PENNSYLVANIA HOUSE OF REPRESENTATIVES' INSURANCE COMMITTEE

House Bill 1663 of the 2023 Session

October 1, 2024

Testimony of Oliver R. Goodenough

Introduction

This written testimony is provided by Oliver R. Goodenough to the Pennsylvania House of Representatives' Insurance Committee. He is a Research Professor at the Vermont Law and Graduate School, an Adjunct Professor at the Thayer School of Engineering at Dartmouth College, and Affiliated Faculty at Stanford University's CodeX Center for Legal Informatics. He is also a director and officer of Skopos Labs, Inc. and an officer of its subsidiary Brooklyn Investment Group, LLC. These companies both apply Artificial Intelligence ("AI") to support investment management and advisory services but are not currently involved in insurance. Professor Goodenough has published in the field of insurance contract automation; some of these publications are listed in the Resources section at the end of this testimony.

This testimony represents Prof. Goodenough's personal views, presented on his own time, and nothing expressed in this testimony should be assumed to represent the views of any of his employers or other affiliates. This testimony will first address AI in general and its application to law and insurance broadly and will then turn to a discussion of aspects of the proposed House Bill 1663 of the 2023 Session (the "Bill"), which is the target of the hearing of the Insurance Committee.

Understanding AI and its Potential in Law and Insurance.

A starting point for discussing AI is agreeing, at least broadly, on what it is referring to. The term is sometimes traced back to a 1956 conference at Dartmouth. Over most of its history as a field, AI had two principal branches: rules-based approaches and machine learning. The rules-based applications centered on establishing specific, deterministic pathways of event and consequence. Given fact X and circumstance Y, result Z is the outcome. This approach has the advantage of certainty and predictability of result; it often is time consuming to establish, but once set – and "debugged" – it turns out predictable and reliable outcomes.

In recent years, however, there has been a revolution in the capacity of machine learning ("ML") to perform useful work. The fundamental principles of ML involve a learning algorithm, i.e. one that can change with experience, a training set of data and linked outcomes, and a lot of repetitions as the algorithm "learns" better and better how to match inputs and outcomes to mimic the training set's outcomes even when presented with a new variation. In the past few years, a version of ML called large language models ("LLMs") has applied a version of this approach to very large bodies of text, learning both associations among words and phrases and how to build on those associations to create responses to questions and tasks posed by a user. These LLMs are being increasingly deployed to answer questions about decisions with legal implications, such as the medical insurance coverage questions that raise the concerns addressed in the Bill.

A recurring problem with LLMs is that often they have been trained to provide a grammatically accurate, natural language answer – which they generally do well – but not necessarily a factually

accurate one – which they can be dreadful at. Early stages of the most popular LLMs were famously prone to hallucinations – confident assertions that were simply and sometimes spectacularly false. More recent LLM versions are improving on this, and some will now consult sources and cite them for the assertions the LLM makes. By their very nature, however, ML generally and LLMs in particular are probabilistic: they learn to make increasingly educated guesses, but without the strict predictability of a well-constructed rules-based approach.

Challenges in the explainability of ML outcomes heightens these concerns. Because of the trial, error and correction basis for much of the learning done by the learning algorithm, it can be very difficult to surface the exact combination of factors that lead to a particular outcome. Progress is being made in developing explainable AI ("XAI"), but processes are still often a "black box". Auditing for accuracy is often better done by running the AI application on a set of test data and evaluating the quality of the outcomes produced than by trying to understand the combination of data an learning the produced the evolved system. More complex ML algorithms are generally harder to explain, and LLMs – being some of the most complex ML algorithms yet devised – are only beginning to be understood at a level that permits some degree of explainability. However, this is an area of active research, and further advances in XAI can be anticipated in the near future.

The two key questions in deploying ML approaches to do useful work are (i) do its algorithmic design and data training approaches lead to sufficient accuracy in the outcomes, and (ii) is the probability of error limited enough to be a permissible flaw in the system in the context of the task, or can the process can be designed with built in guard-rails to mitigate this error? Such guard rails can be technological, but they can also involve keeping a "human in the loop," as a source of supervision and appeal.

To some degree, these guard rails undercut the economic rationale for deploying AI in a large-scale task, such as insurance determinations. Although the design and training process can be expensive, once the AI is applied to decision making that would take considerable human deliberation, efficiency savings can emerge quickly. But putting safeguards, particularly human safeguards, on top of the AI will inevitably erode some of those savings. So where is a good balance struck? Perfection of outcome is too exacting a goal. After all, humans performing the same tasks will make mistakes as well. On the other hand, too much sloppiness, or worse yet, invidious bias, needs to be rectified.

On the subject of bias, that fault can arise in the application of, but it often arises during *training* based on data involving biased human-decided outcomes. Machines do not inherently discriminate. Machines trained to emulate prejudiced humans may perpetuate that prejudice in their outcomes. This is to be guarded against, but it is not an inevitable property of ML. Here reviewing the quality of the data set can be useful, alongside outcome testing, to minimize the risk of importing unwanted bias into the system.

Because of their relatively high decision-making ability, ML and LLMs have already been adopted to provide judgments in many spheres of activity. In the legal context, initial enthusiasm for their application has waned somewhat in the wake of a more thorough understanding of their limitations. The product of ML and LLM queries can provide a good starting point, but human curation and reworking remain critical, particularly where the outcome is generative – i.e. new creativity, like writing a brief, as opposed to evaluative, like finding certain kinds of clauses in a

corpus of contracts or summarizing a body of information. Lawyers are finding outcome prediction helpful – but best if set in expressly probabilistic terms. Automated outcome determination by ML, such as the specter of "robo-judges," is a much harder problem. Many find taking the time to construct a more tightly determined rules-based system preferable if highly reliable outcomes are required.

In insurance, there are a number of advantages that can flow from deploying AI – in both its ML and rules-based forms. These potential benefits do not just help the company – consumers and regulators can benefit as well. The kind of ML-based coverage evaluation process that is the subject of the Bill is only one of a number of approaches being evaluated. Many insurance companies are exploring the rules-based approaches of "computable contracts" as providing a foundation for a wide range of activities, including policy design, claim administration, underwriting and risk management. This is work that I have been engaged in at Stanford's CodeX Center (further information is available at https://law.stanford.edu/codex-the-stanford-center-for-legal-informatics/codex-insurance-initiative/). Many researchers are also exploring ML and LLM-based approaches. Some, such as internal data analytics and policy design aspects, can probably be left to management to implement and assess. When applied to customer-facing determinations, some regulatory oversight may be appropriate. The Bill is an attempt to initiate such oversight.

Deploying AI in business and legal contexts is both a challenge and an opportunity. AI guided processes can provide significant efficiencies, increased product quality, and even improved accuracy, over human dependent systems. AI is the future in many contexts – including insurance – and too much, or poorly designed, regulation can inhibit good uses that can realize that potential. As with deploying AI, regulating AI involves a balance of opportunity and caution. This leads us to the proposed Bill – how does it measure up?

Evaluating House Bill 1663

The intent of the Bill is to provide safeguards for consumers when health insurance providers use ML and LLM approaches to evaluate a specific step in the coverage determination process, specifically a "utilization review" leading to approval or denial for a particular treatment, referral or other course of action. In summary, if an insurer uses an "artificial intelligence-based algorithm" in conducting a utilization review, the law will impose several requirements on the insurer for notice, transparency, and review. While the goal is laudable, I believe that some of the proposed requirements may prove both burdensome and ineffective. More specifically, I offer the following suggestions for changes in some of the provisions of the current bill.

1. The Definition of Artificial Intelligence-Based Algorithms. In Section 2 of the Bill, the definition of "artificial intelligence-based algorithms" appears to be aimed at ML systems rather than at rule-based approaches. If so, making that explicit could be helpful, such as by adding "but not including any system that primarily uses determined, rule-based automation processes."

2. Insurer Requirements. Section 3 is the heart of the Bill, setting out requirements for insurers. The disclosure requirement, set out in subsection (a), appears to be reasonable and not particularly onerous. Subsection (b) on transparency is more problematic. It requires an insurer using AI to "submit the artificial intelligence-based algorithms and

training data sets that are being used or will be used in the utilization review process to the department for transparency." This process may well not achieve the desired result. As discussed above it is often more effective to audit an ML process by seeing how it performs against a test data set than to try to audit based on how it was constructed. Because ML processes are often set to continue learning from their application, the training data set may well be a constantly moving target rather than a one-and-done kind of filing, making the filing requirement problematic. Further, requiring submission of the algorithms and data effectively reveals to competitors expensive processes meant to give the insurer a lead in the marketplace. The requirement may not deliver the hoped for benefits while being a drag on innovation. I suggest that the requirement be reconsidered to provide for periodic outcome testing by the Insurance Department on each insurer's system in its native environment on the servers of the insurer. The test data originated by the Commonwealth can contain examples that test for the points of bias concern listed in Subsection (b), thus providing a direct check on the possibility of concerning outcomes. Certification on care taken to reduce bias arising from flawed *training* data might be left to the insurer itself rather than being added as a burden to the Insurance Department. Finaly, flexibility on the approach might be given to the Insurance Department to deploy new methods of assessment as XAI develops.

3. Specialist Requirements. Section 4 provides a requirement for the presence of a specialist "human in the loop," and a certification by the specialist that specific customer-specific factors in the determination have been opened and assessed, limiting AI to performing an initial review, but not the final determination. Such human-in-the-loop requirements will, of necessity, undercut the efficiency elements of using AI in this context. That may be a good idea if the accuracy of the AI is shown to be below a certain threshold, but I suggest that greater flexibility be built in for allowing the insurer to request and demonstrate the effectiveness of alternatives, such as deploying automated guardrail checks or implementing a human in the loop presence in an appeal process. Requiring human supervision in every case with no possibility of variation will not keep up with the improvements in accuracy and of XAI audit processes that are bound to emerge as these systems are used.

4. Fines and Penalties. The level of fines and penalties is high enough that it may discourage potentially beneficial adoption rather than simply incentivize careful adoption. It may be helpful to make the highest levels of penalty apply to reckless, willful or knowing violations, and to set a lower level for inadvertent and unintentional failures.

Conclusions

The Bill in its current form provides a starting point for seeking to ensure that the deployment of machine learning processes in health insurance does not lead to inaccurate, and even biased, outcomes. As I suggest, however, somewhat different means for testing and review may deliver more useful results while diminishing the drag on innovation that excessively intrusive regulation can impose. I hope that this discussion and my suggestions will prove useful for the Committee as it considers moving forward with the Bill.

Resources

For the outputs of the CodeX Insurance Initiative, see the publications listed at https://law.stanford.edu/codex-the-stanford-center-for-legal-informatics/codex-insurance-initiative-publications/ and the Special Release: Computable Contracts and Insurance with CodeX, the Stanford Center for Legal Informatics of the *MIT Computational Law Report* available at https://law.mit.edu/codex-computable-contracts-and-insurance.

And the following outside publication deals with matters discussed in this testimony:

Goodenough Oliver R. and Carlson Preston J. 2024. Words or code first? Is the legacy document or a code statement the better starting point for complexity-reducing legal automation? *Phil. Trans. R. Soc. A.* 38220230160, available at https://royalsocietypublishing.org/doi/10.1098/rsta.2023.0160.