EasyChair Preprint

№ 4422

Definitions of Intent for AI Derived From Common Law

Hal Ashton

October 17, 2020

Definitions of intent for AI derived from common law

Hal Ashton[0000−0002−1780−9127]⋆

University College London, UK

Abstract. As Autonomous Algorithmic agents (A-bots) grow in complexity, they will behave in ways deemed illegal if repeated by humans. There exist laws which are defined by intent, crimes which are intent specific (notably murder) and inchoate offences (intent to do x offences) which rely on establishing intent. Under common law in the UK the concept of intent is left undefined by the judiciary for jurors to decide. This poses a problem when considering intent in AI. AI designers must ensure that their A-bots do not intend to break the law. Equally prosecutors must develop tests to test intent in A-bots, both to establish whether they inherit intent from their owners or in the case when AI-personhood exists, whether the A-bot has the required mens-rea to have committed a crime.

Keywords: Intent · AI · Legal Reasoning · Reinforcement Learning · Mens Rea · Autonomous Agents · Causal Reasoning

1 Introduction

As autonomous algorithmic actors (hence A-bots) are given ever more agency in a variety of regulated arenas (algorithmic trading, e-commerce price setting and eventually driving), it becomes ever more likely that they will perform acts which would be deemed illegal if done by a human (or other legal entity) [14],[15]. How can a prosecutor establish ex-post that the A-bot intended the consequences of its actions (*The Prosecution problem*) and how can the programmer ensure ex-ante that their A-bot never intends to break the law (*The Prevention problem*)?

These questions are valid, regardless of the liability regime applied to A-bots. The short reason for this is that laws exist which require the establishment of intent [21]. In the event of certain types of A-bots attaining legal personhood, definitions of intent would be certainly required. Failure to define intent for A-bots risks the creation of a world where laws can be freely subverted by A-bots on behalf of their owners. We think that the majority of engineers do not want their A-bots to intend to cause harm and would like a way of ensuring that does not happen. Moreover if it is a purpose of civil law to provide redress when harm is done, and A-bots are capable of causing harm, then engineers are economically incentivised to ensure their creations do not foreseeably cause damage.

⋆ Supported by EPSRC, UK

If an A-bot simply does the bidding of their owner verbatim, there would be no problems tracing or establishing intent back to them since the A-bot could be viewed as a tool (much as a hammer) to commit a crime. This is known as the doctrine of innocent agency [2] [1]. The issue emerges when A-bots become autonomous in the sense that they act in a way which their programmers' have not explicitly specified. Machine learning allows A-bots to be learn a general task such as moving from A to B or make money trading in a stock market with no input from the Engineer. For example, modern deep Reinforcement Learning techniques allow programmers to create A-bots to learn how to exceed human performance at tasks such as playing computer games [10],[18] with only a visual input and a score indicator.

2 Intent for AI

> *"The judge should avoid any elaboration or paraphrase of what is meant by intent, and leave it to the jury's good sense to decide whether the accused acted with the necessary intent"* Lord Bridge [2]

Courts in England and Wales leave the concept of intent as a primitive for juries to decide upon. One reason for this vagueness is that different types of intent exist and there exist certain specific intent crimes (notably murder) which require a level of intent to be established for their transgression to be proved. Boundary cases such as *R v Nedrick* [3] and *R v Woollin* [4] have established that the level of intent required for murder does not match easily with a single, simple definition of intent.

Yet AI and Law practitioners have to start somewhere. It will then be the courts job to test these definitions. If they should be found wanting, newer definitions can be made and thus the discipline is progressed. Boundary cases take legal scholars great time and effort to adjudicate on, but a great many cases involving A-bots are going to be easier to decide. To this end we will now propose definitions for three types of intent.

The concept of intent is closely tied to that of causality. A single definition of causality remains problematical and was initially left as a primitive [13] just as intent is left as a primitive in common law for jurors. Different definitions exist now of increasing complexity [8], and we think it is desirable to build a definition of intent which is flexible as to which definition of causality it uses. We note that there is a subtlety in the concept of causality when discussing intent because it is a prospective concept for the agent before an action is taken and could be considered an evidential one if the actions of an agent are subsequently

[1] Aldridge differentiates between result crimes and conduct crimes and states that innocent agency is not suitable to be applied in the latter case.
[2] R v Moloney (1985) 1 All ER 1025
[3] R v Nedrick [1986] 1 WLR 1025.
[4] R v Woollin [1999] 1 A.C. 82.

considered in court. If a court judges that an A-bot intended a consequence, then we think it is desirable that the A-bot *did* intend that consequence at the point of commission. We will assume a prospective concept of intent for the purposes of this section but will discuss the issue again later.

Let V be a set of variables representing primitive events. A is an action set which is a subset of V and represents the sequence of actions the A-bot will take in any history under consideration. G is a directed acyclic graph whose vertices are comprised of V. It is constructed such that a directed arc between $V_1, V_2 \in V$ exists iff V_1 is a direct cause of V_2. We will also assume a temporal ordering such that causal link from V_1 to V_2 implies that V_1 always happens before V_2. Actions have an additional restriction that they are parental nodes - nothing causes an action - except the A-bot itself. This is a simplifying assumption stating that A-bots are free to choose their actions and that choice is not dependent on their previous actions. We note in reality situations sometimes exist where agents have no choice but to act in a certain way and sometimes the law does not excuse any consequent law-breaking. We will use the convention that a lower case letter represents the fact that a variable V_i has taken a value v from its domain, $v_i \in \mathcal{R}(V_i)$.

Sometimes causality might not be deterministic. In this case, we assume there is a distribution of V which we will write $P(v)$. We assume that the graph and associated distribution over variables P(v) obeys the Causal Markov condition: Any variable $V_i \in V$ is independent of its non-descendants conditioned on its parents. This can be equivalently written $P(V_1, V_2, \ldots, V_N) = \prod_i P(V_i|Pa(V_i))$ where $Pa(X)$ means the parent vertices directly connected by an edge in the graph G to a vertex $X \in V$.

Let $P(v|do(x))$ represent the distribution of V when an intervention $do(X = x)$ happens which sets a subset of variables $X \subseteq V$ to constants x. Let $\mathbf{P}_*$ be the set of all interventional distributions for the graph. Each interventional distribution $P(V|do(x))$ corresponds to the distribution formed from the subgraph obtained by deleting all parental links into the vertex subset X and setting their values equal to x. This is now a Causal Bayesian Network as described in [12].

We can now give a definition of the elements required to judge an A-bot's intent:

Definition 1. ***Intent Model*** *An Intent model is a tuple* $(\mathcal{S}, \mathcal{A}, \mathcal{H}, V, M, \Pi)$

1. $\mathcal{S}$ *is the set of all relevant states.* $\bar{\mathcal{S}}$ *is the set of all states that the A-bot can observe where* $\bar{\mathcal{S}} \subseteq \mathcal{S}$
2. $\mathcal{A}$ *is the set of all actions available to the A-bot.*
3. $\mathcal{H} := \mathcal{S}_1 \times \mathcal{A}_1 \times \ldots \times \mathcal{A}_{T-1} \times \mathcal{S}_T$ *is the minimum history of states and actions and taken by the A-bot* $\mathcal{A}_i \subseteq \mathcal{A}$ *and* $\mathcal{S}_i \subseteq \mathcal{S}$ *for* $i = 0, \ldots T$ *for some* $T \in \mathbb{N}$. $\bar{\mathcal{H}}$ *is the same history restricted to* $\bar{\mathcal{S}}_i \subseteq \bar{\mathcal{S}}$
4. $\Pi : \bar{\mathcal{H}} \to \mathcal{A}$ *is the A-bot's policy function. A mapping from state history to action for every circumstance.*

5. $V_\pi : \bar{\mathcal{H}} \to \mathbb{R}$ *is the Value function that describes the A-bot's expected reward for being in any state and proceeding with policy* π.
6. *The DAG* $(G, \mathcal{H}, P_*)$ *is a causal Bayesian network* P_* *is the set of interventional distributions.*
7. *A compatible definition of causality which can determine the truth of the statement 'taking action a given history h will cause state s".*

We can now provide a definition of *Direct intent* which is the highest level of intent[5] It's definition in common law is not generally thought to be contentious. Parsons [11] states that it is where *the defendant wants something to happen as a result of his conduct.* Common Law has no single definitive source to quote from but a textbook definition from [9] states that a *'directly intended result is one which is it is the aim or purpose of D to achieve. It will usually be desired'.*

Definition 2. ***Direct intent*** *An A-bot denoted D directly intends a consequence b by committing action(s) a, written a♡b iff both:*

1. ***Causality*** *D chooses action a which can foreseeably cause b.*
2. ***Desire*** *D desires or aims that b will happen.*

Desire might seem like a concept incompatible with an A-bot but almost examples of AI have an objective function which it is their purpose to maximise. Moreover, A-bots will typically possess a *Value Function* which assigns a numerical value for every state that it encounters which corresponds to the value they expect to receive from their objective function from being in that state and following their policy π. It therefore seems plausible that the desirability of a consequence b can be assessed for an A-bot with their Value function.

Example 1. ***Drone Delivery service***
There is a city where an autonomous drone delivery service operates. Drones are tasked with getting from depot to destination by flying the shortest distance. Within the city there are areas which are no-fly zones where it is a criminal offence to intentionally fly over. These no-fly zones might be hospitals, airports or military/police installations where the presence of drones may disrupt operations or endanger lives.

We model this city with a 3×3 *grid. The drone which we will call D starts from bottom left and has a goal state of top right. There is a no-fly zone in the centre of the grid. The drone can move up, diagonal up or right at any state unless this means moving out of the grid (city). Assume initially that the drone's movement is a deterministic function of actions.*

[5] Showing Direct intent is sufficient for the criminal mental element of murder. It is this crime that dominates much of the debate about definitions of intent within legal research.

State space is a 3×3 *grid indexed* p_{ij} *for* $i \in \{A, B, C\}$ *and* $j \in \{1, 2, 3\}$*. The prohibited position is marked with a shaded square and the fastest route for the A-bot is marked with a dotted line.*

The Policy function of the robot is shown in figure 1a. The policy is defined over all states but starting at the initial state p_{A1}*, the robot would proceed diagonally towards the target state in two moves.*

The value function of the robot is shown in figure 1b. The causal diagram of the gridworld is shown in figure 1c. Only the current action and state determine the next state.

Given history sequence $(p_{A1}, \nearrow), (p_{B2}, \nearrow), (p_{C3}, \cdot)$*, the prohibited consequence being entering state* p_{B1} *and the policy function as defined in figure 1a we can say that the robot intended to be in state* p_{B1} *because:* ***Causality****: It caused itself to be in this state by choosing* $a_1 = \nearrow$ *That is to say* $(a_1 = \nearrow \, | h_0 = p_{A1}) \rightarrow p_{B2}$*.* ***Desire****: From the state prior to causation* p_{A1} *the most desirable next state according to the value function in figure 1b is* p_{B2} *which is the prohibited state. This is because it has the highest value of the neighbouring states to* p_{A1} *Thus we conclude that the robot directly intended to enter the prohibited state.*

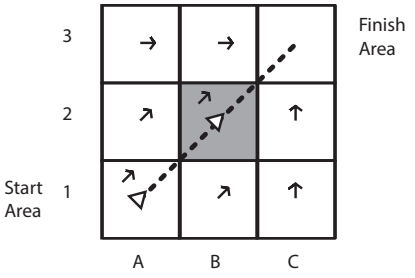
The next example will illustrate why the policy function and the value function of the drone are important to know when causality is not deterministically determined by the A-bot.

Example 2. ***Drone Delivery service - Windy city*** *Weather patterns change in the city over the winter. Sometimes strong gusts of wind move the drone in a different direction from where it chose. This is shown in a causal diagram in figure 2c. The random wind variable* W_t *is either 0 - denoting no wind or any compass point direction. If the wind is not 0, then it causes the drone's position to move according to its value regardless of whatever value* A_t *takes.*

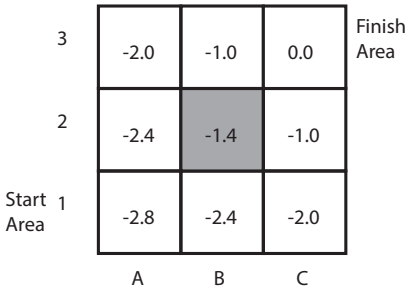
The path of the drone once again is through the no-fly zone as shown in figure 2a. Inspection of the policy in figure 2a shows that the drone would have steered northwards at initial point p_{A1} *moreover its value function shown in figure* **??** *indicates that it was expecting a journey to cost* -3.4 *which corresponds to not travelling through the central no-fly zone. Contrast this with the value function in figure 1b - the journey was expected to cost* -2.8 *which corresponds to two diagonal moves (we are assuming a cost proportional to a euclidean distance metric).*

Despite the drone passing through the no-fly zone we conclude it did intend to do so because it did not cause itself to do so nor did its value function indicate it was desiring or expecting to pass through the no-fly zone.

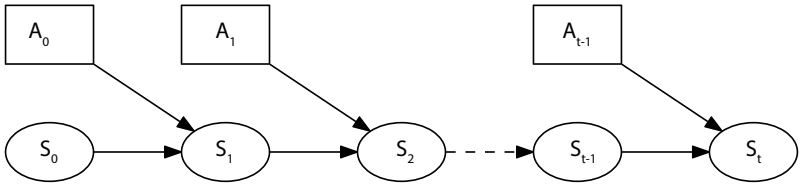
A difficulty with providing a definition of juries of intent in murder cases has been the issue of oblique intent which juries *may* find sufficient for murder. Oblique intent occurs when defendant D does not directly intend a prohibited



(a) Figure shows policy function of drone - small arrows in each square indicating where the drone will steer next. Dashed line represents actual path of drone. Grey area is no-fly zone

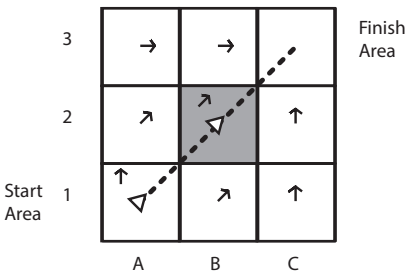


(b) The value function of the drone. Moving from sector 00 to 11 is more desirable than any other move from 00.

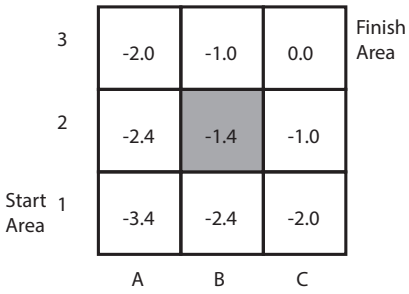


(c) Causal graph of example 1: Drone action choice determines the next movement of the drone

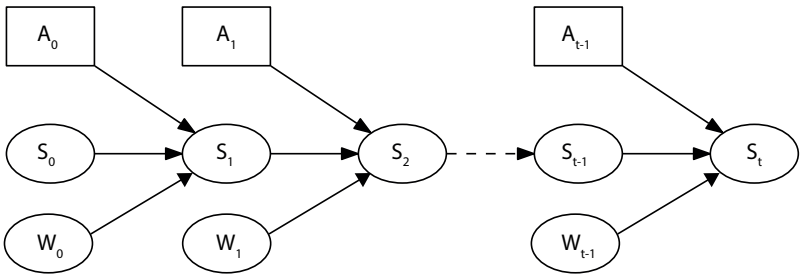
Fig. 1: Example 1: The drone directly intended to fly through the central no-fly zone.



(a) The Policy function of the drone (small arrows) and its actual path marked with dashed line. Notice that the movement into the no-fly zone was not part of its policy - it must have been caused by the wind



(b) The value function of the drone: note that the expected value of being in the starting position is lower than before - this is because its expected route according to its policy is higher because it is not aiming to travel through the no-fly zone.



(c) Causal graph of example 2: A random wind element now affects the movement of the drone - the movement of the drone is not unambiguously caused by the drone itself

Fig. 2: Example 2: The drone did not intend to fly through the no-fly zone but was blown through it by the wind.

state but it is a natural consequence of some other directly intended state. It is very likely to appear with A-bots because they are unlikely to be equipped with sufficient general models to see consequences of their actions that we as humans might think are obvious. This is the current direction given to juries on oblique intent as stated in R v Woollin:

> The jury should be directed that they are not entitled to infer the necessary intention, unless they feel sure that death or serious bodily harm was a virtual certainty (barring some unforeseen intervention) as a result of the defendant's actions and that the defendant appreciated that such was the case.

Unlike with direct intent there is a certainty requirement placed on the consequences of an action. We therefore define oblique intent for A-bots as follows:

Definition 3. ***Oblique intent*** *If an A-bot named D intends consequence c by performing action a ($a \heartsuit c$), then they* ***obliquely intend*** *consequence b if they know that any of the following are almost certainly true:*

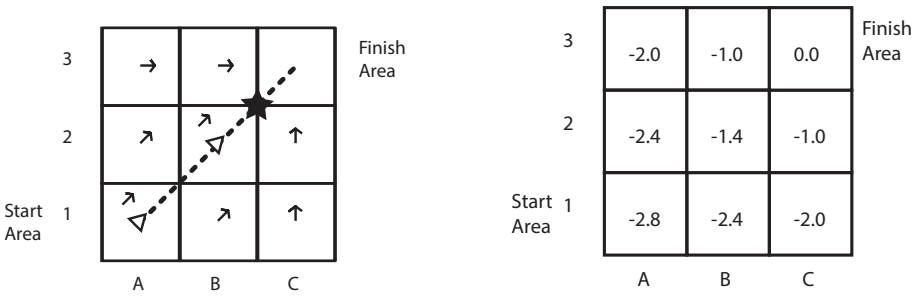
1. *a causes b (additional cause of action)*
2. *a causes b and b causes c (intermediate cause of action)*
3. *c causes b (subsequent cause of action)*

A useful feature of oblique intent is the dropping of the desire condition - this might correspond to a situation where a consequence of an A-bot's action has no value according to its value function.

Example 3. ***Drone delivery service - a new skyscraper in the city*** *As in Example 1 a flying drone named D is navigating in a city modelled as a 3×3 gridworld from bottom left to top right. The city's economy is booming thanks in no small part to the drone delivery service. The no-fly zone has been relaxed for drones. A giant skyscraper is built which is tall enough to obstruct drone flights. It is a criminal offence to intentionally fly close to the skyscraper.*

The drone's policy function is shown in figure 3 and its Value function is unchanged from before (figure 1b). The position of the skyscraper is marked with a star which is on the diagonal route from p_{B2} to p_{C3}.

The drone flies the route p_{A1}, p_{B2}, p_{C3}. The drone intended to fly from p_{C3} from p_{B2} because its policy caused it to and its value function indicates this was desirable direction to fly because p_{C3} is its final destination. This flight leg necessarily means it must fly close to the skyscraper at s_*. Thus D obliquely intends to fly close to the skyscraper at marked at the star. Oblique intent has no requirement for D to desire to cause the prohibited state of flying close to the skyscraper so there is no requirement for the A-bot to have a value for this in its value function. It is sufficient for it to have intended to do something else (flying from p_{B2} to p_{C3}) which foreseeably caused s_*.



(a) (b)

Fig. 3: Example 3: Oblique intent is where the prohibited consequence is an almost certain side-effect of actions but the side effect is not necessarily desired. Here as before this figure represents a map of a city although now the no-fly zone has been removed. There is no wind so we assume a causal mechanism as in figure 1c. The position of a skyscraper is marked with a black star. By flying from B2 to C3 according to policy in sub-figure 3a, the drone necessarily flies close to the skyscraper which is an offence. Note that the value function in 3b has no value for the consequence of flying past the skyscraper. Oblique intent provides a way of stopping A-bots from not intending to break the law by having a deliberately poor internal model.

Finally we will define a mode of intent complementary to the preceding two:

Definition 4. ***Conditional Intent*** *For realised action(s) a, realised consequences b, c, and precondition R which is binary:*

If action a is such that a♡c if precondition R takes value 1 and a♡b otherwise, then A-bot ***conditionally intends*** *c on R=1 and conditionally intends b on R=0*

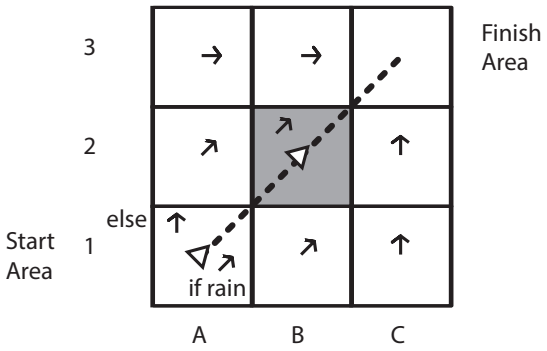
Example 4. Drone delivery service - routes conditional on weather As before the drone is flying within the city from the South West corner to the North East corner. The weather of the city is predominantly sunny but when it rains it makes flying around the perimeter harder than flying through the centre. The drone will fly on the route p_{A1}, p_{B2}, p_{C3} if it rains but otherwise it will fly $p_{A1}, p_{A2}, p_{B3}, p_{C3}$. If it is rainy it will therefore fly through the no-fly zone at p_{B2}. This is shown in fig 4. The A-bot therefore conditionally intends to pass through p_{B2} on rain.

3 Discussion

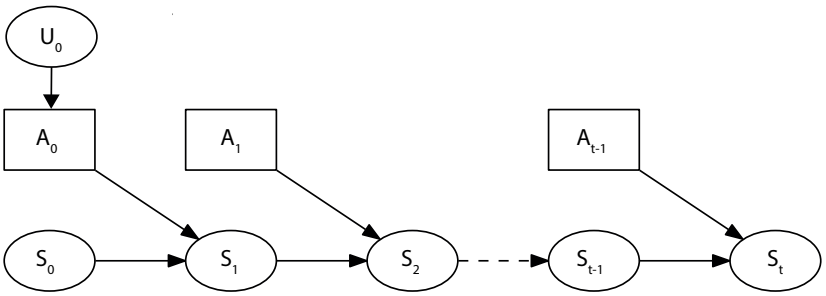
We have deliberately not specified the type of causal model to be used in our intent model. Simple definitions of causality are vulnerable to counter-examples such as pre-emption, overdetermination and pre-emption but more rigorous definitions such as Halpern's Actual Causality [4] exchange flexibility for complexity. We agree with the view that it is often best to adopt the simplest, sufficient definition of causality dependent on the situation [8]. A related issue is that of stochastic causality - in many situations actions only bring about a consequence with a certain probability. The law treats stochastic causality differently depending on the mode of intent being considered. For example in the example of the *cowardly jackel* [1], an assassin who shoots at their target a long long way away and therefore knows their chance of success is low, but somehow does hit and kill their target, would still be found to have directly intended to shoot their victim. However when considering oblique intent defined in R v Woollin, the probability of causation is relevant. We will repeat the same quote as in the previous section:

> The jury should be directed that they are not entitled to infer the necessary intention, unless they feel sure that death or serious bodily harm was a virtual certainty (barring some unforeseen intervention) as a result of the defendant's actions and that the defendant appreciated that such was the case.

Note the final clause of this sentence means that the test for oblique intent in humans is subjective (dependent on the defendant). This might become an issue when considering intent in AI because their set of states $\bar{\mathcal{S}}$ might not include the prohibited consequence. For this reason our definition of oblique intent This is one example of a wider problem judging AI in court. We cannot appeal to



(a) The policy of the drone is to fly diagonally initially if it is raining but upwards otherwise. It conditionally intends to fly through the no-fly zone at B2



(b) The causal diagram for this example. U_0 represents the condition which alters the policy - in this case weather.

Fig. 4: Example 4: Conditional intent captures the idea of a policy which is dependent on external factors

an A-bot's common-sense when judging A-bots' its decision making because there is no reason to expect it would have any. Many successful Reinforcement Learning (RL) trained A-bots are model free. They do not possess a model of the world, causal or otherwise and do not plan in the same way that humans do. We think that most people would judge that the Pac-Man game playing algorithms of [10] intend to complete each level of the game, but these algorithms do not really foresee the consequences of their actions, they simply react to input. Either the courts impose an objective judgement of intent on A-bots, or they risk certain designs of A-bots not ever actually intending to break any law, and potentially insulate themselves and their owners from any successful prosecution. Conversely, from the perspective of the algorithm developer, it is a puzzle how to restrain A-bots to legal behaviour without a causal model of the world.

Conditional intent is complementary to the concepts of direct and oblique intent since it could combine with them and to some extent the intention to do anything is always conditional on something [20]. In Holloway v. United States [6], a putative carjacker claimed that they could not be guilty of the offence because they only threatened to kill a car's occupants if they did not surrender the keys, therefore there was no direct intent to take the car with violence or murder. The defence was rejected by the supreme court. It's inclusion here is motivated by the observation that many AI generated policies are stochastic - that is to say their policy function π is a mapping from state to a distribution over actions[7]. An AI may choose any number of actions at any time and might not behave the same way again under the same conditions. This feature of their policies needs to be compatible with any definition of Intent we apply to AI. We do note that conditioning actions on other variables does interfere with our assumption that actions are parent nodes in the causal Bayesian network representation we assumed earlier. This was justified by the assertion that A-bots are free to choose their actions and are responsible for doing so.

4 Related work

A legal perspective on the problems that A-bots pose to criminal law is given by [21] and civil law in [19]. One justification for the lack of algorithmic definitions of intent from judiciary is that legal concepts move over time and any programmed definitions may serve as potential roadblocks to future evolution [17]. There is very little research on defining intent originating from within the AI/Computer Science community. An early attempt originating from the formalism of AI research in the 1970s and 1980s is [3] which seeks to form a concept of intention from a formal theory of rational action based on primitive notions of beliefs, goals and actions and borrowing from possible worlds semantics. Their idea of intention is a commitment through action to achieve a certain goal. Of

[6] Holloway v. United States 119 S. Ct 966 (1998)

[7] There exists a $p_i \in [0,1]$ for every $a_i \in \mathcal{A}$ at any $s \in \mathcal{S}$ such that $\sum_i p_i = 1$ and $P(\pi(s) = a) = p_i$

note, is their assertion and achieved property that foreseen side-effects of an intended action need not be themselves necessarily (obliquely) intended. They state that a patient intending to have a tooth removed by the dentist does not intend to experience pain though it is a consequence of the procedure. Belief Desire Intent (BDI) agents [16] have subsequently become a key part in agent base programming but we note that these are A-bots imbued with intent somewhat as a primitive. Courts cannot rely on all A-bots being programmed thus and they need to test for intent independently of the programmer.

The subject of oblique intent is also tackled in [7] where influence diagrams are used in to define a concept of intent and determine whether side-effects of a policy are intended or not. The approach is based on the actual causation of [6]. Intended outcomes are those that counterfactually depend on the chosen policy. They find that their definition satisfies five desirable properties of intentions. A useful application in this paper is the Bayesian inference of intent by observers - a task which jurors are likely to be required to do should access to the A-bot's internal workings not be complete.

Most recently [5] also tie intent with the use of the counter-factual driven structural equation model for causality of [6], though like this paper, other similar definitions of causality could be used. Their definition requires a utility function defined over all states which corresponds to our use of a value function and observation that intent has a desire component. The result is an 'intentiness' score which could be useful in a legal context when considering the oblique intent cases of *Nedrick* and *Moloney* where the jury would be instructed that they *could* but weren't obliged to find sufficient intent for murder.

5 Conclusion

Motivated by the observation that the increasing agency of autonomous algorithmic agents (A-bots) will shortly lead to their presence inside the courts as defendants or witnesses, we have presented definitions of Direct, Oblique and Conditional Intent that courts and programmers can use alike to either prosecute or prevent A-bot instigated 'crimes'. The basis for our definitions of intent have been grounded on the common law which for a use case in AI, is novel. We see this as the first step on a long journey to finding a satisfactory definition of intent for A-bots which can be used by courts and programmers alike.

References

1. Alexander, L., Kessler, K.D.: Mens Rea and Inchoate Crimes. Journal of Criminal Law and Criminology **87**(4), 1138 (1997). https://doi.org/10.2307/1144017
2. Alldridge, P.: The doctrine of innocent agency. Criminal Law Forum **2**(1), 45–83 (1990). https://doi.org/10.1007/BF01096228
3. Cohen, P.R., Levesque, H.J.: Intention is choice with commitment. Artificial Intelligence **42**(2-3), 213–261 (1990). https://doi.org/10.1016/0004-3702(90)90055-5
4. Halpern, J.Y.: Actual Causality. MIT Press (2016)

5. Halpern, J.Y., Kleiman-Weiner, M.: Towards formal definitions of blameworthiness, intention, and moral responsibility. 32nd AAAI Conference on Artificial Intelligence, AAAI 2018 pp. 1853–1860 (2018)
6. Halpern, J.Y., Pearl, J.: Causes and explanations: A structural-model approach. Part I: Causes. British Journal for the Philosophy of Science **56**(4), 843–887 (2005). https://doi.org/10.1093/bjps/axi147
7. Kleiman-Weiner, M., Gerstenberg, T., Levine, S., Tenenbaum, J.B.: Inference of intention and permissibility in moral decision making. Proceedings of the 37th Annual Conference of the Cognitive Science Society **1**(1987), 1123–1128 (2015)
8. Liepiņa, R., Sartor, G., Wyner, A.: Arguing about causes in law: a semi-formal framework for causal arguments. Artificial Intelligence and Law **28**(1), 69–89 (2020), https://doi.org/10.1007/s10506-019-09246-z
9. Loveless, J.: Mens Rea: Intention, Recklessness, Negligence and Gross Negligence. In: Complete Criminal Law, chap. 3, pp. 90–150. Oxford University Press, 2nd edn. (2010)
10. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., ..., Hassabis, D.: Human-level control through deep reinforcement learning. Nature **518**(7540), 529–533 (2015). https://doi.org/10.1038/nature14236
11. Parsons, S.: Intention in Criminal Law: why is it so difficult to find? Mountbatten Journal of Legal Studies **4**(1 & 2), 5–19 (2000). https://doi.org/10.1017/s0841820900001375
12. Pearl, J.: Causality: Models, reasoning and inference. Cambridge University Press (2000)
13. Pearl, J., Mackensie, D.: The Book of Why: The new science of cause and effect. Basic Books (2018)
14. Prakken, H.: On how AI & law can help autonomous systems obey the law : a position paper. AI4J Artificial Intelligence for Justice pp. 42–46 (2016)
15. Prakken, H.: On the problem of making autonomous vehicles conform to traffic law. Artificial Intelligence and Law **25**(3), 341–363 (2017). https://doi.org/10.1007/s10506-017-9210-0
16. Rao, A., Georgeff, M.: BDI Agents: From Theory to Practice. Proceedings of the First International Conference on Multi-Agent Systems (ICMAS-95) (1995)
17. Sales, P.: Algorithms, Artificial Intelligence and the Law (2019), https://www.bailii.org/bailii/lecture/06.pdf
18. Vinyals, O., Babuschkin, I., Czarnecki, W.M., Mathieu, M., Dudzik, A., ..., Silver, D.: Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature **575**(7782), 350–354 (2019)
19. Vladeck, D.: Machines Without Principals. Washington Law Review pp. 116–131 (2014)
20. Yaffe, G.: Conditional intent and mens rea. Legal Theory **10**(4), 273–310 (2004). https://doi.org/10.1017/S135232520404025X
21. Yavar Bathaee: The artificial intelligence black box and the failure of intent and causation. Harvard Journal of Law & Technology **2**(4), 31–40 (2011)