Reg. No: 2022/45/N/HS5/00871; Principal Investigator: mgr Andrzej Porębski

Popular Science Abstract
The Understandability Requirement of Machine Learning Systems Used in the Application of Law

Imagine that you have been convicted of a crime. Not good... You would have to find out what determined such an unfavourable verdict. As soon as the sentence has been delivered, you ask for it to be justified. “I can't justify the verdict, because I don't know on what basis the computer system generated it,” you hear. Sounds worrying, doesn't it? I don't think anyone would want to be convicted and not be able to find out what the basis of the sentence was. We feel that where the law is applied, there should be justification for the decisions made.

All right - but what about a situation where certain systems could be used in the judiciary (or any other area of the application of the law): very effective, almost flawlessly determining all the circumstances of cases. They could take over some cases from judges or clerks, giving them more time for other tasks. The systems would have only one disadvantage: it would be impossible to know the motives on the basis of which they reached a decision. If one outlines the issue in this way, it is no longer so obvious: perhaps it is worth giving up the intelligibility of the systems in favour of a possible huge increase in judicial efficiency? But perhaps there should be no compromise on values such as transparency of judgments...?

The issue is difficult because the two scenarios presented could actually be two views of the same situation. Why? Well, the most effective and capable of performing the most complex tasks are often those automatically inferring computer systems, which also have the most complex structure (for example, deep, multi-level, hence the sometimes functioning name of ‘deep learning systems’). Unfortunately, this complex structure means that the rationale for their decisions is often completely unclear to the user. You know they work, you know the quality of their performance (and they can even perform extremely accurately!), but you don't know what made them make a particular decision (even if you are an expert) - what is known as the ‘black box problem’. This problem, combined with the considerable effectiveness of such models, leads to a contradiction between the value of efficiency (e.g. cost reduction) and the value of understanding how the system works (and therefore being able to justify the decisions it suggests). Such a contradiction is particularly evident in the legal context, where, after all, as we intuitively feel, there is often a need to obtain justification, understanding, explanation.

So, as you can see, the two scenarios presented at the beginning are simply extreme highlights of the biggest disadvantage and the biggest advantage of complex computational computer systems that could be used in the judiciary (and in some countries to a limited extent already are: e.g. in the USA or China). We would like the described disadvantage to manifest itself as little as possible. But what should we, as citizens, scientists, lawyers, strive for? What ambiguity in the models are we still able to accept, and what is no longer acceptable? Which system will be compatible with the Polish system of law, and which, due to its unintelligibility, could not be used in it? How important is the value of justifiability of legal decisions, when can we ignore it, and when should we consider it as one of the most important? Which predictive models built from large data sets (i.e. machine learning models) offer the possibility of obtaining the motives behind decisions, and which do not? And further: if the model itself isn’t understandable to humans, are there techniques that can, in a legal context, allow the rationale for a decision to be ‘extracted’ from it? There is no simple answer to these questions, but since the use of computer systems in the application of law is strongly accelerating, research about what standards should be met by machine learning models used (in the future) in the application of law also urgently needs to be accelerated.

This is what I’m dealing with in the project described above. I consider how important the value of decision justification is in law and to what extent it needs to be taken into account in the development of computer systems for the application of law. I examine which classes of models are more and which are less human understandable, and how to make the latter interpretable in some way. I reflect how intelligible the models used in law should be. All this is spiced up by the creation, as part of the research, of example machine learning models that cover the legal context, i.e. that arise using large datasets on the application of the law (e.g. on court decisions). The analysis of such models is intended to allow a better understanding of how their performance can actually be explained within the legal framework of their potential use cases (e.g. prediction of judgements).