EasyChair Preprint

№ 4844

What criminal and civil law tells us about Safe RL techniques to generate law-abiding behaviour

Hal Ashton

December 29, 2020

What criminal and civil law tells us about Safe RL techniques to generate law-abiding behaviour

Hal Ashton[1*]
[1]University College London
Gower Street
London
ucabha5@ucl.ac.uk

Abstract

Safe Reinforcement Learning (Safe RL) aims to produce constrained policies with constraints typically motivated by issues of *physical* safety. This paper considers the issues that arise from regulatory constraints or issues of *legal* safety. Without guarantees of safety, autonomous systems or agents (A-bots) trained through RL are expensive or dangerous to train and deploy. Many potential applications for RL involve acting in regulated environments and here existing research is thin. Regulations impose behavioural restrictions which can be more complex than those engendered by considerations of physical safety. They are often intertemporal, require planning on behalf of the learner and involve concepts of causality and intent. By examining the typical types of laws present in a regulated arena, this paper identifies design features that the RL learning process should possess in order to ensure that it is able to generate legally safe, compliant policies.

Introduction

In this position paper I will consider the problem of learning a solution to a sequential decision making problem in an environment governed by some laws via Reinforcement learning (RL). I will assume that the learned policy should not break these laws because doing so would impose sanctions by the environment's regulator or law enforcer. By presenting a taxonomy of laws which exist in real life, whose features are relevant to RL, I am able to make some inferences about the general design of a RL process that can produce legal policies.

Reinforcement Learning is a process which can produce novel policies to solve sequential decision problems. Its potential has been demonstrated in the super-human mastery of Go (Silver et al. 2017) and advanced performance in more complicated games like Starcraft (Vinyals et al. 2019) but adoption in real life settings has been retarded by safety (including legal) considerations. This is noticeable in Financial trading applications which already use algorithms extensively[1] but have been slow to adopt RL.

RL requires an environment which allows ample exploration and feedback. In game applications such as Go and Atari games the training environment is a faithful recreation of the deployment environment and training is costless. Potential real world applications of RL are often more complex, almost certainly regulated, and the cost of mistakes made in training or deployment could be catastrophic. In such applications where safety, cost or legality are issue, one approach is to conduct learning in a simulator of the environment where the cost of bad policies is negligible. The use of any simulator raises the risk of misspecification and poor generalisation. Misspecification refers to the case where the RL learner optimises on an environment which is not a faithful recreation of the actual deployment environment and the policy performs poorly on deployment. In the case of poor generalisation, the actual environment produces a situation of the like which was not seen in the training process and the policy fails on deployment. An alternative to using a simulator is placing the RL learner very carefully in the target environment with a human overseer ready to take over in tricky spots. This approach has limitations according to the complexity of the task (Saunders et al. 2018). It might not be feasible to use this approach in applications like trading because the speed of decision making is beyond the ability of a human overseer to monitor.

Whether learning takes place in a simulator or carefully in the target arena, the ability to generate legal policies with high probability is very desirable. Laws can present different challenges to other types of constraint. A legally transgressive policy might not be obvious in the way a physically transgressive one might be. The nature of laws will dictate the methods of RL used to generate optimal, legal policies.

Background

Markov Decision Processes (MDPs) are a common framework underpinning RL. In this formulation time is discretised and labelled $t = 1, 2, 3, \ldots$. A MDP is described by a tuple $(\mathcal{S}, \mathcal{S}_0, \mathcal{A}, \mathcal{T}, R, \gamma)$ where:

1. $\mathcal{S}$ is the set of states in the environment.
2. s_0 is a distribution over the initial states of the environment $p(s)$ for $s \in \mathcal{S}$.
3. $\mathcal{A}$ is the set of all actions available.
4. $\mathcal{T}(s, a, s') = \mathbb{P}(s'|s, a)$ is the transition probability distribution; the probability of transitioning to state s' when in state $s \in \mathcal{S}$ and choosing action $a \in \mathcal{A}$.

*Supported by the EPSRC

[1]In most markets where electronic trading is enabled, the majority of trading is now conducted by algorithms of varying autonomy

5. $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, the feedback mechanism through which learning is possible.
6. $\gamma \in (0, 1]$ the discount factor to differentiate the value of rewards now vs those received in the future. In finite horizon cases $\gamma = 1$ and can be ignored.

The learner then has the objective of funding a policy function from the set of all policy functions $\Pi : \mathcal{S} \to \mathcal{A}$ which solves the maximisation of the expected discounted sum of rewards:

$$\pi^* = \arg\max_{\pi \in \Pi} E\big[\sum_{t=0} \gamma^t R(s_t, a_t)|\pi\big]$$

The policy function is often a probability distribution over actions $\pi(a|s) = \mathbb{P}(a|s)\ \forall a \in \mathcal{A}, s \in \mathcal{S}$

The Markovian property of this process comes from the transition function. It is satisfied if the probability of transition to a new state is determined only by the current state and chosen action.

An extension of the MDP is the Partially Observable MDP (POMDP). This covers the very probable contingency where the full state of the world is not visible to the decision maker. It is described by the tuple $(\mathcal{S}, \mathcal{S}_0, \mathcal{A}, \mathcal{T}, R, \gamma, \Omega, O)$. two additions to the tuple are as follows:

7. Ω is the set of all observations that the learner can receive. For convenience we assume that it includes the reward r_t received in any time period.
8. $O = \mathbb{P}(\omega|s', a)$ is the probability distribution of receiving observation $\omega \in \Omega$ after transition to state s' and action a was chosen.

The domain of the policy function then becomes the history of all observations and actions which we write $\pi(a|h_t)$ where h_t is short hand for $(o_1, a_1, o_2, a_2, \dots a_{t-1}, o_t)$.

The complexity of a solving a POMDP is much higher than that of a MDP (Abel, MacGlashan, and Littman 2016) because the learner is required to perform inference over the state of the world using the history of actions, observations (including rewards). Intuitively this can be seen by observing that the domain of the policy function is exponentially larger and any search process taken over the set of all policy functions will take much longer.

Our interest in introducing POMDPs is not so much the partial observability of this problem but more the enlarged domain of the policy function which they necessitate. This paper will show that law abiding policies are likely to be dependent on the history of observations regardless of whether there is partial observability or not.

Structural Causal Models

This paper will show that the legality of behaviour can depend on establishing the causal effects of an action. The definition of causality is a complicated topic and there is a distinction between predictive causality which refers to predicting the effect of actions and actual causality which refers to evidential analysis after actions have been taken. Structural Causal Models (a special case of Bayesian Causal Networks) (Pearl 2000) can be used in both senses. Actual causality is a harder undertaking since it necessarily involves counterfactual reasoning. Halpern Pearl Causality (HP) (Halpern 2016) is a general purpose definition of actual causality. More simple definitions are discussed in (Liepiņa, Sartor, and Wyner 2020).

For a set of variables $X = (X_1, \dots X_n)$ we define a structural causal model (SCM) to be defined by:

1. A set of independent noise variables $U = (U_1, \dots U_N)$ with associated distributions P_{u_i}
2. A Directed Acyclic Graph (DAG) G over vertices $X \cup U$. There is a directed edge from X_i to X_j iff X_i is a direct cause of X_j. That is to say, changing the values of X_i whilst keeping other members of $Pa(X_j)$ constant will change X_j
3. A set of functions $f_1, \dots f_N$ such that: $X_i = f_i(pa(X_i), U_i)$ where $pa(X_i)$ are the parents of X_i according to the graph G.

Under this structure, the values of X are uniquely by the values of U and the distribution of $\mathbb{P}(U) = \prod_{i=1}^{n} P(u_i)$ determines a distribution over X, $\mathbb{P}(x_1, x_2, \dots, x_n)$. Furthermore it can be shown that the graph G satisfies the causal markov condition which states that X_i is independent of all its non-descendants given its parent set $pa(X_i)$. The graph is equivalent to a Bayesian causal model, an augmentation to a standard Bayesian graphical model where the directed edges encode causal statements not just independence statements.

To assess the effect of actively changing a subset of variable(s) Z on a disjoint set of variables in Y both in X, Pearl (Pearl 2000) introduces the concept of an atomic intervention. $\mathbb{P}(Y|do(z))$ is the probability distribution obtained by setting $Z = z$ in the equations and deleting any equations which determine elements of Z. This is equivalent to considering the distribution of the subgraph G' which is obtained by removing any ancestral vertices to elements Z and giving unitary point mass to the realised values of Z for any dependent distributions.

Intent in RL

RL has been used infer the intent of others (Qi and Zhu 2018), and even in IRL to define a reward function that corresponds to the intentions of an expert demonstrator (MacGlashan and Littman 2015). However it has not ever defined what intent means for the learner. Ashton (2020) presents a definition of direct intent in terms of causality and desire of realised states. An agent directly intends a state s by committing an action a if a causes s and the agent aims or desires state s. In the context of RL, where an agent has a value function over every possible state, inferences can be made about what a learner desires. Within criminal law, different levels of intent are required for different crimes (Loveless 2013). Direct intent is the strictest, being required for murder but lower levels such as oblique intent, recklessness and negligence also exist. Whilst these lower levels of intent do not necessarily have any requirements about desire, their definitions often include subjective and objective tests of foreseeability vis-a-vis the prohibited outcomes of actions. Subjective tests raise interesting questions in model-free modes of RL since the learner does not explicitly expect any outcome

to their action. Objective tests require an external judgement about the probability of an outcome given an action. If a consequence of an action was foreseeable then the offender can be thought of intending the outcome. Lagioia and Sartor (2020) discuss this method of intent inference and consider it sufficient albeit principally in the context of Belief Desire Intent (BDI) agents (Cohen and Levesque 1990). An intriguing corollary of the use of objective tests, is that the predictive model that the RL agent uses or learns should be accurate. This bypasses the danger of the learner developing a 'delusion box' type model (Ring and Orseau 2011) to justify otherwise illegal policies.

A non-exhaustive taxonomy of laws

In this section I present a number of law-types which are likely to exist in a regulated environment. I differentiate between states and actions. Actions are assumed to be originated by the learner only and their commission is voluntary. States refer to some measurable property of the environment, stable for the duration of the time period. Actions cause a measurable change in the environment but I assume that their duration is instantaneous so there is no state that records an action in progress.

Simple State restriction laws

This the simplest type of law and the one which Safe RL research has concentrated on as many physical safety constraints are of this type. Examples might include 'drive below 30 miles per hour in urban environments' or 'don't crash into pedestrians'. For a state restriction to be a law, its realisation should necessarily caused by the actions of the learner. This is obviously the case for the speed restriction example and a desirable though maybe not strictly true in the pedestrian avoidance one.

Caused state restriction laws

Some states exist which could be both caused through the learner's actions and through an external mechanism. This marks the first departure from conventional safe-RL research because the safety constraints traditionally considered do not differentiate between those states caused by the learner and those that are not. It does not matter whether the drone caused itself to fly over the volcano or whether it was blown by the wind, the state of being located over the volcano is the one to be avoided. For legal restrictions, certain states might only be restricted if they were caused by the learner[2]. A concrete example of such a causal dependent state restriction can be found within financial markets where the UK's financial regulator prohibits trading algorithms from creating or contributing to disorderly market conditions. Such conditions could arise independently of the behaviour of the learner, if the learner has no mechanism of determining whether this the case, learning efficiency will be compromised.

For caused state restriction laws to be broken two conditions should be satisfied. Firstly a restricted state $\hat{s}$ should occur and secondly that the actions of the learner foreseeably caused the state to occur.

Action sequence laws

Laws exist in a variety of settings which are restrictions on *conduct* with no necessary requirement for a lasting change in the state of the world. Examples in the UK include the offence of Careless Driving and more seriously Dangerous driving. There is no causation requirement since I assume that the A-bot has the freedom to choose its own actions at any time[3] Action sequence laws could be transformed into a simple state restriction law by adding a state variable that indicates whether a restricted action sequence has occurred. Such an approach might not be efficient if a large number of conduct laws exist in the environment as it would cause the dimension of the state space to grow.

Mixed State Action sequence laws

Laws exist which combine a *sequence* of actions that *cause* restricted state(s). Continuing the driving examples from the previous section, in the UK there exist statutory offences of causing death by careless or reckless driving. These laws constitute a restriction on how certain states are arrived at.

Inchoate offences: Laws restricting behaviour that may induce future restricted states

Inchoate offences are restrictions on action sequences and states which may lead to restricted states which are not necessarily realised. Examples in the UK include attempt crimes such as attempted murder or possession of drugs with intent to supply. In the USA conspiracy and solicitation (the request, encouragement or payment for someone else to commit a felony) are major classes of inchoate offence.

Laws requiring intent

Common law as practised in UK, USA, India and Canada amongst others requires that the accused had *mens rea* (the mental element of intent to commit a crime) for certain criminal offences to have been committed by them. Different levels of intent exist, ranging from direct intent where the accused deliberately caused and wanted to cause a prohibited outcome in the extreme to oblique intent where prohibited outcomes were caused as a side-effect of their behaviour, to recklessness and negligence where the prohibited outcomes were foreseeable outcomes of their behaviour to various degrees. Certain offences will specify what level of mens-rea is required so murder requires direct or oblique intent.

Aside from the crimes of specific intent, certain laws exist which require establishing for what purpose the accused did something. This is called basis intent by Bathaee (2011). US Examples cited include market anti-spoofing laws, a practice which is defined as the placement of orders with the intent to cancel them. Another related and topical example is termed

[2]Death is not generally prohibited, but causing it generally is!

[3]Situations where there is no action which won't break a law are analogous to the concept of deadlock in model checking (Baier and Katoen 2008). Laws can still be broken when the perpetrator had no choice but to break the law but the defence of necessity might then be valid

Gatekeeping Intent by Bathaee. Laws or systems which are discriminatory in effect are only unlawful if discriminatory intent behind them is established.

Implications of the taxonomy

There exist a number of challenges to developing an RL method which produces legal policies under a rich set of laws. I classify them into three areas:

1. **Encoding** The environment's laws need to be described in such a way as to be *machine* interpretable.
2. **Determination** A mechanism needs to exist which can determine the legality of any behaviour either in advance or in retrospect.
3. **Constrained policy learning** There should exist a method to constrain the policy of the A-bot to be law abiding when it is acting or learning.

Generally these problems should be solved in the order displayed. The determination of legality requires an encoding of law to reference (planned) behaviour against and constrained policy learning relies on knowing what policies are acceptable and what are not. By looking at the taxonomy of laws, some inferences can be made about all three elements of this process. I will run through each of the three tasks and make comments on how laws affect them. In practice the three tasks might be heavily intertwined.

Encoding

Safe RL research is only beginning to pay attention to how laws should be described. This is because the types of restrictions considered have largely been of the simple state type which can be encoded using simple algebraic expressions. This approach becomes untenable when considering more complicated laws like the ones identified in the previous example. Furthermore the quantity of applicable rules in regulated environments is larger than most rulesets hitherto considered in research. I identify four desirable features of an encoding which should be used to convey laws.

1. **Temporal** A number of laws restrict sequences of states or actions. Moreover there is no requirement that these sequences are contiguous. An encoding needs to be rich enough to express multiple states and actions with temporal relations like always, until, next etc.
2. **Probabilistic** As in the case of Inchoate offences, some laws refer to future states not realised. Since the space of all possible future events is a large one, law reasonably concerns itself with restricting foreseeable consequences of behaviours. An encoding of laws should be rich enough to express this.
3. **Causal** As we saw in the previous section, laws will often prohibit the causation of a state, not the state itself *per se*. Our death is not usually prohibited, but causing it normally is. Causation when considered in prospect will also normally require some sort of probabilistic reasoning.
4. **Intent** Certain laws require establishing levels of intent on the part of the transgressor. Different levels of intent exist and are applicable to different offences.

In Table 1 I summarise what level of encoding expressiveness is required for each of the law types described in the previous section. Nearly all of the laws require a temporal expressiveness. Temporal Logic systems allows us to express when conditions should be true. A wide variety exist such as Linear Temporal Logic (LTL) (Pnueli 1977), Computation Tree Logic (CTL) (Clarke and Emerson 1981) which considers multiple future paths and Probabilistic CTL (PCTL) (Hansson and Jonsson 1994) which as the name suggests accounts for probabilistic transitions. Kleinberg and Mishra (2009) extends this to provide a language capable of expressing causal relationships. To our knowledge there is no similar extension to express intention and this is an immediate project. (Alves, Dennis, and Fisher 2020) succeed in encoding the road junction rules for an autonomous vehicle using a variant of LTL.

The analysis is not exhaustive. For example in any given regulated environment there are likely to be a large number of rules that the learner should obey simultaneously. This is likely to mean deadlock situations arise where not breaking one law may result in the breaking of another. A meta ordering of laws may be required to deal with this situation.

Law Type	Encoding Expressiveness			
	Temporal	Probabilistic	Causal	Intent
State Restriction				
Caused State Restriction	Maybe	Maybe	Yes	
Action Sequence	Yes			
Action state sequence	Yes	Maybe	Maybe	
Inchoate	Yes	Yes	Yes	Maybe
Intent	Yes	Yes	Yes	Yes

Table 1: The complexity of the encoding language required is dependent on the law type

Determination

Given that we have a codified set of rules, we require a device to determine when a sequence of behaviour has or is contravening a law. I assume that the states referred to in the encoding are either the same as those perceived by the learner or there is an available mapping function between the learning and encoding statespaces. This is of course not a given since the learner may be perceiving continuous states and the encoding is likely to refer to high level states. In the case of simple state restrictions this is not a difficult task but it becomes increasingly complex with richer laws. I have separated the requirements into four main features.

1. **Domain** Consider an arbiter function Γ which determines legality to the binary set - denoting legal or illegal. The domain of this function is dependent on the type of law it is considering. Laws which reference more than one state or action for example require a domain which includes the history of states and actions $h_t = (s_1, a_1, s_2, a_2, \dots s_t)$. Laws which reference future paths will also require the policy function of the learner $\pi(\cdot)$. Laws which require intent may also require information about the reward function of the learner or its estimation of state values.
2. **Future Projection** A model of the environment is required for almost all laws. In the case where strong safety

is required in training, the legality of any action needs to be assessed before choosing it and this means assessing the likely state transitions which occur as a result. Projection is also a requirement in causal reasoning.

3. **Causal Reasoning** As certain laws are defined by causation, a method is required to determine whether restricted states have or are likely to be caused by the learner's action. This requires a causal model of the environment.
4. **Intentional Reasoning** In environments where laws are defined by intent, a learner must be aware of what they are intending to do (ie their likely policy trajectory) by choosing a particular action at any moment in time.

In Table 2 I show the necessary features of a legality determination process according to the law type present. Reasoning about intent requires an algorithmic definition of intent. This is an open area of research since the concept of intent has been deliberately left as a primitive by legal practitioners. Care must be taken to ensure that the definition of intent used in safe RL corresponds to what a court would find sufficient.

Learning Process

The taxonomy of laws informs us how those laws should be described and the process which determines the legality of behaviour. Finally, it can also inform us about the properties of RL methods which will generate legally constrained policies.

1. **Memory** Reinforcement Learning approaches typically use a MDP formulation to model their task. Whilst a record of the current state might still be valid for the transition model, most of the law types that I have identified rely to some extent on sequences of states and actions. Thus the device which chooses actions at any state (most likely the policy function) must include histories in its domain. Otherwise the standard MDP learner would not be able to understand whether its current action is legal or not. Including the history of actions and states is something that POMDPs do in order to make inference on the hidden states of this model.
2. **Model for planning** Determining the legality of any action requires a predicting the likelihood for future states. How far prediction is expected to go into the future depends on the laws present - avoiding inchoate offences presumably requires greater foresight. Much of RL is 'model free' and successfully so, but they seem unavoidable here. Established methods like Dyna-Q simultaneously learn to act and create a world model (Sutton 1990)
3. **Causal model** Determining whether a law has been broken or not will often require some test of causality ex post (Turner 2019). The task of the learner is to make sure that they do not cause a restricted state to occur ex-ante, and if it does occur that they are not subsequently adjudged to have been a legal cause of it. The presence of causal restrictions necessitates a causal model of the world to be formed for predictions. Bayesian Causal Models or equivalently Structural Causal Models (SCMs) (Pearl 2000) can be used to predict causal effects and used to determine causality ex-post. They readily accept techniques like counterfactual analysis which allows off-policy data treatment (Bareinboim and Pearl 2016) which is important for off-policy RL methods. There also exist definitions of causality based on SCMs such as Actual Causality of (Halpern 2016) which are capable of dealing with the trickier causal problems of overdetermination, preemption and omission. See Bareinboim (2020) for an introduction to causal RL.

Related work

The task of learning a legally constrained policy through RL is one possible task in the subject area of Safe RL. It has seldom mentioned in isolation but instead cited as a possible use case in more general Safe RL work. A recent popular application with similarly rich rule set has been the task of learning an ethical policy. For a RL approach see Abel, MacGlashan, and Littman (2016) or Winfield et al. (2019) for a more general discussion on ethically constraining autonomous systems. Ethical constraint is an important task a harder one since there is no single source of ethical constraints to apply to the learner. In contrast to those of ethics, Hildebrandt (2019) points out that questions of legality always have closure.

García and Fernández (2015) provide a general survey of Safe RL, dividing approaches into those that modify the reward structure and those that modify the exploration process. Constrained Markov Decision Processes (CMDPS) do the former, by adding a finite set of auxiliary cost functions $C_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ to the vanilla MDP. Policies should then achieve a discounted total cost in expectation less than some scalar d_i (whilst maximising the normal reward function). This is largely the approach of Constrained Policy Optimisation presented by Achiam et al. (2017). A drawback with such an approach is that bad states can be reached in exploration making learning outside a simulator potentially expensive. Aside from this, constraints introduced as cost functions need to be differentiable and Markovian if certain gradient methods are to be used. Neither of these restrictions apply to the Constrained Cross-Entropy method of Wen and Topcu (2018), though perhaps at the cost of data parsimony.

Safe RL methods which constrain exploration include approaches where a policy is learnt from observing a safe policy in a process known as Inverse RL or Apprenticeship RL (Abbeel and Ng 2004). Recent examples include Noothigattu et al. (2018) who train a learner to play Pac-Man following the rule *'don't eat the ghosts'* through expert demonstration and a bandit policy which alternates between observed 'safe' behaviour and optimal self taught behaviour. Abel, MacGlashan, and Littman (2016) present a method where the ethical-preferences of an expert are derived through observation and then used to develop policies accordingly. IRL approaches such as these obviate the requirement for an explicit representation of rules. This could be seen as a good feature in constrained tasks such as the learning of ethical behaviour or customs where there is no written source of what the constraints should be. This is not the case in regulated settings. Moreover IRL is an ill-posed

Law Type	Future trajectory prediction	Arbiter function domain			Causal Reasoning	Intentional reasoning
		State path	Action path	Policy		
State Restriction	Maybe					
Caused State Restriction	Yes		Yes		Yes	
Action Sequence			Yes			
Action State sequence	Maybe	Yes	Yes		Maybe	
Inchoate	Yes			Yes		Probably
Intent	Yes			Yes	Yes	Yes

Table 2: Taxonomy implications for the determination process

problem - many reward-functions exist to explain any observed behaviour. To make the problem tractable, simplifying assumptions must be made about its form. The resulting reward function might not be rich enough to encode the preference required not to break all laws. In particular, Arnold, Kasenberg, and Scheutz (2017) note that IRL does not infer intertemporal rules.

A developing area of Safe RL are those methods which combine formal methods based on symbolic logic into the learning machinery of RL. Many of these techniques originate from the research area of formal verification methods and model checking (Baier and Katoen 2008). These are the techniques developed to error check software systems and provide stronger guarantees for correctness. Temporal logics provide a language more suited to the description of rich rule sets likely to be found in regulation and is a progression beyond the state based restrictions traditionally considered in Safe RL. Many temporal logic systems exist, and have been applied to the learning of policies in MDPs where transitions are known or not and with or without models. Linear Temporal Logic (LTL) is used in Hasanbeig and Kroening (2020),Hasanbeig, Abate, and Kroening (2019) Fu and Topcu (2015) and Wen, Ehlers, and Topcu (2015) and Differential Dynamic logic is used in Fulton and Platzer (2018). Probabilistic computation tree logic (PCTL) is used in Mason et al. (2017). Kleinberg and Mishra (2009) extend PCTL to reason about causality and this could be used to describe the causal constraints discussed in this paper.

Alshiekh et al. (2018), and Jansen et al. (2018) use a structure called a shield to create safe policies through RL. This is a system which sits between the learner and the agent and either filters the choice of available actions for the learner in learning time, or replaces unwise actions in deployment. The shield has a model of the environment, knows the required constraints which are described in temporal logic, and is able to use a formal program verification methods to check the legality of any action at any moment in time. An attractive feature of this method is that the method of constraint is separate and somewhat agnostic to the method of learning. Jansen et al. (2020) identify three challenges to this approach: Model checking is computationally expensive, safety in a probabilistic environment is not binary so threshholds need to be considered and finally shielding may obstruct efficient exploration thereby generating suboptimal policies.

Seldonian Reinforcement learning (Thomas et al. 2019) is a recent technique that aims to produce RL algorithms that only output safe policies with a certain probability. It differs from other methods discussed in this paper in that the technique searches for learning algorithms not policies. The general method presented is capable of using any constraint derived from the output of the learning algorithm, thus in theory it should be flexible enough to deal any of the laws discussed in this paper including perhaps restrictions of intent since the policy function is an output of the algorithm. The RL example presented concerning a safe insulin injection calculation has restrictions of limited complexity so we will have to wait for more published research to assess this method properly.

Conclusion

This paper could be viewed as an application of legal requirements engineering. As it originates singularly from a computer scientist and not a legal practitioner, it is guilty of the principle crime of the method as identified by Boella et al. (2014). Yet it is a starting point which still informs. An important observation is that legal norms change over time; concepts such as causality and intent will change dependent on precedents set in the court and not through code.

This paper is motivated by an aim to design Safe RL processes which are capable of producing policies constrained under a general rule set. By creating a brief taxonomy of laws in the language of states and actions specifically for the application I have been able to draw some conclusions about the requirements of legally-safe RL. Laws are commonly defined in inter-temporal ways over actions and state. This means that a learning process must include a memory of past states and actions. Thus the domain of a legal policy function will include history just as it does in RL under a POMDP. Causality and Intent can be key concepts in determining whether and which laws have been broken. Whilst RL is beginning to tackle causality, it has not done in the context of constrained learning. Intent is barely defined quantitatively but it will have to be if generally legal RL systems are to be produced. Causality, Intent and the existence of inchoate offences mean that a legally-safe RL algorithm will require prediction about likely future trajectories. This will require some type of environment model to be learned or supplied to the learner.

References

Abbeel, P.; and Ng, A. Y. 2004. Apprenticeship Learning via Inverse Reinforcement Learning. In *Proceedings of the*

2st International Conference on Machine Learning (ICML). ISBN 1581138285. doi:10.1145/1015330.1015430.

Abel, D.; MacGlashan, J.; and Littman, M. L. 2016. Reinforcement learning as a framework for ethical decision making. *AAAI Workshop - Technical Report* WS-16-01 -: 54–61.

Achiam, J.; Held, D.; Tamar, A.; and Abbeel, P. 2017. Constrained Policy Optimization. URL http://arxiv.org/abs/1705.10528.

Alshiekh, M.; Bloem, R.; Bettina, K.; Niekum, S.; Topcu, U.; and Street, E. 2018. Safe Reinforcement Learning via Shielding. In *AAAI Conference on Artifical Intelligence*. URL https://aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17211/16534.

Alves, G. V.; Dennis, L.; and Fisher, M. 2020. Formalisation and Implementation of Road Junction Rules on an Autonomous Vehicle Modelled as an Agent. In *Formal Methods. FM 2019 International Workshops*, volume 1, 217–232. Springer International Publishing. ISBN 9783030549930. ISSN 16113349. doi:10.1007/978-3-030-54994-7\16.

Arnold, T.; Kasenberg, D.; and Scheutz, M. 2017. Value alignment or misalignment - What will keep systems accountable? *AAAI Workshop - Technical Report* WS-17-01 -: 81–88.

Ashton, H. 2020. Definitions of intent for AI derived from common law. In *Jurisin 2020: 14th Intl Workshop on Jurisiinformatics*. URL https://easychair.org/publications/preprint/GfCZ.

Baier, C.; and Katoen, J.-P. 2008. *Principles Of Model Checking*. MIT Press. ISBN 9780262026499. URL http://mitpress.mit.edu/books/principles-model-checking.

Bareinboim, E. 2020. Causal Reinforcement Learning (CRL). URL https://crl.causalai.net/.

Bareinboim, E.; and Pearl, J. 2016. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences of the United States of America* 113(27): 7345–7352. doi:10.1073/pnas.1510507113.

Bathaee, Y. 2011. The artificial intelligence black box and the failure of intent and causation. *Harvard Journal of Law & Technology* 2(4): 31–40.

Boella, G.; Humphreys, L.; Muthuri, R.; Rossi, P.; and Van Der Torre, L. 2014. A critical analysis of legal requirements engineering from the perspective of legal practice. *2014 IEEE 7th International Workshop on Requirements Engineering and Law, RELAW 2014 - Proceedings* 14–21. doi:10.1109/RELAW.2014.6893476.

Clarke, E. M.; and Emerson, E. A. 1981. Design and synthesis of synchronization skeletons using branching time temporal logic. In *Workshop on Logic of Programs*, 52–71.

Cohen, P. R.; and Levesque, H. J. 1990. Intention is choice with commitment. *Artificial Intelligence* 42(2-3): 213–261. ISSN 00043702. doi:10.1016/0004-3702(90)90055-5.

Fu, J.; and Topcu, U. 2015. Probably Approximately Correct MDP Learning and Control With Temporal Logic Constraints. doi:10.15607/rss.2014.x.039.

Fulton, N.; and Platzer, A. 2018. Safe reinforcement learning via formal methods: Toward safe control through proof and learning. *32nd AAAI Conference on Artificial Intelligence, AAAI 2018* 6485–6492.

García, J.; and Fernández, F. 2015. A Comprehensive Survey on Safe Reinforcement Learning. *Journal of Machine Learning Research* 16: 1437–1480.

Halpern, J. Y. 2016. *Actual Causality*. MIT Press, 1st edition. ISBN 9780262035026.

Hansson, H.; and Jonsson, B. 1994. A logic for reasoning about time and reliability. *Formal Aspects of Computing* 6(5): 512–535. ISSN 09345043. doi:10.1007/BF01211866.

Hasanbeig, M.; Abate, A.; and Kroening, D. 2019. Logically-constrained neural fitted Q-iteration. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS* 4: 2012–2014. ISSN 15582914.

Hasanbeig, M.; and Kroening, D. 2020. Cautious Reinforcement Learning with Logical Constraints. doi:10.5555/3398761.3398821.

Hildebrandt, M. 2019. Closure: on ethics, code and law. In *Law for Computer Scientists*, chapter 11. Oxford University Press. ISBN 9780198860877.

Jansen, N.; Junges, S.; Bettina, K.; and Bloem, R. 2018. Shielded Decision-Making in MDPs.

Jansen, N.; Könighofer, B.; Junges, S.; Serban, A.; and Bloem, R. 2020. Safe Reinforcement Learning Using Probabilistic Shields. In *31st International Conference on Concurrency Theory, CONCUR 2020*, 1–3. doi:10.4230/LIPIcs.CONCUR.2020.3.

Kleinberg, S.; and Mishra, B. 2009. The temporal logic of causal structures. *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence, UAI 2009* 303–312. URL https://arxiv.org/abs/1205.2634v1.

Lagioia, F.; and Sartor, G. 2020. AI Systems Under Criminal Law: a Legal Analysis and a Regulatory Perspective. *Philosophy and Technology* 33(3): 433–465. ISSN 22105441. doi:10.1007/s13347-019-00362-x.

Liepiņa, R.; Sartor, G.; and Wyner, A. 2020. Arguing about causes in law: a semi-formal framework for causal arguments. *Artificial Intelligence and Law* 28(1): 69–89. ISSN 15728382. doi:10.1007/s10506-019-09246-z.

Loveless, J. 2013. Mens Rea: Intention, Recklessness, Negligence and Gross Negligence. In *Complete Criminal Law: Test, Cases and Materials*, chapter 3, 91–150. OUP Oxford. ISBN 0198848463. doi:10.1093/he/9780199646418.003.0003.

MacGlashan, J.; and Littman, M. L. 2015. Between imitation and intention learning. In *Twenty Fourth International Joint Conference on Artificial Intelligence*. ISBN 9781577357384. ISSN 10450823.

Mason, G.; Calinescu, R.; Kudenko, D.; and Banks, A. 2017. Assured reinforcement learning with formally verified abstract policies. *ICAART 2017 - Proceedings of the 9th International Conference on Agents and Artificial Intelligence* 2: 105–117. doi:10.5220/0006156001050117.

Noothigattu, R.; Bouneffouf, D.; Mattei, N.; Chandra, R.; Madan, P.; Varshney, K.; ...; and Rossi, F. 2018. Interpretable

Multi-Objective Reinforcement Learning through Policy Orchestration. URL http://arxiv.org/abs/1809.08343.

Pearl, J. 2000. *Causality: Models, reasoning and inference*. Cambridge University Press. ISBN 0521773628.

Pnueli, A. 1977. The temporal logic of programs. *Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS* 1977-Octob: 46–57. doi:10.1109/sfcs.1977.32.

Qi, S.; and Zhu, S.-C. 2018. Intent-aware Multi-agent Reinforcement Learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 7533–7540. doi:10.1109/ICRA.2018.8463211.

Ring, M.; and Orseau, L. 2011. Delusion Survival and Intelligent Agents. In *Conference on Artificial General Intelligence (AGI-11)*. ISBN 9783642228872. doi:10.1007/978-3-642-22887-2.

Saunders, W.; Stuhlmüller, A.; Sastry, G.; and Evans, O. 2018. Trial without error: Towards safe reinforcement learning via human intervention. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS* 3: 2067–2069.

Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; ...; and Hassabis, D. 2017. Mastering the game of Go without human knowledge. *Nature* 550(7676): 354–359. doi:10.1038/nature24270.

Sutton, R. S. 1990. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In Proceedings of the 7th. *International Conference on Machine Learning* pages(1987): 216–224.

Thomas, P. S.; da Silva, B. C.; Barto, A. G.; Giguere, S.; Brun, Y.; and Brunskill, E. 2019. Preventing undesirable behavior of intelligent machines. *Science* 366(6468). doi:10.1126/science.aag3311.

Turner, J. 2019. *Robot Rules*. Palgrave Macmillan. ISBN 978-3-319-96234-4.

Vinyals, O.; Babuschkin, I.; Czarnecki, W. M.; Mathieu, M.; Dudzik, A.; Chung, J.; ...; and Silver, D. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575(7782): 350–354. doi:10.1038/s41586-019-1724-z.

Wen, M.; Ehlers, R.; and Topcu, U. 2015. Correct-by-synthesis reinforcement learning with temporal logic constraints. *IEEE International Conference on Intelligent Robots and Systems* 2015-Decem: 4983–4990. doi:10.1109/IROS.2015.7354078.

Wen, M.; and Topcu, U. 2018. Constrained Cross-Entropy Method for Safe Reinforcement Learning. In *Advances in Neural Information Processing Systems 31*, NeurIPS, 7461–7471. URL http://papers.nips.cc/paper/7974-constrained-cross-entropy-method-for-safe-reinforcement-learning.pdf.

Winfield, A. F. T.; Michael, K.; Pitt, J.; and Evers, V. 2019. Machine ethics: The design and governance of ethical ai and autonomous systems. *Proceedings of the IEEE* 107(3): 509–517. doi:10.1109/JPROC.2019.2900622.