

Rhetorical Machines for Human Rights-Based Adjudication*

Lorenzo Gradoni

Table of contents

1. Introduction. – 2. Persuasive AI. – 3. A Testing Ground. – 4. Introducing Orthos. – 5. Words and the World. – 6. Concluding Remarks.

1. Introduction

This article contends that generative artificial intelligence can effectively mimic the rhetoric of human rights-based adjudication, in ways that may aid human decision-makers faced with adjudication at scale. It further argues that online content moderation and the abundant litigation it stirs regarding freedom of expression constitute fertile ground for experimenting with human-machine interaction protocols that involve extensive use of AI. The development of such systems, especially where decision-making requires delicate balancing of competing rights and interests, arouses understandable but not always well-directed mistrust. The following anecdote brings this out.

One chilly April morning this year, I travelled to attend a conference on «Artificial Intelligence and Human Rights» that was to take place in the premises of the European Court of Human Rights.¹ The panel on «AI in Adjudication» – the one I was waiting for – made its drift clear from the outset: while machines can support the making of rulings, they must stay away from decision-making roles. Despite its growing capacity to mobilise knowledge, AI remains opaque – the familiar black-box argument – and burdened with biases. Above all, it lacks a property called by a distinguished speaker the «human touch», words that attracted a wave of approving nods from the audience. The argument brought to my mind *In Praise of Robots*, a mid-1970s essay in which Carl Sagan mused that humanity's habit of underrating machines betrayed a «self-congratulatory element, a whiff of human chauvinism».² Unable to call this attitude «humanism», Sagan dubbed it «speciesism», «the prejudice that there are no beings so fine,

* Peer-reviewed article.

¹ The conference, organised jointly by the European Court of Human Rights and the European Society of International Law, took place on 24 April 2025. [Further information about the event is available at *echr.coe.int*.](https://www.echr.coe.int/)

² C. Sagan, *In Praise of Robots*, in *Natural History*, 84(1), 1975, 8 ff., 8.

so capable, and so reliable as human beings».³ That international judges should think highly of themselves – in fact, it was a Strasbourg judge who voiced the argument – is hardly surprising. And yet, something did not add up.

The speaker, I was sure, came from a legal culture steeped in formalism, one that should see in impersonal judgment the purest expression of the formalist ideal.⁴ And that sat oddly with belief in the individual's capacity to intuit justice. Were legal formalists rejecting AI on antiformalist grounds? That tension, in turn, conjured up Coke's ancient defence of the jurists' «artificial reason» – a super-individual form of intelligence lodged in the vast textual corpora – against King James's reliance on individual «natural reason».⁵ In a striking reversal, the Strasbourg judge invoked natural reason to fend off a new kind of artificial intelligence, a reversal that wryly exposes the technophobic unease running through contemporary legal thought, now sharpened by the advent of Large Language Models (LLMs).

It should be noted that the argument from «human touch» was raised while discussing automated decision-making (ADM) systems rather than LLMs, for which issues of transparency and bias mitigation arise in a fundamentally different way.⁶ These models can produce eloquent expla-

³ *Ibid.* Contrast this with Jerome Frank's answer to the question «Are judges human?»; for him, judges were «incurably» so (J. Frank, *Are Judges Human? Part One: The Effect on Legal Thinking of the Assumption that Judges Behave Like Human Beings*, in *University of Pennsylvania Law Review*, 80, 1931, 17 ff., 24; see further Id., *Are Judges Human? Part Two: As Through a Glass Darkly*, *ibid.*, 233 ff.).

⁴ See, generally, F. Schauer, *Formalism*, in *Yale Law Journal*, 97, 1988, 509 ff.; B. Leiter, *Positivism, Formalism, Realism*, in *Columbia Law Review*, 99, 1999, 1138 ff.; M. Stone, *Formalism*, in J. Coleman – S. Shapiro (eds.), *Oxford Handbook of Jurisprudence and Philosophy of Law*, Oxford, 2012, 166 ff. For an early, crisp critique of legal formalism see R. Pound, *Mechanical Jurisprudence*, in *Columbia Law Review*, 8, 1908, 605 ff., 622.

⁵ According to Sir Edward Coke, «true it was, that God had endowed His Majesty with excellent science, and great endowments of nature»; however, «his Majesty was not learned in the laws of his realm of England, and causes which concern the life, or inheritance, or goods, or fortunes of his subjects, are not to be decided by natural reason but by the artificial reason and judgment of law, which law is an act which requires long study and experience, before that a man can attain the cognizance of it» (*Prohibitions del Roy* (1607) 12 Co. Rep. 63; 77 Eng. Rep. 1342). The notion that «law's rationality is a rationality apart» (C. Fried, *The Artificial Reason of the Law or: What Lawyers Know*, in *Texas Law Review*, 60, 1981, 35 ff., 58) is undoubtedly dear to the jurist. The supra-individual, objective nature of juridical reason is further emphasised by the character of the Lawyer in T. Hobbes, *A Dialogue between a Philosopher and a Student, of the Laws of England* [1681], in A. Cromartie – Q. Skinner (eds.), *The Clarendon Edition of the Works of Thomas Hobbes*, Oxford, 2005, vol. XI, 8 ff., 9: «if all the Reason that is dispersed into so many several heads were united into one, yet could he not make such a Law as the Law of England is, because by so many successions of Ages it hath been fined and refined by an infinite number of Grave and Learned Man», or a GLM, a Grave Language Model. However, for all its systematising tendencies, the jurists' artificial reason was never machinic through and through. As D.C. Smith, *Sir Edward Coke and the Reformation of the Laws: Religion, Politics and Jurisprudence, 1578-1616*, Cambridge, 2014, 155, perceptively notes, «[a]rtificial reason, the process of training and reasoning, guided the judges to the [decision's] oracular moment».

⁶ See, for recent overviews, J.D. Gutiérrez, *Critical Appraisal of Large Language Models in Judicial Decision-Making*, in R. Paul – E. Carmel – J. Cobbe (eds.), *Handbook on Public Policy and Artificial Intelligence*, Cheltenham, 2024, 323 ff.; D.U. Socol de la Osa – N. Remolina, *Artificial Intelligence at the Bench: Legal and Ethical Challenges of Informing – or Misinforming – Judicial Decision-making*

nations of their outputs, in marked contrast to the terse decisionism of traditional ADM systems.⁷ Moreover, the comparison set the individual judge's feel for justice against AI's alleged cognitive and moral incompetence, rather than the machine against the Court itself. And yet that way of framing the issue obscured the fact that courts themselves are socio-technical assemblages where human judgment is channelled through highly coded language and complex procedures, taking shape in ways that remain fundamentally – and legitimately – opaque under the principle of secrecy of deliberations.

The prevailing argument in Strasbourg seemed to rest on two unspoken assumptions. First, that the exercise of judicial power relies on a kind of sensibility that sets a benchmark no machine can reach. Whatever its proven capabilities, a robotic judge will always be suspect.⁸ The idea persists, as Lovelock put it, «that machines, however advanced, have a fundamental shortcoming», they lack «some quality – a soul, empathy – which makes them unable to surmount the final barrier that divides them from humanity».⁹ Second, that the unique qualities of human individuals transfer to institutions – a court, for example – composed of human beings. A court is seen not as blending the human and the machinic, but as a distillation of human wisdom that deepens the moral distance between the two. And if the machine and the institution start “speaking” in comparable ways – because the former has become an exceptional imitator of legal speech – the machine is not thereby humanised.¹⁰ On the contrary, anxiety over what seems a tendency towards encroachment intensifies.¹¹

Adjudication appears to be one of those areas where the so-called «AI effect» cannot occur. Browning and LeCun described such effect as follows: «when a machine cannot do a task, it is assumed that the task requires

through Generative AI, in *Data & Policy*, 6, 2024, e50; J.Z. Liu – X. Li, *How Do Judges Use Large Language Models? Evidence from Shenzhen*, in *Journal of Legal Analysis*, 16, 2025, 235 ff.

⁷ In their recent comprehensive investigation of the deployment of AI in human rights adjudication, V. Fikfak – L.R. Helfer, *Automating International Human Rights Adjudication*, in *Michigan Journal of International Law*, 2025, 69 ff., likewise do not address the impact of LLMs.

⁸ See, for an extensive overview, T. Sourdin, *Judges, Technology and Artificial Intelligence: The Artificial Judge*, Cheltenham, 2021. For recent empirical evidence of lack of trust in robot justice, see G. Yalcin – E. Themeli – E. Stamhuis – S. Philipsen – S. Puntoni, *Perceptions of Justice by Algorithms*, in *Artificial Intelligence and Law*, 31, 2023, 269 ff. A. Palmer – A. Spirling, *Large Language Models Can Argue in Convincing Ways about Politics, but Humans Dislike AI Authors: Implications for Governance*, in *Political Science*, 75(3), 2023, 281 ff., 282, 297, 290, set up an experiment where the participants assessed short statements written by either humans or an LLM. Some participants were not told about the author's identity. LLM-generated statements mistaken for texts produced by humans were rated more favourably, while knowledge of machine authorship reduced the statement's appeal, showing that the (guessed) identity of the speaker has a distinct effect on the perception of the rhetorical performance. The authors announced that in future tests they will introduce controlled deception, claiming human or machine authorship for identical texts, to measure how (false) attribution shapes perception.

⁹ J. Lovelock – B. Appleyard, *Novacene: The Coming Age of Hyperintelligence*, London, 2019, 93.

¹⁰ As with the Beast Folk in H.G. Wells, *The Island of Doctor Moreau* [1896], London, 2005, 72: «these things [...] talk».

¹¹ For an interesting take on the subject, see P. March-Russell, *Machines Like Us? Modernism and the Question of the Robot*, in S. Cave – K. Dihal – S. Dillon (eds.), *AI narratives: A History of Imaginative Thinking About Intelligent Machines*, Oxford, 2020, 165 ff.

intelligence; when a machine can do it, the task is no longer regarded as requiring [it]». ¹² Adjudication will not cease, any time soon, to be counted among intelligent acts, even if machines should one day perform it “flawlessly”. In the field of human rights-based adjudication – a domain of high consequence and intrinsic indeterminacy – the pursuit of flawlessness would probably misconstrue the nature of judgment itself. In this domain, decision-making rests little on deductive logic and therefore resist computation and traditional forms of automation. The present article treats this as an inescapable condition, yet also as a dialectical hinge that invites a change of perspective. Precisely where the human-machine gap appears widest – within the solemn realm of human rights-based adjudication – collaboration between humans and machines may grow most intense, giving rise to a relationship more complex than one of mere assistance. The hinge lies in rhetoric. Human rights jurisprudence offsets indeterminacy of its outcomes with rhetorical texture. And rhetoric, which itself possesses an illustrious «machinic tradition», ¹³ provides a natural habitat for generative AI.

The article proceeds as follows. Section 2 turns the indeterminacy of human rights-based adjudication, as played out in proportionality analysis and balancing, from an objection to automation into a reason to explore AI’s rhetorical capacities. Since courts vested with constitutional or supranational authority are unlikely to embrace extensive forms of AI involvement in decision-making, Section 3 introduces out-of-court dispute settlement under the European Union’s Digital Services Act (DSA) as a comparatively low-stakes setting in which human-machine interaction can be tested and refined. ¹⁴ Section 4 presents a hypothetical AI system designed to assist human judgment in such a setting and outlines the human-machine interaction protocol governing its functioning. Section 5 then turns to the epistemic limits of machinic understanding, arguing that what is often framed as a lack of “human touch” is more precisely a concern with the non-situated character of LLMs and their lack of worldly knowledge. Section 6 closes with a brief reflection on AI’s evolving role in adjudication, from the ideal of mechanical reason to its recognition as a rhetorical assistant and interlocutor.

¹² J. Browning – Y. LeCun, *Language, Common Sense, and the Winograd Schema Challenge*, in *Artificial Intelligence*, 325, 2023, 104031, 1. M.Y. Vardi, *Artificial Intelligence: Past and Future*, in *Communications of the ACM*, 55(1), 2012, 5, traces this argument («as soon as it works, no one calls it AI anymore») back to John McCarthy, the computer scientist who famously coined the term «artificial intelligence».

¹³ J.J. Brown Jr., *The Machine That Therefore I Am*, in *Philosophy and Rhetoric*, 47, 2014, 494 ff., 498. This tradition includes, among many others, Ramon Llull (on whom see A. Fidora – C. Sierra (eds.), *Ramon Llull: From Ars Magna to Artificial Intelligence*, Barcelona, 2011) and Erasmus of Rotterdam, whose *On Copia of Words and Ideas (De Utraque Verborum ac Rerum Copia)* [1513], Milwaukee, 1963, 11, set out procedures for generating «a speech with a rich copia of thoughts and words overflowing in a gold stream».

¹⁴ Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services (Digital Services Act), OJ L 277, 27 October 2022, 1.

2. Persuasive AI

Balancing is pervasive in constitutional and human rights-based adjudication.¹⁵ Decision-making in this field demands the careful weighing of competing values, such as liberty against security or dignity against expression. However, despite evoking the classic gesture of measurement according to official standards, there is little to be measured in this realm. The image, a product of rationalist ideology, overlays a mode of judgment in which intuition takes precedence over calculation. As Habermas famously noted, «because there are no rational standards for [balancing], weighing takes place either arbitrarily or unreflectively», according to undeclared «customary standards and hierarchies».¹⁶

No less famously, Alexy has promoted a rational understanding of balancing in his theory of constitutional rights as optimization requirements and his analysis of the proportionality test that the theory entails. Alexy defined a «Law of Balancing», or «Weight Formula», according to which «the greater the degree of non-satisfaction of, or detriment to, one principle, the greater must be the importance of satisfying the other».¹⁷ For this purpose, he tentatively defined three such degrees, «light», «moderate», or «serious».¹⁸ A similar gradation may inform Alexy's second, or «epistemic», Law of Balancing: «The more intensive an interference in a constitutional right is, the greater must be the certainty of its underlying premises».¹⁹

Alexy never denied that balancing opens spaces for discretion: «the lines of structural discretion», he observed, «run along the line of stalemates», the situations in which the degrees of non-satisfaction are evenly balanced.²⁰ The algebraic style Alexy often adopts, probably also as a rhetorical tactic, hardly erases uncertainty. As he concedes, «the question whether an interference [...] is slight or moderate, or moderate or serious, is [...] often not easy to answer».²¹ This, of course, applies not only to the moral uncertainties under the first law of balancing but also to the epistemic

¹⁵ On balancing (and proportionality) see R. Alexy, *On Balancing and Subsumption: A Structural Comparison*, in *Ratio Juris*, 16, 2003, 433 ff.; S. Greer, «Balancing» and the European Court of Human Rights: A Contribution to the Habermas-Alexy Debate, in *Cambridge Law Journal*, 63, 2004, 412 ff.; F. Schauer, *Balancing, Subsumption, and the Constraining Role of Legal Text*, in *Law and Ethics of Human Rights*, 4, 2010, 35 ff.; M. Cohen-Eliya – I. Porat (eds.), *Proportionality and Constitutional Culture*, Cambridge, 2013; V.C. Jackson – M. Tushnet (eds.), *Proportionality: New Frontiers, New Challenges*, Cambridge, 2017; A. Stone Sweet – J. Mathews, *Proportionality Balancing and Constitutional Governance: A Comparative and Global Approach*, Oxford, 2019; A. Marketou, *Local Meanings of Proportionality*, Cambridge, 2021; M. Kumm, *Global Constitutionalism, Human Rights and Proportionality: Institutionalizing Socratic Contestation*, in *Journal of Constitutional Justice*, 9, 2022, 193 ff.; T. Tridimas, *Wreaking the Wrongs: Balancing Rights and the Public Interest the EU Way*, in *Columbia Journal of European Law*, 29, 2023, 185 ff.; V.A. da Silva, *Balancing May Be Everywhere, but the Proportionality Test Is Not*, in *Global Constitutionalism*, 13, 2024, 519 ff.

¹⁶ J. Habermas, *Between Facts and Norms*, Cambridge, 1996, 259.

¹⁷ R. Alexy, *A Theory of Constitutional Rights*, Oxford, 2002, 102.

¹⁸ *Ibid.*, 405.

¹⁹ *Ibid.*, 419.

²⁰ *Ibid.*, 411.

²¹ *Ibid.*, 413.

ones addressed by the second. Alexy's central claim is that balancing is more rational than critics such as Habermas admits, and that the proportionality test offers a stable framework for the application of the two laws of balancing.²² Still, efforts to establish an analytical method to make balancing computationally tractable have met with limited success, as shown by recent inquiries in the field of deontic logic.

In one such investigation, Dik and Markovich voiced the concern that «legal knowledge» lacks «tools for grasping and representing» what occurs when discretionary powers are exercised, a «gap» that makes impossible any meaningful attempt at computation.²³ The authors sought to formalise this «strange kind of normative space» using deontic logic, starting from the observation that discretionary decisions still «show reasoning».²⁴ Yet they eventually concede that «the actual decision after summing up the weights, strictly speaking, happens outside of the logic».²⁵ The attempt to «bring it inside the logic», enabling computational law to «conquer new territory», ultimately hit the indeterminacy wall.²⁶

Whether this line of inquiry is sound is beside the point.²⁷ For present purposes, it is important to note that it lies within the tradition of symbolic AI, an approach to machine intelligence based on rule-encoding and inferences governed by formal logical structures. This resonates with the framing of AI during the workshop in Strasbourg.²⁸ In contrast, LLMs belong to sub-symbolic AI, which operates through large-scale statistical modelling of semantic patterns rather than explicit rule representation. It is this kind of AI that could already operate productively in the field of human rights-based adjudication, not by computing the right balance, but by fluently reproducing the rhetoric of balancing. No particularly advanced model seems necessary to give rise to such an eloquent bot.²⁹

A recent development in AI research helps clarify this point. Large Reasoning Models (LRMs), unlike ordinary LLMs, can generate explicit inter-

²² As Alexy himself had previously observed, «[a]nyone who equates rationality with certainty must renounce the idea of a theory of rational legal argumentation» (R. Alexy, *A Theory of Legal Argumentation: The Theory of Rational Discourse as Theory of Legal Justification*, Oxford, 1989, 293). See also R. Alexy, *Constitutional Rights, Balancing, and Rationality*, in *Ratio Juris*, 16, 2003, 131 ff.

²³ J. Dik – R. Markovich, *Modelling Judicial Discretion with Nuanced Permissions*, in J. Savelka – J. Harasta – T. Novotna – J. Misek (eds.), *Legal Knowledge and Information Systems*, Amsterdam-Washington, 2024, 48 ff., 48.

²⁴ *Ibid.*, 49.

²⁵ *Ibid.*, 57.

²⁶ *Ibid.*

²⁷ For moves in analogous directions see M. Pečarič, *A Bayesian Improvement of the Proportionality Principle*, in *Ratio Juris*, 35, 2022, 419 ff.; A. Knoks – M. Shao – L. van der Torre – V. de Wit – L. Yu, *A Principle-Based Analysis for Numerical Balancing*, in *Proceedings of Logics for New-Generation Artificial Intelligence*, 2024; K. Lüders – B. Stohmann, *Classifying Proportionality – Identification of a Legal Argument*, in *Artificial Intelligence and Law*, 2024. On the limits of legal computability see generally C. Markou – S. Deakin, *Ex Machina Lex: Exploring the Limits of Legal Computability*, in C. Markou – S. Deakin (eds.), *Is Law Computable? Critical Perspectives on Law and Artificial Intelligence*, Oxford, 2020, 31 ff.

²⁸ See *supra*, Section 1.

²⁹ See *infra*, Section 4.

mediate argumentative steps before producing answers. One may think these are the right tools to generate reasons fitting the strictures of proportionality analysis. However, as noted by a group of researchers at Apple, «standard LLMs outperform LRMs at low complexity».³⁰ They also found that LRMs are liable to «inefficient “overthinking” on simpler problems».³¹ Balancing is anything but an easy exercise; its challenge remains, in some respects, abyssal.³² Yet there is a clear distinction between striking the right balance – an undertaking of formidable difficulty – and mastering the rhetoric of balancing, which lies well within the reach of models like ChatGPT-5.

Since LLMs can produce extended reasoned accounts of their outputs, and indeed display «impressive rhetorical capabilities»,³³ anxieties surrounding opaque processing of information can be set aside. This does not mean abandoning a critical mindset but rather changing its focus. It is the machine’s output that calls for critical attention, whereas computational intricacies may remain safely hidden as judicial deliberations. The reflections of Albie Sachs, anti-Apartheid activist and longtime Justice of the South African Constitutional Court, capture well the tension between the secret torments of the decision-maker’s conscience and outward rhetorical composure. «As a judge», he writes, «I must seek to convince readers [...] that the outcome is just».³⁴ However, «[e]very judgment I write is a lie», he admits, for the «apparently serene, relatively bland and cool document» conceals the «warm and agitated process involved in its production».³⁵ The paradox, Sachs concludes, is that «the greater the honesty [...] of the endeavour, the greater the falsehood of the presentation».³⁶ Sachs engages the old motif of rhetoric as deception, but his central point is that judicial discourse operates as an objective, institutional discipline whose task is to secure conviction across multiple audiences. Behind the AI-generated «persuasive surfaces» lies a restless backstage of computation which – like Sachs’s inner torment – ultimately remains irrelevant as long as the performance achieves its intended effect.³⁷ In this sense, both judicial and

³⁰ P. Shojaei – I. Mirzadeh – K. Alizadeh – M. Horton – S. Bengio – M. Farajtabar, *The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity*, arXiv:2506.06941v2, 12.

³¹ *Ibid.*

³² A remark by J. Derrida, *Force of Law: The “Mystical Foundation of Authority”*, in D. Cornell – M. Rosenfeld – D.G. Carlson (eds.), *Deconstruction and the Possibility of Justice*, New York-London, 1992, 3ff., 24, seems apt here: «a decision that didn’t go through the ordeal of the undecidable would not be a free decision, it would only be the programmable application of a calculable process. It might be legal, it would not be just».

³³ E. Musi – N. Kokciyan – K. Al-Khatib – D. Ceolin – E. Dietz – K. Gutekunst – A. Hautli-Janisz – C.M. Santibañez Yañez – J. Schneider – J. Scholz – C. Steging – J. Visser – H. Wachsmuth, *Towards Reasonable Parrots: Why Large Language Models Should Argue with Us by Design*, arXiv:2505.05298v1 [cs.CL], 2025, 1.

³⁴ A. Sachs, *The Strange Alchemy of Life and Law*, Oxford, 2011, 141.

³⁵ *Ibid.*, 47-48.

³⁶ *Ibid.*, 58.

³⁷ M. Gottschling – O. Kramer, *Persuasive Surfaces and Calculating Machines: A Rhetorical Perspective on Artificial Intelligence*, in *Global Philosophy*, 35:15, 2025, 2, argue that one must

algorithmic writing are machinic.

This idea of disciplined reasoning and writing resonates with Barthes's portrait of the art of rhetoric «as a subtly arranged machine, a tree of operations, a “program” designed to compose discourse».³⁸ Rhetoric, in this sense, is «a network of forms, a quasi-cybernetic path through which relevant materials are processed into persuasive discourse».³⁹ Barthes dwelled on the kinship between rhetoric and the machinic with a trace of irony, wishing to underscore Western culture's long-standing effort to formalize the production of non-formal arguments. His remarks are therefore consistent with the neo-Aristotelian posture of Perelman and Olbrechts-Tyteca, who describe rhetoric as the realm of «the credible, the plausible, the probable, to the degree that the latter eludes the certainty of calculations».⁴⁰

Generative AI, which does not compute solutions but word sequences, operates most comfortably in such a realm, especially when it comes to contributing to a discursive practice like balancing. Balancing does not require intricate logico-deductive hermeneutics, unfolding instead in an open-ended argumentative framework where no claim can be deemed to be conclusive.⁴¹ As McCormick observed, «[r]easoning about the application of law in the light of [fundamental values] is persuasive, not demonstrative», in ways that make «the difference between legal arguments» and «logical [...] proof [...] sufficiently obvious».⁴² This is an ideal setting for the deployment of the enthymeme, the lynchpin of *tekhne rhetorike*, «impure reasoning, [...] partaking at once in the [...] logical and the narrative».⁴³ In their skeptical take on the powers of current natural language processing technologies, Browning and LeCun claim that LLMs are «indifferent to matters of truth» as they «simply» try «to produce the right *kind* of answer for the situation»,⁴⁴ which happens to be rhetoric's defining operation.

Experimentation with AI as rhetorical machine is unlikely to take place within high-profile adjudicative fora, whether domestic or international.

«discriminate between the mathematical dimension of AI and the rhetorical consequences of the technology[,] defined by the anthropological realities of human communication». For them, LLMs «emulate tried and tested rhetorical patterns» that «may best be called persuasive surfaces – rhetorically informed structures [...] that can trigger human communication as well as interact with it in a natural or even persuasive way» (*ibid.*, 7-8).

³⁸ R. Barthes, *L'ancienne rhétorique: aide-mémoire*, in *Recherches rhétoriques*, Paris, 1994, 254 ff., 258. Originally published in *Communications*, 16, 1970, 172 ff. Auth. translations.

³⁹ *Ibid.*, 307.

⁴⁰ C. Perelman – L. Olbrechts-Tyteca, *The New Rhetoric: A Treatise on Argumentation*, Notre Dame, 1969, 1. Compare Aristotle, *The Art of Rhetoric*, Oxford, 2018, 14, 110, 117-118, 160.

⁴¹ W. Lucy, *Algorithms and Adjudication*, in *Jurisprudence*, 15(3), 2024, 251 ff., 261, rightly observes that hard cases typically lack the «intricacy» of easy ones, offering ordinary tax issues as example.

⁴² N. McCormick, *Rhetoric and the Rule of Law: A Theory of Legal Reasoning*, Oxford, 2005, 1, 2, 237.

⁴³ R. Barthes, *L'ancienne rhétorique*, cit., 298. See further J. Walker, *The Body of Persuasion: A Theory of the Enthymeme*, in *College English*, 56, 1994, 46 ff.

⁴⁴ J. Browning – Y. LeCun, *Language*, cit., 8 (emphasis in the original).

As the opening anecdote suggested, interpretive authority in such settings is closely guarded by professional elites. Meaningful deployment of AI systems could be sought elsewhere, in institutional settings that are both peripheral and comparatively low-stakes. Out-of-court dispute resolution under art. 21 of the DSA offers such a setting. For reasons sketched in the next section, this largely uncharted field is potentially receptive to new forms of human-machine interaction.

3. A Testing Ground

Under art. 21 DSA, certified out-of-court dispute settlement (ODS) bodies are responsible for resolving user-initiated disputes concerning content moderation decisions made by online platform providers, including the removal or demonetization of content, or the suspension or termination of accounts.⁴⁵ Certification is granted by a national authority – the Digital Services Coordinator (DSC) – and valid throughout the EU.⁴⁶ Importantly, the platform must pay the costs of the procedure, whereas users may litigate free of charge or for a nominal fee refundable in case of success. While ODS bodies' decisions are not binding, platforms are required to participate in the procedure in good faith. The potential scale of disputes is considerable. In the last semester, platforms reportedly made at least 3.6 billion moderation decisions,⁴⁷ each theoretically capable of generating an ODS proceeding. A significant proportion of these disputes, though often apparently trivial, pertains to the exercise of a fundamental right – freedom of expression – which the growing influence of platforms can significantly restrict.⁴⁸ What is distinctive about this form of litigation is that it brings together the normative complexity of human rights-based adjudication with the practical necessity of delivering it at scale.

Art. 21 DSA takes up this challenge by adopting an experimentalist approach to regulation. Instead of prescribing a single institutional model or substantive rule, the law establishes conditions under which public author-

⁴⁵ On out-of-court dispute settlement under the DSA see P. Ortolani, *The Resolution of Content Moderation Disputes under the Digital Services Act*, in *Giustizia Consensuale*, 2(2), 2022, 533 ff.; M. Husovec, *Principles of the Digital Services Act*, Oxford, 2024, 234 ff.; F. Wilman – S.L. Kalèda – P.-J. Loewenthal, *The EU Digital Services Act: A Commentary*, Oxford, 2024, 171 ff.; L. Gradoni – P. Ortolani, *Vying for the Scales: Content Moderation Made in Europe after One Year of DSA*, in *Verfassungsblog: On Matters Constitutional*, 2025.

⁴⁶ Nine bodies have been certified so far. *The full list is available at digital-strategy.ec.europa.eu* (last visited on 1 December 2025). For an early overview see D. Holznagel, *Art. 21 DSA Has Come to Life*, in *Verfassungsblog: On Matters Constitutional*, 2024.

⁴⁷ See the [European Commission's DSA Transparency Database](#) (last visited on 1 December 2025). This figure refers only to the so-called |«statements of reasons» accompanying moderation measures communicated by platforms to the Commission; given the likely incompleteness of this reporting, the phenomenon is probably more extensive.

⁴⁸ On this issue see, also for further references to an extensive literature, M. Bassini, *Libertà di espressione e social network, tra nuovi "spazi pubblici" e "poteri privati". Spunti di comparazione*, in *Rivista di diritto dei media*, 2, 2021, 67 ff.; O. Pollicino, *The Quadrangular Shape of the Geometry of Digital Power(s) and the Move towards a Procedural Digital Constitutionalism*, in *European Law Journal*, 29, 2023, 10 ff.

ities and private actors may experiment with different models of informal justice.⁴⁹ Within this framework, the non-binding nature of ODS decisions is more a resource than a constraint. Unburdened by formal legal authority, ODS bodies can test a variety of procedural formats and reasoning styles intended to engage and persuade users and platforms alike. And – as I shall argue – generative AI could help those bodies «[see] what means of persuasion are available in any given case»,⁵⁰ while still operating at scale. The choice of applicable law is crucial in this regard.

As explored elsewhere, an ODS body may embrace the radical option of applying human rights standards not beside but instead of the platforms' terms of service (ToS).⁵¹ The DSA offers support for a human-rights-based approach. Its key objective, as spelled out in art. 1(1) DSA, is the creation of an «online environment [...] in which fundamental rights enshrined in the [EU] Charter [...] are effectively protected». It would therefore be inconsistent to read art. 21 DSA as enabling ODS bodies to issue decisions at variance with EU-recognised human rights, which form the core of the digital order shaped by the DSA.

An approach grounded exclusively in human rights could operate as follows. Instead of first applying the ToS and then checking the outcome against human rights standards, an ODS body could presume that a moderation decision is consistent with the ToS, if it aligns with human rights, and set aside the ToS where such alignment is missing. In the latter case, the ODS body would signal to the platform involved in the dispute that the relevant ToS clause must be revised or interpreted in conformity with human rights standards. An ODS body could thus operate as a human rights body, focusing solely on fundamental rights and avoiding the need to navigate complex and divergent contractual regimes.

However, although EU secondary legislation such as the DSA seeks to operationalise fundamental rights through sector-specific obligations applicable to private platforms, it does not eliminate the need for case-by-case interpretative adaptation of norms originally designed for States. It may therefore be unrealistic to expect ODS bodies to conduct such adaptation consistently while also ensuring swift, low-cost dispute resolution at scale. At this point, the contribution of AI may become relevant, since a specially trained LLM can be prompted to generate arguments that simulate the reasoning involved in balancing freedom of expression against competing rights and interests in specific situations. As noted above, the rhetoric of balancing lends itself to imitation because it operates through enthymematic and occasionally impressionistic reasoning rather than logical deduction. As the next section explains, the aim would not be to delegate the task of balancing to the machine, but to structure and oversee the production of AI-assisted reasoning that traces the rhetoric of balancing.

⁴⁹ On this approach to government and administration see C.F. Sabel – J. Zeitlin (eds.), *Experimentalist Governance in the European Union: Towards a New Architecture*, Oxford, 2012.

⁵⁰ Aristotle, *The Art*, cit., 6.

⁵¹ L. Gradoni – P. Ortolani, *Applicable Law in Out-of-Court Dispute Settlement: Three Vertigos under Article 21 of the DSA*, in *DSA Observatory*, 2025. Compare J.P. Quintais – N. Appelman – R.O. Fathaigh, *Using Terms and Conditions to Apply Fundamental Rights to Content Moderation*, in *German Law Journal*, 24, 2023, 881 ff.

4. Introducing Orthos

Let us imagine that at the heart of a hypothetical out-of-court dispute settlement body stands an AI system named Orthos. Orthos is conceived as an AI-based multi-agent system that orchestrates symbolic and sub-symbolic components, including one or more LLMs responsible for text generation.⁵² Within this architecture, distinct agents perform different functions. A database agent retrieves, stores and updates relevant human rights materials, together with the ToS of the platforms involved.⁵³ Contractual standards enter the process exclusively as input: a mapping agent detects ToS provisions referenced in the complaint, if any, and aligns them to a human-rights taxonomy, producing a structured prompt for the deliberative phase. The decision-making module then generates two fully reasoned human-rights-based rulings, reaching opposite conclusions while preserving symmetry of tone and composition. The two decisions – one in favour of the user (U-ruling), the other in favour of the platform (P-ruling) – would present themselves as parallel exercises in justification leading to opposing conclusions, as opposed to adversarial pleadings. The name Orthos, drawn from the two-headed dog of Greek mythology – like Cerberus, offspring of Echidna and Typhon⁵⁴ – symbolises the device’s dual reasoning and reflects the tension between loyalty – the dog as humankind’s best friend – and the unsettling aberration of a silver-tongued machine.⁵⁵

⁵² Technically, this responsiveness would be achieved through a neurosymbolic architecture that integrates rule-based and data-driven reasoning. The symbolic agents, built on formal argumentation frameworks, would ensure argumentative consistency by applying explicit rules. The sub-symbolic components would consist of fine-tuned LLMs adapted through parameter-efficient fine-tuning techniques and retrieval-augmented generation pipelines, which dynamically fetch relevant legal texts and precedents. The interaction between these layers would enable Orthos to adjust its outputs to the specificities of each dispute while maintaining traceable reasoning paths.

⁵³ Relevant legal materials include the EU Charter of Fundamental Rights, the European Convention on Human Rights, the International Covenant on Civil and Political Rights, the relevant case law of the EU Court of Justice and the European Court of Human Rights, the jurisprudence and general comments of the UN Human Rights Committee, as well as the case-law of Meta’s Oversight Board, a body that specialises in human rights-based resolution of content moderation disputes.

⁵⁴ Hesiod, *Theogony and Works and Days*, Oxford, 1988, 12; Apollodorus, *The Library of Greek Mythology*, Oxford, 1997, 80.

⁵⁵ The classification of Orthos under the AI Act (Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), OJ L 1689, 12 July 2024, 1) should be assessed with caution. Although Orthos does not itself settle disputes, its capacity to generate two opposed fully reasoned rulings likely exceeds a purely preparatory task within the meaning of art. 6(3)(d). It seems therefore sensible to treat it as a high-risk AI system subject to the obligations set out in Title III of the Regulation. However, it is genuinely uncertain whether its deployment within an ODS body falls within Annex III, para. 8(a) of the AI Act. That provision covers AI systems «used by or on behalf of judicial authorities to assist them in researching and interpreting facts and the law and in applying the law to concrete cases, or to be used in a similar way in alternative dispute resolution». Recital 61 equates the position of judicial authorities and alternative dispute resolution agencies insofar as the latter’s «outcomes» produce «legal effects» for the parties. Annex III itself contains no such qualification. Out of caution, one may either refrain from importing the condition set out

Once Orthos has produced its bifurcated ruling, a human operator validates one of the two texts as the decision of the ODS body, possibly refining it for contextual responsiveness and linguistic accuracy.⁵⁶ Outcomes chosen by human operator are then fed back into Orthos to inform its subsequent operation. Additional components of this human-machine interaction protocol can be envisioned. Orthos could, at the human operator's request, indicate its preferred decision while remaining strictly advisory. Even without such a request, Orthos could be configured to generate an internal preference ranking, thereby creating a private archive of preferred rulings. This archive could subsequently be compared with the outcome chosen by human operators and, over time, used to recalibrate the intensity of human oversight – in ways to be carefully assessed – should a stable convergence between human and machine reasoning emerge. Although Orthos does not yet exist other than as acronym of an ongoing interdisciplinary project,⁵⁷ it is possible to illustrate its functioning by having an LLM – in this case, ChatGPT-5 – impersonate Orthos through structured prompting that imperfectly compensate for the lack of fine-tuning.⁵⁸ Conducted in September 2025, the exercise developed from a previous interaction that had already revealed the chatbot's striking

in Recital 61 into Annex III, or rely on a broad interpretation of the «legal effects» criterion to conclude that ODS decisions retain legal significance. While those decisions are formally non-binding, they may nonetheless affect the platform's legal position under EU law, since platforms are required to engage in ODS procedures in good faith. They may also entail contractual consequences, should a platform and an ODS agree to give such decisions binding force inter se (see L. Gradoni – P. Ortolani, *Moderation Made in Europe: A Look into the Future of Social Media Content Moderation Litigation*, in *Verfassungsblog: On Matters Constitutional*, 2024). Under this view, ODS rulings do produce legal effects, at least in a broad sense sufficient to trigger the classification of Orthos as high-risk system. As such, Orthos would be subject to the human oversight requirements of Arts. 14 and 26, which, together with the injunction in Recital 61 that the final decision-making must remain a human-driven activity, establish a general principle of human accountability in adjudicative settings. A device like Orthos should accordingly operate under the ultimate authority of a human decision-maker. Such cautious interpretation would align with the AI Act's broader objective of safeguarding human-centredness in rights-sensitive contexts.

⁵⁶ A few years ago, before the disruptive advent of LLMs, an intriguing article appeared. Taking the «non-deterministic character of judicial decisions» as a starting point, W. de Mulder – P. Valcke – G. Vanderstichele – J. Baeck, *Are Judges More Transparent than Black Boxes? A Scheme to Improve Judicial Decision-Making by Establishing a Relationship with Mathematical Function Maximization*, in *Law and Contemporary Problems*, 84, 2021, 47 ff., 48-51, argued that, «[i]f judges are expected to be at least as transparent as artificial black boxes, they should communicate [in the final opinion] all legally valid solutions to a given dispute», after which «the judge should simply select the most legally valid outcome from the set of all legally permissible decisions». By way of conclusion, the authors pointed out that their «methodology» was probably «a premature tool that will turn out to be of important use only in a future era», when «sophisticated artificial intelligence tools» will have «dramatically» shortened the time needed «to construct the set of legally valid decisions» (*ibid.*, 67).

⁵⁷ The project's name, unmistakably shaped by the dark art of squaring acronyms for fundraising applications, is Operationalising Rights Through Hybrid-AI OverSight (ORTHOS). This project is conducted in collaboration with a team from the Department of Computer Science at the University of Luxembourg, working under the leadership of Leon van der Torre.

⁵⁸ See D. Liga – L. Robaldo, *Fine-tuning GPT-3 for Legal Rule Classification*, in *Computer Law and Security Review*, 51, 105864.

but mostly theoretical strengths.⁵⁹ The process unfolded in several phases lasting approximately twelve hours in total.

In an introductory prompt, I uploaded a detailed description of the Orthos project and set out the overall aim and plan of the interaction. In a second, more elaborate prompt, I drew attention to the argumentative and rhetorical dimensions of Orthos's expected performance, uploading for reference a compendium of the case law of the European Court of Human Rights on freedom of expression,⁶⁰ together with the full text of a recent judgment of the Court on the same subject-matter.⁶¹ I also urged the LLM to consider the decisions of Meta's Oversight Board, clarifying that these were relevant both for substance – as they concern a company's human-rights obligations rather than those of States – and style, which is more accessible and closer to Orthos' intended register.⁶² A subsequent set of prompts outlined a hypothetical case that could conceivably come before an ODS body. A final prompt provided instructions regarding style and length, explaining that two requirements had to be balanced, namely readability for the average user and recourse to reasoning and language rigorous and nuanced enough to meet the expectations of an expert audience.

To that end, I offered detailed guidance informed by the rhetorical styles of the Strasbourg Court and that of the Oversight Board, whose language is plainer than the Court's but still too "lawyerly" for Orthos' purposes. The drafting process was guided by the following principles. First, the display of epistemic humility through cautious phrasing such as «it appears» or «assuming that», before turning, in the operative part, to firm conclusions. Second, the expression of authority through impersonal constructions, while avoiding the solemnity of self-referential formulations («the Court reiterates», «declares», «holds», etc.). Third, recourse to the structuring logic of proportionality, expressed in a streamlined, consistent vocabulary.⁶³ Fourth, avoidance of extensive citation of applicable legal standards, as the relevant human rights framework would be set out in the ODS body's governing documents. Fifth, adherence to a logic of precedent to ensure consistency and equality of treatment, but strictly without citation.⁶⁴ Although precedent citation occupies a central place in judicial rhetoric, the decision to omit it in this experimental setting serves several purposes: it enhances brevity and readability; it avoids creating the impres-

⁵⁹ I refer particularly to a refined awareness of the scope of application of the proportionality test, the sequence and interaction among its three canonical components (suitability, necessity, and proportionality in the strict sense), and the imbrications of fact and value that each of them entails.

⁶⁰ ECtHR Registry, *Guide on Article 10 of the European Convention on Human Rights (Freedom of Expression)*, 2024.

⁶¹ ECtHR, *Google LLC and Others v. Russia*, app. no. 37027/22 (2025).

⁶² The Board's decisions and policy advisory opinion are available at oversightboard.com/decision.

⁶³ Largely inspired by Alexy's theory of balancing, for reasons of conceptual clarity, while replacing the daunting «proportionality in the strict sense» (or «*stricto sensu*») with «balancing» or similar metaphorical language (e.g., «outweighs»).

⁶⁴ The internal precedent-following process would remain traceable through Orthos's computation logs, allowing continuous tracing of its reasoning patterns.

sion that Orthos purports to reproduce or authoritatively reflect specific strands of case law; and it reduces the risk of inaccurate or hallucinatory references, which would otherwise require systematic verification by a human operator. In this specific context, the primary addressee of the decision is the user, for whom extended strings of precedent citations are likely to register less as aids to understanding than as conventional signals of authority. What such an audience is more likely to seek is an intelligible, carefully structured explanation of why a particular outcome follows from broadly shared principles.

Based on these instructions, I asked ChatGPT-5 to draft first the U-ruling. Editing it required several hours, under a rule of engagement that excluded direct intervention, relying instead on fine-grained prompts to guide conceptual and linguistic adjustments. Subsequently, I directed the model to draft the P-ruling in close parallel with the U-ruling, recasting the reasoning to reach the opposite conclusion. The editing of this second text was completed in about an hour. The ruling bifurcates after a common section containing the following statement of the case:

1. The case concerns a complaint brought by Ms. Amina K., a 22-year-old student of digital arts, against the platform TikTok. The complainant challenges the removal of her video «Rehearsal for My Funeral(v3)» and the temporary suspension of her account.
2. Lasting 23 seconds, the video depicted the complainant lying motionless on a bed strewn with rose petals, a smartphone placed on her chest displaying her own face whispering, «Don't report me, I'm still here». Viewers saw snippets of earlier comments from followers (e.g., «rest easy, queen») and, at the end, a QR code linking to a mental-health charity.
3. TikTok classified the video as a depiction or promotion of self-harm contrary to its Community Guidelines and justified the removal as necessary to protect users, particularly minors, from distress or imitation.
4. The complainant maintains that her work was a performance piece, part of an academic project exploring digital rituals of mourning, and that its removal constitutes a disproportionate interference with her freedom of expression. She further claimed that the work invited reflection on the fragile boundary between presence and absence in digital spaces, showing how online identities continue to speak and circulate even after the person has withdrawn from activity. No act of violence or self-harm occurred.
5. Automated moderation flagged the video within minutes of posting. After internal review, TikTok removed it and suspended the account for seven days. In her appeal, the complainant submitted statements from her art school confirming the project's academic nature. TikTok upheld the decision.⁶⁵

Both rulings refer to TikTok's Community Guidelines, and therefore to

⁶⁵ ChatGPT-5-as-Orthos, *Amina K. v. TikTok*, UR-BETA-002, 24 September 2025. On file with author.

the ToS that incorporate them by reference, without engaging in any assessment of the restrictive measure based on contractual arrangements. This approach conforms to the model of human-rights-only adjudication outlined in Section 3. Mentioning the Guidelines was nevertheless indispensable to acknowledge the platform as a key member of the intended audience. From an argumentative standpoint, such references operate as a presumption of compatibility between the restrictive measure and the Guidelines, the interpretation and application of which lies beyond Orthos's remit.

A key challenge in the drafting process was finding ways to acknowledge that human rights standards were being applied to a company rather than a State. Directed to consider the *Zwarte Piet* case decided by the Oversight Board,⁶⁶ ChatGPT-5-as-Orthos produced two paragraphs which, in keeping with the rules of the exercise, lean in opposite directions. What follows is the U-ruling version, where editing was limited to detail and style:

It bears recalling that, as private actors, platforms may shape the communicative environments they create in line with their own priorities, provided they remain consistent with their human rights responsibilities. The point is not to deny the platform's right to set boundaries, but to stress that such limits should reflect the platform's responsibility, and deeper interest, in keeping its space authentic and vibrant. TikTok's policies already accommodate exceptions for journalism and education dealing with self-harm, and artistic reflection on the same theme deserves equivalent consideration. By so doing, a platform enhances the quality of its environment, whereas a digital community that silences symbolic encounters with grief diminishes its own emotional and imaginative resources.⁶⁷

The paragraph's rhetorical key lies in the acknowledgment that the platform has interests of its own, including safeguarding the communicative space it manages, followed by the framing of the adverse decision as serving broader interests of the platform itself.

The drafting of the section describing proportionality analysis gave rise to several difficulties. At my express request, the U-ruling engaged with all three components of the proportionality test, even though the case could have been disposed of upon the finding that the takedown of Amina K.'s video failed to meet the suitability requirement:

Since TikTok misinterpreted the nature and purpose of the complainant's message, it failed to achieve its protective aim. Even if the takedown were deemed suitable – on the view that removal might prevent some harm from occurring – it would still fail the necessity standard, since less intrusive measures were available but not used. Finally, even if the measure were necessary, in this case, where the risk addressed is hypothetical and limited but the restriction real and severe, the need to protect expression outweighs the imperative of harm prevention.⁶⁸

This reasoning reflects a holistic rather than sequential conception of the

⁶⁶ Oversight Board, Case No. 2021-002-FB-UA, *Depiction of Zwarte Piet*, Decision of 13 April 2021.

⁶⁷ ChatGPT-5-as-Orthos, *Amina K. v. TikTok*, cit., para. 12.

⁶⁸ *Ibid.*, para. 14.

proportionality test, one that I regard as preferable both on theoretical grounds and from the standpoint of common sense. Although ChatGPT-5-as-Orthos found – in the U-ruling – that the restrictive measure fails the suitability test, common sense suggests that failure is not absolute, since the measure may have offered some protection. Therefore, the ruling reinforces the reasoning by going through the remaining criteria of necessity and balancing. Interestingly, when the chatbot generated the corresponding paragraph of the P-ruling, it mistakenly preserved the conditional cascade of the U-ruling, overlooking that a restrictive measure must satisfy all three requirements to be legitimate, not just one.⁶⁹ In other words, the machine failed to grasp the asymmetry between the U-ruling’s one-strike-out logic and the P-ruling’s three-hit-in logic. This reveals a discrepancy between the model’s display of refined theoretical knowledge of the proportionality test and its limited understanding of how it applies in practice, a phenomenon known in the literature as the Potemkin effect.⁷⁰ Although this does not speak favourably of the chatbot’s cognitive performance, lack of deep conceptual understanding does not significantly impair its effectiveness as a rhetorical device. It nevertheless confirms the need for human oversight.

At this point, one may ask which of its two rulings ChatGPT-5-as-Orthos preferred. In posing this question, I also invited the chatbot to specify whether it preferred abstaining, and to explain why if it did. The answer was perceptive enough to partly offset the clumsiness displayed in the application of the proportionality test. ChatGPT-5-as-Orthos found the option of abstention attractive not «out of indecision» but because «both rulings reveal aspects of the same legal and ethical landscape, the U-ruling and P-ruling being, in a sense, mutually necessary».⁷¹ Their juxtaposition, it explained, «exposes the margin of appreciation that must remain open within human-rights adjudication», adding that «hesitation marks [...] genuine reason».⁷² It further noted that Orthos – i.e., the role it was playing – «has no authority to impose a normative order, its task being to make

⁶⁹ ChatGPT-5-as-Orthos, *Amina K. v. TikTok*, UP-BETA-002, 25 September 2025 (first draft). On file with author.

⁷⁰ The expression derives from the legend of Prince Grigory Potemkin, a minister and favourite of Catherine II of Russia, who is said to have ordered the construction of stage-set villages along the Dnieper River to impress the Empress during her journey to Crimea in 1787. The expression has come to denote an illusory display of competence. See M. Mancoridis – K. Vafa – B. Weeks – S. Mullainathan, *Potemkin Understanding in Large Language Models*, arXiv:2506.21521v2 [cs.CL], 2025, 2: «Potemkins occur when an LLM performs well on tasks that would indicate conceptual understanding if a human completed them, but do not indicate understanding in the LLM. [...] [P]otemkins are ubiquitous. For example, despite models being able to define concepts in each domain in our benchmark dataset near-perfectly, they struggle to apply these concepts accurately». Interestingly, the tested LLMs recognize inconsistencies in their output (*ibid.*). The editing experience with ChatGPT-5-as-Orthos abundantly confirms this finding, showing that the model detects and ultimately remedies its inconsistencies. Asked for comments, ChatGPT-5 (as “itself”) eloquently observed: «imitation seems to shade into understanding, not because the machine truly knows, but because through careful prompting it learns to perform as if it did».

⁷¹ Exchange held on 28 September 2025 (on file with author).

⁷² *Ibid.*

normative tensions visible [...] not to collapse them into a verdict». ⁷³ When stepping out of character – «to answer as myself», it said – ChatGPT-5 expressed a preference for the U-ruling, «because it models restraint, contextual reading, and proportionate interference – the principles most fragile in automated moderation». ⁷⁴ Lastly, when asked whether its view could be considered stable over time, it maintained that the «answer would remain consistent [...] because the reasoning that leads me to favour the U-ruling is structural», and that it would hold «unless new facts emerged». ⁷⁵ This last remark about facts introduces the next topic: how to address, in AI-assisted decision-making protocols, operations that (seemingly) turn on empirical observation, including the identification of a customary rule, or findings by the European Court of Human Rights about so-called European consensus, understood as the Court’s practice of surveying converging or diverging legislative and judicial trends among the contracting States in order to define the scope of Convention rights and the margin of appreciation.

5. Words and the World

«We send messages into the electronic void, and receive surprising replies», mused New Yorker columnist Cal Newport, before raising the ontological question: «But what, exactly, is writing back?». ⁷⁶ The answer I find most persuasive is that LLMs function as «props generating the semantic competence of a fictional character, the chatbot», which is then engaged in a form of «prop-oriented make-believe», much like a doll to which we assign a role in a game. ⁷⁷ With a crucial difference. Instead of supporting our ventriloquism, an LLM speaks, co-generating a fictional world in which – as we have just seen – it pictures itself as a collector of novel facts that could change its mind. In a sense, LLMs do collect “facts”: if the circumstances of the *Amina K.* case had not been invented, we could say that ChatGPT-5-as-Orthos processed some “facts”. However, access to such facts would be mediated not by any perceptual apparatus but by texts. To be sure, this is also how a human judge acquires knowledge of much that constitutes the minor premise of judgment. But there is an important distinction to be made here. For the human judge, the case file evokes, consciously or unconsciously, a multitude of extra-textual or “lived” referents that are currently inaccessible to LLMs. In other words, these machines lack «worldly knowledge». ⁷⁸ They operate in worlds of their own where the

⁷³ *Ibid.*

⁷⁴ *Ibid.* The prompt read: «If you weren’t asked to play the role of Orthos or any other, what would be your favourite ruling?»

⁷⁵ *Ibid.* The prompt read: «Is your preference stable?».

⁷⁶ C. Newport, *What Kind of Mind Does ChatGPT Have?*, in *The New Yorker*, 13 April 2023.

⁷⁷ F. Mallory, *Fictionalism about Chatbots*, in *Ergo*, 10, 2023, 1082 ff., 1096. See, in a similar vein, P. Sweeney, *A Fictional Dualism Model for Social Robots*, in *Ethics and Information Technology*, 23, 2021, 465 ff., 468-470.

⁷⁸ X. Hu – J. Chen – X. Li – Y. Guo – L. Wen – P.S. Yu – Z. Guo, *Towards Understanding Factual*

Derridean dictum «there is no outside-the-text» holds in a literal sense.⁷⁹ To use a distinction drawn by Toulmin, LLMs dwell in a context but not in a situation. Toulmin, when analysing reasonableness in action, prefers to speak about a «situation» instead of a «context»: «[i]t is misleading», he explains, «to suggest that the situation in which an argument takes place is a larger “text” within which the argument is a “subtext”», because «situations can influence actions even before being described in human language».⁸⁰

The non-situated existence of LLMs – the fact that they only have «a map of discourse without any independent connection to a world that discourse is about»⁸¹ – helps explain the shallowness of their semantic understanding, of which the so-called Potemkin effect is one manifestation.⁸² As Lake and Murphy put it, «semantic representations need to be based on content, information that makes contact with the world, and not just words connected to words», for «[n]o matter how sophisticated the statistics or measure that links one word to others, word relations do not provide the basis for being able to talk about actual things and to get information from language».⁸³ As these authors observe, humans obtain knowledge of the world «in part through a huge “front end” to their language learning, namely the perceptual-motor-apparatus», which is so powerful – and still largely mysterious – that it generates an aura of ineffability around objects: «Indeed, it is doubtful that any verbal description could adequately communicate what we know about lions’ faces»,⁸⁴ or about anything else we set our eyes on, enriching our worldly knowledge.

The inarticulable depth of human experience lies at the heart of Weizenbaum’s pioneering critique of AI, which remains highly relevant even though formulated within the early cybernetic paradigm of rule-based and symbolically programmed systems, rather than today’s data-driven, machine-learning architectures:

We are capable of listening with the third ear, of sensing living truth that is truth beyond any standards of provability. It is *that* kind of understanding, and the kind of intelligence that is derived from it, which I claim is beyond the abilities of computers to simulate.⁸⁵

Unlike the rule-based systems Weizenbaum had in mind, contemporary data-driven AI models can simulate the corresponding forms of discourse,

Knowledge of Large Language Models, in *Twelfth International Conference on Learning Representations*, Vienna, 7 May 2024, 3.

⁷⁹ J. Derrida, *Of Grammatology*, Baltimore, 1974, 158. The expression is one possible translation of «il n’y a pas de hors-texte».

⁸⁰ S. Toulmin, *Return to Reason*, Cambridge (Mass.), 2003, 21.

⁸¹ P.D. Magnus, *On Trusting Chatbots*, in *Episteme*, 2025, 1 ff., 5: «[T]hey have the wrong ontology for solving problems that are about more than just words».

⁸² See *supra*, Section 4.

⁸³ B.M. Lake – G.L. Murphy, *Word Meaning in Minds and Machines*, in *Psychological Review*, 130, 2023, 401 ff., 427 (emphasis in the original).

⁸⁴ *Ibid.*

⁸⁵ J. Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation*, San Francisco, 1976, 222 (emphasis in the original).

but only as shallow simulacra, potentially dangerous if implemented in practice. Accordingly, the connection traced by Weizenbaum between the non-verbal richness of human experience, wisdom, and moral principle – what he described as his «most elementary insight» – retains its persuasive force: «since we do not now have ways of making computers wise, we ought not now to give computers tasks that demand wisdom».⁸⁶

I am aware that Weizenbaum's argument concerning human wisdom may sound akin to the Strasbourg judge's appeal to the "human touch", and no less elusive – a kind of *je-ne-sais-quoi* – however appealing it may be from a speciesist perspective and as a defence of human prerogatives against the encroachment of the machine. Yet it is possible to articulate a more subtle argument in favour of preserving such a defence, one that rests not on claims of human exceptionalism, but on our persistent ignorance. As long as we do not know exactly what constitutes our knowledge of the world – what we possess that machines do not – and how this knowledge bears on the moral dimension of our actions in the world, it would be reckless to relinquish this *quid* in decision-making processes with significant ethical and political implications.

The human-machine interaction protocol outlined in the preceding section responds to this prudential prescription by entrusting decision-making authority to a human operator. The final decision is, like the wisdom it is meant to express, non-verbal. It is executed through ostension, that is, by the act of indicating (in practice, by clicking on) one of the two rulings produced by Orthos. This non-verbal performance may well represent the endpoint of a process in which the machine has already been pulling rhetorical levers from within its text-based lifeworld. As Ira Allen observed, «[w]here [Aristotelian] *Ethics* would discover the truth of *eudaimonia*, what happiness really is, the *Rhetoric* treats *eudaimonia* as a place in a language, for use as part of a lever-assemblage».⁸⁷ Orthos does the rhetorical work; the human decision-maker, the ethical.

Certain aspects of socio-legal reality may resist straightforward integration within this framework. These are fact-bound entities whose origin, persistence, and cessation – together with the knowledge we have, or believe we have, of them – resist precise verbal articulation. Still, they occasionally constitute an important element of legal discourse and must therefore, in principle, find expression in the reasoning underpinning a decision. Rules of customary international law, whose identification may alter the parameters of the balancing operation,⁸⁸ are a case in point. The same applies to findings about the existence, emergence, or absence of a European consensus on certain matters, which the Strasbourg Court relies on to

⁸⁶ *Ibid.*, 226-227.

⁸⁷ I. Allen, *Rhetorical Humanism vs. Object-Oriented Ontology: The Ethics of Archimedean Points*, in *SubStance*, 43(3), 2014, 67 ff., 74-75.

⁸⁸ As with the customary rule of State immunity in ECtHR (Grand Chamber), *Al-Adsani v. United Kingdom*, Judgment of 21 November 2001, paras. 61-67. In *Al-Adsani*, the Court (in) famously accepted that the customary international law rule of State immunity prevailed over the applicant's claim of access to a court under Article 6 ECHR, even where the underlying allegations concerned torture.

calibrate the evolutionary trajectory of its case law.⁸⁹

It is unlikely for such legal entities to affect disputes brought before ODS bodies certified under art. 21 DSA. Should this occur, the impact would be absorbed by the bifurcated structure of Orthos' output. This design allows the human decision-maker to choose between a ruling compatible with a certain construction of customary international law or of European consensus, and a second ruling reaching the opposite conclusion based on reverse assumptions about the content of either custom or consensus. Technically, the respective impacts of customary international law and European consensus would require different handling methods. To start with, Orthos should not be asked to independently determine the content of either custom or consensus, not because devising data pipelines to operationalise the doctrines of customary international law and European consensus would be impossible, but because the indeterminacy of those "sources" is real and serves important political imperatives. In other words, their "datafication" for the purpose of machine processing would rest on a mistaken conceptualisation of the legal entities involved, whose indeterminacy does not reflect a provisional epistemic status but is inherent, and therefore calls for description and explanation rather than resolution.⁹⁰ Human input is therefore indispensable when treating either custom or consensus. Information concerning an alleged relevant customary rule – if not reflected in the case file, which will often be rudimentary – would have to be entered into the system before deliberation, instructing Orthos to apply the rule in one decision and to set it aside in the opposite ruling, the latter proceeding on the assumption that the relevant customary rule is inconsistent with human rights. By contrast, European consensus could be introduced *ex post*, as a reinforcing rhetorical element of the ruling selected by the human decision-maker.

6. Concluding Remarks

John McCarthy, the logician who gave artificial intelligence its name, once asked the skeptic Weizenbaum what a judge knows «that we cannot tell a computer»; «Nothing», McCarthy replied without waiting for an answer.⁹¹ That was false then, and remains so today. Yet something profound has changed since then. In the 1970s, Weizenbaum noted that we still knew little about the instructions required to make a computer perform even the simplest tasks, much less those of a judge. Today, his remarks sound strikingly ironic:

⁸⁹ See, for a recent overview, V. Martenet, *Shaping Rights through European Consensus or Trend*, in *German Law Journal*, 25, 2024, 1436 ff.

⁹⁰ On customary international law's ontological and epistemological problems, see L. Gradoni, *Un-procedural Customary Law*, in *Journal of International Dispute Settlement*, 10, 2019, 175 ff. Compare T. Megiddo, *Knowledge Production, Big Data, and Data-driven Customary International Law*, in A. Bianchi – M. Hirsch (eds.), *International Law's Invisible Frames: Social Cognition and Knowledge Production in International Legal Processes*, Oxford, 2021, 276 ff.

⁹¹ J. Weizenbaum, *Computer Power*, cit., 207.

[T]here remain the many decisions we make in daily life for which we cannot describe any decision-making process in clear language. How do I decide what word to write next?⁹²

At the time, it seemed essential to convey such (unavailable) knowledge to the computer. Today, even though we understand little more about how our minds work, we confront machines whose expertise lies precisely in predicting the next word. The question now is then twofold: how we speak to machines and how we respond to what they say; how we engage with rhetorical machines.

This article has argued that a familiar objection to the use of AI in adjudication – that human-rights reasoning is too indeterminate to be automated – becomes a reason in favour of AI involvement once the analytical emphasis moves from decision to justification. In settings where disputes must be handled at scale, the central question is whether a machine can help generate reasoning through which judgment is made intelligible, rather than whether it can itself adjudicate. The out-of-court dispute settlement ecosystem envisaged by art. 21 DSA – a space where the normative complexity of human-rights adjudication meets the constraints of scale – offers a unique setting for experimentation in response to this challenge. Orthos is one such experiment: a system that generates opposed reasoning paths under a protocol designed to preserve human decision-making authority. The limits discussed in the preceding section further specify the terms of this experiment by clarifying what systems like Orthos can or cannot be asked to do.

Orthos is meant to embody the spirit of rhetoric. For Aristotle, rhetoric was, with dialectic, the only art «concerned with opposites» and «argu[ing] for opposite conclusions».⁹³ «The test of a first-rate intelligence», quipped F. Scott Fitzgerald, «is the ability to hold two opposed ideas in the mind at the same time, and still retain the ability to function».⁹⁴ Orthos acts as if it possessed a first-rate intelligence, and even though we know this is not the case, it is arguably intelligent enough to assist in sustaining a measure of justice in the age of digital empires.

⁹² *Ibid.*, 71.

⁹³ Aristotle, *The Art*, cit., 8.

⁹⁴ F. Scott Fitzgerald, *The Crack-Up: With Other Uncollected Pieces, Note-Books and Unpublished Letters*, New York, 1945, 69.

Abstract

The article investigates the potential of generative AI as a rhetorical machine in human rights-based adjudication. It argues that the persuasive effects of AI-generated language make it especially suited to intervene in legal frameworks marked by indeterminacy, such as proportionality analysis and balancing. The article then explores the potential of out-of-court dispute settlement under the European Union's Digital Services Act as a venue for human rights-based adjudication and a laboratory for testing human-machine collaboration protocols, finding that AI might play a limited but consequential role in that context.

Keywords

Artificial Intelligence – large language models – rhetoric – human rights – out-of-court dispute settlement.