On how AI & law can help autonomous systems obey the law: a position paper

Henry Prakken

Abstract. In this position paper I discuss to what extent current and past AI & law research is relevant for research on autonomous intelligent systems that exhibit legally relevant behaviour. After a brief review of the history of AI & law, I will compare the problems faced by autonomous intelligent systems with the problems faced by lawyers in traditional legal settings. This should give insights into the extent to which AI & law models of legal problem solving and decision support can be applied in the design of legally well-behaving autonomous systems.

1 Introduction

Increasingly, computer systems are being employed in practice with some degree of autonomy. Their behaviour is not fully specified by the programmer but is the result of the implementation of more general cognitive or physical abilities. Such artificially intelligent software can do things that, when done by humans, are regulated by law. To give some example, self-driving cars have to obey the traffic laws. Online information systems that decide whether a system of person can be given access to privacy-sensitive data have to comply with data protection law. Actions of care robots that help sick or elderly people can damage property or the health of the person (spilling coffee over an iPad, failing to administer medication on time). Intelligent fridges that can order food or drinks when the supplies run out have to obey contract law. Autonomous robot weapons have to comply with the of war, with its three principles that soldiers should distinguish between the civilian population and combatants, that an attack is prohibited if the expected civilian harm is disproportional to the expected military benefits, and that military force must be necessary to the explicit purpose of defeating an adversary.

When such autonomous systems are being used, legal rules cannot any more be regarded as regulating human behaviour, since it is not the humans but the machines who act. This raises the problem of how the autonomous systems can be designed in such a way that their behaviour complies with the law. Note that this question needs to be asked irrespective of the question whether machines can assigned responsibility in a legal sense. Even if a human remains legally responsible or liable for the actions of the machine, the human faces the problem of ensuring that the machine behaves in such a way that the responsible human complies with the law.

One solution to the problem is to design the system in a way that guarantees that the system will not exhibit unwanted behaviour. This is the conventional solution when non-autonomous machines, tools or systems are used. [16] called this regimentation. A similar approach has been proposed for autonomous systems, such as in the Responsible Intelligent Systems project at Utrecht University, which proposes to verify the behaviour of systems off-line with so-called model-checking techniques. However, when systems are increasingly autonomous, their input and behaviour cannot be fully predicted, so that regimentation or advance off-line testing are impossible or of limited value. How can we then ensure that autonomous systems comply with the law? This position paper discusses to what extent the fruits of AI & law research are relevant for solving this problem. (For a related discussion from a more legal perspective and specifically for robots see [18]). To this end, I will first briefly review the history of AI & law research and then compare the problems faced by autonomous intelligent systems with the problems faced by lawyers in traditional legal settings.

2 A brief history of AI & law

The 1970s and 1980s were the heydays of research on knowledge-based systems, such as the influential MYCIN system for diagnosis and treatment of infection diseases [6]). For long computer scientist could in these days easily think that in the legal domain knowledge-based systems can be much easier developed than in the medical and similar domains. While medical knowledge needs to be acquired from human medical experts who are not always aware how they solve a medical problem, legal knowledge would simply be available as rules in written texts, such as statutes and case law reports. And such rules can easily be represented in a rule-based system like MYCIN, after which their application to the facts of a case would be a simple matter of logic. On this account, once a legal text and a body of facts have been clearly represented in a logical language, the valid inferences are determined by the meaning of the representations and so techniques of automated deduction apply.

However, this mechanical approach leaves out most of what is important in legal reasoning, as every lawyer knows. To start with, legislators can never fully predict in which circumstances the law has to be applied, so legislation has to be formulated in general and abstract terms, such as ‘duty of care’, ‘misuse of trade secrets’ or ‘intent’, and qualified with general exception categories, such as ‘self defence’, ‘force majeure’ or ‘unreasonable’. Such concepts and exceptions must be interpreted in concrete cases, a process which creates room for doubt and disagreement. This is reinforced by the fact that legal cases often involve conflicting interests of opposing parties. The prosecution in a criminal case wants the accused convicted

\(^{1}\) Department of Information and Computing Sciences, Utrecht University and Faculty of Law, University of Groningen, The Netherlands, email: H.Prakken@uu.nl

\(^{2}\) Presented at the ECAI 2016 Workshop on AI for Justice.


\(^{4}\) Some parts of his section are adapted from [22] and [23].
while the accused wants to be acquitted. The plaintiff in a civil law suit wants to be awarded compensation for damages, while the defendant wants to avoid having to pay. The tax authority in a tax case wants to receive as much tax as possible, while the taxpayer wants to pay as little as possible. Both aspects of the law, i.e., the tension between the general terms of the law and the particulars of a case, and the adversarial nature of legal procedures, cause legal reasoning to go beyond the meaning of the legal rules. It involves appeals to precedent, principle, policy and purpose, as well as the consideration of reasons for and against drawing conclusions. Another problem is that the law often gives considerable freedom of judgement to the judge, for example, when determining the extent of financial compensation for a tort or when determining the sentence in a criminal case. Although judges are supposed to decide like cases alike, in these matters there are no clear rules, since cases are never fully alike. In all this, it is relevant that the law is not just a conceptual or axiomatic system but has social objectives and social effects, which must be taken into account when applying the law. A final problem is that determining the facts of a case is often hard, since it requires vast amounts of commonsense knowledge of the world, and giving the computer common sense is a recognised hard problem in AI [7].

In sum, law application is not just logically applying statute and case law rules to the facts of a case but also involves common sense, empathy and a sense of justice and fairness. Modelling these aspects in a computer program has so far proved too hard.

However, this does not mean that AI cannot be usefully applied to the law. Deductive techniques have been practically successful, especially in the application of knowledge-based systems in large-scale processing of administrative law, such as social benefit law and tax law. Such systems apply computational representations of legislation to the facts as interpreted by the human user. The use of such systems has been proved to greatly reduce two major sources of errors in the processing of social benefit applications by ‘street-level bureaucrats’: their incomplete knowledge of the relevant regulations and their inability to handle the often complex structure of the regulations, with complex boolean combinations of conditions, numerical calculations and cross-references [14, 29]. The computer is, of course, ideally suited for retrieving stored information and for handling syntactic and numerical complexities. Deductive rule-based systems have therefore been applied in public administration on a considerable scale. Such systems leave it to the user (the official with the authority to make a decision) to decide whether to accept the system’s recommendation or to deviate from it on non-statutory grounds. Thus these systems do not automate legal judgement but the logic of regulations [15, 14].

The deductive model of legal reasoning has been refined with means to express rule-exception structures and hierarchies of regulations. Two common structural features of legal regulations are the separation of general rules and exceptions, and the use of hierarchies over legislative sources to resolve conflicts between different regulations. AI and law has dealt with these features with so-called non-monotonic logics. Such logics have been shown useful in modeling legislative rule-exception structures and legislative hierarchies [10, 24, 13, 32], and in modeling legal presumptions and notions of burdens of proof [25, 11, 12]. Nevertheless, although nonmonotonic techniques technically deviate from deductive logic, their spirit is still the same, namely, of deriving consequences from clear and unambiguous representations of legal rules, rule priorities and facts. More often, conflicts arise not from competing norms but from the variety of ways in which they can be interpreted. A real challenge for deductive accounts of legal reasoning is the gap between the general legal language and the particulars of a case. Because of this gap, disagreement can arise, and it will arise because of the conflicts of interests between the parties.

These observations can be illustrated with the famous Riggs v. Palmer case discussed in [9], in which a grandson had killed his grandfather and then claimed his share in the inheritance. According to the applicable inheritance law, the grandson was entitled to his share, but every lawyer understands that he killed his grandfather is a reason not to apply this law. And indeed the court denied the grandson his claim on the grounds that nobody should profit from his own wrongdoing. A deductive or nonmonotonic rule-based system cannot recognise this, unless the exception is already represented in concrete terms in the knowledge base. Adding an explicit exception like ‘unless the heir would profit from his own wrongdoing by inheriting’ to the relevant legal rules would not solve the problem, since the system cannot recognize that inheriting from one’s grandfather after killing amounts to profiting from one’s wrongdoing, unless this is explicitly represented in the system’s rule base.

Nevertheless, AI offers more to the law than systems based on deductive or nonmonotonic logic. To start with, when for an interpretation problem the relevant factors are known, and a large body of decided cases is available, and these cases are by and large consistently decided, then techniques from machine learning and datamining can be used to let the computer recognize patterns in the decision and to use these patterns to predict decisions in new cases. One example is [8]’s statistical model for predicting whether a job offered to an unemployed is ‘suitable employment’, in which case refusal of the job offer should lead to a reduction of the employment benefit (see [31] for a neural-network application to the same data and [2] for a similar application to UK social security law). Another example is the sentencing system of [20], which could give judges deciding on sentences for street robberies insight into sentences assigned in similar past cases. On sentencing see also [28].

In spite of the good level of performance of such AI techniques, their practical usefulness in the legal domain is limited, for two main reasons. First, not many legal interpretation problems meet all three requirements for successful use of these techniques: a known and stable set of relevant factors, many decided cases, and little noise among or inconsistency between these cases. More importantly, these techniques are notoriously bad in explaining their output. They are essentially black boxes, which give no insight into how they relate their input to their output. Needless to say that for judges this is a major obstacle to using these systems.

These limitations are addressed in AI & law research on legal argument. This research has led to many important theoretical advances, all based on the idea that legal reasoning is about constructing and critically evaluating arguments for and against alternative solutions of a case. Detailed models have been provided of the role of cases, principles, values and purpose in legal reasoning, of analogical reasoning of reasoning about evidence and of the role of procedure and burden of proof in legal reasoning. For overviews see e.g. [27, 4, 26, 23]. While some of this research has been purely theoretically motivated, others ultimately have practical aims. For instance, [1] sketched a vision of a system which could support an advocate charged with preparing a case at short notice. The system would be able to accept the facts of the case and then generate arguments for the two sides to the case and counterarguments to them, together with the precedents on which they are based. However, such a system is not yet in practical use at any law firm. A main problem with AI & law’s proof-of-concept systems has so far that they are critically dependent on the possibility of acquiring a large amount of knowl-
edge and representing it in a form which can be manipulated by the system. This is an instance of the well known ‘knowledge acquisition bottleneck’, which has proved a major barrier to the practical exploitation of intelligent techniques in many domains.

The most recent development in AI & law research is a revitalisation of research on information retrieval by the recent spectacular developments in such areas as deep learning, data science, and natural-language processing, combined with the availability of huge amounts of unstructured legal information on the internet. This has put new topics such as information integration, text mining and argument mining on the research agenda. With IBM’s Watson system available, the holy grail for many in legal informatics is not an argumentation assistant as described in [1] but a legal research assistant in the form of an application of Watson, which can efficiently find, summarise and integrate information relevant to a case.

Nevertheless, there is some hope that this recent research can also make an argumentation assistant within reach. A very recent application of text mining called ‘argument mining’ has become popular [21, 33, 17] and IBM’s Watson team has already experimented with a ‘debater’ function, which can find arguments for and against a given claim. The fruits of this research can perhaps be combined with AI & law’s argumentation models in such a way that these models can finally be scaled up to realistic size, without the need for formal knowledge representation.

3 Is obeying the law always desirable?

Before discussing how autonomous systems can be made to obey the law, first another question must be discussed: it always desirable to obey the law? In part this is still a legal question, since (parts of) legal systems have general exception categories like the exception concerning self-defence and other ones in criminal law, a general exception in Dutch civil law that statutory rules concerning creditor-debt relations shall not be applied if such application is unreasonable, and so on. Consider the case of the autonomously driving Google car, which was stopped by the California police for driving too slowly. Google had for safety reasons set the car’s maximum speed for roads in a 35mph zone at 25mph and one of its cars was causing a big queue of traffic while driving 24mph.

From a technical legal point of view this is not a case of undesirable norm obedience, since the relevant traffic regulation contains the following general exception clause:

No person shall drive upon a highway at such a slow speed as to impede or block the normal and reasonable movement of traffic, unless the reduced speed is necessary for safe operation, because of a grade, or in compliance with law.

However, there is still a practical problem, since general exception clauses like these introduce vagueness and uncertainty. Human drivers are generally good at determining when their speed is to slow by applying their experience and common sense. However, can autonomous cars be given the same kind of common sense? For a preliminary proposal see [19].

One step further are cases in which behaviour is from a technical legal point of view illegal but still socially acceptable. For example, slightly speeding in a queue of cars that all drive a few miles above the maximum speed; waiting for a red pedestrian crossing light at night with no traffic within eyesight; admitting a student to a university course who missed the strict admission deadline for some stupid reason. Here the reasoning problem is logically the same as with general exception clauses: determining whether particular behaviour satisfies some general exception category to a behavioural rule. That the exception is now for social instead of legal acceptability is irrelevant for the kind of reasoning involved.

This all means that the behaviour of autonomous systems should not be seen as rule-governed but as rule-guided. Legal rules are just one factor influencing socially optimal or permissible behaviour. Other factors are e.g. social conventions, individual or social goals or simply common sense. And sometimes these other factors override the legal factors. There has been some research on such norm-guided behaviour in the NORMAS community of International Workshops on Normative Multi-Agent Systems. See, for instance, [5].

4 The classic AI & law problems vs. the new challenge

For several reasons the above story about the practical applicability of AI & law research does not automatically apply to the problem of making autonomous systems obey the law. First, as we saw above, AI & law research has traditionally focused on support tools for humans carrying out legal tasks. With autonomous systems this is different: they do not support humans in their legal tasks (although they may support humans in other tasks) but they have to decide about the legal status of their own actions. In many cases it will be impossible for humans to check or override the system’s decision.

Moreover, the tasks supported by traditional AI often concern the application of the law to past cases, to determine whether some past behaviour or some existing state of affairs is lawful or induces legal consequences. With autonomous systems this is different, since they have to think about the legal status of their future actions. Among other things, this means that in contrast to in traditional legal settings, autonomous systems do not face evidential problems in the legal sense. Even when traditional AI & law supports legal tasks with an eye to the future, such as deciding on benefit or permit applications, drafting regulations or contracts or designing tax constructions, there are differences with autonomous systems. While traditionally supported future-oriented task concern behaviour in the non-immediate future and often contain classes of actions (as with contract or with regulation design), autonomous systems have to ‘run-time’ consider individual actions in the immediate future.

Another difference, as explained in Section 3, is that the tasks supported by traditional AI & law are usually strictly legal while autonomous systems have to balance legal considerations against other considerations. This is not a black-and-white difference since, as explained in Section 2, law application also involves considering the social context and issues of fairness, common sense and the like. However, in the law, this is always done in service to the overall problem of classifying behaviour into legal categories. With autonomous systems this is different, since they do not have as their sole or primary aim to stay within the law.

Yet another difference is that the legal tasks supported by traditional AI & law tools require explanation and justification of decisions. With autonomous systems there is no need for this: all that counts is that legally acceptable behaviour is generated. Of course, when an autonomous system does something legally wrong, its behaviour might have to be explained in a court case. However, this does not require that the system itself can do that; it may suffice to have a log file recording the system’s internal actions.


\(^6\) \url{http://icr.uni.lu/normas/}, accessed 30 May 2016.
Finally, one may expect that the bulk of the cases encountered by an autonomous system will from a legal point of view be standard, mundane cases. For example, autonomous cars will not have to determine the legal responsibility for car accidents but will have to decide about driving from A to B in a way that respects the traffic regulations. While processing legislation in public administration also usually concerns standard cases, in the court room this is different.

5 Implications for applicability of AI & law research

What are the implications of the similarities and differences between the ‘traditional’ and new settings for the applicability of AI & Law research? The discussion here has to be speculative, since the answer depends on the type of autonomous system, how advanced it is, how safety-critical it is, and so on. Moreover, presently, there are still only few autonomous systems in practical use that have to take legally relevant decisions in a non-trivial way. Nevertheless, the technological developments go fast. Just 10 years ago, recent advances like IBM’s Watson system and autonomously driving vehicles seemed unthinkable for the near future. Therefore, thinking about these issues cannot be postponed to the future.

Essentially, there have so far been three kinds of successful AI & law applications: decision support for large volumes of routine decision tasks (as in public administration); retrieval, summary and integration of legal information; and prediction of outcomes of decision problems in narrowly-defined factor-based domains.

Does the ‘standard’ nature of many cases faced by autonomous systems mean that the techniques for routine decision support as used in public administration can be applied to autonomous systems? This is not likely, since the traditional rule-based systems crucially rely on humans for preprocesses the input facts in legal terms and for overriding if necessary the system’s decisions.

Can Watson-like legal research agents that retrieve, summarise and integrate information support autonomous systems? Here a similar problem arises, since the effective use of retrieved, summarised and integrated information still crucially relies on human judgement. Moreover, it remains to be seen whether the currently available legal information will be useful for the mundane and future-oriented normative decision problems faced by autonomous systems.

Are nonmonotonic reasoning techniques useful as a way to deal with exceptions and conflicting regulations? Not really, since such techniques do not offer ways to recognise the need for an exception to a legal rule or to recognize the best way to resolve a conflict between regulations, unless this has been programmed into the system in specific terms. Moreover, if the rules contain general exception clauses or the regulations contain general conflict resolution principles, the classification and interpretation problem will be too big.

Can machine-learning techniques as applied to factor-based domains support autonomous systems in classification and interpretation problems? Perhaps to some extent but there is room for caution here, since in the law these techniques have so far only worked for narrowly defined domains with a large amount of relatively consistent data. And the law does not have many of such domains. Moreover, even when the data has to come from case law, a problem is that the cases may not be standard future-oriented cases of the kinds faced by the autonomous system. On the other hand, the ‘traditional’ drawback that these systems cannot justify or explain their output does not apply for autonomous systems, which are only meant to generate legally correct behaviour, not to explain or justify it.

Finally, there is the question whether an autonomous system should be designed to reason about how to behave lawfully or whether it can be trained to do so with machine-learning techniques applied to a large number of training cases. In the first approach there is the need for explicit representation of legal information in the system and for giving the system explicit reasoning and decision making capabilities. This is still somewhat similar to the traditional AI & law systems for supporting human decision making, except that the human is taken out of the loop. An important issue then is whether the mundane nature of cases faced by the autonomous system can reduce the complexity of the classification and interpretation problems to such an extent that the machine can fully take over. On the other hand, the reasoning can, unlike in the traditional settings, be opaque in that there is no need for explaining or justifying why the behaviour is legally correct. Incidentally, the latter combined with the run-time and forward-oriented setting with mundane cases, makes that the current research strands on evidential legal reasoning and sophisticated legal argument will likely be less relevant here.

The other approach is that the ability to behave legally correctly is acquired implicitly by training. For very advanced autonomous systems, like robots operating in daily life, this might be equivalent to solving the notorious AI common-sense problem, but for more modest systems this approach might do. One interesting question is how autonomous vehicles classify on this scale. [18] discuss some interpretation and classification problems in Dutch traffic law that are relatively easy for humans but seem very hard for the current generation of autonomous vehicles. The ‘training’ approach does not necessarily avoid the need for explicit representation of legal rules and regulations. They must now be represented as part of the design specification. One issue here is whether these specifications should be machine-processable in the same way as when designing explicit legal reasoning (as in the methods proposed by [3, 30]). It seems likely that at least some form of semi-formal representation is required, for purposes of verification and maintainability.

6 Conclusion

This position paper has been motivated by the rapidly increasing prospects of practically used autonomous artificial systems performing legally relevant tasks. The aim was to discuss how the current fruits of AI & law research on supporting human legal decision making can be used for making autonomous artificial systems behave lawfully. To this end the problems faced by human lawyers were compared to those faced by autonomous systems. The main similarity is that in both cases there is automated application of norms to facts. However, main differences are that the legal problems faced by autonomous systems have to be solved run-time and are future-oriented instead of past-oriented. Moreover, while in traditional legal settings being lawful is the main goal, for autonomous systems it is only one of the concerns, to be balanced against, for example, social and individual goals. On the other hand, the legal problems faced by autonomous systems are, unlike those faced by lawyers in traditional settings, usually standard, mundane cases. Moreover, unlike lawyers in traditional settings, autonomous systems will usually not have to explain why their behaviour is lawful.

Because of the similarities, research on designing legally well-behaving autonomous systems can profit from the fruits of current AI & law research. However, because of the differences, applying these fruits in the new contexts is not trivial and requires extensive further research. In this position paper I have tried to create some awareness of the need for such research and pointed at some possible research directions.