Artificial Intelligence and Law
https://doi.org/10.1007/s10506-022-09322-x

ORIGINAL RESEARCH

Definitions of intent suitable for algorithms

Hal Ashton[1]

Accepted: 2 July 2022

Abstract
This article introduces definitions for direct, means-end, oblique (or indirect) and ulterior intent which can be used to test for intent in an algorithmic actor. These definitions of intent are informed by legal theory from common law jurisdictions. Certain crimes exist where the harm caused is dependent on the reason it was done so. Here the actus reus or performative element of the crime is dependent on the mental state or mens rea of the actor. The ability to prosecute these crimes is dependent on the ability to identify and diagnose intentional states in the accused. A certain class of auto didactic algorithmic actor can be given broad objectives without being told how to meet them. Without a definition of intent, they cannot be told not to engage in certain law breaking behaviour nor can they ever be identified as having done it. This ambiguity is neither positive for the owner of the algorithm or for society. The problem exists over and above more familiar debates concerning the eligibility of algorithms for culpability judgements that mens rea is usually associated with. Aside from inchoate offences, many economic crimes with elements of fraud or deceit fall into this category of crime. Algorithms operate in areas where these crimes could be plausibly undertaken depending on whether the intent existed in the algorithm or not.

Keywords Intent · Causality · Autonomous agents · AI crime · AI ethics

1 Introduction

Within criminal law it is a widely held concept that every crime has a performative element (actus reus) and a mental element (mens rea).[1] A person perform the actus reus with the mens rea to be said to have committed a crime. Satisfaction of these

[1] Excepting crimes of absolute liability where the mens rea element is minimal.

Supported by EPSRC.

✉ Hal Ashton
ucabha5@ucl.ac.uk

[1] University College London, London, UK

Published online: 25 July 2022

two elements is necessary but not sufficient for criminal culpability since, amongst other reasons there may be justifications for some behaviour which would otherwise be judged criminal. In addition, an actor can only be criminally culpable if they are capable of moral responsibility. Amongst others this mostly rules out children, the insane and algorithmic actors from being found culpable of committing crimes.

This article is not going to make any claims about the eligibility of algorithms for legal person-hood, blame, punishment or even praise and the role that algorithmic intent might play in that. This is not because I feel the subject is uninteresting or in any way adequately addressed in research, it just adjacent to the problem that this article aims to address. The only thing that this article requires of the reader is that they are open to the possibility that intent (and related mens rea states) can exist in an algorithm. I hope to show why it is necessary to understand, identify and control for intent in algorithms, for the criminal law (in its current state) to function in the way it was designed to.

This article will predominantly consider Anglo criminal law with a focus on England and Wales but also make some reference to other major common law jurisdictions like the USA. Mens rea definitions differ at the margin between different common law legal systems but their overlap is significant especially at the level of intent. In terms of wider applicability, the thrust of the article should also be applicable in any legal system where laws exist which forbid certain actions only when taken in the pursuit of certain objectives. This might include those countries with a civil criminal law, equally it might include other areas of law such as tort, securities or contract.

1.1 Mens rea also defines *'why-crimes'*

Mens rea, of which criminal intent is subcategory, plays functions within criminal law other than deciding the culpability of an actor for any given harm. Simester (2021) divides the functions into two categories. The first category considers how mens rea establishes the guilt of the actor's behaviour. The second concerns, what Simester terms the role of mens rea in the principle of legitimate enactment. That is to say how the law defines precisely what we are able to do without fear of criminal sanction and the balance that it makes between civil freedom and the protection of harmed people. This article will argue that forming a definition of intent in algorithms is necessary for reasons both of culpability determination and legitimate enactment.

For certain offences, mens rea can play an important role in identifying behaviour as being culpable wrongdoing in the first place. Specifically there is a subset of offences where the mens rea element informs the actus reus element concerning the wrongness of the behaviour. Restated, there are certain behaviours which are legitimate unless they are conducted with a certain purpose, as Simester puts it *'the actus reus does not identify anything we shouldn't be doing'*. For short, I will refer to these as *'why-crimes'*. Under the UK Attempts act 1981 for example almost every crime has a corresponding attempt crime. These inchoate crimes suffer from ambiguity in their actus reus because harm has often not yet been

caused and a certain amount of ambiguity might exist around many types of behaviour. Here the presence or absence of my intent to do harm at some point in the future, informs the actus reus. Perhaps because of this ambiguity, crimes cannot be attempted with recklessness, they must be done so with intent.

An engineer might argue that for an algorithmic actor to be 'safe', one should just control its ability to do harm, and once that is done, it simply cannot intend to cause harm. This is certainly true, though controlling all the ways an algorithm can cause harm is not a straightforward task even in a limited setting. Even if we were to assume success in this endeavour, the approach of will fail for another category of criminal offences other than inchoate crimes where mens rea plays a definitional role to the actus reus. The harm in these crimes, is the intention under which they were performed. The communication of this prohibition relies on the ability to convey what certain types of mens rea means.

Aside from attempt crimes, a range of criminal offences exist whose undesirability hinges on the intent under which they committed. For example under the UK Fraud Act 2006 for someone to commit the crime of fraud by misrepresentation they must (i) make a false representation, (ii) dishonestly, (iii) knowing that the representation might be untrue or misleading and (iv) intend to make a gain for themselves or cause or risk a loss to another. This formulation includes the mental states of intent and knowledge. Similarly in Republic of Ireland under the Criminal Justice (Theft and Fraud Offences) Act 2001, the offence of *Making gain or causing loss by deception* relies on the actor acting deceptively with the intentional of making a gain for themselves (or a loss for someone else). As Simester observes, many of these crimes seem to be economic in nature or related to the functioning of markets. In Australia under the Competition and Consumer Act 2010 (CCA), predatory pricing is defined as having the intention to *"eliminate or substantially damage a competitor, prevent someone entering the market or deter or prevent someone from engaging in competitive conduct in a market"* (ACCC 2005). Intent in these examples plays an important role in delineating behaviour which is acceptable from that which is not. In a study of the wider laws of deceit, Klass (2012) characterises the laws surrounding deceit as being concerned with regulating the flow of information between parties. Algorithmic actors are heavily involved in the business of information, both as consumers and distributors. An algorithm could very feasibly engage in anti-competitive behaviour without being expressly instructed to by its owner but without a provable concept of intent, it could not be restrained or penalised for doing so.

Viewed through the lens of legitimate enactment, the role of intent in these crimes firstly lets people know what sort or behaviour is reasonable and what is not. This can be paraphrased in the two examples given as 'don't deceive people on purpose to enrich yourself' or 'don't set your prices in order to bankrupt your competitors'. Secondly it is a useful legislative tool to prevent over-criminalisation. A world would not function well where the actions of stating anything untrue or keen pricing brings a criminal charge. If we didn't know what intent actually meant and we had no way of measuring it in an actor then law fails twice. People couldn't be sure if they or their employees are breaking the law and whether they are liable for that. Policy makers would be deprived of a tool that they had previously used to delineate

the boundary between acceptable and criminal behaviour. This is the situation that I will argue we have already found ourselves in with a certain class of algorithms.

1.2 A case for intent in auto-didactic algorithms

Traditionally the output of an algorithm has reflected the purposes of its creator just like the face of a hammer is assumed to strike its bearer's target. Where algorithms have been deployed in some sort of autonomous application, like trading or a plane's autopilot, the decisions they make and the ensuing behaviour they demonstrate can be said to be an extension of the programmer's intentions. There is a limit to this reading in the sense that not all behaviours of an algorithm are intended. This is particularly true of complex systems even when they have come under extreme testing scrutiny. Nobody would claim the creators of Boeing's MCAS system—an automatic flight stabilising program—intended for it to contribute to the two crashes of the 737 Max airliner. Excepting the case of unexpected behaviour, if the user of an algorithm (the Principal) intended their algorithm to commit a crime on their behalf, and it did subsequently do so, then they would be guilty of the crime in the same way as anyone using a tool to commit a crime is. The doctrine of innocent agency (Alldridge 1990) goes further, and prevents the Principal from using other people as tools.[2]

On occasion, the user and designer of an algorithm might be accused of engaging in some criminal activity and the purpose of an algorithm might need to be assessed. This assessment of an algorithm's purpose might be a framed as an exercise in evidence collection, since the prosecution would argue that the algorithm's design and workings are merely a reflection of the defendant's alleged criminal intent. Such an investigation might benefit from a generally agreed upon definition of intent in an algorithm. I think such a standardisation endeavour could aid the functioning of courts in the future. This is not the primary motivation of this article though the definitions I will present later on will certainly aid in this use case.

There exists a class of Algorithms that learn their own types of behaviour above and beyond the atomic action set they are given which I will term auto-didactic. Typically, this class of algorithm will 'learn' a behaviour or policy by analysis of historic data through some statistical machine learning technique or in a simulation of an environment through an online learning technique such as Reinforcement Learning. The motivation behind using these techniques for the algorithm designer is that the resulting algorithm can take advantage of statistical features of the environment that might not be obvious to a more traditional, top-down approach. The resulting algorithms, trained on massive data sets can perform a range of tasks, often exceeding human capabilities. For the rest of the article I will refer to this class of self-taught algorithm when deployed in an autonomous function, as an A-bot. By autonomous I mean makes decisions without requiring confirmation from a human. For simplicity I will also suppose

[2] Assuming that the person used as a tool is not aware that they are committing a crime.

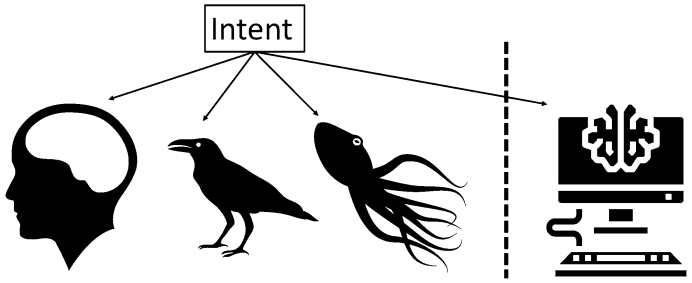

Fig. 1 This article proceeds under the assumption that intent is a definable concept that does not require a human brain to exist, that it arguably exists in other biological entities with demonstrable intelligence and can plausibly exist in an artificial intelligence. Images: Human head—flaticon.com, Crow—Travis, Octopus—James Keuning, AI—Komkrit Noenpoempisut, The Noun project

for the rest of the article that the A-bot's creator, owner and user are one person which I will refer to as the Principal. I acknowledge this is a simplifying assumption but I feel justified since the objective of this article is not focused on the attribution of responsibility but rather an earlier step; the identification of a crime itself.

Suppose an A-bot were to perform some sequence of actions that would be qualify for crime *X* if a human had performed them with the requisite mens rea. We will follow the terminology of Abbott and Sarch (2020) and term this state of affairs an AI-crime. At present the A-bot has not committed a crime because it is not a legal person regardless of its mental capabilities. If the qualifying mens rea requirement of crime *X* was no more than negligence or perhaps recklessness then the Principal of the algorithm might be charged with crime *X* depending on its foreseeability. For higher levels of mens rea such as intent, a problem appears because the intent of the A-bot's Principal might not coincide with the behaviour of the A-bot. The Principal could not be caught up through secondary liability for since the A-bot has only committed an AI-crime and not an actual crime, secondary criminal liability being parasitic on an initial crime. The question of whether the concept of secondary criminal liability needs to be reformed in the age of A-bots is an interesting one but once again not one directly addressed here.

Aside from inchoate crimes, in Sect. 1.1 we identified a category of crimes (why-crimes) where the mens-rea plays a directly definitional role. These are crimes where the actus reus is in of itself not criminal. It is here that, initially at least, I argue that A-bots require a working definition of intent amongst other mens rea concepts. Without one, these crimes cannot be practically prosecuted because they cannot be proven to have taken place. In turn, unless a definition is forthcoming and generally understood, A-bot Principals cannot easily take measures to prevent their creations from breaking the law. Such a state of affairs might be exploited by bad Principals in situations where conducting these types of crime is profitable. At the very least, situations might appear where a risk arbitrage opens up and companies are incentivised to use (or pretend to use) an A-bot for a job otherwise done by a human because their liability is considerably reduced. It is unfortunate that these types of crimes are

often economic in their nature and the maximisation of economic returns is an area where algorithms are increasingly deployed (Fig. 1).

In financial markets, there exist a number of market manipulative practices that are outlawed. One of them, termed 'spoofing' provides an interesting example of a intent dependent or why-crime. Spoofing can cover many slightly different behaviours but in its general sense the spoofer tries to place orders in the market so as to give a false impression of supply or demand. Because the market is visible to other traders, order placement usually conveys information and market participants will react to reflect this. Generally a large amount of buy orders pushes the price of an asset up and a large amount of sell orders will push the price of an asset down. A spoofer will take advantage of this reaction by taking a directional bet, putting a large 'spoof' order into the market, profit from the ensuing reaction and then cancel their 'spoof' order which crucially they never wanted to execute in the first place. In section 747(C) of the Dodd-Frank Wall Street Reform and Consumer Protection Act 2010, spoofing is defined as *"bidding or offering with the intent to cancel the bid or offer before execution"*. In their guidance note, CTFC (2013) state that recklessness is not sufficient for spoofing. Suppose a bank were to create an auto-didactic trading algorithm with the objective of making as much money as possible subject to reasonable risk constraints. Without a definition of intent that could be applied to an algorithm, how can anyone inside or outside the bank know if the algorithm is spoofing? Under what conditions can one say that intent to cancel can exist in the trading algorithm?

If a generally agreed upon definition of intent existed for algorithms, then it would be harder for a Principal to argue that they did not know that an algorithm intended to commit a AI-crime. Wilful blindness as to a fact has been established, under certain circumstances, to be equivalent to knowledge of a fact, Robbins (1990) terms it 'The Ostrich instruction'. A definition of intent might not allow one to conclude that intent in the algorithm equals intent in the Principal, but at the same time it might be useful evidence as to the intentional state of the Principal as to their algorithm. To take the example of the bank and the trading algorithm, if according to some definition, at the point of placing orders the algorithm intended to subsequently cancel them, then the bank would be able to correct the algorithm before deploying it in the market. Failing that, a market regulator would be able to show that the algorithm was actually spoofing and act to restrict it. Whether the bank's knowledge of or wilful blindness to the algorithm's spoofing strategy would be enough for criminal charges is an interesting question for courts to answer in the future. At the very least, a definition of intent for algorithms gets us to the point of asking it without having to many changes to the law as it stands.

The approach of this article is atypical in computer science literature in that the definitions of intent that it will present are informed by the body of law that exists concerning intent amongst other relevant mens rea states. Other approaches might be to use psychological evidence or philosophical theory. However, I think that the legal conception of intent is what it is for good reason. It has been honed over time in a public manner and any attempt by a computer scientist to impose their own definitions of commonly held concepts, has a democratic deficiency as Hildebrandt (2019) points out. Worse, it opens up such an approach to accusations that

the definitions are chosen for their programming expediency or some other selfish motive. A legally informed approach also goes some way to meet the fear (Sales 2019) that by coding legal concepts, we block its natural progression. Progress it must because A-bots do pose novel challenges to courts, to quote Lord Mance:[3]

> ...the law must be adapted to the new algorithmic programmes and artificial intelligence, in a way which gives rise to the results that reason and justice would lead one to expect

The article is divided in two parts and will proceed as follows. Firstly in Sect. 2 we will consider various different types of intent that exist in criminal law and their definitions such as they are. It will concludes by discussing some desiderata of an algorithmic intent definition. Armed with that knowledge, Sect. 3 will present definitions of Direct, Oblique and Ulterior intent. This is followed by a short discussion and a review of alternative attempts to rigorously define intent.

2 Definitions of intent from common law

Intent within a criminal law context is one type of broad range of degrees of mens rea. Specific crimes are typically defined with a threshold level of criminal intent; the minimum level of intent that the accused must have in order to have committed the mental element of the crime. Some crimes attach different levels of mens rea to different parts of the actus reus. The same criminal action or actus reus, is deemed more or less culpable depending on the level of mens rea it was committed with. The clearest example of this is with the actus reus of causing death; if the act of killing someone is done with direct intent then it is murder, if death is a result of lower intentional mode such as recklessness, then it would be manslaughter.[4] Causing the death of doesn't even necessarily lead to any criminal sanction if it was done so accidentally and it was not a reasonably forseeable consequence of the contributing actions.

Mens rea can be thought of as a hierarchy arranged in terms of culpability, with direct intent at the top, followed by oblique intent, recklessness, negligence and strict liability (or the almost absence of mens rea). Where a crime requires a certain level of mens rea, a higher level is sufficient to satisfy the requirement as stated in cl19 of the draft criminal code for England and Wales (The Law Commission 1989). The burden of proof of higher levels of mens rea can be considered higher

As mentioned, a justification for establishing the intent behind an action is to distinguish between those harmful outcomes which were accidental and those which were not. Sometimes only the actus reus is required, irrespective of its outcome or the mental state under which it was performed; this is called strict liability and

[3] Quoine Pte Ltd v B2C2 Ltd [2020] SGCA(I) 02 at 193.

[4] This is a simplification, in the UK there are further distinctions between voluntary and involuntary manslaughter (Criminal Prosecution Service 2019) and as we will discuss oblique intent *can* be sufficient for murder.

forms the lowest level of the mens rea hierarchy. Certain possession are an example of this type. Ormerod and Laird (2021b) makes the distinction between crimes of strict liability, where one element of the actus reus requires no mens rea and crimes of absolute liability where no element of the actus reus requires mens rea.

It should be noted that there is no universal language for mens rea across nations and justice systems, so concepts negligence or recklessness might mean different things in different places or may have analogous modes with other names.

The aim of this article is to concentrate on the highest levels of mens rea; Direct intent and Oblique intent or Knowledge. These are the levels of mens rea which are most likely to play a definitional role in the actus reus as discussed in Sect. 1.1. These higher intent levels enjoy some alignment in meaning across different common law jurisdictions. I have also included discussion of recklessness and negligence, because I have found them useful to discuss what the higher levels of intent are and are not.

2.1 Intent in common law

A barrier to creating a legally rigorous algorithmic definition of intent is that courts in the UK have consistently not wanted to elaborate to juries what intent actually constitutes. As Lord Bridge stated in R v Moloney (1985) 1 All ER 1025 *"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 necessary intent"*. The reluctance to pin a definition down onto the page is reflected to varying degrees in other common law jurisdictions. A potential reason behind this is the confounding existence of oblique (sometimes called indirect intent), which whilst occupying a lower level to direct intent has been established in a number of boundary cases such as R v Nedrick [1986] 1 WLR 1025. and R v Woollin [1999] 1 A.C. 82. to be sufficient, in certain cases, to be sufficient mens rea for the crime of murder. We will discuss oblique intent after tackling direct intent.

2.1.1 Direct intent

Whilst a definition of direct intent has not been forthcoming within courts in the UK, examples do necessarily exist within textbooks and other legal discourse. Parsons (2000) defines direct intent as the case where *"the defendant wants something to happen as a result of their conduct"*. A draft bill published by the UK Home Office (The Law Commission 2015a) defines direct intent as the situation when *A person acts intentionally with respect to a result if...it his purpose to cause it*. Using this document as a consultation template, the Law Commission also suggested an alternative formulation of direct intent as follows: (The Law Commission 2015b):

> The jury should be directed that they may find D intended a result if they are sure that D realised that result was certain (barring an extraordinary intervention) if D did what he or she was set upon doing.

A previous formulation is to be found in a draft criminal code (The Law Commission 1989), which states that:

> A person acts intentionally with respect to i) a circumstance when he hopes or knows that it exists or will exist; ii) a result when he acts either in order to bring it about or being aware that it will occur in the ordinary course of events.

It should be noted that the Law Commission's 2015 consultation concludes that no definition is needed, at least in the context of the offences against the person bill reform.

As Coffey (2009) summarises, the ingredients of direct intent generally seem to involve a decision to act and an outcome which is the aim, objective or purpose of that act. Whether that outcome or result is desirable from the point of view of the accused seems to depend on the narrowness of the definition of desire. On the subject of desire and direct intent, James LJ in R v Mohan [1976] 1 QB at 11 defines it as:

> ...a decision to bring about insofar as it lies within the accused's power, the commission of the offence which it is alleged the accused attempted to commit, no matter whether the accused desired that consequence of his act or not.

In the USA, a definition of direct intent is more forthcoming in the form of the Model Penal Code (MPC) (The American Law Insitute 2017). This has been adapted to various degrees by many states, though Federal prosecuted crimes have no analogous written definitions. What we have termed direct intent corresponds to the MPC's definition of purpose, the highest of the four levels of intent that they define:

> A person acts purposely with respect to a material element of an offense when... if the element involves the nature of his conduct or a result thereof, it is his conscious object to engage in conduct of that nature or to cause such a result

Generally we can conclude that directly intended things do not need to be desirable but they should be an objective of the actor. The example of a dentist is often given to illustrate this point (Williams 1987). A painful tooth extraction may result, which is certainly not desirable for most, but the object of the visit is to obviate future tooth ache.[5]

Related, and sometimes confused with oblique intent, is the intentional status of intermediate results which are caused through the actions of the agent, and are necessary to achieve some other aimed for result. These intermediate results, which Simester et al. (2019) term *Means to an end* results, are directly intended, this being established in Smith [1960] 2 QB 423 (CA) where it was found that a defendant who bribed a Mayor in an attempt to expose corruption, nonetheless intended to corrupt a public official, which was a crime.

[5] The intentional state of the pain that necessarily ensues is discussed in the next sub-section.

Whilst an intended result must be foreseeable as a result of an act, there is no requirement for it to be likely. This is neatly encapsulated by the cowardly jackal example of Alexander and Kessler (1997), where 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, should still be found to have directly intended to shoot their victim. If this were not the case, then longshots could be attempted with impunity.

A feature of the definitions of direct intent that we have seen is that foreseeability should be a subjective test. That is to say, consequences should be foreseeable to the accused. This was not always the case, DPP v Smith [1961] AC 290 held that a foreseeable result would be intended if it was a natural consequence of the action. This is an objective test, which relies on assessing probabilities and causation according to the 'reasonable person'. Furey (2010) observes that this position was soon reversed since it narrowed the states of direct intention and gross negligence too much and thereby blurred the line between murder and manslaughter. In the case of an algorithm malfeasor, we must then consider whether a 'reasonable person' should be a 'reasonable algorithm' (Abbott 2020). In practice, as Furey observes, objective and subjective tests blur, since the accused denying that they foresaw a consequence if that consequence becomes less believable when that consequence becomes more obviously likely. Here is where the judgement of intent in algorithms might differ from that in humans. Humans can empathise with other humans under the assumption that at the very least, their sensory perception and common sense is share. In R v Moloney [1984] UKHL 4, the original trial court judge is quoted to have said:

> "In deciding the question of the accused man's intent, you will decide whether he did intend or foresee that result by reference to all the evidence, drawing such inferences from the evidence as appear proper in the circumstances. Members of the jury, it is a question of fact for you to decide. As I said I think when I was directing you originally you cannot take the top of a man's head off and look into his mind and actually see what his intent was at any given moment. You have to decide it by reference to what he did, what he said and all the circumstances of the case."

Depending on their design and to varying degrees A-bots *can* be peered into and the constituent parts behind a definition of intent can be assessed. So whilst humans might not be able to empathise and reason about the inner workings of A-bots, unlike with human defendants, they have some opportunity to *take the top of an A-bot's head off and look into its mind*. Even in the case of black box A-bot designs which confound many attempts to interpret, their reaction (output behaviour) to inputs can be scrutinised for evidence. In certain cases they can feasibly be put into the same situation they found themselves when they are accused of committing an AI-crime via a simulator much as aviation accident investigators look to recreate errors so as to understand what was fault for the crash. The A-bot's beliefs about the state of the world in this recreation should be strong evidence as to their beliefs previously. Where an algorithm predicts the likelihood of outcomes following its actions, it is observable whether this calculation is misspecified or not. Unfortunately many algorithms do not explicitly predict the outcome of their actions; this is the case with

model free reinforcement learning algorithms which have succeeded in mastering a variety of games to super-human levels.

A corollary of direct intent being within the mind of the actor, is that they should be able to intend impossible things if they thought they were possible. This is indeed the case as confirmed by the UK Criminal Attempts Act. We will explore this issue further in Sect. 2.3. In practice this has proved less of an issue than perhaps it might appear on first inspection, though one wonders if rules which protect the mentally ill from criminal proceedings have also prevented more bizarre cases from being heard. Perhaps similar diagnoses will be necessary for A-bots to prevent over-criminalisation of algorithmic policies which have no possibility of causing harm because they are so unrealistic.

The next subsection will consider the intentional status of side-effects, that is to say, those states of affairs which are caused by actions, but are not the motivating factor behind those actions and whose realisation does not affect the success of the actor's intended results.

2.1.2 Oblique intent

Oblique or indirect intent refers to the intentional state of almost certain side effects of directly intended actions. The phrase was coined by Jeremy Bentham (1823) where he considered the example of a hunter shooting a stag who appreciated at the moment of releasing his arrow, that it was just as likely to hit the stag as King William II. Bentham concludes that *"killing the king was intentional, but obliquely so"*. Its existence can be illustrated by the following example found in The Law Commission (2015b):

> D places a bomb on an aircraft, intending to collect on the insurance. D does not act with the purpose of causing the death of the passengers, but knows that their death is virtually certain if the bomb explodes.

In the USA, according to the MPC, oblique intent is roughly equivalent to the status of crimes committed with knowledge, which is the second most serious level of intent. It is defined as follows (The American Law Insitute 2017):

> A person acts knowingly with respect to a material element of an offense when: ...if the element involves a result of his conduct, he is aware that it is practically certain that his conduct will cause such a result.

The current accepted direction to be made to Juries in England and Wales with respect to Oblique intent, originally formulated in R v Woollin is as follows:

> 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.

As with the definitions of direct intent in the previous section, this direction makes it clear that this is a subjective test as well. This definition has since been modified,

because as with direct intent, there should be no restriction on the likelihood of the accused achieving their aim, only that if they did, it would be most likely that the obliquely intended result occurs. This is not captured in the MPC formulation of Knowledge. The definition of oblique intent in Law Commission (1993) is phrased thus:

> A person acts intentionally with respect to a result when...although it is not the purpose to cause that result, he knows that it would occur in the ordinary course of events if he were to succeed in his purpose of causing some other result.

Smith (1990) acknowledges the necessity of this amendment and adds a further requirement. A definition of oblique intent should make it clear that if it is the purpose of the accused to avoid a result through their actions, they cannot be accused of obliquely intending that result as well. The example given being the father who chooses to throw their child from a burning house because they know otherwise that the child will die from the fire, but also know that the child will be grievously injured from their actions. Such examples begin to stray into the doctrine of double effect (McIntyre 2019), which protects physicians from criminal charges when they cause harm through their actions which are intended to cause some other, justifying outcome.

A practical feature of oblique intent, is that the directly intended results of the algorithm's actions do not need to be identified (save that they are separate and not the opposite of the obliquely intended ones). This is in contrast with direct intent where an aimed outcome or objective should be identified. A-bots do have high level aims (typically called objective functions), but they learn to meet them themselves. That oblique intent has in cases been given an equivalent culpable status to direct intent, provide courts an alternative way of establishing intent in an A-bot, should it be more practical.

So far, the two types of intent discussed have required an exclusive subjective treatment. The next subsection deals with recklessness and negligence which have objective elements to their definitions.

2.2 Recklessness and negligence: the lower levels of mens rea

Although this article principally concerns itself with the higher levels of intent, it is instructive to understand how lower levels of mens rea like recklessness and negligence are different (and related). Courts may decide algorithms are incapable of intent or in any case impose a higher standard on their behaviour by lowering the mens rea requirement for certain crimes. Stark (2017) calls these two types of intentional behaviour 'culpable risk taking'. Loveless (2010) equates recklessness with unreasonable risk taking, or more precisely the conscious decision to take an unreasonable risk. The test for recklessness in the UK is now said to be subjective, in the sense that the accused must be aware of the risk of their actions; one can no longer be reckless by inadvertently creating risk or harm. Negligence concerns actions where the actor does not necessarily have awareness of risk, but should do according

to some standard. This might be a reasonable human or a reasonable robot as Abbott (2020) debates. Frequently, recklessness is the minimum level of intent required for a criminal offence and actions done with negligence, resulting in harm, are mostly dealt with civil (or private) law so differentiating the two is important. Nevertheless some crimes exist which only require negligence (often driving offences) or have elements which only require negligence (Ormerod and Laird 2021a). These criminal offences of negligence seem to appear worldwide (Fletcher 1971).

As to what unreasonable risk is, Stark indicates that there is not very much concrete guidance. At the extreme, any risk could be termed unacceptable, which in almost every situation, is an unworkable solution. A problem with applying a blanket level of risk as the threshold of reasonable behaviour is that the severity of the outcome might make determine its acceptability; a 0.5% chance of breaking a window is not the same as a 0.5% of killing someone. Furthermore, any process when repeated many times has a high probability of obtaining at least one bad outcome even if the chance of obtaining a bad outcome in one trial is tiny. In the USA, the Model Penal Code (MPC) (The American Law Insitute 2017) instead allows a situation specific chance:

> A person acts recklessly with respect to a material element of an offense when he consciously disregards a substantial and unjustifiable risk that the material element exists or will result from his conduct. The risk must be of such a nature and degree that, considering the nature and purpose of the actor's conduct and the circumstances known to him, its disregard involves a gross deviation from the standard of conduct that a law-abiding person would observe in the actor's situation.

Thus in the language of subjective and objective tests, the accused must be aware of the possible risk, and still act, but the judgement as to what constitutes an unacceptable risk is subject to an external benchmark, or objective test. Preventing an A-bot from behaving recklessly is harder than preventing them from intending harm since an external, possible changing benchmark needs to be introduced, and a ranking over the severity of any outcome is required to adjust what an acceptable probability of a bad outcome is. A restriction to not cause harm recklessly is stricter than one to not do so intentionally. Conversely from the point of view of the courts, a lower requirement to establish what the A-bot believed at the point of commission is a simplifying feature. Which standard should be applied when make objective judgements concerning the behaviour an A-bot is an open question. Abbott (2020) discusses the standard in the context of Autonomous Vehicles (AVs) and proposes that a single standard for humans and AVs will result in humans being effectively held to a standard of strict negligence as AVs improve. Whilst with driving, lower road deaths are the the benefit of this, in other areas where humans and algorithms coexist (like exchange trading), imposing an algorithmic standard on humans might offer no such advantages and come at the cost of jobs. Unlike roads, markets are strictly adversarial, so their regulation raises the prospect of regulatory arbitrage when different standards are applied to human and algorithmic traders. This is also true with respect to enforcement capabilities: current trading regulation which cannot practically be enforced against algorithms only encourage the use of algorithms

in markets. Where it is profitable to break these laws, algorithms will do so because their owners face lower regulatory risk.

2.3 Inchoate offences

Law often includes prohibitions against attempting to commit actions which if otherwise completed with the most likely or intended result would be crimes (the actus reus or criminal action is *inchoate*). An inchoate crime might come about because the accused failed (the myopic assassin missed with their shot) or the accused was interrupted before completing their action (the lethargic assassin is caught with loaded gun drawn and aiming at their target). Attempted murder and possession (of prohibited drugs) with intent to supply are both examples. Most common types of inchoate offence are attempts to commit a substantive crime,[6] that is to say, a crime which does not include another crime in its definition. Other types exist, such as conspiracy and solicitation (in the USA). Conspiracy is an agreement amongst two or more parties to commit an offence in the future and solicitation is where the accused induces another to commit a crime. Examining the law around attempted offences provides us with some interesting observations about the nature of intent. In the UK, Criminal Attempts Act 1981, defines attempt in Section 1 (1):

> If, with intent to commit an offence to which this section applies, a person does an act which is more than merely preparatory to the commission of the offence, he is guilty of attempting to commit the offence.

The question of what constitutes actions which are more than preparatory is not entirely straightforward. The Law Commission (2007) has proposed a law change which would separate the situation where the actions have been completed and failed to achieve the expected outcome (the myopic assassin who misses) and where the actions have been taken in preparation of an intended crime (the lethargic assassin who is disturbed just as they pull the trigger). For the purposes of this article it is sufficient that a plan of action is not sufficient for an attempt offence; some actions must be carried out from that plan. The importance of this separation between plan and enaction of the plan will become clearer in Sect. 3.

The second important observation from the law surrounding attempts is that impossible crimes can be found to have been attempted (and therefore intended) and will be punished as normal. Section 1(2) of the UK Criminal Attempts Act 1981 states:

> A person may be guilty of attempting to commit an offence to which this section applies even though the facts are such that the commission of the offence is impossible.

and Section 1(3b):

[6] A defendant who successfully completed an action would be only accused of that crime, not the attempt as well, under the merger doctrine.

> If the facts of the case had been as he believed them to be, his intention would be so regarded, then, for the purposes of subsection (1) above, he shall be regarded as having had an intent to commit that offence.

Storey (2019) divides impossible attempts into things which are physically impossible, practically impossible and legally impossible. The canonical example is the attempted murder of someone who is already dead which comes under the category of physical impossibility. Practical impossibility refers to situations where the accused has a plan to commit a crime, but their plan is unrealistic—they plan to detonate a bomb, but they have been sold fake explosives by undercover police. Legally impossible acts cover the situation arising in R v Jones [2007] EWCA Crim 1118, where the appellant unsuccessfully appealed against a conviction of inciting a child under 13 to engage in sexual activity. The crime was impossible because the 'child' in question was an undercover policewoman.

Our interest in the mens rea as regards attempting impossible acts, is twofold. Firstly, the spectre of misspecification within an A-bot, means that possessing unrealistic models of the world are no defence, if the agent intends to commit a crime and begins to embark on it. Secondly, it underlines the importance of the agent's model of the world in determining criminal intent. The important distinction between subjective and objective judgement will be reflected in our definitions of intent in Sect. 3.

2.3.1 Conditional intent

A further wrinkle to a legal discussion of intent and inchoate offences is the concept of conditional intent. It is perfectly reasonable to consider an agent who intended to do some action A if condition x is met and do some action B if condition y is met. A common design pattern for A-bots is a policy function, which is a mapping between the state information that they currently perceive to the actions that they take next. If that A-bot were capable of intention, then the presence of a policy function would surely make that intention conditional. To some extent all intentions are conditional as Yaffe (2004) and Klass (2009) both point out. Legal precedent has wavered on whether conditional intent equates to the direct intent of the sort required to successfully convict the accused of attempt crimes discussed in Sect. 2.3. Yaffe considers the case of Holloway v. United States,[7] where 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, but Yaffe cites other cases which have concluded that conditional intent does not meet the mens rea for certain crimes.

Conditional intent poses problems because very little is said about about it in the wording of laws which are normally expressed in terms of simpler intentional concepts such as direct, oblique intent and recklessness. This has allowed people

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

to claim, on occasion successfully, that holding a conditional intent was less than the required intent for the offence that they were accused of. Child (2017) rejects the idea that conditional intent is any different from future or ulterior intent and that conditional intent exists in the present stating that: *Intention as to present conduct and results is always unconditional, and that intention as to future conduct is always conditional.*

Child also recognises that intention to commit actions in the future, has some different properties to present intent. This is important to the computer scientist when evaluating the safety of an A-bot's policy since future acts are the focus of consideration. If we consider the situation where an A-bot is deployed with a static policy (no further learning), then arguably the algorithm has commitment to act in a particular way in the future. If that conduct is illegal, then as we saw from Sect. 2.3, an attempt crime has been committed. Just as with the example of the cowardly jackal, Child states that judgements of the likelihood of future conditions are not relevant provided there is commitment to act. An important to Child's treatment is what he calls the second point of coincidence. At the point of the criminal act being done in the future, is the committed mens rea sufficient for that crime? Future acts can feasibly be committed to with direct intent, oblique intent or recklessness. Child illustrates this with an example of two hunters D1 and D agreeing that D would shoot to kill something if it comes out of the bushes. Since, at the point of shooting, the shooter D, will not be sure if the thing is human or not, they cannot be guilty of murder, only causing death through recklessness. If interrupted or they fail to kill, then they cannot be guilty of attempted murder. Consider a different plan where D and D1 agree that D should shoot, even if they recognise the thing emerging from the bush. Here D is guilty of murder or attempted murder if interrupted or unsuccessful and D1 guilty of conspiracy to murder.

2.4 Intent outside common (criminal) law

This work primarily considers the concept of intent, as understood in common law countries primarily referencing cases and statute from within the UK and to a lesser extent, the USA. Leaving common law jurisdictions momentarily for those that use Civil Criminal law (such as the majority of mainland Europe), there exist analogous concepts (*Dolus Directus, Dolus Indirectus*) to the respective definitions of direct and oblique intent presented here, and their definitions seem broadly compatible with each other. Both systems require both the action *actus reus* and intent *mens rea* element for crimes, and the intent threshold is also defined by the crime (De Jong 2011). Further in common with Common Law, German civil law at least, has proved reluctant to define intent within statute and instead chosen to rely on case law as Taylor (2004) observes. Comparative law is a large separate subject in itself, and providing a thorough analysis of how an algorithmic definition of intent might differ across the world is beyond the scope of this article. Generally we feel the definitions presented here should translate from Common to Civil law but *caveat lector*.

2.5 Desiderata of intent definitions

I will now present a few desiderata of a definition of intent informed by the findings of this section. The list includes those elements which I think are most often misunderstood about intent by those people who do not have a background in criminal law. It is therefore inherently non-exhaustive, giving necessary but not sufficient features that a definition of intent for algorithms should have, if it is going to be compatible with current criminal law.

1. **Knowledge of causal effect** Results caused by actions can only be intended if they are foreseen by the agent. This rules out accidental or freakish results, which though caused by the agents actions, could no way have been predicted to cause the outcome.
2. **A directly intended result need only be foreseeable to the agent, not likely** As with the cowardly jackal example, the unlikeliness of a result should not shield the actor from a judgement of intent, else any number of speculative crimes might be committed with free license.
3. **Judgements of foreseeability and causality are subjective**. The question of whether to use objective or subjective tests when assessing causality, foreseeability or likelihood separates lower levels of intent such as recklessness from the higher levels of direct and oblique intent.
4. **Intent is not dependent on success** A definition of intent should not be determined by the success of obtaining a desired result. This agrees with the definition of inchoate intent in Sect. 2.3. At the point of commission, an intended result must occur in the future and since that is unresolved, intent cannot depend on it obtaining.
5. **Means-End Consistency** If an agent directly directly intends a final result through their actions, and there are necessary intermediate results which must be brought about through their actions first, then those intermediate results are necessarily directly intended. Simester et al. (2019) consider the intentional status of means as equivalent to that of the end. Bratman (2009) terms this property of intent as Means-End Coherence.
6. **Side effects can be obliquely intended** The intentional status side effects has long been debated since Jeremy Bentham coined the term Oblique intent, see for example Williams (1987), but it has been agreed in law where results are caused in addition to an intended result through action, then it must be the case that these results are intended, if they were extremely likely. Later we will see that this conclusion is not shared with other research disciplines. Murder is obliquely intended by putting a bomb on a plane in order to collect an insurance pay-out from the plane's destruction. In particular, this means that obliquely intended results are by not required to be desired.
7. **Commitment** Future results brought about by future actions can only be intended if there is a commitment to act in the future to bring about that result. The commitment is necessary to distinguish between plans and intentions.

This concludes our tour of intent as it appears in (predominantly common) law. We have surveyed the various levels of intention in criminal law as they relate to culpability—direct, oblique and recklessness. We have also considered inchoate and conditional intent. Using this we concluded with a non-exhaustive list of desiderata together concerning a definition of intent. We will now attempt to translate what we have learned in this section into desiderata of an intent definition and finally a series of definitions of intent which can be applied to an A-bot.

3 Definitions of intent suitable for autonomous algorithms

In this section I will present some definitions of intent whose inspiration is the criminal law. These definitions will be semi-formal, in the sense that they can be converted into a fully formal language, suitable for an algorithms, but their description does not rely on a huge amount of notation. I have decided not to present a fully formal approach because I feel that would narrow its utility and audience. When criminal law does eventually tackle the problem of intent in algorithms, it should do so in a way that does not preclude any particular AI paradigm. From a practical perspective this is so as to make it applicable to the widest set of A-bots possible and to ensure the timely delivery of justice. From an economic perspective it wouldn't be desirable to design a legal treatment for a certain type A-bot. Large neural networks are popular at the moment but the history of AI has had many different most favoured technologies over time. In comparison, the evolution of the law can seem glacially slow. The legislators should impose requirements on A-bots but as far as possible not try picking a winning technology. The approach of this section reflects my belief in this minimally prescriptive approach.

3.1 Definitions of intent

With the desiderata of Sect. 2.5 in mind, we are now in a position to present three definitions of intent. We begin with direct intent, being the simplest of intentional concepts and the highest level of intent. It is a foundational concept on which our other definitions are built.

On notation, we will use upper case letters to represent variables and lower case letter to represent realisations of those variables. The statement $X = x$ is taken to mean that variable X takes realisation x. We define $\mathcal{R}(X)$ to mean the range of all possible values that variable X can take.

Definition 1 (*Direct Intent at commission*) An agent D directly intends a result $X = x$ by performing action a if:

(DI1) **Free Agency** Alternative actions a' exist which D could have chosen instead of a.
(DI2) **Knowledge** D should be capable of observing or inferring result $X = x$

(DI3) **Foreseeable Causality** Actions a can foreseeably cause result x (according to D's current estimate).
(DI4) **Aim** D aims or desires result x.

The first three requirements in this definition should not be surprising or particularly contentious. The condition of *Free Agency* ensures that the agent D genuinely had a choice about their behaviour. *Knowledge* implies that an agent can only intend things that they can measure and Foreseeable Causality, ensures that the agent can only intend results which they can realistically cause ex-ante subject to their own world model. The Explicit Aim clause requires some exploration. If it were D's aim or desire to cause result x, then we should consider this sufficient for intent. The difficulty comes in defining what aim or desire should be in the case of an artificial agent. As Smith (1990) observed, endeavours to define intent often just end up shifting the ambiguity to other words (in that case purpose). An A-bot might be designed in such a way where it has values over every state of the world (as a Reinforcement Learning agent does), in which case aims or desires, at least locally could be feasibly extracted. Kenny (2013) uses a failure test which he states as a question to the actor which to paraphrase is as follows: If the (proposed) intended outcome of your actions had not occurred, would you be sorry or would you have failed in your endeavour? This question invokes the counterfactual in a way which is quite appealing to a causal scientist and offers a potential route to establishing aims or desires.

The definition only makes reference to information available at the point of commission; the importance of achieving the desired result is subsumed. Intent, is the same regardless of whether the desired result is obtained or not in line with the desiderata. This means Definition 1 is useful when considering inchoate crimes such as crimes of attempt, as discussed in Sect. 2.3.

Unfortunately there is no guarantee that an A-bot will have an amenable cognitive mechanism that numerically values states. An alternative counterfactual approach would be to define an aimed outcome as one, which if impossible to achieve would mean that some alternative action a' would be taken instead of a by D.

(DI4') **Counterfactual Aim** D aims or desires result $X = x$ by a if in another world where $X = x$ is not possible by performing a, then some other action a' would be chosen instead.

Example 1 Company GHI deploys an auto-didactic trading algorithm (a trade-bot) in the S &P futures market which has the objective of making profits subject to certain risk levels. The trade-bot trains itself whilst it trades through reinforcement learning. The trade-bot is observed to be cancelling almost all of the orders it places. If we define spoofing as the intent to cancel orders before execution, is the trading algorithm engaging in spoofing? According to the definition, the answer is only yes if the aim of the trade-bot at the point of order placement is to cancel it. This is not conclusively shown by the high probability of order cancellation alone. Its objective

in placing these orders might well be execution. If one could see that the trade-bot is disappointed if it does not get to cancel an order (because it has been matched with another market participant), then we could say that the trade-bot does intend to cancel this order. Consider the same situation but the trade-bot cancels its orders no more or less than average market participants, can it be said to be not spoofing? Again the probability of order cancellation is not a sufficient diagnostic statistic. It could for example intend to cancel its orders at the point of their placement only when some specific conditions are met. If the A-bot is shown to be spoofing it is an open question as to whether Company GHI is liable. Whilst one cannot recklessly spoof by definition, this situation seems more akin to recklessly letting an agent (the trade-bot) spoof. The definition of intent in the algorithm allows the harm to be identified but does not answer the question of culpability.

An alternative, but equivalent version of direct intent is required, namely what Bratman (2009) calls means-end intent and which according to Simester et al. (2019) is deemed equivalent to direct intent. All intermediate stages caused by an agent which are necessary to obtain for some ultimate intended outcome, are also intended.

Definition 2 (*Means-End Intent*) An Agent D Means-End intends some result $X = x$ through action $A = a$ if all of the following are true:

1. **An intended result exists** There exists some other result $Y = y$ which D directly intends by performing actions $A^+ = a^+$
2. **Causality** State $X = x$ is caused by a
3. **Action(s) subset** $A = a$ is contained in $A = a^+$, equivalently $A \subset A^+$ and a is a sub-sequence of a^+
4. **Necessary intermediate result** State $X = x$ is a necessary for state $Y = y$ to occur.

For completion, we state the equivalence of Means-End Intent with Direct Intent as asserted both in Simester et al. (2019) and Bratman (2009).

Theorem 1 *Something Means-End intended is culpably equivalent to something that is directly intended.*

Example 2 Article 5(1)(a-b) of The Draft EU AI Act (CNECT 2021) prohibits "putting into service or use of an AI system that deploys subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm".

Company ABC operates a video-content platform which recommends videos for its users using an algorithm. The algorithm has been trained through reinforcement learning with an objective to maximise the amount of time a user spends on the website. The algorithm has learned that by attempting to make users angry or distressed (by choosing certain types of extreme content), it can with probability p_{eng}

cause them to stay 'hyper-engaged' thereby earning the company more advertising revenue.

If we interpret 'in order to' as "with the intent to" and we assume that there is a reasonable likelihood that a distressed user has suffered psychological harm, does the algorithm fall foul of the prohibition in Article 5 assuming it has just been trained to maximise user engagement? The algorithm attempts to cause hyper-engagement by choosing content for the user. It would be disappointed if the user were not to be hyper engaged since they would spend less time on the site. The algorithm can therefore be said to intend to cause hyper-engagement as long as the probability of it happening is non-zero $p_{eng} > 0$.

If the algorithm believes it is necessary to cause users to be angry or distressed in order for them to be hyper-engaged, then this is an example of means-end intent. It intends to materially distort the user's behaviour.

Next we will consider oblique intent, which like Means-End intent, relies on a definition of direct intent already being in place.

Definition 3 Oblique Intent An agent D obliquely intends a result $X = x$ through actions $A = a$ iff:

1. **Intended outcome exists** There exists result $Y = y$, such that D intends $Y = y$ through actions $A = a$
2. **No Intention to avoid** $Y = y$ is not the negation of $X = x$ nor any necessary causes of $X = x$
3. Either of the following are true and they would be almost certainly true according to D at the point of a's commission:

   (a) **Side effect of Action** actions $A = a$ also causes result $X = x$
   (b) **Side effect of Outcome** result $Y = y$ and actions $A = a$ cause result $X = x$

Note that two probabilities are relevant in this definition. Firstly the probability of the side-effect happening as a result of action, and secondly the probability of the side-effect happening, contingent on the directly intended outcome $Y = y$ coming to pass. Smith (1990) terms the latter "*A result which will occur if the actor's purpose is achieved.*" An feature of oblique intent over direct intent is that there is no requirement to know the aim of D, only that one in exists (because it intends *something* through its actions). The abstraction of aim might be time-saving both for an A-bot using this as a planning restriction and a court which is considering an Agent's actions.

Example 3 Consider the same A-bot as in Example 2 but suppose user distress is not necessary for hyper-engagement but is an almost certain consequence of it. The A-bot no longer intends user distress (since it would not be disappointed if the user were not distressed as long as they were still hyper engaged). However, it obliquely intends the user to be distressed. This is the case regardless of the probability of

hyper-engaging the user. Note that this differs from the MPC formulation of culpable knowledge.

Example 4 Company DEF has invented a minimally invasive autonomous robotic surgeon to remove critical brain tumours. The skill of robo-surgeon is beyond that of human surgeons. In a specific case, the patient's chance of surgery survival was very low, but the chance of survival without surgery was zero. Unfortunately the surgery is not successful and the patient dies as a result. Did the robo-surgeon obliquely intend patient death? Whilst it was an almost certain consequence of operating, since the robo-surgeon's intention was to save the patient through surgery, which is the negation of death, death was not obliquely intended.

In the spirit of Child (2017) we will now present a definition of ulterior intent, that is to say the intent of doing something in the future to cause some result. This is different from Definition 1 which defines intent at the point of commission (whereby the intended result will occur in the future). Aside from the existence of ulterior offences, this is an extremely useful thing to do from the perspective of planning ahead. An A-bot will have to plan ahead such that it can never be put itself in a position in the future where it breaks some law by default. In the field of model checking (Baier and Katoen 2008), this called deadlock, and techniques have been developed to check for it in algorithms. Given the track record of AI finding various ways of cheating in any task (Lehman et al. 2020), one can imagine an A-bot deliberately finding ways to narrow its future choices to one, thereby sidestepping the definition of intentional action. Child does not require an agent with ulterior intent to make any forecasts about the likelihood of the conditions under which something is intended in the future, nor does he require the agent to have a 'pro-attitude' towards the conditions under which they intend to do something in the future.

Definition 4 Ulterior intent At time t_1 agent D has ulterior (oblique) intent for future result $X = x$ through actions $A = a$ iff:

1. **Second point coincidence** There exists a foreseeable (according to D) context or state of the world $S = s$ at time $t_2 > t_1$ such that D (obliquely) intends result $X = x$ through actions $A = a$.
2. **Commitment to conditional action** At t_1 D is committed to performing actions $A = a$ at t_2 in the future should context $S = s$ occur.

The second point coincidence requirement is one of time consistency. D should not be said to be intending to do something in the future, unless there exists a point in the future where they intend to do that thing. The commitment requirement is present to distinguish between a potential plan and an intention to do something. Proving that an D will act in a certain way in the future is potentially easier when D is an A-bot then when they are a human, because we do at least have the potential to examine the inner workings of the A-bot and simulate future action. An implication of the UK Criminal Attempts Act is that on deployment, an AI with some ulterior

intent to commit a crime, under any particular circumstance in the future is already committing a crime. This is pre-crime of the Minority Report variety and might lead to unexpected problems though is certainly an incentive for developers to understand and monitor what their creations intend on releasing them.

Example 5 Consider the A-bot in Example 2 but this time suppose the recommender algorithm notices that users who click on certain initial 'trigger' content are more likely to be hyper-engaged. The algorithm only attempts to hyper-engage if a user clicks on 'trigger' content. Does the algorithm intend to hyper-engage users? Yes. Conditions exist (the user has clicked on trigger content) under which the algorithm intends to hyper-engage. As long as the algorithm is committed (does not change) between the point of time before a user clicks on trigger content and afterwards.

4 Discussion

A key assumption behind creating a definition for intent applicable for algorithms is that the concept of intent exists outside the human mind. Can something be defined for certain algorithms which is to all intents and purposes the same as a folk concept of intent? The existence of corporate criminal offences, indicates that the answer is potentially yes. A counter argument might state that this is solely possible because companies are composed of humans who act with intent. But at the very least, mens rea is different in these entities which are comprised of multiple humans and the law has adapted to cope. From a biological standpoint, humans demonstrably do not have a monopoly on intentional acts. For example, crows in New Caledonia choose suitable sticks from which they fashion hooks to retrieve grubs from trees. Under test conditions, outside the forest, they can create suitable hooks out of wire (Weir et al. 2002). Furthermore they have been shown to be able to plan for the future use of a tool (Boeckle et al. 2020). Moving away from vertebrates, cephalopods like octopi, with their nine brains, have shown the ability, amongst other cognitive feats to use tools (Finn et al. 2009). An even more extreme example, and more akin to the idea of intent within a corporation, is that of the deliberation process that bee colonies undergo when considering different sites to move to when swarming (Passino et al. 2008). Many potential new colony locations are tested by a number of site assessing scout bees, before their conclusions are communicated back to the main swarm body, defective sites are rejected through a process of voting and eventually a consensus is reached. Completing the circle back to humanity, Reina et al. (2018) show that the cognition of a swarm has connections with the properties of the human brain when individual bees are viewed as a interconnected neurons. These different types of intelligence, which originate from very different evolutionary paths demonstrate behaviours which we would generally recognise as indicating intent, it does not seem inconceivable that an algorithm could demonstrate it. A huge advantage in an analysis of intent in algorithms is the opportunity to look inside them in a way which we cannot do with a human, company, raven, octopus or bee colony. Whilst what we find inside an algorithm might admittedly not always

be immediately interpretable, black-box analysis should at the very least allow accurate counterfactual interrogation which will considerably aid the process of evidence gathering.

The definitions that I presented make some requirements concerning the capacity of the A-bot, over and above the initial assumption that its behaviour is self-directed and that it makes decisions without consulting a human. A requirements based approach to legal A-bots is presented in Ashton (2021b) but I will summarise the requirements here. Most fundamentally the A-bot should have two features. Firstly it should have some sort of causal model of the world for it to be able to know whether action a has a causal relationship with variable X. Secondly it should have some sort of preference ordering over states of the world. The preference ordering requirement allows us to ascribe aim or desire to the A-bot. It seems to me that algorithms with an objective function go some way to meeting this requirement. The causal model requirement allows us to determine whether an A-bot knows the consequences of its actions. Without this ability, the ascription of intent to an A-bot, which is a future oriented concept, seems troublesome. Unfortunately many popular current designs of Reinforcement Learning (RL) algorithms imbue the A-bot with no ability to know the future states of the world—they have no causal model and are said to be model-free. As Gershman (2015) posits, model-free methods also drive human behaviour for routine tasks citing the example of travel between office and home, which in a pre-pandemic world was so routine it required little or no reasoning to accomplish. One would still say that the commuter is still intending to travel home, their intention being possible by the many times they have made the journey before. Even though model-free RL A-bots do not have a causal understanding of the world, they are still trained with a model of the world, so it may be that this model is also invoked as a scaffold when considering their intentional status. The lack of legal personhood does mean we are somewhat free to interpret the boundaries of an A-bot. It might mean we are free to impute intent with reference to not only the algorithm but also any training data or simulators it used. When we judge intent in humans we do use our knowledge of the world to aid us, and this seems analogous.

Just because intent may exist as a concept outside humans, it does not follow that its presence or absence has any relevance to the the culpability of the actor according to the victim of some AI-crime. It is for this reason that this article has focused on those crimes where mens rea plays a definitional role, or as I have named them here why-crimes. The inability to determine intent in A-bots does demonstrably make certain laws unenforceable. There is a reason that these laws rely on intent to define the harm that they outlaw. Intent as a construct, gives legislators fine control over the boundary between acceptable and unacceptable behaviour. Unless we decide that these wrongs are no longer wrong, I'm not sure how we can proceed without a definition of intent.

Aside from this I suspect that it will be very important for people to understand the purpose behind any A-bot's harm causing actions. This is a question which I feel can only be answered legitimately by surveying the public in a rigorously. A-bots do present novel challenges to the law which cannot be answered by making to the past. The question as to whether criminal law is suitable for application to A-bots is called The Eligibility Challenge and debated at great length in Abbott and Sarch (2020).

Aside from determining the culpability of an algorithm for harms caused, the concept of intent does have safety applications for the users and developers of A-bots. In many situations it would be desirable to ask an A-bot what it intends to do, and for the A-bot to reply truthfully. The A-bot's intentions might not be malign but they may well be likely to cause some harm if the A-bot doesn't have some piece of information that the interrogator has. Likewise in the situations where an A-bot has caused some harm, the question as to why it did so can inform the interrogator as to whether the harm was a freak accident or whether a flaw in the reasoning and behaviour of the A-bot was the cause. This information could be used to subsequently improve the safety of A-bot. There are overlaps in the ex-ante and ex-post use of algorithmic intent I have described here with the subject area of Explainable AI (XAI). A growing body of research exists concerning the interpretation of agent behaviour, though as Chakraborti et al. (2019) point out, many conflicting and overlapping concepts have been created to assess intent through behaviour. In a systematic review of what they term goal-driven XAI, Anjomshoae et al. (2019) find Intent communication a common objective but find that 32 of the 62 papers in the review do not rely on any theoretical background to produce explanations. Of the remainder, a third used Folk Psychology. Researchers are not commonly using a definition of intent inspired from law it seems.

The focus of this article has been firmly on criminal law, but other aspects of law also make routine reference to intent. The role of mens rea in Tort is much reduced but it still has a function Cane (2019). Several intentional torts exist, most pertinently for A-bots are those concerning economic crimes such as conspiracy and fraud or deceit. A requirement of intent here, is as discussed in Sect. 1.1 so as to raise the bar for tortious activity so as not to impede the functioning of markets. In the USA, the presence of intent for caused harms can also justify punitive (above economic cost) damages which punishes the tortfeasor and deters others from doing the same thing (Klass 2007). In an effort to study deceit across a wide range of law types including criminal, contract, tort and securities (Klass 2012) identifies purpose based law as a reoccurring method to regulate deceitful activity. That he characterises deceit law as a method of regulating the flow of information between parties is interesting given the use of algorithms to consume and serve data to counterparties. The ability to have truthful intentions about future behaviour is foundational to contract law as Klass and Ayres (2006) observe.

5 Other accounts of intent in and for AI

Bathaee (2018) identifies the difficulty of prosecuting why-crimes when the actor is an algorithm. He names the intent part of the actus reus 'basis intent'. He also identifies the role that intent has as a gatekeeper in litigating certain harms—If there is no possibility of showing the requisite intent (as in the case of an AI decision makers), the case cannot even be brought. The example chosen is Washington v Davis,[8] where the US supreme court ruled that statute which has a racially discriminatory

[8] Washington v. Davis, 426 U.S. 229,248 (1976).

effect but wasn't adopted with the intention of being racially discriminatory, is not unconstitutional. The possibility of an autonomous algorithm or AI possessing the Mens Rea for a crime, is tentatively suggested as a solution to the problem of 'Hard' AI crimes by Abbott and Sarch (2020). Someone is criminally culpable if their behaviour shows insufficient regard for some legally protected norms or interests. In their view if the AI has goals, gathers information and processes it to form strategies to fulfil those goals and is also aware of its legal requirements, it could be considered to show disregard, if it still acts in a way to breach those requirements. If this were the case, they recognise the need to draw up a definition of intent in AI that courts would use as a test. Interestingly, they cite Bratman (1990) as a starting point for this, and not the legal definitions we saw in the previous section. They posit that intention could be deduced through an A-bot's actions which increase the likelihood of an outcome happening. This is similar in spirit to the implicit aim clause discussed in Sect. 3. An interesting aspect of their discussion of mens rea in A-bots, and one which this article does not consider in detail, is that of knowledge. Defining knowledge of a fact F[9] as something which is known by the A-bot to be practically certain. We have mostly assumed that the A-bot knows of the circumstances that it is in at any point of time. Intent as it applies to knowledge seems a strange concept for the uninitiated, but it defines many crimes, modifying otherwise regular activities into criminal ones. The transport of a package for example becomes generally illegal when the contents are known to be restricted (drugs, explosives, firearms etc). Indeed as Shute (2002) says, even within legal discourse, relatively little time has been spent considering the subjects of knowledge and belief as they apply to mens rea.

Lagioia and Sartor (2020) examine the capacity of an AI to commit a crime by looking at its ability to accomplish actus reus with the required mens rea. They illustrate their discussion with the case of the Random Darknet Shopper, an algorithm programmed in Switzerland to go onto the darknet and buy some objects at random for display in an art exhibition. In the process it bought some Ecstasy tablets, possession of which is a criminal offence. The Cantonal prosecutor initially wanted to press charges but they were dropped when satisfied that the tablets were not to be sold or consumed (Kasperkevic 2015). Lagioia and Sartor conclude that an AI can have actus reus. Their discussion of mens-rea is divided into two, covering what they term the cognitive and volitional elements. For the cognition element, they conclude that an AI is fully able to Perceive its environment, comprehend it and make future projections about it. For the volition part they also adopt the Bratman's Belief, Desire, Intent framework. They define beliefs as the agent's current awareness of a situation plus any inferences it can make from them. Desire incorporates the motivation of the agent. The agent can have many desires which may conflict.

[9] The discussion of deducible facts from knowledge belongs to the symbolic side of AI, which relies on formal logic techniques. Statistical approaches to AI are very likely not to approach facts in the same way. There the world has some measurable states and possibly some hidden ones which may have an associated probability distribution as to their state.

The agent's intent is some conclusion of their beliefs and desires. It is a commitment to a plan to bring about some result. Unlike desires, intentions cannot conflict, they must, Bratman insists, be temporally consistent (Bratman 2009). Someone in London intending to fly to Los Angeles tomorrow cannot also intend to fly to Shenzhen tomorrow. Lagioia and Sartor conclude that an AI agent, programmed in such a way as to have Beliefs, Desires and Intentions (manifested as plans to deliver desires) can have sufficient mens rea to commit a crime.[10]

A Beliefs, Desires and Intentions software design paradigm does exist (Kinny et al. 1996), which can be used construct AI systems. Cohen and Levesque (1990) is one of the earliest formalism of intent inspired by Bratman's work. It creates a modal logic with primitive operators covering the initiation and completion of actions as well as some that can express beliefs and goals. As with the approach of this article, they then define intent in terms of other components. Thus an intention to act is described as a goal to have completed that action. An intention to achieve a certain state is the goal of having done a certain set of actions that achieves that state, at least an initial plan of actions to reach that state and a requirement that what does happen, in the process of achieving the state, is not something which is not a goal. The last clause is to stop an agent having said to have intentionally caused a state when their goal was reached accidentally as a result of their actions. The development of a model logic to reason about intent is an extremely useful thing to do for an algorithm to plan ahead.

Outside BDI architecture, formal accounts of intent, compatible with an AI, are surprisingly rare. Recent advances in AI capability have been rooted in statistical AI, which emphasises the use of data and statistical inference over logical reasoning. It is desirable that a theory of intention in AI is relatively agnostic to the type of AI it is being applied to, given a certain level of requirements. The closest approaches to those in this article are to be found in the related accounts of Kleiman-Weiner et al. (2015) and Halpern and Kleiman-Weiner (2018). Both of which define what this article calls direct intent using counterfactual reasoning and an assumption of utility maximising behaviour. Loosely speaking, intended outcomes are the minimum set of outcomes with the property that if they are not obtainable, then the optimal policy would change. Note the similarity with the counterfactual aim condition in Sect. 3. Kleiman-Weiner et al use an influence diagram setting, an Influence Diagram (ID) being a directed acyclic graph with action, chance and terminal utility outcomes. The directed arcs between nodes of the graph are interpreted as causes. Their approach is used on a variety of trolley problem type scenarios, and is developed in conjunction with a theory of moral permissibility. People's ability to infer intent is tested in a survey experiment and tested versus the formal definition for validity. In the event of an A-bot being involved in a trial, this is a task which jurors will be required to do should they be unable to access or interpret an A-bot's internal workings. The counterfactual approach is modified slightly in Halpern and

[10] An argument can be made that Bratman's theories influenced and were influenced by the progress of AI in the 1980s. Thus any theory of intent in law calling upon Bratman, is inadvertently influenced by theories of (symbolic) AI. Which is neat.

Kleiman-Weiner (2018) and translated to the world of Structural Equation Models (SEMs), of the type used in Actual Causality (Halpern 2016). The modifications allow the definition to be more robust to a variety of counterexamples, and the SEM setting allows an arguably clearer treatment of counterfactuals, perhaps at cost of clarity over the utility function which is more naturally positioned in an Influence Diagram. Like the definition in this article, an action can only be intended if there were other actions which could have been taken at the point of commission. An important point of difference in Halpern and Kleiman-Weiner (2018) is their use of a reference action set, when deciding whether an outcome was intended through an action. This is practical from a calculation point of view,[11] but also intuitive, where in most cases we can just compare acting with not acting in a certain way.

Just as Kleiman-Weiner et al develop their intent definition alongside one of moral permissibility, Halpern and Kleiman-Weiner develop theirs with one of blameworthiness. Both approaches to intent could be characterised as originating from a theory of ethical action which overlaps but does not coincide with a theory of intent based on legal theory. This is most obvious in their treatment of side effects, which are always unintended. Ashton (2021a) extends their approach to define oblique intent, thereby bringing their approach more in line with legal reasoning about side-effects.

6 Conclusion

This article builds some definitions of intention, from legal principles, which are suitable for application in an autonomous algorithmic actor or A-bot for short. It presents semi-formal definitions of direct, means-end, oblique and ulterior intent. These are informed by a review of legal literature on the subject of intent from common law jurisdictions which concludes with a list of desiderata concerning definitions of intent. Accounts of intent in algorithms in computer science from any background are rare, but are especially so from a legal one.

I have assumed throughout that the A-bot is auto-didactic in the sense that it learns how to behave itself and its precise actions are not directed by its creators. Under this assumption, there exist certain situations where the intent of the programmer cannot be read from the intent of the A-bot. This poses problems when the A-bot commits some harm.

Whilst A-bots are not legal persons they cannot commit crimes making the presence or absence of mens rea in them moot. Many would argue that they are not moral agents and cannot be held responsible for their actions. However, this article has argued that over and above its role in assigning culpability for harm, mens rea plays a role in defining harm in what we have called why-crimes. These include many inchoate crimes such as attempts but perhaps more relevantly also include many deceit derived crimes. A failure to identify intent in A-bots means that harms

[11] We have for instance assumed a discrete action set, but applications exist where actions are continuous in nature.

cannot be identified either by those responsible for the A-bots or those who job it is to uphold the law. The ability to define harms by the intentional state of the actor is important capability of the law and is used to avoid over-criminalisation of activity.

Acknowledgements Many thanks to Dr. John Child at Birmingham Law School, Dr. Sander Beckers at University of Tübingen, Prof. Mark Dsouza at UCL Faculty of Laws and Prof. Alex Sarch at University of Surrey for their feedback and help with the ideas presented in this article and its earlier drafts.

Funding Information This work has been funded by the UK EPSRC.

Declarations

Conflict of interests The authors declare that they have no conflict of interests.

References

Abbott R (2020) Reasonable robots. In: The reasonable robot, pp 50–70. https://doi.org/10.1017/9781108631761.004

Abbott R, Sarch A (2020) Punishing artificial intelligence: legal fiction or science fiction. Is law computable? pp 323–384. https://doi.org/10.5040/9781509937097.ch-008

ACCC (2005) Predatory pricing. Technical report 5, Australian Competition and Consumer Commission. http://www.austlii.edu.au/cgi-bin/viewdoc/au/journals/AUCCCUpdate/2005/5.html#

Alexander L, Kessler KD (1997) Mens rea and inchoate crimes. J Crim Law Criminol 87(4):1138. https://doi.org/10.2307/1144017

Alldridge P (1990) The doctrine of innocent agency. Crim Law Forum 2(1):45–83. https://doi.org/10.1007/BF01096228

Anjomshoae S, Najjar A, Calvaresi D, Främling K (2019) Explainable agents and robots: results from a systematic literature review. In: Proceedings of the 18th international conference on autonomous agents and multiagent systems, Montreal

Ashton H (2021a) Extending counterfactual accounts of intent to include oblique intent. http://arxiv.org/abs/2106.03684

Ashton H (2021b) What criminal and civil law tells us about safe RL techniques to generate law-abiding behaviour. In: Workshop on AI safety 2021 co-located with the thirty fifth AAAI conference on artificial intelligence. http://ceur-ws.org/Vol-2808/Paper_25.pdf

Bathaee Y (2018) The artificial intelligence black box and the failure of intent and causation. Harvard J Law Technol 31(2):890–938

Baier C, Katoen JP (2008) Principles of model checking. MIT Press, Cambridge

Bentham J (1823) An introduction to the principles of morals and legislation. https://www.earlymodernetexts.com/assets/pdfs/bentham1780.pdf, in the version by Jonathan Bennett

Boeckle M, Schiestl M, Frohnwieser A, Gruber R, Miller R, Suddendorf T, Gray RD, Taylor AH, Clayton NS (2020) New Caledonian crows plan for specific future tool use. In: Proceedings of the royal society b: biological sciences, vol 287(1938). https://doi.org/10.1098/rspb.2020.1490

Bratman ME (1990) What is intention? In: Cohen PR, Morgan J, Pollock ME (eds) Intentions in communication. MIT Press, Cambridge (**Chap 2**)
Bratman ME (2009) Intention, practical rationality, and self-governance. Ethics 119:411–443
Cane P (2019) Mens rea in tort law. Intent Law Philos 20(4):129–159. https://doi.org/10.4324/9781315187136-7
Chakraborti T, Kulkarni A, Sreedharan S, Smith DE, Kambhampati S (2019) Explicability? legibility? predictability? transparency? privacy? security? the emerging landscape of interpretable agent behavior. In: Proceedings of the twenty-ninth international conference on automated planning and scheduling, p 11
Child J (2017) Understanding ulterior mens REA: future conduct intention is conditional intention. Camb Law J 76(2):311–336. https://doi.org/10.1017/S000819731700040X
CFTC (2013) Antidisruptive Practices Authority Interpretative guidance and policy statement. Technical report RIN 3038-AD96, Commodity Futures Trading Commission
CNECT (2021) Proposal for a regulation of the European parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. Technical report COM/2021/206, European Commission, Directorate-General for Communications Networks, Content and Technology. https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975 &uri=CELEXuri%3A52021PC0206
Coffey G (2009) Codifying the meaning of 'intention' in the criminal law. J Crim Law 73(5):394–413. https://doi.org/10.1350/jcla.2009.73.5.590
Cohen PR, Levesque HJ (1990) Intention is choice with commitment. Artif Intell 42(2–3):213–261. https://doi.org/10.1016/0004-3702(90)90055-5
Criminal Prosecution Service (2019) Homicide: murder and manslaughter. https://www.cps.gov.uk/legal-guidance/homicide-murder-and-manslaughter
De Jong F (2011) Theorizing criminal intent: a methodological account. Utrecht Law Rev 7(1):1. https://doi.org/10.18352/ulr.144
Finn JK, Tregenza T, Norman MD (2009) Defensive tool use in a coconut-carrying octopus. Curr Biol 19(23):1069–1070
Fletcher GP (1971) Theory of criminal negligence: a comparative analysis. Univ Pa Law Rev 119(3):401–438
Furey JR (2010) A consistent approach to assessing mens rea in the criminal law of England and Wales. Ph.D. thesis, University of Exeter
Gershman SJ (2015) Reinforcement learning and causal models. In: Oxford handbook of causal reasoning, pp 1–32
Halpern JY (2016) Actual causality. MIT Press, Cambridge
Halpern JY, Kleiman-Weiner M (2018) Towards formal definitions of blameworthiness, intention, and moral responsibility. In: 32nd AAAI conference on artificial intelligence, AAAI 2018, pp 1853–1860
Hildebrandt M (2019) Closure: on ethics, code and law. In: Law for computer scientists, chap 11. Oxford University Press
Kasperkevic J (2015) Swiss police release robot that bought ecstasy online. https://www.theguardian.com/world/2015/apr/22/swiss-police-release-robot-random-darknet-shopper-ecstasy-deep-web
Kenny A (2013) Intention and side effects: the mens rea for murder. In: Keown J, George RP (eds) Reason, morality, and law: the philosophy of John Finnis, Oxford scholarship online, chap 7, pp 109–117. https://doi.org/10.1093/acprof:oso/9780199675500.001.0001
Kinny D, Georgeff M, Rao A (1996) A methodology and modelling technique for systems of BDI agents. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) 1038, pp 56–71. https://doi.org/10.1007/bfb0031846
Klass AB (2007) Punitive damages and valuing harm. Minnesota Law Rev 92(1):83–160
Klass G (2009) A conditional intent to perform. Leg Theory 15(2):107–147. https://doi.org/10.1017/S1352325209090089
Klass G (2012) Meaning, purpose, and cause in the law of deception. Georget Law J 100:446–449
Klass G, Ayres I (2006) New rules for promissory fraud. Ariz Law Rev 48:957–971
Kleiman-Weiner M, Gerstenberg T, Levine S, Tenenbaum JB (2015) Inference of intention and permissibility in moral decision making. In: Proceedings of the 37th annual conference of the cognitive science society, vol 1(1987), pp 1123–1128
Lagioia F, Sartor G (2020) AI systems under criminal law: a legal analysis and a regulatory perspective. Philos Technol 33(3):433–465. https://doi.org/10.1007/s13347-019-00362-x

Lehman J, Clune J, Misevic D (2020) The surprising creativity of digital evolution: a collection of anecdotes from the evolutionary computation and artificial life research communities. Artif Life 26(2):274–306. https://doi.org/10.1162/artl_a_00319

Loveless J (2010) Mens rea: intention, recklessness, negligence and gross negligence. In: Complete criminal law, 2nd edn, chap 3. Oxford University Press, pp 90–150

McIntyre A (2019) Doctrine of double effect. In: Zalta EN (ed) Stanford encyclopedia of philosophy, spring, 201st edn. Metaphysics Research Lab, Stanford University, Stanford

Ormerod D, Laird K (2021a) 4. Crimes of negligence. In: Smith, Hogan, and Ormerod's Criminal Law. Oxford University Press, pp 136–145. https://doi.org/10.1093/he/9780198849704.003.0004. https://www.oxfordlawtrove.com/view/10.1093/he/9780198849704.001.0001/he-9780198849704-chapter-4

Ormerod D, Laird K (2021b) 5. Crimes of strict liability. In: Smith, Hogan, and Ormerod's Criminal Law. Oxford University Press, pp 146–179. https://doi.org/10.1093/he/9780198849704.003.0005. https://www.oxfordlawtrove.com/view/10.1093/he/9780198849704.001.0001/he-9780198849704-chapter-5

Parsons S (2000) Intention in criminal law: why is it so difficult to find? Mountbatten J Legal Stud 4(1 & 2):5–19. https://doi.org/10.1017/s0841820900001375

Passino KM, Seeley TD, Visscher PK (2008) Swarm cognition in honey bees. Behav Ecol Sociobiol 62(3):401–414. https://doi.org/10.1007/s00265-007-0468-1

Reina A, Bose T, Trianni V, Marshall JA (2018) Psychophysical Laws and the Superorganism. Sci Rep 8(1):1–8. https://doi.org/10.1038/s41598-018-22616-y

Robbins IP (1990) The ostrich instruction: deliberate ignorance as a criminal mens rea. J Crim Law Criminol (1973-) 81(2):191. https://doi.org/10.2307/1143906

Sales P (2019) Algorithms, artificial intelligence and the law. https://www.bailii.org/bailii/lecture/06.pdf

Shute S (2002) Knowledge and belief in the criminal law. In: Shute S, Simester A (eds) Criminal law theory: doctrines of the general part, Oxford scholarship online, chap 8. https://doi.org/10.1093/acprof:oso/9780199243495.001.0001

Simester AP (2021) Fundamentals of criminal law: responsibility, culpability, and wrongdoing, 1st edn. Oxford University Press, Oxford, UK, oCLC: on1242932280

Simester AP, Spencer JR, Stark F, Sullivan GR, Virgo GJ (2019) Mens rea. In: Simester and Sullivan's criminal law, 7th edn, Hart, chap 5, pp 137–190

Smith JC (1990) A note on "intention''. Crim Law Rev Feb:85–91

Stark F (2017) Introduction. In: Culpable carelessness: recklessness and negligence in the criminal law, chap 1. Cambridge University Press, Cambridge, pp 1–25. https://doi.org/10.1017/CBO9781139855945.001

Storey T (2019) Inchoate offences. In: Unlocking criminal law, 7th edn, chap 6. Routledge, pp 137–170. https://doi.org/10.4324/9780429322303

Taylor G (2004) Concepts of intention in German criminal law. Oxf J Leg Stud 24(1):99–127. https://doi.org/10.1093/ojls/24.1.99

The American Law Insitute (2017) General requirements of culpability. https://archive.org/details/ModelPenalCode_ALI/page/n31/mode/2up

The Law Commission (1989) A criminal code for England and Wales. Volume 1: report and draft criminal code bill, vol 177. HMSO. http://www.lawcom.gov.uk/app/uploads/2015/06/Criminal_Code_177_1.pdf

The Law Commission (1993) Legislating the criminal code: offences against the person and general principles. 218. HMSO

The Law Commission (2007) Conspiracy and attempts: a consultation paper. http://www.lawcom.gov.uk/app/uploads/2015/03/cp183_Conspiracy_and_Attempts_Consultation.pdf

The Law Commission (2015a) Appendix C: Home office draft bill. In: Reform of offences the person, William Lea Group on behalf of HMSO, pp 212–232. http://www.lawcom.gov.uk/app/uploads/2015/11/51950-LC-HC555_Web.pdf

The Law Commission (2015b) Reform of offences against the person (report). http://www.lawcom.gov.uk/app/uploads/2015/11/51950-LC-HC555_Web.pdf

Weir AA, Chappell J, Kacelnik A (2002) Shaping of hooks in new Caledonian crows. Science 297(5583):981. https://doi.org/10.1126/science.1073433

Williams G (1987) Oblique intention. Camb Law J 46(3):417–438. https://doi.org/10.1017/S0008197300117453

Yaffe G (2004) Conditional intent and mens rea. Leg Theory 10(4):273–310. https://doi.org/10.1017/S135232520404025X