Thoughts on the Black Box: Getting to Cooperative Intelligence in Public Administration

Jacob Livingston Slosser 2022-04-01T14:07:20

Introduction

The requirement of explanation for administrative decisions can be found, in one guise or another, in most legal systems. In Europe, it is often referred to as the "duty to give reasons" ("begrundelse" in Danish, "Begründung" in German, and "motivation" in French). This requirement is a positive obligation on decision-makers in public administrative bodies (among others) to provide the legal basis for their decision. With the continuing growth of artificial intelligence/machine learning technologies being used to streamline administrative decision-making, providing for a right to explanation from black box algorithmic decision-making systems is not a straightforward endeavor.

In a recent publication, we put forward that while the right to an explanation is a bedrock principle in public law, the comparison between humans and machines making decisions that require explanations is better understood through the requirements of legal explainability rather than by a causal model. That is, when we ask what a requirement for an explanation is meant to accomplish, the comparative weakness of machines to explain decisions in a „meaningful" way should be understood as an ability for that explanation to function in the legal apparatus rather than an ability to provide some type of causal reasoning. Our point in emphasizing this is, in our present legal system, we do not demand causal explanations from human decision-makers. The legal requirement of explanation does not require a description of how the architecture of the decision-maker's brain produced that decision, what thought components mattered most, or how the decision-makers have sought to overcome their own biases, etcetera. In existing legal practice, the operation of the human brain is not addressed at all, when it comes to the requirement set out for explanations in administrative law. The black box lives on in both human and AI decision-makers.

Why Legal Explainability Is Not Causal Explainability

Many of the same claims now made against algorithmic decision-making systems (ADMs) could very easily be leveled at the kind of explanations offered by human decision-makers in public bodies in the granting (or more likely the denial) of permits, licenses, social benefits, or the like. The only reasoning that is required (if at all) is reasoning that establishes the legal basis for the decision and the usefulness of those reasons for reapplication or an appeal process. It is, in our view, the ability to

challenge, appeal, and assess decisions against their legal basis, which ensures citizens of protection. Though we consider transparency a desirable ethic of AI development, unpacking the black box to its mathematical functions for the purposes of explanation is antithetical to the function which explanation is supposed to serve in the context of law.

In a simplified version of a real-life scenario (let's say an application for a social benefit), a caseworker at the relevant public institution will be trained on past cases, gather current data on the applicant, sort the relevant criteria and make a prediction about whether or not, given the current information, the applicant is more like the decisions in the past that were approved or more like the decisions from the past that were denied. The basic premise for this *modus operandi* is the principle of equality before the law. Like cases should be treated alike. This takes the form of a categorization game that machines are particularly good at. What would be meaningful information to the denied applicant? If we focus on the process of appeal and contestation, it would (usually) include: a boilerplate text that would accompany the denial, stating the rules on which the decision is made, and a statement relaying that the information provided did not fit the criteria for approval or some similarly vague language. It would also likely include instructions for reapplication, avenues for contact, and inform the applicant of their right to complain/appeal.

Retaining the existing human standard for explanation, rather than introducing a new standard devised specifically for AI-supported decision-making, has the additional advantage that the issuing administrative agency remains fully responsible for the decision, no matter how it has been produced. From this also follows that the administrative agency issuing the decision can be queried about the decision in ordinary language. This then assures a focus on the rationale behind the explanation being respected, even if the decision has been arrived at through some algorithmic calculation that is not mathematically transparent. Requiring algorithmic transparency in legal decisions that rely on AI-supported decision-making would be a failure to address the explanation requirement at the right level. It is extremely uncommon for reason giving to include a blow-by-blow recantation of the weights of individual criteria. We believe that this is the bar that ADM needs to hit. The human standard of reason giving, no more, no less. And there is a simple way to test if it passes that bar.

Opacity, of a Kind

Now mainly a cliché of computer science and philosophy, Alan Turing's original „Turing Test" – to see whether a machine is intelligent – provides a design template for what we believe to be the most appropriate system to incorporate ADM systems into public administrative bodies. For the uninitiated, the test he devised consisted of a set up in which (roughly explained) two computers were installed in separate rooms. One computer was operated by a person, the other was operated by an algorithmic system (a machine). In a third room, a human "judge" was sitting with a third computer. The judge would type questions on his computer and the questions would then be sent to both the human and the machine in the two other rooms for them to read. They would then in turn write replies and send those back to the judge.

If the judge could not identify which answers came from the person and which came from the machine, then the machine would be said to have shown ability to think.

Akin to this, an administrative body could implement algorithmic decision support in a way that would imitate the set-up described by Turing. This could be done by giving it to both a human administrator and an ADM. Both the human and the ADM would produce a decision draft for the same case. Both drafts would be sent to a human judge (i.e. a senior civil servant who finalizes and signs off on the decision). In this set-up, the human judge would not know which draft came from the ADM and which came from the human, but would proceed to finalize the decision based on which draft was most convincing for deciding the case, and providing a satisfactory explanation to the citizen. This final decision would then be fed back to the data set from which the ADM system learns.

Rather than concerning itself with whether or not the machine is thinking in terms of Turing's original test, our administrative Turing test is about ensuring that a hybrid system can pass the requirements of explanation, as they exist in different legal settings. Using a hybrid set up ensures the oft-called for „human-in-the-loop" model, with one specific addition: author blindness. Though it may be counter-intuitive, this specific feature of opacity might be the key to functional transparency of decision-making, particularly if the hybrid set up is built around the idea of a continuous human-machine interaction. Relying on this model makes it possible to develop ADM systems that can be introduced to enhance the effectiveness and consistency (equality), without diminishing the quality of explanation. The advantage of our model is that it allows ADM to be continuously developed and fitted to the legal environment in which it is supposed to serve. Furthermore, such an approach may have further advantages. Using ADM for legal information retrieval allows for analysis across large numbers of decisions that have been handed down across time. This could grow into a means for assuring better detection of hidden biases and other structural deficiencies that would otherwise not be discoverable. This approach may help allay the fears of the black box.

There are many caveats we can make here and remaining issues to tackle.

First, we believe our model only really concerns a small set of legal decisions, based on written preparation and past case retrieval. These are areas where a large number of similar cases are dealt with and where previous decision-making practice plays an important role in the decision-making process (e.g. land use cases; consumer complaint cases; competition law cases; procurement complaint cases, applications for certain benefits, etc.). It is not so clear that explanation can function the same way when it comes to harder cases. There are large swaths of decisions that do require a higher bar. However, it is not unreasonable to assume that as the growth in algorithmic competence continues, some of those use cases can be addressed.

Second, there is a remaining issue of wider trust in the decisions. Though focusing on the functional (legal) aspect of explanation anticipates a blindness of the caseworker to the true author of a decision, it relegates the experience of the applicant to a level of trust many might not be willing to give to a hybrid or fully

automated system. If the functional aspect is taken care of, the psychological need for a human in the loop or a reasoned explanation still remains. While not a legal issue as such, there is definitely a behavioral obstacle to be overcome. A boilerplate explanation might be satisfying enough if one knows that it comes from a fellow human being whom one can trust made the decision in an unbiased and objective manner, but a replicated, machine-born explanation – even if reproduced verbatim – may not be enough to satisfy the affective aspect of the experience. After all, many of these decisions matter a great deal to those they affect, and many of these systems require buy-in and use to be justified for public expenditure. Developing a trustworthy system is often about more than just the legal basis or questions of due process.

The search then is not just for hybridity in decision-making designs, but cooperation. A cooperative intelligence would engender both the legal values of procedural administrative safeguards while ensuring the trustworthiness of the decision. It is still quite unclear exactly what this might entail. For example, the European Parliament's recent proposal of an Artificial Intelligence Act (AIA) characterizes regulating AI through a risk-based approach aimed at developing "an ecosystem of trust by proposing a legal framework for trustworthy AI", posing human arbiters as risk assessors and stop gaps to a varying landscape of risky implementations of AI. This approach precludes approaches that marry the strengths of both actors in a decision-making system. Some have argued for the need to establish the "[…] scientific study of intelligent machines, not as engineering artefacts, but as a class of actors with particular behavioural patterns and ecology." Perhaps we are not quite at the intelligence parity point to require machines to be seen as full actors, but the imperative to approach the challenges ADM presents as a study of the ecology of the decision-making environment would go a long way to developing a system of non-rivalling intelligences – a cooperative intelligence. Introducing ADM in public administration is not necessarily a matter of implementing a fully automated decision-making system; instead, ADM can be combined in various hybrid systems with manual (i.e. human) case-work, taking into account alternatives to decisions suggested by ADM. Developing the framework to support an evidence-based appraisal of these relationships will help devise solutions that can make administrative decision-making in public organizations aligned and efficient – while ensuring agency and autonomy for the citizens that are subject to these decisions that is fundamental for ensuring the legality of their implementation.

A third and final caveat is that the reduction of the issue of ADM explainability to solely its legal foundation relies on accepting the premise that the explanations given by a decision maker are meaningful enough as they are given presently. Setting the bar where it currently stands may concretize institutions' reluctance to go into detail regarding their decisions. Our focus may lock in a lost opportunity for improving systems that could be more forthcoming in their explanations.

Of course, our insistence on the primacy of the legal basis of explanation rather than the causal explanation reflects our training as legal academics. That bias notwithstanding, we see the opportunity for implementing ADMs in real life scenarios as integral to the prospect of adding value and expediency to administrative

decisions without a loss of legality. As for the affective consequences, we are currently developing a forum for more interdisciplinary work to be focused on administrative decision-making by cooperative intelligence. Our first event will focus on how different disciplines might approach the affective problems of opaque machine decisions without requiring full explainability (XAI) that are often detrimental to accurate decision models. It is only with disciplinary cooperation that we can envision a true cooperative intelligence.
