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

Frank Pasquale and 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 set of claims as particularly like known, meritorious claims, without understanding the meaning of any of these legal texts. Score-based natural language processing (SBNLP) may expand the range of claims that may be categorized as urgent and obvious, but its practitioners may not be able to offer a narratively intelligible rationale for why it does so. However, practitioners may now use generative AI to attempt to fill this explanatory gap, offering a rationale for decision that is a plausible imitation of past, humanly-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 adverse 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 based on large language models, there is a good case for entirely excluding these automated outputs in administrative and judicial decision-making settings. Nevertheless, SBNLP may be established parallel to or external to justification-based legal proceedings, for humanitarian purposes.

Generative AI, Explainability, and Experiments in Benefits Administration

Frank Pasquale and Gianclaudio Malgieri

Outline

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

Frank Pasquale and Gianclaudio Malgieri

For a casual observer of the legal system, the adoption of legal technology (legaltech) might appear to be an automatic force for fairness and access to justice. The American Bar Association regularly celebrates "legal rebels" who are shaking up the industry with apps and software. The entrepreneur Joshua Browder has claimed that, "By moving the justice system to a software-first approach, we can improve transparency, automate tedious processes, and in some cases, even avoid the need for expensive lawyers altogether."[1] Some claim that big data and law-as-code may even resolve the classic tension between rules and standards, with AI-driven "microdirectives" that apply the right rule, in the right way, to the right persons, automatically, potentially unleashing a "legal singularity."[2]

However, the scholarly response to these claims has been largely skeptical when it comes to automation of state functions. As Danielle Citron and Ryan Calo have recently argued, the many failures of initiatives to automate even basic processes of the administrative state have culminated in a "crisis of legitimacy" for such projects.[3] Even worse, there appears to be a disparate negative impact of automation on many disadvantaged groups. While simple processes, like the passport control checks of the Transportation Safety Administration's "Global Entry" program, have accelerated some basic conveniences for the relatively wealthy, other forms of computation have put whole communities under suspicion.[4] Virginia Eubanks has cataloged case study after case study of claimants denied or delayed benefits by either incompetently or malevolently designed automated systems embedded in torpid bureaucracies.[5] In one of her most moving studies, a

[1] Joshua Browder, *Law as Code*, at https://future.a16z.com/law-as-code/; *but see* Jacob Silverman, *Angry Users Want DoNotPay to Pay Up: Troubles Accumulate for the Robot Lawyer That's Not a Robot Or a Lawyer*, https://www.jacobsilverman.com/p/angry-users-want-donotpay-to-pay (2023).

[2] Casey & Niblet, *Death of Rules & Standards*; Abdi Aidid and Benjamin Alarie, The Legal Singularity: How Artificial Intelligence Can Make Law Radically Better (2023).

[3] Danielle K. Citron and Ryan Calo, The Automated Administrative State: A Crisis of Legitimacy, Emory L.J. (2021).

[4] Rashida Richardson, Jason M. Schultz, and Kate Crawford, Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice, 94 N.Y.U. L. Rev. Online (2019); Ferguson, *The Rise of Big Data Policing*; Elizabeth E. Joh, *Private Security Robots, Artificial Intelligence, and Deadly Force*, 51 U.C. Davis L. Rev. 569 (2021).

[5] Virginia Eubanks, *Automating Inequality* (2018).

woman battles a highly automated state Medicaid office for months to obtain coverage, only to die on the day she wins her case.

At this point, an emerging scholarly consensus has largely refuted the assumption that legaltech *must* be an emancipatory, egalitarian project. There are too many examples of biased data, incompetently coded directives, and malfunctioning software, to believe in that any longer. Nevertheless, the search for an emancipatory legaltech agenda continues. Advancing AI to accelerate legal determinations that help the disadvantaged is one way for legaltech to atone for its past record of "automating inequality."[6] For example, the Social Security Administration has developed "Compassionate Allowance" and "Quick Disability Determination" processes to speed benefits to persons with particularly urgent and obvious claims. These processes quickly identify meritorious claims that might have been unduly delayed.

Scoring methods may also expand the range of claims that may be categorized as urgent and obvious.[7] For example, if nearly all claimants with six serious diseases were awarded benefits in the past, a point system (that grants benefits to someone with over 110 points) might simply assign that six-disease combination as meriting 115 points. Alternatively, such a system may assign 20 points per serious disease. In either situation, the scoring system would help accelerate claims that were unduly delayed in the past. The promise of machine learning is that more complex versions of such scoring could identify combinations of factors that always led to awards in the past.[8] We call this hypothetical approach to processing benefits application "score-based natural language processing" (SBNLP), and predict it will become an increasingly tempting expedient wherever decisionmakers confront staffing limitations and a great deal of applications to decide on.

From a rule of law perspective, though, there may be a key problem with such a process: a lack of explainability. A sufficiently complex scoring system will not demonstrate how, say, a given combination of features gave rise to a finding of disability. It only identifies *that* it has done

[6] Virginia Eubanks, *Automating Inequality* (2018).

[7] Desmet et al., *Development of Natural Language Processing Tools to Support Determination of Federal Disability Benefits in the U.S.*, LT4Gov 1 (May 2020); Supplemental Security Income Program - FY 2021 Congressional Justification, 43 (2021).

[8] Indeed, many machine learning approaches would not assign any numerical value at all, assessing similarlity 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.

so, albeit in a mathematical way. The normative value of such legal analytics is uncertain here.[9] 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.

The rise of generative AI provides new hope, though, for an entirely automated decisionmaking process. Just as SBNLP can derive a score from a given set of filings, a chatbot based on an LLM 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 rationalize the claimant's success, or highlighting factual dimensions of the present case that are similar to factual aspects in past precedents. Eugene Volokh proposed that such a system could replace appellate judges, and the rise of legal applications of ChatGPT and similar applications has given new relevance to his proposal.[10]

This article explores whether such an LLM-based opinion writer would be a valuable adjunct to SBNLP of benefits claims. Part II begins with a description of proposals for automating eligibility for and receipt of benefits. The U.S. Social Security Administration has developed some programs of computational administration of benefits. Other automatic or near-automatic benefits are also a part of other federal agencies, as well as some state agencies. Simple word-matching algorithms and document authentication software may eventually lead to more advanced forms of natural language processing, including 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 decisionmaking authority to such scoring, given that it is already at one remove from the language-based application of rules to fact patterns. To the extent it is sufficiently non-monotonic, it is also likely to lack convincing narrative explanation as well.

Part III explains how chatbots based on large language models may offer some apparent solutions to this problem (IIIA), and why these approaches should be rejected (IIIB). Given a large enough data set of past cases with authoritative opinions, a chatbot may be fine-tuned with specific legal training in order to produce justifications based on the fact patterns and relevant legal

[9] Genevieve Vanderstichele, *The Normative Value of Legal Analytics. Is There a Case for Statistical Precedent?*, at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3474878
[10] Eugene Volokh, 'Chief Justice Robots' 68 Duke Law Journal 1135.

precedents of cases that have been decided via score based natural language processing. A complementary program may also suggest meaningful information about the scoring process itself. However, this would not be is sufficient explanation for a decision ultimately made on other grounds. Moreover, several legal commentators have cautioned against outsourcing core judicial functions to automated processes. The justification of an outcome is just as important to judicial legitimacy as the outcome itself.

Part IV reflects on where the discussion in Part III leaves score based natural language processing. It is important to note that such a procedure could lead to much more rapid allocation of benefits if deployed to find patterns that match past beneficiary claims that have been deemed the neediest and most meritorious cases. Therefore, it would be wise to suspend verbal justification requirements for those cases, rather than being forced into the dilemma of either banishing score based natural language processing because of 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 within them warping their integrity.

II. From Computational Decision Support to Score-Based Natural Language Processing

Legal automation is advancing in many practice areas. In private practice, coders and lawyers are working together to promote automatic contracting, monitoring, and dispute resolution.[11] 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 landmark report published in 2019, law professors identified numerous opportunities for automation of several dimensions of the administrative state.[12] They identified scenarios of "mass justice" as being particularly promising targets for AI. And justice does not get more "mass" than the adjudication of Social Security Disability Insurance (SSDI) applications, which now number over two million per year.

[11] Firms like DoNotPay market chatbots to solve pressing problems of access to justice, providing users with easy to fill and file forms for an array of simple claims. Venture capital in the "legaltech" space is betting on numerous "turnkey solutions" to discovery, venue selection, and even legal search and argumentation.

[12] Daniel Engstrom et al., "Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies," at https://www-cdn.law.stanford.edu/wp-content/uploads/2020/02/ACUS-AI-Report.pdf.

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.[13] For purposes of this article, the key determination includes the following steps:

> (i) At the first step, we consider your work activity, if any. If you are doing substantial gainful activity, we will find that you are not disabled. …(ii) At the second step, we 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 in § 404.1509, 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.[14]

Multiple layers of appeal mean that a significant percentage of claimants will wait months or even years for benefits they are 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.[15] This means that over

[13] 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.").

[14] 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. An individual's statement as to pain or other symptoms shall not alone be conclusive evidence of disability as defined in this section; there must be medical signs and findings, established by medically acceptable clinical or laboratory diagnostic techniques, which show the existence of a medical impairment that results from anatomical, physiological, or psychological abnormalities which could reasonably be expected to produce the pain or other symptoms alleged and which, when considered with all evidence required to be furnished under this paragraph (including statements of the individual or his physician as to the intensity and persistence of such pain or other symptoms which may reasonably be accepted as consistent with the medical signs and findings), would lead to a conclusion that the individual is under a disability. Objective medical evidence of pain or other symptoms established by medically acceptable clinical or laboratory techniques (for example, deteriorating nerve or muscle tissue) must be considered in reaching a conclusion as to whether the individual is under a disability.").

[15] SOC. SEC. ADMIN., ANNUAL STATISTICAL REPORT ON THE SOCIAL SECURITY DISABILITY PROGRAM, 2017 158–164 tbls.60–63 (2018), https://www.ssa.gov/policy/docs/statcomps/di_asr/2017/di_asr17.pdf. As ABT Associates note, "Totals and percentages reflect national data for disabled workers with an initial application in 1998." DANIEL GUBITS, SARAH PRENOVITZ, CARA SIERKS, & ZACHARY EPSTEIN, ABT ASSOCIATES, CLAIMANT REPRESENTATIVE

200,000 applicants were likely to have been delayed in accessing benefits they were due, some by many months or years.

The third step, based on "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."[16] Therefore, a finding of a listed impairment ends the disability determination at the third step, before the fact-intensive process of determining whether a putatively disabled person could take on some kind 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.[17]

The SSA has recognized the importance of identifying the neediest cases—the most vulnerable--among those who are likely to be found to have a listed impairment. This is particularly pressing because so many beneficiaries die while waiting for their claim to be processed.[18] A Compassionate Allowance (CAL) initiative identifies a subset of the most pressing listed impairments.[19] 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."[20] There is a specific list of CAL conditions.[21] Once deployed, the "CAL selection software identifies cases for CAL processing based solely on the claimant's alleged impairments listed on" the disability report filed by the claimant."[22]

DEMONSTRATION TECHNICAL EXPERTS MEETING: FINAL REPORT (2019), https://www.ssa.gov/disabilityresearch/documents/Claimant_Representative_Demo_TEP_report_508a%20final_Final.pdf.

[16] 20 C.F.R. § 404.1525(a) (2017); 20 C.F.R. § 404.1520(d) (2012).

[17] David Rajnes, *"Fast-Track" Strategies in Long-Term Public Disability Programs Around the World*, SOCIAL SECURITY BULLETIN, Feb. 2012.

[18] Rasch EK, Huynh M, Ho PS, et al.. 'First in line: prioritizing receipt of Social Security disability benefits based on likelihood of death during adjudication,' *Med. Care.* 2014;52(11):944–50 (estimating the deaths of waiting recipients in the thousands).

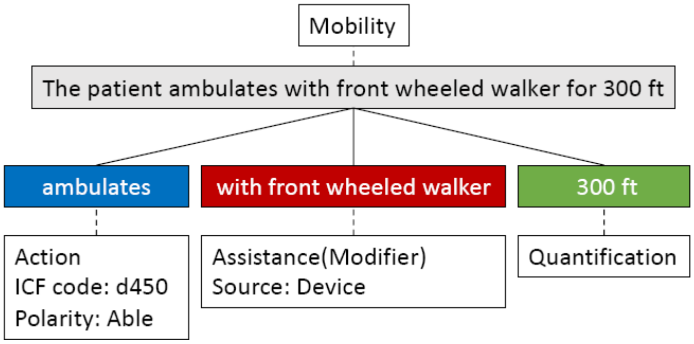
[19] Social Security Administration. Processing Compassionate Allowances (CAL) in the Field Office (FO). Baltimore, MD: Social Security Administration (https://secure.ssa.gov/poms.nsf/lnx/0411005604 (2015).

[20] Kenneth Abbott, Yen-Yi Ho, and Jennifer Erickson, 'Automatic health record review to help prioritize gravely ill Social Security disability applicants,' *Journal of the American Medical Informatics Association*, 24(4), 2017, 709–716 doi: 10.1093/jamia/ocw159.

[21] *Compassionate Allowances Conditions*, https://www.ssa.gov/compassionateallowances/conditions.htm (last visited Jun. 11, 2021).

[22] 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.")

There is a long-term project to develop even more advanced NLP for SSA's Disability Evaluation Process.[23] This project has been part of a collaboration between the National Institute for Health ("NIH") and SSA, ongoing since at least 2013.[24] This NLP would take on even the more complex aspects of the disability determination process than CAL, TERI, and XR: the determination of the "residual functional capacity" of claimants who do not have a listed impairment, but still claim 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 disability. The example below, an image from the cited article by Desmet *et al.*, illustrates how such coding may work, based on the International Classification of Functioning ("ICF"), a standardized medical vocabulary published by the World Health Organization:



As Desmet et al. explain, "The four polarity values in our annotation schema are able, unable, unclear, and none."[25] By coding certain strings of words as either indicative of disability or ability, as above, their NLP may ultimately aggregate scores or other quantitative measures of similarity between current claims and past claims. This is already done in immigration contexts, where, for example, being able to speak both English and French gives an aspiring emigrant a certain number of points toward the quantity of points necessary to qualify to immigrate.

[23] Desmet et al., *Development of Natural Language Processing Tools to Support Determination of Federal Disability Benefits in the U.S.*, LT4Gov 1 (May 2020); Supplemental Security Income Program - FY 2021 Congressional Justification, 43 of PDF (2021).
[24] *See* Pengsheng et al., *Development of a Computer-Adaptive Physical Function Instrument for Social Security Administration Disability Determination*, Archives of Physical Medicine and Rehabilitation (2013), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4017369/.
[25] Desmet et al., *Development of Natural Language Processing Tools to Support Determination of Federal Disability Benefits in the U.S.*, LT4Gov 1 (May 2020).

For example, stipulate that, in a given set of cases, one needs 100 points to achieve disabled status. In the chart above, 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 typically lead to 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 kilometers per day. So, the point deduction may be reduced to, say, only 25 points, given that modification.

The hypothetical just described is quite easily explicable. 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 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. Indeed, one way of modeling these models, so to speak, 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 may be able to find a subset of such persons with an interlocking set of characteristics very similar to those of past, successful applicants. 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."[26] Given its dependence on past data sets and future prediction, such machine learning may ultimately be closer to historical inquiry and futurology than natural science. And adversely affected users will likely demand an explanation for it, 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.

[26] Laura Spinney, *Are we Witnessing the Dawn of Post-Theory Science?*, Guardian, Jan. 9, 2022.

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

A. Explanation and Administration

Reaching adequate explanations has been one of the main efforts of socio-technical research behind AI (the so-called "Explainable AI" or "XAI" field).[27] 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.[28] 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.[29]

In the administrative justice field, there is a compelling need for explanations to be relevant and accurate. Adequate explanation of the decisions that the public administration takes (e.g., in the field of public welfare and benefit allocations) is required by the principle of due or fair process (across different legal systems).[30] The intersection between AI explanation and administrative due

[27] Andrew D Selbst and Solon Barocas, 'The Intuitive Appeal of Explainable Machines' (2018) 87 Fordham Law Review 1085; Marzyeh Ghassemi, Luke Oakden-Rayner and Andrew L Beam, 'The False Hope of Current Approaches to Explainable Artificial Intelligence in Health Care' (2021) 3 The Lancet Digital Health e745; Ronan Hamon and others, 'Impossible Explanations? Beyond Explainable AI in the GDPR from a COVID-19 Use Case Scenario', *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2021) <https://doi.org/10.1145/3442188.3445917> accessed 27 May 2021; Federico Cabitza *et al.*, 'Quod Erat Demonstrandum? - Towards a Typology of the Concept of Explanation for the Design of Explainable AI' (2023) 213 Expert Systems with Applications 118888.

[28] For a concrete account of the different "levels of explanation" that may be relevant here, see Frank Pasquale, *Data Access and Explainability* (forthcoming, 2024).

[29] Tim Miller, 'Explanation in Artificial Intelligence: Insights from the Social Sciences' (2019) 267 Artificial Intelligence 1; Cabitza and others (n 27).

[30] 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 2016) <https://doi.org/10.1093/acprof:oso/9780198788386.003.0009> accessed 28 October 2023.

process has already been explored in the literature.[31] 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 were itself mediated by another automated system's analysis of the administrative automation.[32] 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 would seem to be near the center of that core.[33]

However, the success of Generative AI (GenAI) at generating fluent texts can pose an unprecedented challenge to this discussion.[34] GenAI may produce texts that have all the external qualities of an explanation, even though GenAI itself does not understand the world in which the situation which necessitated the explanation occurred. It 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. This case has already generated a discussion in the legal field and public opinion, with grave concerns being raised about the potential use of such technology in juridical explanation-giving.[35] Presumably these concerns would be heightened even more in a situation where a chatbot was "explaining" another automated process, rather than simply suggesting text for a rationale for a decision that the judge fully understood.

Past research should also inform the current legal debate on the benefits and risks of using AI to write opinions. 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 shoveler.[36] On Volokh's approach, if an AI program can generate rationales that are indistinguishable from human judges' writing (based

[31] See, e.g., Margot E Kaminski and Jennifer M Urban, 'The Right to Contest AI' (2021) 121 Columbia Law Review 1957; Aziz Huq, 'Constitutional Rights in the Machine Learning State' (2020) 105 Cornell Law Review 1875.

[32] Kiel Brennan-Marquez and Stephen Henderson, 'Role Reversible Judgment.'

[33] Joe McIntyre and Anna Olijnyk, 'Public Law Limits on Automated Courts' in Janina Boughey and Katie Miller (eds) The Automated State: Implications, Challenges and Opportunities for Public Law (Federation Press, 2021) 89, 89.

[34] Giovanni De Gregorio, 'The Normative Power of Artificial Intelligence' (2023) 30 Indiana Journal of Global Legal Studies 55.

[35] Juan David Gutiérrez, 'ChatGPT in Colombian Courts: Why we need to have a conversation about the digital literacy of the judiciary' [2023] Verfassungsblog <https://verfassungsblog.de/colombian-chatgpt/> accessed 28 October 2023.

[36] Volokh (n 10).

on the judgment of expert human judges), 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, there are some important challenges to its reasoning. 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 come out, or at least how the opinion should be written. For example, if the judge's snap judgment of the case rested in part on a fuzzy understanding of a statute or precedent, the clarified understanding of the scope of action available to the judge based on direct reading of the statute or precedent may lead the judge to alter their decision or opinion. The better an automatic legal writing tool becomes, the more likely it is to short-circuit such a reflective process by burying whatever doubts and reflection that might have arisen out of the writing process, in analysis supportive of the judge's original position. This is one reason why the use of AI systems in courts is already considered illegal in some EU Member States,[37] and is considered at "high risk" in the draft AI Act.[38]

B. The Special Case of Explaining Scoring

In the administrative benefits landscape, the potential integration of Score-Based Natural Language Processing (SBNLP) signals another potential shift towards efficiency at the cost of narrative explanation. Compared to traditional methods, often burdened by extensive manual documentation reviews, 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 favorably in the past, where "compelling" is defined by some

[37] Gianclaudio Malgieri, 'Automated Decision-Making in the EU Member States: The Right to Explanation and Other "Suitable Safeguards" in the National Legislations' [2019] Computer Law & Security Review 105327.

[38] 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 Computer Law Review International 97. See also Samuel Dahan, et al., Lawyers Should not Trust AI: A call for an Open-source Legal Language Model (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). See also emerging worries regarding data poisoning attacks (Nightshade) or other hacks (CatGPT). The Nightshade method, if translated to text databases, could sabotage LLMs to promote certain desired outcomes or interpretations.

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, it might too speed consideration or approval of applications extraordinarily divergent from the norm.[39]

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 intricate cases that demand more profound judgment. At the very least, it may speed benefits to some claimants who are in dire need.[40] 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 personalization in a literal sense: an authoritative human applying law to facts.[41] Can a legal system effectively remove some decisions to be made extra-legally?

We believe so, based on long-standing theories of emergency as a rationale for suspending otherwise sacrosanct legal requirements.[42] The delay of much-needed assistance to a person or family because of constraints on legal resources is an urgent situation. Law in many instances responds to such vulnerability, but sometimes in order to do so must itself give way to a more expedient mode of action (such as politics, auctions, markets, or, as here, technology). A great deal of ink has been spilled in order to 'square the circle' here, by attempting to assimilate emergency processes and purely technical decisionmaking into the rule of law. But the wiser course is to acknowledge that some areas of decision-making will be subject to summary decisionmaking or automation outside the legal system, rather than distorting our conception of law in order to accommodate them.

[39] For an example of such aneurysm detection, see Christina Jewett, 'Doctors Wrestle With A.I. in Patient Care, Citing Lax Oversight,' *N.Y. Times*, Oct. 30, 2023 ("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.").

[40] Frank Pasquale, *Automated Grace: Toward More Humane Benefits Administration via Artificial Intelligence,* University of Melbourne Centre for AI and Digital Ethics, July, 2022, https://www.unimelb.edu.au/caide/news-media-and-events/online-seminar-with-frank-pasquale (PowerPoint on file with authors).

[41] For a collection of rationales for this position, see Guido Noto La Diega, Against the Dehumanization of Judgment, JIPITEC, https://www.jipitec.eu/issues/jipitec-9-1-2018/4677; Frank Pasquale, Foreword: The Resilient Fragility of Law, in *Is Law Computable,* Simon Deakin and Christopher Markou, eds., 2019.

[42] See, e.g., William E. Scheuerman, The Economic State of Emergency, 21 Cardozo L. Rev. 1869 (2000) ("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.").

The administrative due process principle is rooted in the foundational belief that citizens deserve transparency, fairness, and the ability to challenge decisions that are adverse to them. Therefore, automation should never be used to deny benefits once an application has crossed a very low threshold of plausibility. However, the grant of benefits *vel non* is something no one can rationally contest.[43] 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 see it expand 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, likely to be presented as a potential salve to this quandary. With its capacity to produce detailed, coherent, and seemingly unassailable simulations of justifications, Generative AI appears to be an ideal tool to bridge the gap between the swift decisions 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-3 and similar models—they carry inherent risks that cannot be ignored. The primary concern is the authenticity of these explanations. 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 reveal the genuine reasoning that led to it.[44] “Result-based reasoning” is a formidable epithet in law for a reason. When introducing Score-based Natural Language Processing (SBNLP) into the domain of

[43] 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, 85 N.Y.U. L. Rev. (2010).

[44] Competition & Markets Authority (CMA) of the United Kingdom, ‘AI Foundation Models: Initial Report’ (2023) 81, 82.

administrative benefits, the expectations set by the commitments of justification acquire newfound complexity. Such expansion threatens to undermine individual agency and justifiable governance.

Artificially constructed justifications, though appearing robust on the surface, may not be grounded in factual accuracy.[45] The mode of action of LLMs is next word prediction, not reasoned understanding of the world, or normative evaluation of situations.[46] 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 contest these decisions meaningfully is undermined. This could inadvertently erode trust in the administrative system, leading citizens to view these justifications with skepticism, if not outright disbelief. 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.

Furthermore, the reliance on Generative AI for explanations risks creating a veneer of transparency without substance. The 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 adhering to the requirement for explanations, the process might still violate the spirit of administrative due process. Interestingly, the proposed 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."[47] High-risk systems are obliged to follow some requirements in terms of design, data governance, 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

[45] Minderoo Centre For Technology And Democracy, 'Policy Brief: Generative AI' (Minderoo Centre for Technology and Democracy 2023) 25 <https://www.repository.cam.ac.uk/handle/1810/358089> accessed 28 October 2023. See also Johanna Okerlund et al., What's in the Chatterbox? Large Language Models, Why They Matter, and What We Should Do About Them, at https://stpp.fordschool.umich.edu/research/research-report/whats-in-the-chatterbox (2022).

[46] Cite Stephen Wolfram on next-word-prediction; Timothy B. Lee on ways of explaining how LLMs operate.

[47] Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative actsBrussels, 21.4.2021, COM(2021) 206 final 2021 Annex III (5).

decisions. This means that additional strictures should be proposed now to govern the use of generative AI by decisionmakers.

In sum, while SBNLP offers promising efficiencies in the administrative benefits landscape, and Generative AI seems a tempting solution for providing requisite explanations, strict limits on their use are 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.

IV. 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 individualized, potentially artificial 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. [48] A comprehensive justification statement detailing the structural functioning of SBNLP can provide a clear overview of its operations and methodologies.[49] Moreover, allowing affected individuals the avenue for post-contestation human-driven decisions can instill confidence, ensuring that citizens retain the ultimate power to challenge and seek redress. By integrating these measures, we can uphold the spirit of administrative due process while harnessing

[48] Heleen L Janssen, 'An Approach for a Fundamental Rights Impact Assessment to Automated Decision-Making' International Data Privacy Law <https://academic.oup.com/idpl/advance-article/doi/10.1093/idpl/ipz028/5788543> accessed 10 April 2020; Alessandro Mantelero, *Beyond Data: Human Rights, Ethical and Social Impact Assessment in AI: 36* (2022); Alessandro Mantelero and Samantha Esposito, 'An Evidence-Based Methodology for Human Rights Impact Assessment (HRIA) in the Development of AI Data-Intensive Systems' [2021] Computer Law & Security Review <https://papers.ssrn.com/abstract=3829759> accessed 27 May 2021; Atoosa Kasirzadeh and Damian Clifford, 'Fairness and Data Protection Impact Assessments', *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Association for Computing Machinery 2021) <https://doi.org/10.1145/3461702.3462528> accessed 19 June 2022; Margot E Kaminski and Gianclaudio Malgieri, 'Algorithmic Impact Assessments under the GDPR: Producing Multi-Layered Explanations' [2020] International Data Privacy Law <https://doi.org/10.1093/idpl/ipaa020> accessed 1 February 2021.

[49] Gianclaudio Malgieri, '"Just" Algorithms: Justification (beyond) Explanation Od Automated Decisions under the GDPR' (2021) 1 Law and Business; Gianclaudio Malgieri and Frank Pasquale, 'Licensing High-Risk AI: Towards Ex Ante Justification of a Disruptive Technology' (2023) (forthcoming) Computer Law & Security Review <https://papers.ssrn.com/abstract=4346120> accessed 28 October 2023.

the efficiencies of SBNLP to speed benefits to those most deserving of them, striking a harmonious balance between innovation and justice.