, 'Strengthening Legal Protection Against Discrimination by Algorithms and Artificial Intelligence' (2020) The International Journal of Guman Rights 1573.

[6] The term "mediated" is used here, on the one hand, to reflect a materialistic and human-centered view of technology, beyond any metaphysic speculation. On the other hand, once the assumption of algorithmic neutrality is deconstructed, a factual equivalence emerges between human and machine-generated discriminatory effects. This leads to a behavioristic perspective, whereby the machine may operate as a discriminatory agent in the same way as a human. On this basis, the legal evaluation of discriminatory treatment remains inherently open and subject to diverse interpretations. For further discussion of the relationship between neutrality, agency, and the interaction between technology and law, as well as for the relevant bibliography, see: G Battaglino, 'La fairness algoritmica: incontro tra sviluppo socio–tecnologico ed esigenze dell'ermeneutica giuridica', (2025) 1 B.L. Journal 633-639.

[7] For a comprehensive overview of the general properties of large language models, see WX Zhao and others, 'A Survey of Large Language Models' [2023], arXiv:2303.18223. In this paper, a *summa divisio* is drawn between formal systems and stochastic systems (such as Large Language Models). Consistent with the principle of precision for readability, the analysis does not aim to conceptualise or technically reconstruct these systems from an informatics perspective. Rather, it seeks to clarify their practical legal relevance and operational use within juridical evaluation: Zuiderveen Borgesius (n 4) 1574.

Within the scholarship on algorithmic discrimination, a provocative thesis is advanced by computer scientist Jon Kleinberg.[8] He argues that algorithms, rather than merely obscuring discriminatory behaviour, may assist in its legal evaluation.

The difficulty, he suggests, is epistemic:[9] legal interpreters often struggle to identify intent behind a discriminatory outcome, whether it arises from human implicit bias or from algorithmic processes: while human decision-making is subject to cognitive opacity,[10] algorithms may offer a form of computational transparency.[11] Kleinberg thus poses a provocation: when assessing discriminatory outcomes, are legal interpreter equipped to interrogate deeply hidden cognitive states, or to navigate computational processes?[12]

Against this backdrop, this paper examines the potential of Kleinberg's epistemic paradox as a meaningful analytical tool for understanding and evaluating AI-mediated contractual discrimination?[13]

The analysis begins by establishing how discriminatory treatment is evaluated.[14] From this legal scaffolding, the paper builds two conceptual models through which the legal evaluation of discrimination is commonly articulated: one grounded in legal formalism, the other in substantive functionalism. Both frameworks acknowledge the epistemic difficulty involved in assessing discriminatory outcomes.[15]

The analysis now turns to a critical evaluation of their explanatory power and normative coherence: the paper supports the second approach, the substantive-functional model.[16]

---

[8] J Kleinberg and others, 'Discrimination in the Age of Algorithms' (2019) 10 J. Leg. Anal 113 ff.

[9] In this paper, the term "epistemology" is used, regardless of any broader interpretations, to refer specifically to the capacity to make a legal choice intelligible and understandable to the human subject who is expected to evaluate it, an aspect that remains crucial regardless of the nature of the agent making the decision: P Moro, 'Algoritmi e pensiero giuridico. Antinomie e interazioni' (2019) 2 Media Laws 20-22.

[10] In fact, individuals may lie, dissemble, or be unaware of their own biases. On this sense, see Consiglio (n 4) 84-86.

[11] J Kleinberg (n 7) 116: «Deciding whether there has been discrimination is difficult for one obvious and one less obvious reason. The obvious reason is generic to any legal system: people dissemble, obfuscate, and lie. The less obvious reason is that people may not even know themselves. A large body of research from behavioural science (...) tells us that people themselves may not know why and how they are choosing – even (or perhaps especially) when they think that they do».

[12] ibid, 116: «The opacity of the algorithm does not prevent us from scrutinizing its construction or experimenting with its behaviour, two activities that are impossible with humans. (…) In contrast, a well-regulated process involving algorithms stands out for its transparency and specificity: it is not obscured by the same haze of ambiguity that obfuscates human decision-making. (...) In a nutshell: For the legal system, discovering "on what basis are they choosing?" and "why did they pick those factors?" becomes much more feasible». Accordingly, in the first case (discrimination caused by a human agent), the epistemic problem is rooted in human ontology; in the second (discrimination caused by a machine), the epistemic problem is systemic and operational in nature. In this sense, see: G Gometz, 'Intelligenza artificiale, profilazione e nuove forme di discriminazione' (2019), in F. Mancuso, V. Giordano (eds.), *Ombre del diritto, Materiali dei seminari el Prin 2017, The Dark Side of Law',* 1 ff.

[13] In order to assess the relevance of Kleinberg's epistemic paradox to the legal evaluation of contractual discrimination, this paper adopts a facet recommendatory methodology, which involves distinct descriptive and evaluative phases: L Kestemont, *Handbook on Legal Methodology: From Objective to Method* (Intersentia 2018) 63-73.

[14] On this point, see section 2, which focuses on the EU and Italian normative framework and the relevant scholarship that explicitly engage with the discriminatory contractual wrongdoing.

[15] See infra, section 4.

[16] See infra, Section 3, which focuses through a combination of historical, doctrinal, jurisprudential and logical reasoning.

In this light, Kleinberg's paradox offers limited analytical utility within a substantive-functional evaluation of contractual discrimination.[17]

However, in interpretive practices the paradox regains significance for a more formalistic interpreter. In this sense, the paper expects to propose the machine as a moral mentor, capable of guiding legal interpretation beyond rigid formalism in the evaluation of contractual discrimination. [18]

2. How Discriminatory Treatment Is Evaluated.

To establish the epistemic centrality of evaluating discriminatory conduct, it is necessary to outline when and how the legal system recognises it.

Legislation[19] tend to evaluate discriminatory conduct on the basis of three core indicators:[20] (a) the existence of a detrimental or unfavourable treatment, broadly construed, suffered by an individual or a group; (b) the presence of a protected characteristic, as defined by law, that causally underpins such treatment; (c) the absence of a reasonable justification. [21]

What matters at this stage is to understand how the first two elements, symptomatic indicators of discrimination, are weighed in the legal evaluation.

On this point, it is possible to outline two main interpretative approaches:[22]

---

[17] See infra, Section 4.

[18] See infra, Section 5.

[19] The legislative acts, although the legislator does not possess absolute semantic authority, and language alone is not the sole object of legal science, inevitably serves as a starting point in the reconstruction and interpretation of the legal system. On this point, see P Femia, *Interessi e conflitti culturali nell'autonomia privata e nella responsabilità civile* (Jovene 1990) 11–12, 20–23; P Perlingieri, 'Produzione scientifica e realtà pratica: una frattura da evitare' (1969) 9 Riv. dir. comm. 457–458. Therefore, it must be questioned whether statutory provisions alone can exhaustively define the indicators and criteria required both to assess the discriminatory nature of a given conduct and to attribute legal responsibility for it. on this point see specifically par 3

[20] On this point, with particular reference to elements (a) and (b) identified above, see Article 2 of Directive 2000/43/EC, Directive 2000/113/EC, and Article 43 of the Testo Unico sull'Immigrazione (d. lgs. 25 luglio 1998, n. 286 The first directive seeks to prohibit all forms of discrimination based on race or ethnic origin, particularly in the context of employment (see Article 3, Dir 2000/43/EC). The second addresses discrimination based on sex in the access to and supply of goods and services (see Article 3, Dir 2000/113/EC). The third provision (Article 43) targets discrimination grounded in race, colour, descent, or national or ethnic origin, as well as religious beliefs and practices, regarding the equal enjoyment or exercise of human rights and fundamental freedoms in political, economic, social, and cultural life, as well as in all other areas of public life. Regarding point (c), reference should be made to Article 2(2)(b) of both Directives, which address the concept of indirect discrimination (on which, see infra). Under these provisions, a difference in treatment does not constitute unlawful discrimination if it is objectively justified by a legitimate aim, and if the means of achieving that aim are appropriate and necessary. Both Directive 2000/43/EC and Directive 2000/113/EC have been duly and comprehensively transposed into the domestic legal systems of all EU Member States. In Italy, the implementation of Directive 2000/43/EC was carried out primarily through Legislative Decree No 215 of 9 July 2003, which also established the National Office Against Racial Discrimination (UNAR). Directive 2000/113/EC was implemented by Legislative Decree No 196 of 11 April 2006, which introduced specific provisions ensuring gender equality in access to goods and services. See: Decreto Legislativo 9 luglio 2003, n. 215, Attuazione della direttiva 2000/43/CE per la parità di trattamento tra le persone indipendentemente dalla razza e dall'origine etnica, G.U. n. 173 del 28 luglio 2003; Decreto Legislativo 11 aprile 2006, n. 196, Attuazione della direttiva 2000/113/CE sulla parità di trattamento tra uomini e donne nell'accesso a beni e servizi, G.U. n. 115 del 19 maggio 2006.

[21] The latter element functions as an epistemic exception clause and, as such, its centrality will be addressed in the fourth section.

[22] These models do not reflect rigid doctrinal traditions or established academic labels; rather, they are proposed here as analytical constructs adopted for explanatory clarity and operational effectiveness.

(1) Formalist-legistic:[23] These positions presuppose that the express identification of legal indicators and protected categories is a necessary condition for qualifying a given conduct as discriminatory: when the conduct falls squarely within these predefined legal classifications can the legal status of the affected individual be recognised and protected.

(2) Substantive-functional:[24] under this model, discrimination is understood as a multifactorial phenomenon. Legal Indicators serve as symptomatic markers but are not exhaustive determinants. The evaluation, therefore, goes beyond formal classification and considers the broader structural and contextual effects of the conduct.

3. (Follow) The Reasonable Approach to the Evaluation of Discriminatory Treatment.

Before establishing the epistemic centrality (point c) common to both models, it is necessary to clarify why the substantive-functional approach is preferred. The reasons can be summarised as follows:

(a) Historical contingency**.** Protected indices and categories are context-dependent and evolve over time; they cannot be treated as fixed or exhaustive.[25]

(b) Methodological shortcomings**.** Recognising legal positions solely on statutory signifiers ignores the concrete circumstances of cases and the interests at stake.[26] Rigid reliance on such signifiers risks a tyrannical close of value[27], whereby dominant majorities or stronger contracting parties dictate the framework.[28]

---

[23] Thus, for example, the evaluation of discriminatory conduct is often framed through a twofold analytical distinction based on intersecting dimensions: direct versus indirect discrimination, and intentional versus unintentional discrimination. The first of these distinctions appears to be expressly reflected in Article 2(2) of both Directives previously mentioned, which codify the differentiation between direct and indirect forms of discrimination. On this pint see, Steege (n 4) 60-61. The main distinction between direct and indirect discrimination lies, in fact, in the requirement to establish intentionality in the latter in order for it to o be ascribed legal responsibility. On this point, see: Zuiderveen Borgesius (n 4) 1576-1578. Appears to align with this approach, both in the inquiry into intentionality and in the assessment of a causal nexus between the conduct and the harm, Nardocci (n 4) 2367 ff. The assessment of discrimination as either direct or indirect appears to find a terminological counterpart overseas in the so-called *disparate treatment* and *disparate impact*. On this point, see Kleinberg (n 9) 6-9.

[24] On this approach, see, among others, but most notably Femia (n 4) 498 ff.

[25] Indeed, even prior to the enactment of specific anti-discrimination legislation, racial hatred, religiously motivated sexism, and homophobia were already regarded as morally reprehensible and socially intolerable. This highlights that the very indefinability and fluidity of the legal indicators is not a limitation but rather an integral feature of the concept of discrimination itself. On this point, see Femia (n 4) 505-521; Zuiderveen Borgesius (n 4) 1584-1586; R Post, 'Prejudicial Appearances: The Logic of American Antidiscrimination Law' (2000) 88 Cal L Rev 1, 8–9; F Bagenstos, '"Rational Discrimination," Accommodation, and the Politics of (Disability) Civil Rights' (2003) 89 Va L Rev 825, 846–47. On this point, see Cons. Stato 9 gennaio 2017 n. 30, in *Foroplus.it*

[26] P Perlingieri, 'Relazione conclusiva', in P Perlingieri and L Ruggeri (eds), *Diritto privato comunitario*, II (2nd ESI 2008) 393–399;

[27] On the theory of the "tyrannical value" and its implications for legal methodology and interpretation, see especially Battaglino (n 6) 645-649., along with the bibliography cited therein.

[28] P Femia, 'Segni di valore' [2014] in www.*Civilistica.com*; P Perlingieri, 'Valori normativi e loro gerarchia: una precisazione dovuta a Natalino Irti' [1985] Rass. dir. civ. 298. The very idea of democracy is founded on a compromise which never crystallises into value absolutism. No single value can be elevated to an absolute; instead, it necessarily presupposes the coexistence of diverse opinions and interests, all of which must be

(c) Instrumental nature of indices. Legal indicators are tools, not ends; their application in practice determines whether treatment is discriminatory, justified[29], or belongs "merely" to the domain of morality[30].

On those bases, the second approach appears preferable: discrimination must be understood as a retrospective and complex phenomenon, dependent on circumstances, competing interests, and justifications. The ideological risk, rather than the technical one, is that of a corrupted egalitarianism in which every need becomes a legal entitlement.[31] However, by the same token, denying protection outright simply because the relevant index or category is not represented in a statutory rule or contract[32] would be equally untenable.[33]

In conclusion, indices remain no more than indicators: useful to the interpreter, but not sufficient to exhaust the evaluative task. Rather, the inquiry necessarily shifts from an abstract (and presumptively neutral) assessment to one that is retrospective and concrete.[34]

---

considered. See also H Kelsen, *I fondamenti della democrazia* (Il Mulino 1966) 168–170; F. Rimoli, 'Laicità', in *Enciclopedia Giuridica Treccani* (XVIII, 1990) 8–9.

[29] The criterion, in itself, is always neutral; indeed, more often than not, as in the case of so-called positive discrimination, it is the legal norm itself which requires reliance on that criterion in specific circumstances, without thereby producing an effect that can structurally be regarded as discriminatory. On this point, see D Maffeis, 'Libertà contrattuale e divieto di discriminazione' (2008) 1 Riv. trim. dir. proc. Civ. 425-426.

[30] In certain contexts, however, the use of artificial intelligence systems, similar to those employed in dating platforms such as Tinder, falls outside the scope of legal scrutiny regarding discrimination. For all its sadness, if a woman rejects a man because an AI-based matching system highlights that he is "too short" or "too thin", such a decision remains an expression of personal and affective choice. As such, it is not amenable to judicial review under anti-discrimination law, since the sphere of intimate and emotional preferences is generally regarded as lying beyond the reach of legal regulation. This example is inspired by J Gardner, 'On the Ground of Her Sex(uality) (1998) 18 OJLS 167 ff.

[31] D Maffeis (n 27) 409. Part of the legal scholarship, in order to avoid such an egalitarian drift, where the prohibition of discrimination would unduly restrict contractual freedom, argues that the former should apply only in cases involving a public offer, and not to negotiations between private parties. Although this view is not endorsed here, it may nonetheless be observed that, in the situations under discussion, even if one is not technically faced with a public offer, those who employ LLMs in legally relevant decision-making are invariably enterprises. This lends to the thesis of discrimination advanced here a kind of elliptical value. For the approach that is not endorsed here, see D Maffeis (n 4) 163 ff. By contrast, see Ciancimino (n 4) 678-677.

[32] In this sense, discrimination must be addressed with every instrument that civil law makes available, and not solely through legislative or political mechanisms, as a more positivistic strand of legal scholarship would suggest. This line of reasoning also emphasises the absence of an explicit prohibition against discrimination on the basis of artificial intelligence, a lacuna which could pose a problem at the legislative level. It is as if AI were not regarded as an *instrumentum* that merely mediates and amplifies discriminatory dynamics already present, rather than constituting an autonomous ground of discrimination in itself: Nardocci (n 4) 2367 ff. On this point, moreover, the very criteria themselves, as has already been argued, are historically contingent and carry different weight, at least insofar as their legislative translation is concerned. Thus, for example, while history has led to the enactment of explicit statutory protections against discrimination on the grounds of race or sex, it is also clear that, aside from any personal considerations, the list of protected characteristics may evolve to include new categories (such as LGBT rights). Conversely, other potential indices are excluded precisely because they are not regarded as "appropriate" or "normal". Yet this does not preclude the possibility that, in a given case, reliance on such criteria may still reveal a discriminatory treatment: Shin (n 4) 11-15.

[33] On this point, see Femia (n 4) 517; B Parenzo, 'Profilazione e discriminazione. Dal GDPR alla Proposta di Regolamento sull'IA' (2023) 1 Tecn. Dir. 114.

[34] Moreover, it remains unclear why, if indices and classes are assumed to be all-encompassing of every possible form of discriminatory treatment, thus generating a kind of seriality of discriminatory phenomena, there should be any need to employ an LLM-based system or, more generally, a framework that goes beyond mere formal logic, as if it were itself a formal system. This is particularly evident when considering the availability of formal, interpretable and explainable algorithms, which can already provide a structured and transparent assessment without the epistemic opacity inherent in large-scale machine learning models. On this point, see Battaglino (n 6) 854-856.

4. Justification of the Alleged Discriminatory Choice and Epistemic Centrality.

It is now necessary to turn to point (c) of the framework outlined above: the absence of justification for discriminatory treatment. This element concerns less the structure of evaluation than the attribution of responsibility to an agent. Indeed, even where statutory indices or protected categories are expressly violated in the concrete case, an additional stage of inquiry into the existence of a discriminatory intent (so called: *animus discriminandi*) is often invoked in order to impute the discriminatory treatment to a specific subject. It is precisely here that Kleinberg's epistemic paradox becomes relevant, since the difficulty lies in uncovering the motives behind discrimination.[35]

As noted, para. 1, both human agents and machines, with their different cognitive properties, may conceal the underlying causes or "reasons" for the alleged discriminatory treatment, making evaluation particularly difficult. [36] This epistemic deficit may translate in civil law into a form of de-responsibilisation of the discriminating agent, whether the act is performed directly or mediated by an LLM. In this sense, where discrimination is carried out directly by a human agent, one might conclude that the absence of *intent* to discriminate leads to a finding of no fault. Conversely, in the case of an LLM-based system, the discriminating agent may invoke the machine's autonomy in order to evade liability, thereby appealing to a logic of the kind: "It was not I, but the machine; it discriminated (it is at fault, or in any case endowed with an autonomy that renders me blameless); I merely switched it on." [37] In both cases, therefore, what emerges is a dissimulation of the discriminatory treatment.

In this sense, Kleinberg's proposal proves to be highly attractive, as it may appear preferable to focus on operational capacity rather than on the ontology of the human mind and consciousness (see para. 1). Nevertheless, when transposed into the field of civil law, such a conclusion requires argumentative refinements. Indeed, if, as widely recognised in doctrine, case law, and even within legislation itself, there is no necessary correlation between intentionality and agency for the juridical attribution of a discriminatory effect. [38]

---

[35] On what is meant by the term *epistemic* in the present discussion, see n. 9.

[36] If the discriminatory cause or motive is attributable to a human agent, its concealment (whether conscious or unconscious) remains hidden *in interiore hominis*. By contrast, when it derives from a machine, presupposing the absence of consciousness, the hateful motive is not concealed (the machine does not hate); rather, what remains obscured is the justification of the decision, which is veiled by the system's own technical and operational complexity. In both phenomena, therefore, the "smoking gun" is absent when it comes to assessing the justification of the discriminatory treatment: J Kleinberg, 'Discrimination' (n 9), 27.

[37] On this point, see P. Femia, 'Essere norma. Tesi sulla giuridicità del pensiero macchinico', in P Perlingieri, S Giova, I Prisco (ed.), *Il trattamento algoritmico dei dati tra etica, diritto ed economia* (Esi, 2020), 69.

[38] More generally, civil liability has long since moved beyond the consideration of the agent's intention for the purpose of attributing responsibility under civil law. Intention remains a cornerstone only within the domains of criminal law and/or ethics (C. Maiorca, voce *Responsabilità (teoria gen.)*, in *Enciclopedia del diritto* (XVIII, 1990) 8–9. On this point, the bibliography is vast; it suffices here to refer to those texts that address discrimination directly and which do not treat intention as a necessary condition for attribution: Parenzo (n 27) 105; Zuiderveen Borgesius (n 4) 1573-1574; Shin (n 4) 28; Femia (n 4) 503-504; Harvey (n 4) 55.
On the overcoming of intention as a requirement for the attribution of civil liability, and its relegation to the domain of criminal law, see: Cass. Civ., 21 December 2003, n 46783, C-188/15 *Asma Bougnaoui and Association de défense des droits de l'homme (ADDH) v Micropole SA* (Grand Chamber, 14 March 2017) ECLI:EU:C:2017:204. For further bibliography on this point, see Battaglino (n 6) 645-649.

On this view, Kleinberg's paradox risks losing practical utility, since legal evaluation must ultimately focus on the concrete effects of discrimination rather than the motives behind it.[39]

5. The Old Problems of Technique and Liberal Rhetoric in the field of Contractual Liability. The Machine as Moral Mentor and a Revival of Kleinberg's Thesis.

Following the substantive-functional approach, which conceives indices and categories as multifactorial and does not require intentionality for the attribution of responsibility, Kleinberg's thesis may appear of limited appeal. Nevertheless, part of doctrine and case law still favours a formalist perspective, where both a precise index and a clearly defined class are required to structure a claim, often with an undue emphasis on the actor's intent (illicit motivation).[40] In this sense, Kleinberg notes that human agents tend to conceal unlawful motives behind formal compliance, whereas machines, unburdened by fear of sanctions, may within their operational complexity «reveal our own biases».[41]

His provocation should not be read in absolute terms, but as a conceptual tool to expose formalistic dynamics. Reframed substantively, it frees the interpreter from strict reliance on indices and categories or on the supposed necessity of intentionality.[42]

To illustrate: suppose a hypothetical example where a recruiter refuses to hire a candidate out of a strong aversion towards smokers.[43] Two scenarios can be envisaged:[44]

(a) Discrimination by a human agent. The illicit motive remains hidden *in interiore hominis*, and the recruiter may even claim that smoking is not expressly a protected ground.[45]

(b) Discrimination by an AI system. By contrast, stochastic LLMs or probabilistic deep learning models[46] acquire and refine their behaviour through "experience"[47] in real or simulated environments, building models upon an

---

[39] As discussed earlier according to this approach, the assessment of a discriminatory phenomenon must instead concentrate on its concrete effects, considering, as already discussed, not only the variability of indices and classes but also the very absence of intentionality.

[40] Kleinberg (n 7) 28; C R Lawrence III, 'The Id, the Ego, and Equal Protection: Reckoning With Unconscious Racism', (1987) 39 Stan Law Rew. 317.

[41] Kleinberg (n 7) 37.

[42] In this sense, according to the behaviourist perspective previously outlined, a factual equivalence is drawn between the conduct of the human agent and that of the machine agent. By contrast, to believe that the use of the latter could put an end to millennia-old problems, such as the very definition of discrimination, is a technophilic and fallacious misconception and, as in the case of liability, more often than not a fraudulent one: E Morozov , *To save everything, click here: Technology, solutionism, and the urge to fix problems that don't exists* (Public Affairs, 2013).

[43] A similar example has been advanced in the literature regarding a recruiter who discriminates against a candidate based on the applicant's zodiac sign: P S Shin, 'Is there an unitary concept', (n 4), 17-18.

[44] This analysis does not engage with those doctrines which maintain that no discriminatory treatment arises where the criterion invoked by the subject is one of their own characteristics and is attributable to their own fault. On this point, see E Pariotti, **'** "Vulnerabilità e diritto" '(2019) 2 Ars Interpr. 13.

[45] If one assumes the necessary correlation between a illicit motive or action and a legal index Consiglio (n 4) 84.

[46] Z Ghahramani, 'Probabilistic machine learning and artificial intelligence' (2015) 521(7553) Nature 452.

[47] Here the noun *experience* is employed in an improper sense. Machines are in fact regarded as lacking phenomenal consciousness. On this distinction, see N Block, 'On a Confusion about a Function of

enormous number of parameters that render the resulting decisions potentially irreproducible and unpredictable.[48] Such phenomena can be linked to reinforcement learning techniques,[49] which enable the system to update its inferences based on new information, thereby emphasising correlations that were not anticipated at the design stage. Within this framework, a system even when trained to comply with anti-discrimination norms, for instance by excluding legally protected sensitive features may nonetheless deviate from them in practice.[50] In a manner analogous to the human case, such a system might identify a statistical correlation between reduced work efficiency and smoking, inferred from yet another correlation between smoking and the worker's geographical location.[51] However, unlike the human agent, a sort of zero-sum game emerges: while the human will almost certainly dissemble the illicit intent, the machine may instead expose its bias and expressly exclude the smoker.

In this latter example, the more formalistic interpreter must evaluate the empirical and factual singularity within which it occurs, moving beyond the rigid reliance on indices and classes in evaluation as well as the supposed necessity of intention.

From this perspective, LLMs may become a cultural *leitmotiv* that destabilises the positivist methodology; a sort of "moral mentors": [52] not because machines surpass human ethics, but because they expose the fractures of a positivised system that merely reproduces market interests in legally relevant decisions. [53]

Beyond this expectation or rhetorical suggestion, the underlying problem, however, remains primarily cultural rather than technological. [54] Liberal rhetoric has long confined strict liability to exceptional cases, labelling it anachronistic,[55] when the presence of causation and intent is required.[56] As also

---

Consciousness' (1995) 18 Behavioral and Brain Sciences 227, where the author, in criticising computational theories of consciousness, distinguishes between phenomenal consciousness and access consciousness: the former consists of specific experiential mental states, that is, the qualitative properties of experience that render a situation consciously "lived." Nonetheless, at least from the perspective of civil law, theorising machine consciousness reflects merely an anthropomorphic anxiety, without altering the framework of responsibility. On this point, see Battaglino (n 7) 663-666.

[48] W Xin Zhao and others, 'A Survey of Large Language Models' [2023] arXiv:2303.18223, 1.

[49] T Lesort and others, 'Understanding Continual Learning Settings with Data Distribution Drift Analysis' [2022] arXiv:2104.01678, 1.

[50] D Tafani, 'Sulla moralità artificiale. Le decisioni delle macchine tra etica e diritto' (2020) 1 Riv. Fil 81.

[51] G Gometz, 'Intelligenza artificiale' (n 27) 27-28.

[52] The idea of the machine as a moral mentor has now ignited a debate in philosophy, is no longer confined merely to the realm of literature (for those who distinguish between the two). One may think, for example, of *Solaris* by Stanisław Lem, or of the relationship between Wagner and the homunculus in Goethe's *Faust*, to cite a more erudite reference. For relevant bibliography on this topic, see F Fossa, 'Artificial moral agents: moral mentors or sensible tools?' (2018) 20(1) Ethics Inf. Technol. 1 ff.

[53] But there is more. Certain discriminatory phenomena protected under anti-discrimination law, such as harassment or exploitation, require a bio-mechanical component which, at present, machines do not appear to possess. This problem may therefore be relegated either to technophobia or to the realm of science fiction.

[54] P Femia, 'Minus ut valeat: l'impossibilità di una regolazione giuridica metaculturale', in VV.AA *I rapporti civilistici nell'interpretazione della Corte costituzionale nel decennio 2006-2016* (ESI, 2020) 357 ff.

[55] On the individualistic and anachronistic conception that confines the application of strict liability to an exception, compare Battaglino (n 6) 636.

[56] Nardocci (n 4) 2372.

becomes evident in the evaluation of indices, this approach does not primarily presuppose substantive protection.[57]

This formal logic equally extends to the sphere of contractual liability, where an expansion of liability would raise no problem of causation or intention.[58] Here too, this rhetoric operates through subsumption, employing the historical liberal techniques of binarization and confinement:[59] whatever is not expressly included in the contract is deemed legally irrelevant, while what is included becomes valid by default: *qui dit contractuel dit juste*.[60] In this way, contractual liability is often reduced to a residual category vis-à-vis tort liability[61]; and, where the latter is invoked, liberal theory tends to cling to the moralisation of the agent, insisting on the necessity of fault, of a material causal nexus, and on the exceptional character of strict liability. Modern theory, however, has long since overcome the identification of the obligatory relationship with the mere duty of performance, recognising instead the relevance of additional and ancillary obligations. In this perspective, the "contractual" qualification of liability operates as a synecdoche: what is truly at stake is a conceptual shift from the paradigm of breach of the contract-as-act to that of breach of the obligation tout court.[62]

---

[57] Insisting on subsumption into indices in the name of mere legal certainty and predictability On the relationship between subsumption, positivism, legal certainty and liberal culture, reference may be made to the bibliography in Battaglino (n 6) 616-628.

[58] On this point, see C Castronovo, 'Swinging malpractice Il pendolo della responsabilità medica' (2020) 3 883 ff.

[59] In this sense see Femia (n 4) According to this view, *binarisation* is the condition by which multilaterally complex social interaction is reduced to the iteration of two contracting parties; *confinement*, in turn, represents its functional corollary, whereby social relations are confined within what is deemed to be valid or invalid solely on the basis of what is contractually stipulated. Within this framework, anything that does not conform to contractual relativity, to the logic of party/third party, or more generally to a network of simple, confined, and binary units, remains excluded from the contractual sphere and thereby relegated to regulatory indifference within the cognitive operations of the contractual system. In this respect, within common law jurisdictions, the attribution of contractual liability to an act of will, or, more strongly, to the presence of consideration, beyond which lies only tort, amounts to a form of binarisation.

[60] It has by now become a mere capitalist slogan, despite the meaning originally intended by Alfred Jules Emile Fourié.

[61] *Contra* C Castronovo, 'La relazione come categoria essenziale dell'obbligazione e della responsabilità contrattuale' (2011) 1 Eur. Dir. priv., *passim.*

[62] Liberal theses which exclude contractual liability when discrimination occurs "outside the contract" are grounded in a formal and static conception of contractual responsibility. Within such an ordinary configuration, certain cases would be excluded from liability for the absence of a contract as a formal act and, even more so, for the lack of a consideration, its constitutive nucleus, as provided since the Romanist tradition. Modern theory, however, precisely through a substantive approach, has long since moved beyond the identification of the obligatory relationship with the sole obligation of performance, recognising instead the significance of duties other than prestation. From this perspective, the "contractual" characterisation of liability functions as a synecdoche, as what is at stake is a conceptual shift: from the violation of the contract-as-act to the violation of obligation *tout court*. It follows that contractual liability may arise even in the absence of a formal act, provided that an obligation, whose source may differ from the contract itself, has been violated. On this point, see for Example: V Cuffaro, *Responsabilità precontrattuale,* in *Enciclopedia del Diritto*, XXXIX, Milano, 1988, 1270 e C Castronovo, *La responsabilità civile,* cit., 524. C Castronovo, *Sul significato di materia contrattuale,* cit., 518-519 Similarly, even in markedly liberal systems, though some fifty years later than German doctrine, the traditional individualistic conception of contract has been surpassed in favour of a vision of social solidarity, whereby contractual liability may be invoked even beyond the traditional barriers of privity and consideration: EH Hondius, *Precontractual Liability: Reports to the XIIIth Congress, International Academy of Comparative Law, Canada, 18–24 August 1990* 5 ff. It follows that contractual liability may also arise in the absence of a formal act, provided that an obligatory bond has been breached, the source of which does not necessarily coincide with the contract itself. See, inter alia, C Castronovo, *La nuova responsabilità civile*, cited above, *passim.* Similarly, even in markedly liberal legal systems – albeit with a delay of approximately fifty years compared to German doctrine – the traditional

It follows that the problem does not lie in the absence of a new legislative prohibition, such as a statute expressly banning discrimination by machines or regulating their use as such; rather, it remains cultural and hermeneutic. A substantive and functional approach to interpretation is capable of setting meaningful limits to discriminatory phenomena, regardless of whether the agent is human or artificial.

Rather, it is the interpreter, and not politics or formal law alone,[63] who must take up the task of confronting the injustice of discriminatory phenomena.

In conclusion, as suggested by Kleinberg, the machine may play a constructive role: whether in extra-contractual or contractual liability, it may function as a heuristic trigger that prompts a retrospective and genuinely multifactorial assessment of discrimination across all stages of the facts.[64] In this way, processes, rule-systems, practices, or institutional arrangements , even when apparently neutral and impersonal, such as anti-discrimination law itself, may, once considered in their concrete effectiveness, reveal a treatment whose absolute constant may be identified in the violation of human dignity.[65] And if civil law is to commit itself, by every available means, to combating an injustice, then its remedies must not appear anachronistic in relation to a phenomenon whose complexity grows in step with economic and technological development.[66]

individualistic conception of contract has been superseded in favour of a vision grounded in social solidarity, by virtue of which contractual liability may also be invoked beyond the traditional barriers of privity and consideration. In this regard, EH Hondius, *Precontractual Liability: Reports to the XIIIth Congress, International Academy of Comparative Law*, Canada, 18–24 August 1990, 5 ff. For further discussion, see the relevant bibliography in G Battaglino, L Pandolfelli, 'Per una responsabilità contrattuale dei marketplace' (2025) 3 Riv. crit. dir. priv. 369-385.

[63] On this point, see Nardocci (n 4) 2369-2372.

[64] In legal scholarship, there is also debate as to the temporal moment at which the prohibition of discrimination should operate within contractual dealings. Some argue for its application only after negotiations have taken place and, therefore, once the contract has been concluded. Yet, once again, those who deny that contractual liability may apply at every stage of the contractual process reproduce the liberal vision of binarisation and confinement. On this point, see the bibliography cited in note 4.

[65] Gentili (n 4) 216.

[66] The market, in fact, as an artificial construct, does not appear to reduce the phenomenon of discrimination in any meaningful way., SJ Schwab, 'Un approccio di law and economics alla teoria della Discriminazione nei rapporti di Lavoro', in P Inchino (ed), *Lezioni di diritto del lavoro. Un approccio di labour law and economics* (Giuffrè 2004) 347 ff.