JOURNAL OF
CROSS-DISCIPLINARY RESEARCH
IN COMPUTATIONAL LAW [CRCL]

# Generative AI, Explainability, and Score-Based Natural Language Processing in Benefits Administration

Frank Pasquale[*] Gianclaudio Malgieri[**]

## Abstract

Administrative agencies have developed computationally-assisted processes to speed benefits to persons with particularly urgent and obvious claims. One proposed extension of these programs would score claims based on the words that appear in them (and relationships between these words), identifying some sets of claims as particularly like known, meritorious claims, without understanding the meaning of any of these legal texts. This score-based natural language processing (SBNLP) may expand the range of claims categorized as urgent and obvious, but as its complexity advances, its practitioners may not be able to offer a narratively intelligible rationale for how or why it does so. At that point, practitioners may utilize the new textual affordances of generative AI to attempt to fill this explanatory gap, offering a rationale for decision that is a plausible imitation of past, human-written explanations of judgments in cases with similar sets of words in their claims.

This article explains why such generative AI should not be used to justify SBNLP decisions in this way. Due process and other core principles of administrative justice require humanly intelligible identification of the grounds for administrative action. Given that 'next-token prediction' is distinct from understanding a text, generative AI cannot perform such identification reliably. Moreover, given current opacity and potential bias in leading chatbots – which are based on large language models – as well as deep ethical concerns raised by the databases they are built on, there is a strong case for excluding these automated outputs from administrative decision-making. Nevertheless, SBNLP may be established parallel or external to justification-based legal proceedings for humanitarian purposes.

**Keywords:** Generative AI, Explainability, and Score-Based Natural Language Processing in Benefits Administration, AI, Due Process, Administrative Procedure, Simulation, AIA, Benefits

*Journal of Cross-disciplinary Research in Computational Law*

DOI: pending

www.journalcrcl.org

[*] Professor of Law, Cornell Law School and Cornell Tech.

[**] Associate Professor of Law & Technology, eLaw - Center for Law and Digital Technologies, University of Leiden; Co-Director of the Brussels Privacy Hub, Free University of Brussels (VUB).

## Generative AI, Explainability, and Score-Based Natural Language Processing in Benefits Administration

Legal technology has a long record of 'automating inequality.'[1] Advancing AI to accelerate legal determinations that help the disadvantaged is one way to atone for these negative impacts. For example, the U.S. Social Security Administration has developed 'Compassionate Allowance' and 'Quick Disability Determination' processes to speed the delivery of benefits to persons with particularly urgent and obvious claims. These processes quickly identify meritorious claims and avoid undue delays.

Scoring methods may also expand the range of claims that may be categorized as urgent and obvious. For example, if all claimants with a set of medical conditions were awarded benefits in the past, a point system (that, say, grants benefits to someone with over 110 points) might simply assign 115 points to that set of conditions and end the inquiry once it had been identified. Alternatively, such a system may assign 20 points per serious disease. In either situation, the scoring system would help accelerate claims prone to undue delays in the past. However, it might also errantly grant benefits in cases where circumstances, other than the presence or absence of that set of conditions, would exclude the claim from statutorily specified categories. For example, in the U.S. disability determination context, someone who is substantially gainfully employed at a job would be ineligible for such disability payments, even if they met all medical criteria.

Nevertheless, the promise of machine learning is that more complex versions of such scoring could identify combinations of factors (and not just medical conditions) that always led to awards in the past.[2] We call this hypothetical approach to processing benefits applications 'score-based natural language processing' (SBNLP) and predict it will become an increasingly tempting expedient wherever decision-makers confront staffing limitations and the processing of myriad applications.

From a rule of law perspective, SBNLP's lack of explainability is deeply problematic. A sufficiently complex scoring system will not demonstrate how, say, a given combination of factors gave rise to a finding of disability. It only identifies *that* it has done so, similar to machine learning processes based on correlations rather than accounts of causation.[3] The normative value of such legal analytics is uncertain here.[4] In legal systems that demand some explanation of the basis of state action, this lack of truly meaningful information about the nature of the information processing could prove an insuperable barrier to such scoring systems' adoption.

While such problems seemed insurmountable just a few years ago, the rise of generative AI provides new hope for an explainable (or at least rationalized) automated decision-making process. Just as SBNLP can derive a score from a given set of filings, a chatbot based on a large language model (LLM) and fine-tuned to past authoritative written opinions in the benefits field (and the underlying filings in such cases) may be able to generate an opinion rationalizing the score's result. This may involve finding past precedential holdings to simulate an explanation of the claimant's success, or highlighting factual dimensions of the present case that parallel similar factual aspects of supportive, past precedents. Eugene Volokh proposed that such a system could replace appellate judges, and the rise

[1] See Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (St Martin's Publishing Group 2018). *Automating Inequality* is a riveting, emotionally compelling story of vulnerable lives turned upside down by bad data, shoddy software, and bureaucrats too inept or corrupt to make things right. All too often, systems billed as a way to protect the vulnerable in fact, do just the opposite, trapping them in a modern-day star chamber. See also Sofia Ranchordás, 'Empathy in the Digital Administrative State' (2022) 71(6) Duke Law Journal 1341.

[2] Indeed, many machine learning approaches would not assign any numerical value at all, assessing similarity on the basis of, say, a 'bag of words' comparison. Nevertheless, since the most plausible version of this type of short-cut pattern matching would be a scoring system, we will focus on this possibility in this intervention.

[3] Frank Pasquale and Glyn Cashwell, 'Prediction, persuasion, and the jurisprudence of behaviourism' (2018) 68(supplement 1) University of Toronto Law Journal 63.

[4] See Geneviève Vanderstichele, The Normative Value of Legal Analytics. Is There a Case for Statistical Precedent? (3474878, 30 August 2019).

[5] Eugene Volokh, 'Chief Justice Robots' (2019) 68(6) Duke Law Journal 1135.

in legal applications of ChatGPT and similar tools has given new relevance to his proposal.[5]

This article explores whether such an LLM-based opinion writer would be a valuable adjunct to SBNLP of benefits claims. Section 'From Computational Decision Support to Score-Based Natural Language Processing' begins with a description of proposals for automating eligibility for and receipt of benefits. The U.S. Social Security Administration (SSA) has developed some paths toward computational determination of benefits eligibility. Simple word-matching algorithms and document authentication software may eventually lead to more advanced forms of natural language processing. This may include the scoring of words, phrases, and even sentences and paragraphs, for likelihood of a positive result. However, there will be resistance to the allocation of decision-making authority to such scoring, given that it is already at one remove from the language-based application of rules to fact patterns. As the number of factors and interactions among factors influencing a decision increases, the less likely the decision is to be explained coherently and substantively.[6]

Section 'The ChatGPT Solution? Proposals for LLM-Based Opinions and Their Flaws' explains how chatbots based on LLMs may offer some apparent solutions to this problem ('Explanation and Administration'), and why these approaches should be rejected ('Explaining Scoring?'). The provision of a large enough dataset of past cases with authoritative opinions may allow a chatbot to be fine-tuned with sufficient "training" so that it can generate outputs that are similar to human-written explanations. These outputs could invoke similarities between the fact patterns of the case decided via SBNLP and past precedents. They may also suggest meaningful information about the scoring process itself. However, this would not be sufficient explanation for a decision, since LLMs are language models, not knowledge models. They are mere next-word-predictors, incapable of the reasoned application of law to facts that is the hallmark of legitimate legal decision-making. This is one of many reasons that experts have cautioned against outsourcing core judicial functions to automated processes.[7] The justification of an outcome is just as important to judicial legitimacy as the outcome itself.

The conclusion reflects on the normative implications of the discussion in the previous section for SBNLP. SBNLP deployed to find patterns that match the neediest and most compelling claims of the past could lead to the much more rapid allocation of benefits to the most urgent appeals for assistance. Given the emergency character of many such applications, it would be acceptable to suspend written justification requirements in favour of expediency. This avoids the dilemma of either banishing SBNLP due to its explanatory deficits, or kludging together *post hoc* explanations for it. SBNLP can in this way be an exception, outside of or parallel to justification-based legal proceedings, rather than a force warping their integrity from within. Unlike the exception theorized in the work of Carl Schmitt, this algorithmic exceptionalism is less disruptive of the nature of legal authority than a technical appropriation of a small part of law's domain.

## From Computational Decision Support to Score-Based Natural Language Processing

Examining governmental services, experts in administrative law have also seized on the promise of automation using natural language processing (NLP), artificial intelligence (AI), and machine learning (ML). In a report published in 2019, law professors identified numerous opportunities for automation of several dimensions of the administrative state.[8] They identified scenarios of 'mass justice' as being particularly promising targets for AI. Justice is rarely more 'mass' than in the adjudication of Social

[6] As social theorists have observed, coherent narration depends on the pruning of detail. BC Han, *The Crisis of Narration* (D Steuer tr, Polity Press 2024); Mark Andrejevic, *Automated Media* (Routledge 14 October 2019).

[7] Joe McIntyre and Anna Olijnyk, 'Public Law Limits on Automated Courts' in Katie Miller and Janina Boughey (eds), *The Automated State - Implications, Challenges and Opportunities for Public Law* (The Federation Press 17 June 2021).

[8] David Freeman Engstrom and others, 'Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies' [2020] SSRN Electronic Journal.

Security Disability Insurance (SSDI) applications, which now number over two million per year.

The Social Security Disability Determination process is complex and, for a high number of claimants, lengthy. As the SSA explains, many steps of evaluation are required.[9] For the purposes of this article, the key determination includes the following steps:

> "[W]e consider the medical severity of your impairment(s). If you do not have a severe medically determinable physical or mental impairment that meets [the] duration requirement . . . or a combination of impairments that is severe and meets the duration requirement, we will find that you are not disabled. . . . (iii) At the third step, we also consider the medical severity of your impairment(s). If you have an impairment(s) that meets or equals one of our listings in appendix 1 of this subpart and meets the duration requirement, we will find that you are disabled."[10]

Multiple layers of appeal mean that a significant percentage of claimants will wait months or even years for benefits they are ultimately deemed to have deserved at the time of their application. These appeals are time-consuming and predictably delay a large number of claimants who are ultimately successful. In a study of one year of SSDI claims from 1,041,383 applicants, only 36% (374,376) of claims were initially successful, while 59% were ultimately allowed.[11] This means that over 200,000 applicants were likely to have been delayed in accessing benefits they were due, some by many months or years.

Qualification for benefits via 'listed impairments' offers a particularly important opportunity to streamline the disability determination process. Listed impairments are 'severe enough to prevent an individual from doing any gainful activity, regardless of his or her age, education, or work experience.'[12] Therefore, a finding of a listed impairment ends the disability determination. This also avoids fact-intensive determinations of whether the applicant could take on some form of work that is available in the national economy. The statutory and regulatory category of 'listed impairments' is, therefore, a good foundation for 'fast-track' disability determination processes.[13]

The SSA has recognized the importance of quickly identifying those applicants with a listed impairment. This is particularly pressing because so many beneficiaries die while waiting for their claim to be processed.[14] SSA's Compassionate Allowance (CAL) initiative has identified a subset of the most pressing listed impairments.[15] As disability determination experts Kenneth Abbott, Yen-Yi Ho, and Jennifer Erickson explain, cases typically 'receive CAL designation because SSA text-matching software finds reasonably accurate spellings of qualifying diseases, such as glioblastoma multiforme, in a specific field on the electronic disability application.'[16] There is a specific list of CAL conditions.[17] Once deployed, the 'CAL selection software identifies cases for CAL processing based solely on

[9] 20 C.F.R. § 404.1520(a)(4) (2012). ('If we can find that you are disabled or not disabled at a step, we make our determination or decision and we do not go on to the next step. If we cannot find that you are disabled or not disabled at a step, we go on to the next step.')

[10] 20 C.F.R. § 404.1520(a)(4) (2012). This is a binding interpretation of the relevant statute, 42 U.S.C.A § 423(d)(5)(A). ('An individual shall not be considered to be under a disability unless he furnishes such medical and other evidence of the existence thereof as the Commissioner of Social Security may require.')

[11] United States Social Security Administration, *Annual Statistical Report on the Social Security Disability Insurance Program, 2017* (2018) (analysing applications from 1998).

[12] 20 C.F.R. § 404.1525(a) (2017); 20 C.F.R. § 404.1520(d) (2012).

[13] David Rajnes, '"Fast-track" strategies in long-term public disability programs around the world' (2012) 72(1) Social Security Bulletin 79.

[14] Elizabeth K Rasch and others, 'First in Line: Prioritizing Receipt of Social Security Disability Benefits Based on Likelihood of Death During Adjudication' (2014) 52(11) Medical Care 944, 944 (estimating the deaths of waiting recipients in the thousands).

[15] Social Security Administration, 'SSA - POMS: DI 11005.604 - Processing Compassionate Allowances (CAL) in the Field Office (FO) - 05/24/2023' (2015) ⟨https://secure.ssa.gov/poms.nsf/lnx/0411005604⟩ accessed 28 December 2023.

[16] Kenneth Abbott, Yen-Yi Ho, and Jennifer Erickson, 'Automatic health record review to help prioritize gravely ill Social Security disability applicants' (2017) 24(4) Journal of the American Medical Informatics Association 709.

[17] Social Security Administration, 'Compassionate Allowances Complete List of Conditions' (2021) ⟨https://www.ssa.gov/compassionateallowances/conditions.htm⟩ accessed 11 June 2021.

[18] Social Security Administration, *Program Operations Manual System* § DI 23022.010 (2018). ('If the claimant alleges a medical condition (by name, synonym, or abbreviation) that is on the CAL list, the selection software identifies the case for CAL processing.')

the claimant's alleged impairments listed on the disability report filed by the claimant.'[18]

There is a long-term project to develop more advanced NLP for SSA's Disability Evaluation Process.[19] This project has been part of a collaboration between the National Institutes and Health ('NIH') and SSA.[20] This NLP would analyse more complex aspects of the disability determination process than those identified via CAL: the determination of the 'residual functional capacity' of claimants who assert that their disability prevents them from taking on work. For example, software may be programmed to code certain language as either indicative or not indicative of residual functional capacity. The example below, an image from the cited article by Desmet and others, illustrates the first steps toward such coding, based on the International Classification of Functioning ('ICF'), a standardized medical vocabulary published by the World Health Organization:

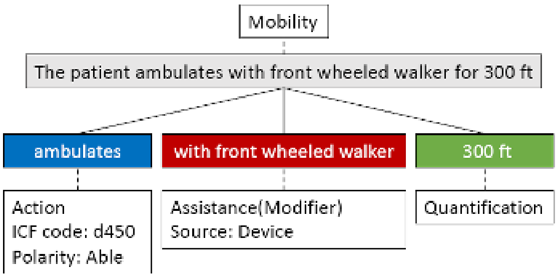

**Figure 1:** *Potential decomposition of language in medical records to facilitate disability evaluation, from the article* 'Development of Natural Language Processing Tools to Support Determination of Federal Disability Benefits in the U.S.'

The language annotation scheme is based on 'four polarity values': 'able, unable, unclear, and none'.[21] These initial binaries are then modified; for example, in Figure 1, the polarity of 'able' connoted by the blue-boxed word "ambulates" is limited by the red-boxed term 'with front wheeled walker', which indicates an inability to walk unaided. The green-boxed term '300 feet' might be thought of as a limitation of the limitation, if the typical person who needs a front wheeled walker to ambulate is not able to do so for, say, more than 200 feet. The green-boxed term may intensify the limiting modification if, say, the typical person who needs a front-wheeled walker to ambulate can do so for 500 feet; then this claimant would be less 'able' than the typical claimant who needs a front wheeled walker to ambulate.

Such language processing could be reduced to quantification, based on benchmarking some set of claims against existing approved and disapproved applications in varied contexts. We recognize that such scoring would be subject to critique and contestation, given the difficulty of comparing multifarious capacities, and we do not endorse it here; we only wish to delineate this possibility in order to provide analysis of its legal implications. By adopting such a quantified polarities-based approach after coding certain strings of words as indicative of either disability or ability, NLP may ultimately result in an aggregation of scores or other quantitative measures of similarity between current claims and past claims. Just as fluency in English or French earns an aspiring immigrant to Canada a certain number of points toward the qualification threshold for residency or citizenship, disability-coded language could earn an applicant points toward qualification for benefits.

For a concrete example, consider the hypothetical possibility that, in a given set of cases, an applicant needs 100 points to achieve a residual functional capacity so low as to qualify for disability benefits. In such cases, the ability to walk (ambulate) counts against an applicant, indicating ability to do some aspect of gainful employment. Unmodified, such an attribution of ability may lead to, say, a deduction of 50 points from an applicant. However, the modification here indicates the applicant needs a walker, and even with that assistance, can only travel 300 feet. This would, for example, eliminate the possibility of working in an Amazon warehouse as a gatherer of merchandise, for such employees routinely walk many kilometres per day.

[19] Bart Desmet and others, 'Development of Natural Language Processing Tools to Support Determination of Federal Disability Benefits in the U.S.' (Doaa Samy, David Pérez-Fernández, and Jerónimo Arenas-García eds, European Language Resources Association May 2020).

[20] Pengsheng Ni and others, 'Development of a Computer-Adaptive Physical Function Instrument for Social Security Administration Disability Determination' (2013) 94(9) Archives of Physical Medicine and Rehabilitation 1661.

[21] Desmet and others (n 19).

In which case, the point deduction may be reduced to, say, only 25 points, given that modification.

The preceding hypothetical is quite easily explicable, but also unmoored from the full panoply of statistics and past cases that would give it the meaning necessary to guide or make a decision. However, the promise (and threat) of machine learning is to use massive data sets to find exceptions, and exceptions to exceptions, in varied cases. It may turn out that, in order to best match the corpus of past training precedents, the system assigns not a 25-point deduction, but instead a 20-point deduction in a set of cases deemed A, and a 30-point deduction in a set of cases deemed B (where A and B are conditions ascertainable from the data collected about the applicant). The exceptions can continue on indefinitely, both at the first level (regarding possible conditions C, D, E, F, *ad infinitum*), and the second (regarding, say, situations A1, A2, A3, A4, *ad infinitum*). Indeed, one way of modelling these models is to think of them as finding local exceptions to general rules. For example, the general theory or rule may be that a person who ambulates is not disabled, but SBNLP will likely find subsets of such persons with an interlocking set of characteristics very similar to those of past, successful applicants.[22] Given its dependence on past data sets and future prediction, such ML may ultimately be closer to historical inquiry and futurology than natural science. Adversely affected users will likely demand an explanation, intuitively sensing that the machine learning system that reached a negative decision in their case was only one of many possible ways of processing the data. Moreover, the public at large will have justified concerns even about successful applications – leading to demands for justification even when the decisions are positive.

## The ChatGPT Solution? Proposals for LLM-Based Opinions and Their Flaws

Reaching adequate explanations of automated decision-making has been a critical concern the ‘Explainable AI’ or ‘XAI’ field.[23] Explaining automated decisions is complex, as they tend to encompass many variables. There are also multiple audiences to consider, including the explainer (the entity who adopted the automated decision and offers the explanation), and the explainee (the person receiving the explanation). Their relative understanding of the technology involved, and the circumstances of the case, may be quite disparate. The AI-driven decision (the object of the explanation) may entail many dimensions of complexity, particularly as more variables are permitted to influence the ultimate outcome.

The result is a recurring demand for flexible and tailored forms (and levels) of explanation.[24] The explainee’s level of understanding, competencies and needs can strongly influence the definition of an adequate explanation. The type of AI that led to a specific answer can strongly influence the comprehensibility of a decision-making process and the level of details of certain explanations. The specific position of the explainer, their interests, and intellectual property claims, can also play a significant role in the design of a meaningful explanation.[25]

### Explanation and Administration

In the administrative justice field, there is a compelling need for explanations to be relevant and accurate. The principle of due or fair process requires adequate explana-

[22] As one journalist explains: ‘The bigger the dataset, the more inconsistencies the AI learns. The end result is not a theory in the traditional sense of a precise claim about [a domain], but a set of claims that is subject to certain constraints. A way to picture it might be as a branching tree of “`if` ... `then`”-type rules, which is difficult to describe mathematically, let alone in words.’ Laura Spinney, ‘Are we witnessing the dawn of post-theory science?’ [2022] The Guardian.

[23] Andrew D Selbst and Solon Barocas, The Intuitive Appeal of Explainable Machines (3126971, 2 March 2018); Ronan Hamon and others, ‘Impossible Explanations? Beyond explainable AI in the GDPR from a COVID-19 use case scenario’ (FAccT ’21, Association for Computing Machinery 1 March 2021); Federico Cabitza and others, ‘Quod erat demonstrandum? - Towards a typology of the concept of explanation for the design of explainable AI’ (2023) 213(PA) Expert Systems with Applications: An International Journal.

[24] For a concrete account of the different ‘levels of explanation’ that may be relevant here, see Frank Pasquale, *Data Access and AI Explainability* (forthcoming 2025).

[25] Tim Miller, ‘Explanation in artificial intelligence: Insights from the social sciences’ (2019) 267 Artificial Intelligence 1; Cabitza and others (n 23).

tion of the decisions that the public administration takes.[26] The intersection between AI explanation and administrative due process has already been explored in the literature.[27] For example, Joe McIntyre and Anna Olijnyk have argued that AI's 'role should never extend to the core business of judicial determinations', and the writing of an explanation for a decision is near the centre of that core.[28] By and large, scholars have assumed that a person would be needed to write (or otherwise express) an explanation for an automated administrative decision, even if the explanation itself were mediated by another automated system's analysis of the facts.[29]

However, the success of generative AI at generating fluent texts can pose an unprecedented challenge to this assumption.[30] Generative AI may produce texts that have all the external qualities of an explanation. Yet generative AI cannot understand the world or the situation which necessitated the decision. Nor can it understand the decision warranting that explanation. Nevertheless, generative AI has already been deployed in at least one authoritative juridical context. In Colombia, a judge used ChatGPT to write part of an opinion in a case involving the fundamental right to health of a minor diagnosed as being on the autism spectrum.[31] This case has already generated controversy, with grave concerns raised about the potential use of such technology in judicial reasoning. Such concerns would be heightened even more in a situation where a chatbot was 'explaining' another automated process, rather than simply suggesting explanatory text for a decision that the judge fully understood.

Nevertheless, some scholars have proposed that there are ways of legitimating AI-written decisions. For example, Eugene Volokh has proposed a 'Modified John Henry Test', evoking the classic competition between a human and steam-powered mechanical shoveler.[32] On Volokh's approach, if an AI program can generate rationales that are indistinguishable from the writing of human judges' (with sufficient expertise in these areas), they can be inserted legitimately into judicial processes to rationalize the decisions made by human judges. Though written before the rise of models like Chat-GPT-3, Volokh's article expertly anticipated them.

Nevertheless, as we learn more about the actual political economy of LLM development, we also gain clarity regarding their limits. The work of a judiciary evolves over time, so it is unclear how long any particular 'Modified John Henry Test' should remain valid. Human judges may need to continually review and validate AI models, potentially undermining efficiency gains. Moreover, the writing of an opinion can lead a judge to modify their own understanding of how the case should be resolved, or at least how the decision should be justified. For example, where the judge's decision rests on a snap judgment and a fuzzy understanding of a statute or precedent, the need to engage directly with primary legal sources may, in turn, clarify their understanding of the scope of action available to the judge and alter their decision or opinion. As automatic legal writing tools become better at drafting analyses supportive of the judge's original position, they become increasingly likely to short-circuit the traditional reflective process that arises when drafting decisions. This may bury opportunities for reflection or doubt which are critical to the decision-making process. This is one reason why the use of AI systems in courts is already considered illegal in

[26] Giacinto della Cananea, 'Administrative Due Process as a General Principle of Public Law' in Giacinto della Cananea (ed), *Due Process of Law Beyond the State: Requirements of Administrative Procedure* (Oxford University Press 22 September 2016).

[27] See, e.g., Margot E Kaminski and Urban Jennifer M, 'The Right to Contest AI' (2021) 121(7) Columbia Law Review 1957; Aziz Z Huq, 'Constitutional Rights in the Machine-Learning State' (2019) 105(7) Cornell Law Review 1875.

[28] McIntyre and Olijnyk (n 7).

[29] Kiel Brennan-Marquez and Stephen E Henderson, 'Artificial Intelligence and Role-Reversible Judgment Criminal Law' (2019) 109(2) Journal of Criminal Law and Criminology 137.

[30] Giovanni De Gregorio, 'The Normative Power of Artificial Intelligence Digital Constitution: On the Transformative Potential of Societal Constitutionalism' (2023) 30(2) Indiana Journal of Global Legal Studies 55.

[31] Juan David Gutiérrez, 'ChatGPT in Colombian Courts' en [2023] Verfassungsblog: On Matters Constitutional.

[32] Volokh (n 5).

some EU Member States,[33] and is considered a 'high risk' in the draft AI Act.[34]

## Explaining Scoring?

In the administrative benefits landscape, integration of Score-Based Natural Language Processing (SBNLP) would shift decision-making toward more efficiency at the cost of narrative explanation. Compared to traditional methods, often burdened by extensive manual documentation reviews, complex and unexplainable SBNLP offers a more streamlined, albeit lawless, approach. It might be best deployed as a way of finding, outside the requirements of law, a set of claims that are similar to the most compelling cases decided favourably in the past. 'Compelling cases' may be defined by some combination of attributes susceptible to scoring. Just as AI might rapidly recognize a deadly aneurysm in the thousands of scans that can swamp a radiology department, SBNLP might speed consideration or approval of the most unambiguously meritorious applications.[35]

This capacity is particularly valuable when there is a vast influx of benefit claims that make manual review cumbersome. This efficiency might allow for better human resource deployment, focusing personnel on cases that demand judgment. At the very least, it may speed benefits to some claimants who are in dire need.[36] But such scoring does raise something of a jurisprudential paradox. We have called it extra-legal and are steadfast in our insistence that a legal process demands personalisation in a literal sense: an authoritative human applying law to facts.[37] Can a legal system effectively remove some decisions to be made extra-legally?

We believe it can do so, based on long-standing theories of emergency as a rationale for suspending otherwise sacrosanct legal requirements.[38] The delay of much-needed assistance to a person or family because of constraints on legal resources is an urgent situation. In many instances, law responds appropriately to such vulnerability. However, sometimes law itself gives way to a more expedient mode of action (such as politics, auctions, markets, or, as here, technology). Traditionalists may be anxious to 'square the circle' here, by attempting to assimilate emergency processes and purely technical decision-making into the rule of law. But the wiser course is to acknowledge that some allocations of benefits will be subject to summary decision-making or automation outside the legal system, rather than distorting our conception of law in order to accommodate them.

The administrative due process principle is rooted in the foundational belief that citizens deserve transparency, fairness, and the ability to challenge adverse decisions. Therefore, automation should never be used to deny benefits

[33] Gianclaudio Malgieri, 'Automated decision-making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations' (2019) 35(5) Computer Law & Security Review 105327.

[34] Michael Veale and Frederik Zuiderveen Borgesius, 'Demystifying the Draft EU Artificial Intelligence Act — Analysing the good, the bad, and the unclear elements of the proposed approach' (2021) 22(4) Computer Law Review International 97. See also Samuel Dahan and others, Lawyers Should Not Trust AI: A call for an Open-source Legal Language Model (4587092, 28 August 2023) (for an emphasis on the current transparency limitations that make it difficult to trust that the underlying data (used to train legal applications built on LLMs) in generalized AI like ChatGPT is actually representative of the state of the law).

[35] For an example of such aneurysm detection, see Christina Jewett, 'Doctors Wrestle With A.I. in Patient Care, Citing Lax Oversight' [2023] The New York Times, ('The image went to Greensboro Radiology, a Radiology Partners practice, where it set off an alert in a stroke-triage A.I. program. A radiologist didn't have to sift through cases ahead of [the patient's] or click through more than 1,000 image slices; the one spotting the brain clot popped up immediately. The radiologist had [the patient] transferred to a larger hospital that could rapidly remove the clot. He woke up feeling normal.')

[36] Frank Pasquale, 'Automated Grace: Toward More Humane Benefits Administration via Artificial Intelligence' (powerpoint on file with authors, University of Melbourne Centre for AI and Digital Ethics, July 2022).

[37] See Guido Noto La Diega, 'Against the Dehumanisation of Decision-Making – Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information' (2018) 9(1) JIPITEC 3 Law 3 para 1 (for a collection of rationales for this position) ;Frank Pasquale, 'The Resilient Fragility of Law' in Simon Deakin and Christopher Markou (eds), *Is Law Computable?: Critical Perspectives on Law and Artificial Intelligence* (Hart Publishing 2020).

[38] See, e.g., William E Scheuerman, 'The Economic State of Emergency Symposium - Carl Schmitt: Legacy and Prospects - An International Conference in New York City: Exception and Emergency Powers' (2000) 21(Issues 5-6) Cardozo Law Review 1869, ('the "motorization of the lawmaker" accurately described by Schmitt is best explained with reference to a compression of time that some contemporary social theorists see as essential to ongoing changes in the capitalist economy').

once an application has crossed a low threshold of plausibility. However, the grant of benefits in a small percentage of cases is hard to contest given its likely negligible impact on public finance, inflation, and the rest of the applications that are to be decided in a traditional manner.[39] It therefore may fairly fall out of the general protection of the rule of law itself, lest such 'protection' entail the harm of those it is intended to help.

Of course, advocates of SBNLP will likely want to expand it beyond a benefit-granting function. SBNLP's allure lies in its ability to expedite the evaluation of a vast array of cases, a boon to administrative efficiency. However, with this rapidity comes a pivotal challenge: ensuring that adverse decisions, now made at an accelerated pace, are accompanied by clear and cogent explanations. Such explanations are vital not only for upholding the integrity of the process, but also for allowing citizens to understand and, if necessary, contest a potential harm done to their interests. Enter generative AI once again, likely to be presented as a potential solution to this quandary. With its capacity to produce detailed simulations of justifications, generative AI appears to be an ideal tool to bridge the gap between the swiftness of SBNLP and the due process mandate for comprehensible explanations of adverse decisions.

However, diving deeper, we encounter murky waters. While the explanations provided by generative AI might check the boxes of formality–likely being consistent, comprehensive, and comprehensible in some prospective advance beyond ChatGPT-4 and similar models–they carry inherent risks that cannot be ignored. So-called hallucinations (less anthropomorphically styled 'fabrications') abound in LLM-generated content.[40] Bias concerns have been well-documented, and will persist as long as the industry remains as opaque as it is today.

But even if these concerns about fabrications and bias could be addressed, the lack of authenticity of these explanations is an insuperable normative hurdle. They might be artfully crafted and technically sound, but there is a necessary disconnect between the rationale provided and the true underpinnings of the decision, when the rationale is the result of mere next-token prediction. Unmoored from direct observation and empathy, such explanations would be misleading post hoc constructs, designed to fit the outcome rather than shape and reveal the genuine reasoning that led to it.[41] 'Result-based reasoning' is a formidable epithet in law for a reason. 'Simulation of reasoning' is even worse.

Artificially constructed justifications, though appearing robust on the surface, may not be grounded in factual accuracy.[42] The mode of action of LLMs is next word prediction, not reasoned understanding of the world, or normative evaluation of situations.[43] This misalignment poses significant challenges to the ethos of administrative due process. If citizens receive explanations that, while polished, are not rooted in the actual decision-making process, their ability to receive truly reasoned rationales for a decision has been compromised. Rather than arising organically out of reasoning, the 'on-demand' explanation ornaments a decision that may well have been taken on entirely different grounds. This could erode trust in the administrative system, leading citizens to view these justifications (and perhaps even justification in general) with scepticism, if not outright disbelief. As Rob Horning has observed regarding LLM outputs:

> "Automation [often] deskills the tasks [persons] are required to perform, making them more rote, depleting, and mind-numbing. There is no reason to suppose that generative AI will do something

[39] To be sure, if there are negative collateral consequences of such an award of benefits, the decision may be termed not entirely positive, and in that way inappropriate for SBNLP. On collateral consequences, see Michael Pinard, Collateral Consequences of Criminal Convictions: Confronting Issues of Race and Dignity (1435320, 2010).

[40] Matthew Dahl and others, Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models (2024); Sayash Kapoor, Peter Henderson, and Arvind Narayanan, Promises and Pitfalls of Artificial Intelligence for Legal Applications (4695412, 15 January 2024).

[41] Competition and Markets Authority (CMA) of the United Kingdom, *AI Foundation Models: Initial report* (2023) (p 81, p 82).

[42] Minderoo Centre for Technology and Democracy, *Policy Brief: Generative AI* (report, Minderoo Centre for Technology and Democracy 2023). See also Johanna Okerlund and others, *What's in the Chatterbox? Large Language Models, Why They Matter, and What We Should Do About Them* (report, University of Michigan Science, Technology, and Public Policy Program 2022). See Annex III, point 8 of the EU AI Act.

[43] On the lack of meaning and intentionality in texts generated by generative AI, see Dan L Burk, Asemic Defamation, or, the Death of the AI Speaker (4667410, 18 December 2023).

> different to language-oriented tasks. It will make them less meaningful to us as we have to do more of them. (Think of the piece workers clicking yes or no on an endless series of decontextualized language fragments to train tomorrow's AIs.) It will inculcate people with the idea that language use itself — the effort to communicate at all — is a hassle, something more and more difficult to initiate with any expectation of good faith, given that so much more of the language we encounter will have been generated to stupefy and deceive us. (The preponderance of advertising may already have accomplished this.) Of course if you don't care about speaking in good faith, AI will be very helpful to you."[44]

Questions of power and meaning are paramount here: the power of the state to effectively steamroll the claims of its subjects, without investing in the personnel necessary to fully comprehend the nature, impact, and consequences of such rationalizations. This further parallels remarkable levels of disregard for labour, history, and citation at the core of contemporary commercial generative AI. As Eryk Salvaggio has observed:

> "The imaginary worlds of generative AI feel bleaker for me every day. A surrealism without a subconscious, rendered with the aesthetic predictability of its training data: advertisements and clip art fused with atrocity footage and family snapshots. All of the images are extensions of the visual mélange, hypothetical images based on all images prior. Paired with a sense that the origins do not matter, that labor does not matter, that any obligation to citation or history do not matter."[45]

These infirmities (with respect to identifying and crediting labour, robustly citing sources closest to the generated output, and releasing audits of data origins and treatment) would need to be corrected before policymakers could even begin to contemplate the use of generative AI in automatically rationalising SBNLP-based decisions. But even with such reforms, reliance on generative AI for explanations risks creating a veneer of transparency without substance. These generated explanations, no matter how comprehensive, could act as a smokescreen, obfuscating the real workings and potential biases of the SBNLP process. Thus, while technically generating the type of artifact (a series of comprehensible and relevant words) required for verification of explanations, the process might still violate the spirit of administrative due process.

Reflecting such concerns, the EU AI Act has classified as 'high risk' the AI systems 'intended to be used by public authorities or on behalf of public authorities to evaluate the eligibility of natural persons for public assistance benefits and services, as well as to grant, reduce, revoke, or reclaim such benefits and services'.[46] High-risk systems are obliged to follow some requirements in terms of design, data governance, and risk management that could be very beneficial in this case. However, it is important to remember that this provision would apply only to SBNLP itself, but not to generative AI–driven explanations of SBNLP decisions. Additional strictures should be proposed now to govern the use of generative AI by decisionmakers.

In the administrative benefits landscape, SBNLP offers promising efficiencies while generative AI seems a tempting solution for providing requisite explanations. Nonetheless, strict limits on their use will be necessary. The true essence of administrative due process–genuine transparency, fairness and accountability–must remain at the forefront of any integration of such technology into benefits management.

[44] Rob Horning, 'Have you heard the word' (21 December 2023) ⟨https://robhorning.substack.com/p/have-you-heard-the-word⟩ accessed 28 December 2023.

[45] Eryk Salvaggio, 'The Hypothetical Image' (29 October 2023) ⟨https://www.cyberneticforests.com/news/social-diffusion-amp-the-seance-of-the-digital-archive⟩ accessed 28 December 2023.

[46] See Annex III, point 8 of the most updated text of the EU AI Act, "Corrigendum to the position of the European Parliament adopted at first reading on 13 March 2024 with a view to the adoption of Regulation (EU) 2024/... of the European Parliament and of the Council laying down harmonised rules on artificial intelligence", P9_TA(2024)0138.

## Conclusion: SBNLP to Identify the Most Meritorious Claims, Parallel or External to Justification-Based Legal Proceedings

In light of the challenges posed by introducing generative AI-driven explanations of SBNLP processes, it may be prudent to consider an exemption to the principle of individual explanation for positive SBNLP decisions. Instead of mandating artificial simulations of justifications, emphasis could shift to broader, more systemic accountability, fairness, and transparency measures. Regular audits can ensure the SBNLP algorithms function as intended, without bias. Periodical impact assessments can gauge the real-world ramifications and fairness of decisions.[47] A comprehensive justification statement detailing the structural functioning of SBNLP can provide a clear overview of its operations and methodologies.[48] By integrating these measures, we can uphold the spirit of administrative due process while harnessing the efficiencies of SBNLP to speed benefits to exceptionally deserving applicants, thereby striking a harmonious balance between innovation and justice while recognizing the distinctness and integrity of each.

## Acknowledgements

We wish to thank Lyria Bennett Moses for her excellent comments on a draft presented at the conference 'Computational "Law" on Edge', as well as Laurence Diver and Pauline McBride for their expert research and organizational work at the affiliated COHUBICOL project. We also thank Mireille Hildebrandt for her intellectual leadership at the conference and in so much of computational law.

## References

Abbott K, Ho Y.-Y, and Erickson J, 'Automatic health record review to help prioritize gravely ill Social Security disability applicants' (2017) 24(4) Journal of the American Medical Informatics Association 709.

Andrejevic M, *Automated Media* (Routledge 14 October 2019).

Brennan-Marquez K and Henderson SE, 'Artificial Intelligence and Role-Reversible Judgment Criminal Law' (2019) 109(2) Journal of Criminal Law and Criminology 137.

Burk DL, Asemic Defamation, or, the Death of the AI Speaker (4667410, 18 December 2023).

Cabitza F and others, 'Quod erat demonstrandum? - Towards a typology of the concept of explanation for the design of explainable AI' (2023) 213(PA) Expert Systems with Applications: An International Journal.

Cananea G della, 'Administrative Due Process as a General Principle of Public Law' in G della Cananea (ed), *Due Process of Law Beyond the State: Requirements of Administrative Procedure* (Oxford University Press 22 September 2016).

Competition and Markets Authority (CMA) of the United Kingdom, *AI Foundation Models: Initial report* (2023).

Dahan S and others, Lawyers Should Not Trust AI: A call for an Open-source Legal Language Model (4587092, 28 August 2023).

Dahl M and others, Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models (2024).

De Gregorio G, 'The Normative Power of Artificial Intelligence Digital Constitution: On the Transformative

[47] Heleen L Janssen, 'An approach for a fundamental rights impact assessment to automated decision-making' (2020) 10(1) International Data Privacy Law 76; Alessandro Mantelero, *Beyond data: Human rights, ethical and social impact assessment in AI* (Springer Nature 2022); Alessandro Mantelero and Samantha Esposito, An Evidence-Based Methodology for Human Rights Impact Assessment (HRIA) in the Development of AI Data-Intensive Systems (3829759, 22 March 2021); Atoosa Kasirzadeh and Damian Clifford, 'Fairness and Data Protection Impact Assessments' (AIES '21, Association for Computing Machinery 2021); Margot E Kaminski and Gianclaudio Malgieri, 'Algorithmic impact assessments under the GDPR: producing multi-layered explanations' (2021) 11(2) International Data Privacy Law 125.

[48] Gianclaudio Malgieri, '"Just" Algorithms: Justification (Beyond Explanation) of Automated Decisions Under the General Data Protection Regulation' (2021) 1(1) Law and Business 16; Gianclaudio Malgieri and Frank Pasquale, Licensing High-Risk AI: Towards Ex Ante Justification of a Disruptive Technology (4346120, 1 July 2023).

Potential of Societal Constitutionalism' (2023) 30(2) Indiana Journal of Global Legal Studies 55.

Desmet B and others, 'Development of Natural Language Processing Tools to Support Determination of Federal Disability Benefits in the U.S.' (Samy D, Pérez-Fernández D, and Arenas-García J eds, European Language Resources Association May 2020).

Engstrom DF and others, 'Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies' [2020] SSRN Electronic Journal.

Eubanks V, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (St Martin's Publishing Group 2018).

Gutiérrez JD, 'ChatGPT in Colombian Courts' en [2023] Verfassungsblog: On Matters Constitutional.

Hamon R and others, 'Impossible Explanations? Beyond explainable AI in the GDPR from a COVID-19 use case scenario' (FAccT '21, Association for Computing Machinery 1 March 2021).

Han BC, *The Crisis of Narration* (Steuer D tr, Polity Press 2024).

Horning R, 'Have you heard the word' (21 December 2023) ⟨https://robhorning.substack.com/p/have-you-heard-the-word⟩ accessed 28 December 2023.

Huq AZ, 'Constitutional Rights in the Machine-Learning State' (2019) 105(7) Cornell Law Review 1875.

Janssen HL, 'An approach for a fundamental rights impact assessment to automated decision-making' (2020) 10(1) International Data Privacy Law 76.

Jewett C, 'Doctors Wrestle With A.I. in Patient Care, Citing Lax Oversight' [2023] The New York Times.

Kaminski ME and Malgieri G, 'Algorithmic impact assessments under the GDPR: producing multi-layered explanations' (2021) 11(2) International Data Privacy Law 125.

Kaminski ME and Urban Jennifer M, 'The Right to Contest AI' (2021) 121(7) Columbia Law Review 1957.

Kapoor S, Henderson P, and Narayanan A, Promises and Pitfalls of Artificial Intelligence for Legal Applications (4695412, 15 January 2024).

Kasirzadeh A and Clifford D, 'Fairness and Data Protection Impact Assessments' (AIES '21, Association for Computing Machinery 2021).

Malgieri G, 'Automated decision-making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations' (2019) 35(5) Computer Law & Security Review 105327.

— '"Just" Algorithms: Justification (Beyond Explanation) of Automated Decisions Under the General Data Protection Regulation' (2021) 1(1) Law and Business 16.

Malgieri G and Pasquale F, Licensing High-Risk AI: Towards Ex Ante Justification of a Disruptive Technology (4346120, 1 July 2023).

Mantelero A, *Beyond data: Human rights, ethical and social impact assessment in AI* (Springer Nature 2022).

Mantelero A and Esposito S, An Evidence-Based Methodology for Human Rights Impact Assessment (HRIA) in the Development of AI Data-Intensive Systems (3829759, 22 March 2021).

McIntyre J and Olijnyk A, 'Public Law Limits on Automated Courts' in K Miller and J Boughey (eds), *The Automated State - Implications, Challenges and Opportunities for Public Law* (The Federation Press 17 June 2021).

Miller T, 'Explanation in artificial intelligence: Insights from the social sciences' (2019) 267 Artificial Intelligence 1.

Minderoo Centre for Technology and Democracy, *Policy Brief: Generative AI* (report, Minderoo Centre for Technology and Democracy 2023).

Ni P and others, 'Development of a Computer-Adaptive Physical Function Instrument for Social Security Administration Disability Determination' (2013) 94(9) Archives of Physical Medicine and Rehabilitation 1661.

Noto La Diega G, 'Against the Dehumanisation of Decision-Making – Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information' (2018) 9(1) JIPITEC 3.

Okerlund J and others, *What's in the Chatterbox? Large Language Models, Why They Matter, and What We Should Do About Them* (report, University of Michigan Science, Technology, and Public Policy Program 2022).

Pasquale F, 'The Resilient Fragility of Law' in S Deakin and C Markou (eds), *Is Law Computable?: Critical Perspectives on Law and Artificial Intelligence* (Hart Publishing 2020).

— 'Automated Grace: Toward More Humane Benefits Administration via Artificial Intelligence' (powerpoint on file with authors, University of Melbourne Centre for AI and Digital Ethics, July 2022).

Pasquale F, *Data Access and AI Explainability* (forthcoming 2025).

Pasquale F and Cashwell G, 'Prediction, persuasion, and the jurisprudence of behaviourism' (2018) 68(supplement 1) University of Toronto Law Journal 63.

Pinard M, Collateral Consequences of Criminal Convictions: Confronting Issues of Race and Dignity (1435320, 2010).

Rajnes D, '"Fast-track" strategies in long-term public disability programs around the world' (2012) 72(1) Social Security Bulletin 79.

Ranchordás S, 'Empathy in the Digital Administrative State' (2022) 71(6) Duke Law Journal 1341.

Rasch EK and others, 'First in Line: Prioritizing Receipt of Social Security Disability Benefits Based on Likelihood of Death During Adjudication' (2014) 52(11) Medical Care 944.

Salvaggio E, 'The Hypothetical Image' (29 October 2023) ⟨https://www.cyberneticforests.com/news/social-diffusion-amp-the-seance-of-the-digital-archive⟩ accessed 28 December 2023.

Scheuerman WE, 'The Economic State of Emergency Symposium - Carl Schmitt: Legacy and Prospects - An International Conference in New York City: Exception and Emergency Powers' (2000) 21(Issues 5-6) Cardozo Law Review 1869.

Selbst AD and Barocas S, The Intuitive Appeal of Explainable Machines (3126971, 2 March 2018).

Social Security Administration, 'SSA - POMS: DI 11005.604 - Processing Compassionate Allowances (CAL) in the Field Office (FO) - 05/24/2023' (2015) ⟨https://secure.ssa.gov/poms.nsf/lnx/0411005604⟩ accessed 28 December 2023.

— 'Compassionate Allowances Complete List of Conditions' (2021) ⟨https://www.ssa.gov/compassionateallowances/conditions.htm⟩ accessed 11 June 2021.

Spinney L, 'Are we witnessing the dawn of post-theory science?' [2022] The Guardian.

United States Social Security Administration, *Annual Statistical Report on the Social Security Disability Insurance Program, 2017* (2018).

Vanderstichele G, The Normative Value of Legal Analytics. Is There a Case for Statistical Precedent? (3474878, 30 August 2019).

Veale M and Borgesius FZ, 'Demystifying the Draft EU Artificial Intelligence Act — Analysing the good, the bad, and the unclear elements of the proposed approach' (2021) 22(4) Computer Law Review International 97.

Volokh E, 'Chief Justice Robots' (2019) 68(6) Duke Law Journal 1135.