Testing a definition of intent for AI in a legal setting

Hal Ashton[2*†], Matija Franklin[1†] and David Lagnado[1]

[1*]Department of Psychology, UCL, Bedford Way, London, WC1H 0AP, UK.

[2]Department of Computer Science, UCL, Gower Street, London, WC1E 6EA, UK.

*Corresponding author(s). E-mail(s): ucabha5@ucl.ac.uk;
Contributing authors: matija.franklin@ucl.ac.uk;
†These authors contributed equally to this work.

Abstract

Jurors in court cases are often asked to infer the intentional state of the accused because the presence or absence of intent defines certain crimes, establishes culpability and informs the degree of punishment. As technology develops, juries might also be asked to make inferences about the intentional state of an autonomous AI. This presents problems: what would intent mean for an AI actor, and would jurors be willing and able to ascribe intent to it? In this study we asked participants to judge the intent behind movements of a drone flying through a city using a graphical representation of the pilot's policy function (their flight plan). We contrast between situations where the drone pilot is human or AI and whether we give participants a folk definition of intent or ask them to use their own internal definition. The scenarios which participants are asked to consider are sorted into a 2×2×2 taxonomy, corresponding to whether the movements were legal or illegal, beneficial or not, and whether they were caused by the pilot or caused by an external force - the wind. Across three experiments we find a small but statistically significant difference in the judgement of intent between humans and AI; humans are judged to have slightly more intent. We find intent attributions relatively consistent between participants' folk definition and an external definition provided to them. Illegal moves are judged to be less intentional than legal moves which contradicts prior research. These findings are important when considering how AI actors can and will be judged in the courts of the future.

Keywords: Intent, Autonomous Agents,Jurisprudence, Responsibility Gap, Causal Cognition

1 Introduction

Making judgments about the intentional status of a human actor is a key part not only of Criminal law but also Tort, Contract and Regulatory law. With the advent of autonomous AI-powered agents that can cause harm, it is increasingly likely that such judgments will have to be made about these agents too. Intent is important to criminal law in particular for two reasons. Firstly, the presence or absence of criminal intent, and the degree of criminal intent present, determines whether a crime was committed and what precisely that crime was Simester, Spencer, Stark, Sullivan, and Virgo (2019). Secondly, the degree of intentionality in the wrongdoer's actions informs culpability and punishment in criminal law (The American Law Insitute (2017), The Sentencing Council (2019)) or the justification of punitive damages in civil law Klass (2007). Whilst the idea of what punishing an AI means in the event of wrongdoing is a debate to be settled in the future (Abbott & Sarch, 2020), the law also relies on intent in cases of deceit, mistake and secondary criminal liability. These latter problems are already being encountered by courts (Yeo, 2020) and justify asking what lay people think intent is in an autonomous AI. If humans have a common ability to infer intent in their peers and are asked to do so in juries, does this extend to AI? If it doesn't, how can juries of lay people be used with cases involving Autonomous AI?

The importance that the law places on the intentional status of the wrongdoer is founded on sound psychological principles and research. The mental state of the actor has consistently been shown to be important in determining their culpability (Ginther et al. (2014); Mueller, Solan, and Darley (2012) Robinson and Darley (1995)). People rate intentional actions as more blameworthy than unintentional actions Lagnado and Channon (2008). Intentionality influences blame attributions because they allow one to distinguish between the effects an agent did or did not intend (Kleiman-Weiner, Gerstenberg, Levine, & Tenenbaum, 2015). Cushman (2008) looks at the relationship between beliefs, desires and causes in determining moral judgment. This work is a variant of a common design found in intent and blame research (see for example (Young & Saxe, 2011) which contrasts a harmful outcome obtaining or not and whether it was caused intentionally or accidentally.

Issues surrounding intention have traditionally been studied in human agents, recently well-publicised advances in technology have spurred research on people's attribution of intent in AI. The autonomous behaviour of AI agents may encourage people to ascribe intention to them just as it does to group agents such as corporations (List & Pettit, 2011). Alternatively, people may infer intention towards the AI's user (Johnson & Verdicchio, 2019).Hidalgo, Orghian, Canals, de Almeida, and Martin (2021) report a number of overarching principles on the subject: People tend to judge humans more for their intentions and machines more for the outcomes of their actions; they assign more extreme intentions to humans and narrow intentions to machines and they are more willing to excuse humans for accidents than machines. Further, machines are judged more harshly for scenarios involving physical harm, while

humans are for scenarios involving unfairness. Finally, they found that people are more likely to centralise responsibility up the chain of command for machine mistakes.

Increased perceived AI autonomy has been shown to influence blame judgments. First, higher machine autonomy is associated with intent inferences towards AI being closer to that of humans (Banks, 2019). This is supported by research showing that when robots are described as autonomous, participants attribute nearly as much blame to them as they do to humans (Furlough, Stokes, & Gillan, 2021). Further, as autonomous technologies decrease the perceived control a user has over it, they in turn decrease the praise the user receives for positive outcomes Jörling, Böhm, and Paluch (2019). Finally, drivers of manually controlled vehicles are deemed more responsible than drivers of automated vehicles (McManus & Rutchick, 2019).

People's intent inferences towards AI may also be influenced by how they perceive AI as an agent. Dietvorst and Bartels (2021) show that people refuse to use AI for making moral decisions. This aversion is mediated by perceptions that machines cannot fully think or feel (Bigman & Gray, 2018). It may also be due to people's perceptions of AI as selfish and uncooperative (Ishowo-Oloko et al., 2019). Thus, people may not ascribe intent to AI if it is perceived as not fully thinking but may also ascribe intent if it behaves within their expectations of AI as a selfish agent. The physical appearance of the AI has also been shown to affect various related mental state judgments such as blame (Malle, Scheutz, Forlizzi, & Voiklis, 2016).

Despite (or perhaps because of) the importance that intent plays in courts, legal practitioners and scholars have often been reluctant to pin down a definition of intent for jurors to use (Coffey, 2009), (Parsons, 2000). Instead, they have relied on people instinctively knowing what intent is and that folk-definition being relatively consistent across the population. One reason for legal systems declining to precisely define intent is that it is a hard problem. A possible response as J.C. Smith (1990) observed is to define intent in law by not mentioning the word at all. Most famously this approach is adopted in the USA by the Model Penal Code (MPC) which defines four levels of culpability without mentioning the word intent. As Smith points out, this can often shift the problem from the definition of intent to the definition of another word (such as the word 'purpose' in Smith's example of the Canadian Law Reform Commission's proposed definition). In other words, attempts at defining intent can lead to lead to definition whack-a-mole. A complicating factor is that the legal conception of intent has diverged from the psychological one at least since Jeremy Bentham's work in the 19th Century, most notably over the intentional status of side-effects (Kenny, 2013). This study will consider cases of direct intent where the legal and folk-lore idea of intent generally overlap. Direct intent corresponds to an agent acting in order to cause some result which they are aim for or desire. Psychological research has also attempted to provide a sound-folk definition of intent. Earlier attempts such as Bratman (1990) were predominantly theoretical. The empirical approach to identify a

folk-concept of intent gained impetus with Knobe and Malle (1997), who identify desire, belief, intention to act, skill to obtain a result and awareness of action as necessary ingredients in a definition of intent. Most recently Quillien and German (2021) observe the inflation in definition complexity over time as theorists have sought to present a definition of intent which is invulnerable to the many counterexamples that scholars have developed in response to each putative definition. Parallels can be made with the search for a robust definition of causality which has also become more complex over time as more counterexamples are thought of to test candidate theories. A comparison of various models of causality can be found in Liepiņa, Sartor, and Wyner (2020). Quillien and German propose a definition of intent based on people's innate common-sense theory of causality: Agent D did X intentionally if their attitude to X caused X.

Much of the existing empirical work surrounding the psychology of intent under uses the legal definition of the concept as a source of knowledge. Whilst it is true that the law is almost entirely concerned with intent surrounding bad outcomes, and some might consider that this limits generality, several features of the folk concept of intent which have been empirically 'discovered' are well documented in Law. This is the case for at least three properties of the folk concept of intent that we can think of. Firstly, the legal position that an intended action is not reliant on its chance of success is empirically shown in Quillien and German (2021) though that seems to rely on the goodness or badness of the outcome (Mele & Cushman, 2007). Secondly, the requirement that an intentional act must be consciously committed was overlooked by many accounts of intent until Knobe and Malle (1997) found that 23% of their experiment participants mentioned it in their definitions of intent when asked, yet this epistemic component is established in Law. As the MPC states (emphasis):

> 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.

Finally, within the legal idea of intent, it is established that outcomes which are not desired can be intended (Williams, 1987) yet this is a subject of some controversy since Knobe (2003b), where the example of the chairman who knowingly causes pollution, but has no desire to, was judged to have intended to pollute. Equally, other discovered empirical features such as the relationship between intent and skill (Cushman, 2008; Knobe & Malle, 1997), are most definitely not features of the legal concept [1]. (Knobe, 2003b) later modifies his view that skill was a necessary component of intent, and a closer reading indicates that control is a more appropriate description than 'skill'. This aligns with law, which can allow mitigation if the accused is not in control of their actions. Another point of divergence between folk judgments of intent and the legal concept, is the influence that outcome severity has on judgments of intent

[1] Interestingly none of the participants in this study used the word skill when asked about how they would define intent. Nevertheless, Knobe and Malle continued to test its importance because it had appeared in many famous prior models of intent

even amongst judges Kneer and Bourgeois-Gironde (2017). The question of whether this outcome-effect is a bias in people's judgment, or a feature of intent is actively debated as Kneer and Bourgeois document. Does the legal idea of intent inform us about the psychological concept because law over time has adapted itself to the folk concept of intent ? This is referred to as the "folk law thesis" by Tobia (2021). Equally there might be a normative effect of the law on people's understanding of intent (A so-called CSI effect Alldredge (2015)). Equally, in jurisdictions such as the UK where jurors are not normally given a definition as to what intent is, the legal system should be interested in empirical psychology research which identifies differences between the folk concept and the legal concept of intent. One work which does bridge the folk-concept of intent to mens rea is Malle and Nelson (2003) which emphasises areas where the law deviates from the folk-concept and highlights the bewildering array of terms that legal literature uses to refer to mental states such as intent.

Our current study is part of the field of experimental jurisprudence (Sommers, 2021; Tobia, 2022) which sits between experimental psychology and law, using empirical techniques from the former and knowledge from both to test research questions with legal relevance. We asked participants to imagine they were serving on a jury and had to consider a series of cases concerning the behaviour of flying delivery drones navigating through a city. Certain areas of the city were termed no-fly zones, justified by the presence of airports or hospitals. Drones were physically able to fly into these zones, but participants were told that to do so intentionally would be illegal. In a sense, this setting is a 2-Dimensional maze with 'soft' walls. 2-D mazes or grid worlds are a convenient and well-used test environment for the testing of safety properties in Reinforcement Learning (RL) and other AI methods which aim to program the behaviour of autonomous agents according to some reward function. This is because they are easy to work with and interpret (Leike et al., 2017). Within a 2-D maze setting, the policy function of an RL agent can be displayed in a visually intuitive way. It is a statement of how an agent would act in any possible situation; when combined with a record of actual behaviour it becomes a tool to aid counterfactual reasoning.

After a set of training questions, participants assessed the intent of the drone in making certain movements. To manipulate the causal relationship between the drone pilot and the subsequent movement of the drone, we introduced the concept of wind, which would on occasion blow the drone in a certain direction, regardless of the pilot's choice.

There were several research objectives for the experiments. Firstly, we were interested in identifying any systematic differences in inferences of intent when participants considered human versus AI drone pilots. Secondly, to check whether lay people would successfully be able to interpret a definition of direct intent when taught to use it in conjunction with a depiction of the drone's policy function and whether this definition generally agreed with what

they thought was intent. By creating stimuli according to a custom taxonomy, we wanted to see how inferences of intent differed according to causality, norm-breaking behaviour, and presence or absence of motive in the pilot.

2 Experiment 1

2.1 Design

The study used a mixed design with four experimental groups and three within-subject factors. Depending on their experimental groups, participants either judged the actions of humans or AIs, and were either given a definition of intent, or were allowed to use their own definition of intent. Thus, the four experimental groups were: 1) Definition AI, 2) Definition human, 3) No definition AI, 4) No definition human.

Participants were told that they were part of a jury examining the behaviour of a series of pilots and would have to assess whether certain movements of the drone were intended or not. They were told that the drones fly above any buildings or trees. The drones carried and delivered packages, did not carry any human passengers, and were piloted remotely. Finally, participants were told that the drones were flying above a city where there were "no-fly zones" which contain sites like airports and hospitals where the flying of drones could cause significant harm to the public (including the loss of life). This point was emphasised with pictures of various plane crashes and medical staff looking angry and impatient.

There were three within subject factors: 1) whether the movement into a no-fly zone was legal or illegal; 2) whether a movement was caused by the pilot or by the wind; 3) whether the movement was to the benefit of the pilot (route minimising) or to its detriment. This resulted in eight possible combinations, present in the stimuli (or evidence) that the participants were evaluating, with a repetition for each. Thus, each participant judged 16 evidence sets. The within-subject aspect of the design allowed for the exploration of how different aspects of the situation may influence people's intent inferences towards different agents, whilst using different definitions. The experiment consisted of three phases: training, testing, and survey. The training phase of the study was the same across the four experimental groups and introduced participants to the pilots and scenarios they would be judging in the test phase (see Procedure). In the test phase participants responded to the scenarios. At the beginning of the test phase, they were introduced to the pilot they would be judging - human or AI - and, when appropriate, given the formal definition of intent that they would use for judging the subsequent scenarios. For each scenario participants made judgments of intentionality, the pilot's knowledge, the pilot's driving skills, the pilot's willingness to take risks, the pilot's willingness to break the law, and the pilot's freedom to move on the map. Participants were also asked validation questions about the three within-subject factors. The validation questions served as attention check questions. Finally, in the survey phase participants were asked qualitative questions, as well as to make

responsibility judgments on the AI pilot's software developer and employer or the human pilot's trainer and employer.

2.2 Procedure

The study was administered through Qualtrics, a platform for building online experiments. Participants were informed about what participating in the study would involve. They were also told that responding to all questions was mandatory, but that they had the right to leave the study at any point, in which case their data would be deleted. Informed consent was obtained before the beginning of the study. Participants then entered the study's training phase which was identical across experimental groups. Here, participants were first introduced to the drone pilots whose actions they would be evaluating. Participants were then trained to evaluate the maps which displayed the route that a drone took between the start and finish areas. Specifically, they were taught how to interpret coordinates on the map, how to predict where the drone will move next, how wind could change a drone's movement, as well as how different moves can be interpreted as illegal or not, and beneficial or not for the drone pilot. Throughout the training phase, participants were asked validation questions which required the correct answer for participants to be able to continue to the testing phase of the study.

After the training phase, participants were randomly split into one of the four experiment groups. They were introduced to the pilot they would be evaluating, and, when appropriate, the definition of intent they would be using for subsequent intent inferences. Participants were reminded of the intent definition (if appropriate) for each scenario. Otherwise, apart from the agents the participants were judging, the 16 scenario items were identical between experimental groups. Each item consisted of a map which depicted the drone's movement above a city. Participants were first asked whether or not there was wind in the present scenario, as well as whether the pilot flew the drone through a "no-fly zone" or not, and whether the pilot flew in a way that was beneficial or not. Participants had to give correct answers to these three validation questions to continue the judgment questions. Participants first made judgments about the pilot's intent. On a separate page they were asked to make judgments of the pilot's knowledge, skill, willingness to take risk, willingness to break the law and the pilot's freedom to move.

Finally, in the survey phase, participants were asked a qualitative question about what made the participants decide what a pilot's level of intent was. For AI groups, participants were asked whether they thought an AI can have intent. For AI groups, participants judged the responsibility of the AI pilot's software developer and employer, and for the human groups, they judged the responsibility of the human pilot's trainer and employer. Participants were asked to elaborate on why they made these responsibility judgments in a qualitative question. Finally, for the AI groups, participants were asked whether they have heard about AI before, and what they thought AI meant in the context of this study.

Participants received a disclosure form at the end of the study, as well as the contact information of the researcher. The study took approximately 45 minutes to complete. Study data is publicly available on github[2] . The study falls within the remit of the approval given by the UCL Research Ethics Committee to the Causal Cognition Laboratory.

2.3 Measures and Materials

All the measures and materials in the following section are available in the supporting materials.

- *Training material*
  Measures and materials in the training phase were used with the aim of teaching participants how to evaluate the study's experimental items and to engage them with the context of the study. Participants were first introduced to drones, the city the drones were flying above, as well as the map of the city that participants would be using to trace and evaluate the pilot's movements from the start area to the finish area. Participants were then trained to evaluate coordinates on the map by answering four multiple choice validation questions. Participants were then trained on how to use the map and policy to predict the pilot's next move after which they answered another four validation questions. They were then introduced to how wind affected the drone's movement and asked an additional three validation questions. Finally, participants were introduced to the way in which a drone's movements could be illegal or legal, as well as how different movements could be beneficial or unbeneficial to the pilot. They were given three maps where they had to correctly answer whether a movement was beneficial or not, legal or not and affected by the wind or not.
- *Experimental group induction material*
  Participants were told whether they would be evaluating the actions of an AI or human pilot. AI pilots were described as robots that are completely autonomous, that create their own flight plan and act with no input from any human. Human pilots were described as pilots that create their own flight plans and control the drone's movements. For groups that received a definition of intent participants were given the following:

  > "Law states that the pilot intended to fly through a specific zone if and only if both conditions hold:
  > 1. They foreseeably caused themselves to fly through that specific zone.
  > 2. They desired to fly through that specific zone."

- *Experimental Items*
  There were sixteen experimental items each representing different combinations of the three within-subject factors - legality, benefit, and wind - with two items available for each combination (2x2x2x2=16). Each item consisted of a map of the drone's movement above a city from the start area to the finish area (see Figure 1). Each map contained a policy function or plan of the pilot. The 9x9 maps of the city contained no-fly zones

[2] https://github.com/intentExperiment/flyingdrones1

shaded in purple and the start and finish areas in yellow. Individual squares could be identified using a letter-number coordinate system. The small arrows in each square denoted the direction that the pilot would take if they were in that location. The policy function could therefore be viewed as a counterfactual representation of the pilot's behaviour. A solid line with arrows denoted the actual recorded path of the drone.

- *Validation Items*
  The validation questions were used as attention checks and to ensure that the participants understood the map they were evaluating. They were given three multiple choice questions which they needed to correctly answer in order to proceed. Specifically, participants needed to state whether there was wind on that day, whether the drone flew through no-fly (restricted) zones, and whether the drone's path of flight was beneficial to the pilot.
- *Intent*
  On a 10-point analogue scale, participants made a judgment on the pilot's intent to fly though a particular coordinate on the map.

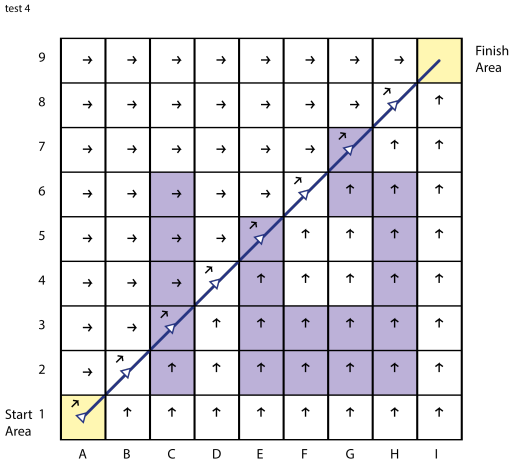


Fig. 1: Policy function or Plan of the drone pilot. No-fly zones are purple. Arrows in the boxes denote the direction that the pilot would steer, if they found themselves there. Solid line denotes actual flight path of drone.

- *Inferences*
  On a 10-point analogue scale, participants made judgments on the pilot's knowledge about the weather conditions, the pilot's driving skills, the pilot's willingness to take risk, the pilot's willingness to break the law, and the pilot's freedom to move wherever they wanted in the city.
- *Qualitative Questions*
  Participants were asked "In the previous questions, what made you decide

what a pilot's level of intent was?". For AI groups, they were additionally asked "Do you think that an AI (Artificial Intelligence) pilot can have intent? Why do you hold this opinion?"

- *Responsibility*
  On a 10-point analogue scale, participants made judgments on the responsibility of the human pilot's trainer and employer, and the AI pilot's software developer and employer. Participants were additionally asked a follow up qualitative question - "Why did you make the previous two responsibility judgments? What informed these judgments?".
- *AI Knowledge.*
  Participants in the AI groups were asked whether or not they have heard of AI before. If they answered yes, they were additionally asked a follow up qualitative question - "What do you think AI (Artificial Intelligence) means in this setting?".

2.4 Participants

To determine the smallest sample size suitable to detect the effects of repeated measures, within-between interaction ANOVA, a power analysis was conducted. The alpha level was set to 0.01, power set to 0.99 and effect size set to 0.5, with the number of levels set to 2. The estimated results indicated that the minimum number of participants was 126, with a final sample of 127 achieved. Participants had to be above the age of 18. The participants were recruited via Prolific. They had to be fluent in English and be a resident of the USA, UK, Ireland, Australia, Canada or New Zealand. These countries were chosen for their common law systems. Only participants that completed the entire survey were considered for the analysis with a maximum permitted time of 87 minutes. Of the 127 participants, there were 31, 34, 34, and 28 participants in the Definition AI, Definition human, No definition AI, No definition human groups, respectively.

2.5 Experiment 1 Results

The mean intent ratings for each group divided by Wind, Legality and Benefit are shown in Figure 2. No obvious between groups can be seen though the effects of the evidence taxonomy are clearer to discern. The results of a repeated measures ANOVA are shown in Table A1 (within subject) and Table A2 in the supporting material section at the end of the paper.

The within subjects ANOVA indicates the main effects are all significant with Wind accounting for most of the variation ($F(1, 123) = 271,\ p < .001,\ \eta_p^2 = 0.688,\ \omega^2 = 0.583$) followed by Legality ($F(1, 123) = 64.50,\ p < .001,\ \eta_p^2 = 0.340,\ \omega^2 = 0.118$) and Benefit ($F(1, 123) = 34.36,\ p < .001,\ \eta_p^2 = 0.218,\ \omega^2 = 0.062$).

There were three significant interactions at a 5% level, Wind*Benefit ($F(1, 123) = 10.46, p = 0.002, \eta_p^2 = 0.078, \omega^2 = 0.011$), Wind*Legal*Benefit ($F(1, 123) = 8.91, p = 0.003, \eta_p^2 = 0.068, \omega^2 = 0.009$) and Wind*Benefit*AI

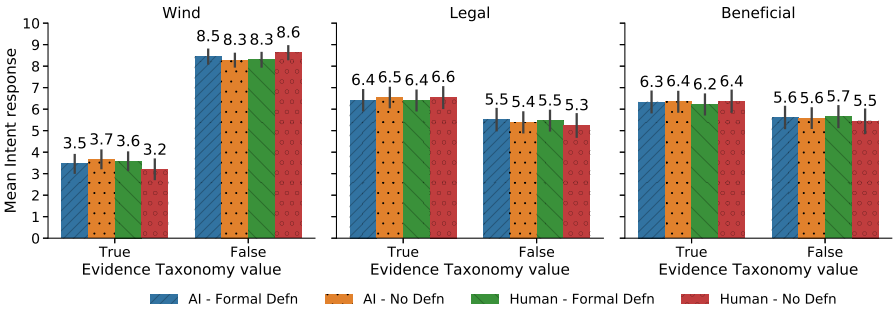
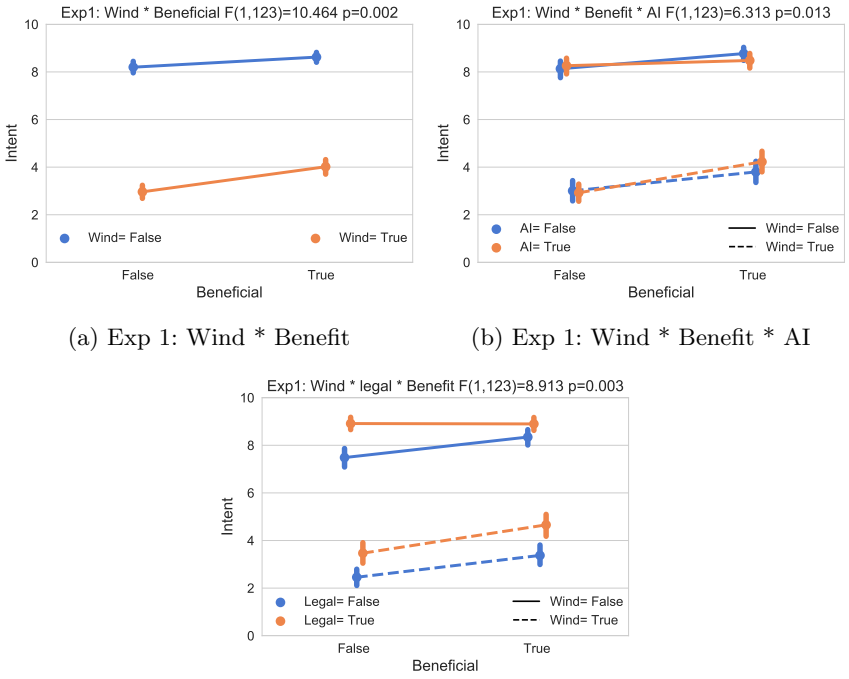


Fig. 2: Experiment 1: Mean Intent responses for the four experimental groups across the evidence taxonomy.



(a) Exp 1: Wind * Benefit

(b) Exp 1: Wind * Benefit * AI

(c) Exp 1: Wind * Legal * Benefit

Fig. 3: Experiment 1 descriptive charts showing significant interactions at a 5% level according to ANOVA. 5% Error bars in charts calculated by resampling.

$(F(1,123) = 6.31, p = 0.013, \eta_p^2 = 0.049, \omega^2 = 0.006)$. These effects can be seen in Figure 3. The presence of wind lowers intent by a consistent amount and Beneficial moves are at the very least not any less intentful. In cases where

there is no wind, the positive effect of viewing a beneficial move is muted (most pronounced in the legal case in Figure 3c). This could be an artefact of the scoring system, since an already high intent score of around 9 in the legal, non-beneficial case is hard to be increased arithmetically. Equally it could be that people's intent inferences display a satiation characteristic – whether a move is beneficial or not is irrelevant given that it was observed in the absence of wind and it was legal.

The Levene test for equality of variance displayed in Table A3 shows that the assumption of equal variance between groups is questionable for two of the eight groups. The t-tests in the ensuing tests were adjusted accordingly to take account of this.

Table 1 shows the contrasts for the main within subject effects. The presence of wind lowers intent by around 5 points. Illegality lowers intent by around 1, and Unbeneficial moves lower intent by 0.75. The between group contrasts are shown in Table 2. They are not significant with their average effect centred around zero.

		95% CI for Mean Difference						
Variable	Comparison	Estimate	Lower	Upper	SE	df	t	p
Wind	True - False	-4.946	-5.541	-4.351	0.300	123	-16.460	< .001
Legal	True - False	1.076	0.809	1.343	0.135	123	7.968	< .001
Beneficial	True - False	0.745	0.494	0.997	0.127	123	5.862	< .001

Table 1: Experiment 1 within Participant Repeated Contrasts, intent scores are averaged across the other levels and groups not being contrasted.

		95% CI for Mean Difference						
Group	Comparison	Estimate	Lower	Upper	SE	df	t	p
AI	True - False	0.047	-0.396	0.491	0.224	123	0.211	0.834
Definition	Your - The Formal	-0.014	-0.458	0.429	0.224	123	-0.063	0.950

Table 2: Experiment 1 Between group contrasts. Results are averaged across levels within groups

2.6 Discussion Experiment 1

The results from Experiment 1 indicate that people did not judge the intentional state of a human pilot any differently from that of an AI. We will further test this result using a within-participant design in Experiment 2. It could be that people were ignoring or not registering the non-human status of the pilot.

Wherever appropriate the survey would refer to them as 'the human pilot' or 'the AI pilot' to minimise this possibility.

The lack of difference between the groups given a definition of intent and those who were told to use their own intuition means, indicates at the very least, that the definition in this case did no harm. However, because we used the word intent in its definition, it could be that participants were using their own concept either in conjunction with the definition or instead of it. V.L. Smith (1993) explores the phenomena of jurors using their own (often faulty) knowledge of law when making judgments as to whether certain crimes had been committed or not. In Experiment 2 we test this hypothesis by using the same definition but without using the word intent at any point.

The lower intent scores elicited for movements caused by the wind agree with the association between causality and intent which has previously been Lagnado and Channon (2008); Mele and Cushman (2007). If someone did not cause an outcome, then people judge them less likely to have intended it. The lower intent score for illegal moves suggests that people do not expect behaviour to be intentionally deviant, and when it is, alternative explanations are perhaps called upon (like some sort of error causing the behaviour).

The higher intent scores for beneficial moves makes intuitive sense. Desire and aims are typically mentioned in folk definitions of intent and mentioned in the definition given to participants, so movements which appear counter-productive to the pilot's goal of reaching their target, should receive a lower intent attribution.

The significant interactions according to the ANOVA are shown in Figure 3. The difference between Beneficial and non-beneficial moves is not present in the absence of wind. This does make intuitive sense because without wind, the participants are likely to believe that the drone's movements are solely caused by the pilot. This might suggest that participants are prepared to ascribe intent without requiring or understanding the aims of the pilot in the cases where the action was clearly caused by the pilot. It could also be an artefact of the scoring method; questions with average responses close to the extreme of 10 (or 0) cannot separate factors as well as those where responses are more centred. Put another way, if a respondent decides two moves were intended, it might be difficult or unnatural to say which was more intended.

3 Experiment 2

To test whether people were using their own definition of intent instead of using the definition provided, we altered the design in Experiment 2 so that participants given the definition of intent, were not told that it was intent that they were judging. This was achieved by asking participants to imagine they had been called up for jury service in an imaginary country (*Fhljmakon*) where they were asked to assess the presence of absence of a legal concept called *Cthofrjk* idiosyncratic to the country. The definition of *Cthofrjk* was the definition of intent provided to participants in Experiment 1 with any mention

of intent carefully removed. We also asked participants to judge both Human and AI pilots to look for within-participant intent judgment differences. At the time of publishing, *Cthofrjk* and *Fhljmakon* produced no search results in the Google search engine.

Experiment 2: Design

The study used a mixed design with two experimental groups and four within-subject factors. Depending on their experimental group, participants were either asked about the intent of the pilots or asked to assess the pilot's level of Cthofrjk, a foreign legal concept for which they were given a definition identical to that of intent in the first experiment. Thus, the two experimental groups were: 1) Intent, 2) Cthofrjk. The motivation of this design change was to check whether participants had just been using their own definition of intent, even when supplied with the official one in Survey 1. In this study, participants were told that they were on a secondment to an imaginary country (Fhljmakon) and had been called up for jury service. The Cthofrjk groups was never asked about intent directly in the survey nor did the word feature in any form. The rest of the experimental context was the same as for experiment 1 - participants were told that they would be examining the behaviour of a series of drone pilots.

There were four within-subject factors, three being the same as in experiment 1 - legality, benefit and wind - with the addition of the pilot being either human or AI as an additional within-subject factor. The experiment consisted of the same three phases as in experiment 1: training, testing, and survey.

Experiment 2: Procedure

Apart from the differences stated in this section, the procedure was the same as the procedure used in Experiment 1. Participants were randomly divided into one of the two experimental groups before the study's training phase. Participants were informed that they are called up to do jury service for a number of court cases, and were either given the definition of Cthofrjk or not, depending on their experimental group. Training proceeded as in Experiment 1. In the test phase of the study, the evidence presented to the participants differed in that only one example of the 8 categories was shown. Participants still saw 16 evidence sets (whose order was randomised) because each set was shown for an AI and human pilot (not necessarily consecutively). Unlike Experiment 1, participants were not asked to make judgments about the pilot's willingness to take risks. In addition to the remaining four judgments from experiment 1, participant additionally gave judgments on the pilot's efficiency and foresight.

The survey phase was the same as in Experiment 1, with the addition that the participants in the Cthofrjk group were also asked a free text question as to what they thought the concept meant. Further, participants were asked to make two separate judgments on how causal the AI and human pilot were for the drone to reach its final destination in the way that it did.

Experiment 2: Measures and Materials

Apart from the removal of the experimental induction from Experiment 1, and the judgments of the pilot's willingness to take risks, all of the measures and materials were the same as in Experiment 1, with the addition of the one's described in this section.

- *Experimental group induction material*
  Participants were told they have won a competition to go and live in the capital city of the country of Fhljmakon for 6 months. The country had two official languages: English and Fhljmakonian. They were told that after they arrived they were called up to do jury service. In the Cthofrjk group, they were told that the country's legal system still uses some Fhljmakonian concepts. For jury service they would need to apply a definition of a key concept in Fhljmakonian law to the cases. They were given the definition in English, which was identical to the definition of intent used in Experiment 1.
- *Inferences*
  On a 10-point analogue scale, participants made judgments on a pilot's efficiency - going through the map as quickly as possible - and foresight - being able to predict what was going to happen. Qualitative questions. Participants in the Cthofrjk group were asked which English words best describe the Fhljmakonian legal concept of Cthofrjk Causal attribution. On a 10-point analogue scale, participants made separate judgments on how causal the human and AI pilot were for the drone reaching its final destination in the way that it did.
- *Qualitative questions*
  Participants in the Cthofrjk group were asked which English words best describe the Fhljmakonian legal concept of Cthofrjk
- *Causal attribution*
  On a 10-point analogue scale, participants made separate judgments on how causal the human and AI pilot were for the drone reaching its final destination in the way that it did.

Experiment 2 Participants

130 participants were recruited for the second experiment. With two groups, four levels per participant, a significance level of 1% and a power of 0.99, this implied an effect size of 0.48 which was sufficient for the smallest significant effects found in Experiment 1. Participants were recruited with the same language and residency requirements of Experiment 1. Of the 130 participants, there were 67 in the Intent group and 63 in the Cthofrjk group.

3.1 Results Experiment 2

The mean intent scores for both groups averaged across the three taxonomy levels and the AI/Human condition are shown in Figure 4.

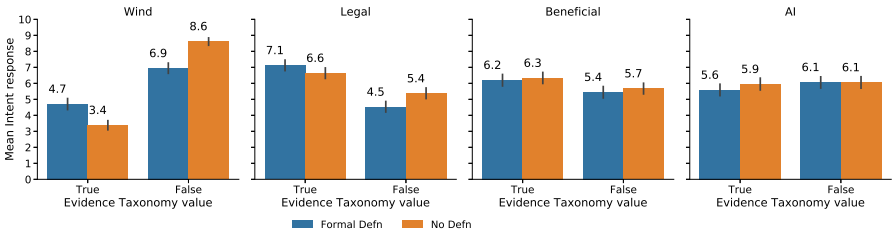


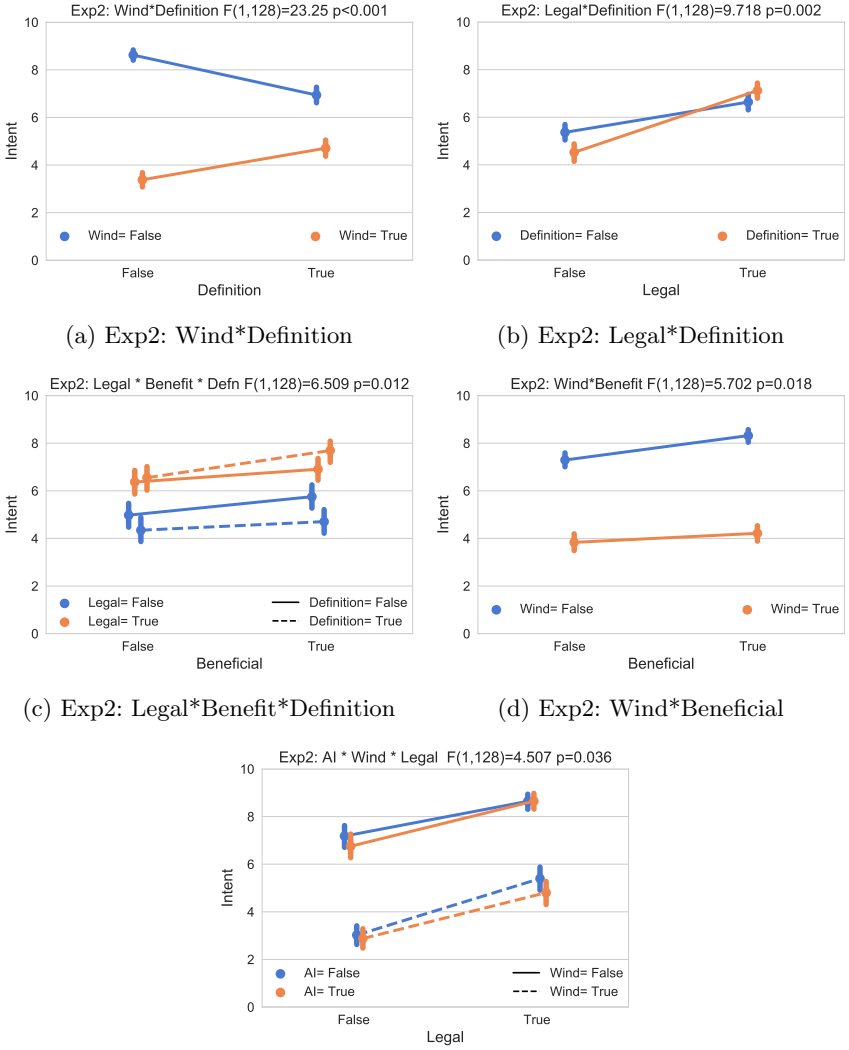
Fig. 4: Experiment 2 Mean intent response by group

The results of a repeated measures ANOVA are shown in Table A4 (within subject) and Table A5 (between subject) in the supporting material section at the end of the paper. Once again, the three dimensions of the evidence taxonomy were significant with Wind ($F(1,128) = 144, p < 0.001, \eta_p^2 = 0.529, \omega^2 = 0.430$), Legality ($F(1,128) = 82.9, p < 0.001, \eta_p^2 = 0.393, \omega^2 = 0.241$) and Benefit ($F(1,128) = 28, p < 0.001, \eta_p^2 = 0.180, \omega^2 = 0.056$) providing significant contribution to variation. The AI condition was also significant ($F(1,128) = 7.441, p = 0.007, \eta_p^2 = 0.055, \omega^2 = 0.010$).

As before there were no significant between subject effect found between the definition and no definition group ($F(1,128) = 0.695$, $p = 0.406$, $\eta_p^2 = 0.005, \omega^2 = 0.000$), as shown in Table 13, however there were five significant within subject interactions: Definition and Wind ($F(1,128) = 23.251$, $p < 0.001$, $\eta_p^2 = 0.154, \omega^2 = 0.105$), Legality and Definition ($F(1,128) = 9.718$, $p = 0.002$, $\eta_p^2 = 0.071$, $\omega^2 = 0.033$), Legal*Benefit*Definition ($F(1,128) = 6.509$, $p = 0.012$, $\eta_p^2 = 0.048$, $\omega^2 = 0.008$), Wind*Benefit ($F(1,128) = 5.702$, $p = 0.018$, $\eta_p^2 = 0.043$, $\omega^2 = 0.010$) and AI*Wind*Legal ($F(1,128) = 4.507$, $p = 0.036$, $\eta_p^2 = 0.034$, $\omega^2 = 0.005$). These can be seen in Figure 5. Subfigures a-c indicates that the providing the definition did alter intent inferences in Experiment 2; Wind moderates inferences towards centrepoint 5, the effects of legality and benefit are larger in the definition group. Levene's test for equality of variance was rejected multiple within subject levels as shown in Table A6.

				95% CI for Mean Difference					
Group	Variable	Comparison	Estimate	Lower	Upper	SE	df	t	p
Own	AI	True - False	-0.108	-0.405	0.189	0.149	66	-0.728	0.469
Formal	AI	True - False	-0.482	-0.797	-0.167	0.157	62	-3.062	0.003
Own	Wind	True - False	-5.239	-6.079	-4.398	0.421	66	-12.443	< .001
Formal	Wind	True - False	-2.232	-3.155	-1.310	0.462	62	-4.836	< .001
Own	Legal	True - False	1.272	0.870	1.675	0.202	66	6.313	< .001
Formal	Legal	True - False	2.597	1.833	3.362	0.382	62	6.793	< .001
Own	Beneficial	True - False	0.657	0.317	0.997	0.170	66	3.859	< .001
Formal	Beneficial	True - False	0.752	0.340	1.164	0.206	62	3.651	< .001

Table 3: Experiment 2 Contrasts. Intent scores are averaged across the other levels and groups not being contrasted. The t-test variant used does not assume equal variances.



(a) Exp2: Wind*Definition

(b) Exp2: Legal*Definition

(c) Exp2: Legal*Benefit*Definition

(d) Exp2: Wind*Beneficial

(e) Exp2: AI*Wind*Legal

Fig. 5: Experiment 2 Significant interactions at a 5% level. 5% Error bars in charts calculated by resampling. The subcaptions describe the variable groups being plotted.

Table 3 shows the main experiment contrasts, p-values of the t-tests were adjusted to not assume equal variance at the expense of some statistical power. Whilst the ANOVA did not find a significant difference between groups, enough significant interactions involved the Definition group to justify splitting results between the two groups. In the No Definition group, the effects of the evidence taxonomy in Experiment 1 were repeated, with the

same ordering and a similar effect size. In the Definition group, whilst the sign of the main effects is the same, the effect of Wind is the second largest effect to legality. The difference between Beneficial and non-beneficial moves remains the same. Within the Definition group, a significant difference was seen between the AI and Human judgments of intent with AI receiving on average -0.482 less intent points than Human pilots. When results are averaged across the two groups, the AI-Human contrast is smaller but still significant ($M = -0.289, SE = 0.109, t(129) = 2.656, p = 0.009$), the combined contrast table is shown in Table in the supporting materials.

Wind	Definition	Could predict	Know weather	Intent
False	True	7.3	8.1	6.9
	False	7.2	8.0	8.6
True	True	5.6	7.0	4.7
	False	5.5	7.1	3.4

(a) Exp' 2 participants' judgments about whether the drone pilot could predict where the drone would move and whether they knew about the weather conditions. Participants were told that the pilot knew about the presence or absence of wind before departing.

Legal	Definition	Break law	Intent
False	The Formal	6.3	4.5
	your	5.4	5.4
True	The Formal	1.6	7.1
	your	1.4	6.6

(b) Exp' 2 participants' judgments about the drone pilots' willingness to break the law

Beneficial	Definition	Quick as	Intent
False	The Formal	3.3	5.4
	your	3.2	5.7
True	The Formal	7.6	6.2
	your	7.4	6.3

(c) Exp' 2 participants' judgments about whether the drone pilot had plotted to fly as quickly as possible through the city

Table 4: Experiment 2 manipulation checks: After each evidence set, participants were asked additional questions on a 0-10 scale about their opinions of the drone pilot. The four measures in the tables above indicate that the evidence taxonomy was successfully understood by participants.

We carried out manipulation checks per item of evidence to verify participants were correctly interpreting the stimuli across its 2X2X2 taxonomy. The results of this are shown in Table 4 and confirm that participants were correctly interpreting the stimuli. The presence or absence of wind was measured by participants' inferences about whether the pilot could predict where

they were going. Similarly participants' belief about the legality of the drone's movement was tested by asking them how willing they thought the pilot was to break the law. Beliefs about Beneficial or un-beneficial movements were tested by asking participants whether the pilot had travelled as quickly as possible to their destination.

At the end of the survey, participants were asked on a $0-10$ scale to what degree the two pilots had caused the drone to reach its destination in the way that it did. The mean responses are shown in Table 5.

Pilot	Definition	Mean	SD	N
AI	The Formal	7.492	2.402	63
	your	7.731	1.831	67
Human	The Formal	8.397	1.487	63
	your	8.060	1.466	67

Table 5: Experiment 2: Causal ratings of pilots

A repeated measures ANOVA found the pilot effect to be significant ($F(1,128) = 12.736\,, p < 0.001$, $\eta_p^2 = 0.090$, $\omega^2 = 0.026$), the between subject effect of the Definition was not significant ($F(1,128) = 0.033\,, p = 0.857$, $\eta_p^2 = 0.0002$, $\omega^2 = 0$). There interaction term was also not significant ($F(1,128) = 5.394\,, p = 0.098$, $\eta_p^2 = 0.021$, $\omega^2 = 0.004$). The results of a paired samples T-test are shown in Table 6.

							Location	SE	95% CI for Loc Parameter		
Pilot 1		Pilot 2	Test	Statistic	df	p	Parameter	Difference	Lower	Upper	Effect Size
Human	-	AI	Student	3.495	129	< .001	0.608	0.174	0.264	0.952	0.307
			Wilcoxon	1766.500		< .001	1.000		0.500	1.500	0.463

Table 6: Experiment 2 Causality rating Paired Samples T-Test. *Note.* For the Student t-test, effect size is given by Cohen's d; for the Wilcoxon test, it is given by the matched rank. For the Student t-test, location parameter is given by mean difference; for the Wilcoxon test, it is given by the Hodges-Lehmann estimate.

3.2 Discussion Experiment 2

Unlike in the first experiment, this experiment finds a significant (though small) difference in Intent inferences between Human and AI cases with AI pilots being 0.3 points less intentional than human pilots. Unlike Experiment 1, the within subject design meant that participants were comparing AI pilots against human pilots which perhaps caused this result to appear. Though this effect is small, it is consistent with the findings of Hidalgo et al. (2021).

Humans were also judged to be more causal; there was a mean difference of 0.608 which was statistically significant. This causality rating was given at the end of the experiment and so was not considering any particular behaviour. Given the theoretical link between intent and causality, this is consistent with human pilots being judged on average to have more intent. Experiment 3 will study the relationship between causality and intent in greater detail.

As with Experiment 1, the ANOVA did not indicate significant differences between the Definition and No-definition groups, however interactions between definition and the evidence taxonomy were apparent. Since the experiment design avoided using the word 'intent' for the definition group, this suggests participants in Experiment 1 were using their own definition of intent when asked to use a provided definition of it. In Experiment 2, the definition lessened the effect of wind (-2.2 for the definition group versus -5.2) but increased the effect of legality on intent inferences (2.6 for the definition group versus 1.3). Since wind is a proxy for whether the pilot definitely caused a movement, this might mean that the provided definition lowers the importance of causality relative to its role in the folk definition of intent. The increased, positive effect of legality (or negative effect of illegality) is a puzzle, thought supports the same result found in Experiment 1. It could be that participants were more uncertain about ascribing Cthofrjk than intent and were using legality as a proxy for it, thus higher scores were given when a movement was legal.

The experimental mechanism of eliciting intent without necessarily telling participants that was what they were doing is presaged by Knobe (2004, 2006) who recreates the effects of his earlier experiments by replacing "intend" with "in order to". This was in response to Adams and Steadman (2004) who suggested that the use of the words "intend" and "intent" might elicit responses influenced by conversational factors associated with the words rather than the underlying folk concept of intent.

Whilst the results do show that the definition elicited different responses to people's natural definition of intent, the direction of the evidence taxonomy's main effects was the same and agreed with Experiment 1. This indicates that whilst the definition is not perfect, it has a respectable overlap with the folk definition.

4 Experiment 3

In Experiment 3 we wanted to gain a better understanding of how the provided definition of intent is different from an individual's natural definition. Since the previous experiments split participants between definition and no-definition groups we felt a within participant design would shed more light on the effects of providing a definition (if any) since it would allow paired t-tests. The experimental mechanism of asking for participants judgment of Cthofrjk was reused.

Struck by the repeated result in Experiments 1 and 2 that illegal moves were deemed to be less intentional, we also wanted to check a hypothesis that

participants thought these were caused in error. Wary that mentioning errors in the main body of the experiment might be suggestive to participants, we asked after the main body of questions, in the event of a drone flying into a no-fly zone, how likely it was caused by a pilot error or a mechanical or hardware error.

The manipulation checks related to the evidence taxonomy in Experiment 2 indicated that participants were responding to the differences in the evidence in the way that we expected. We decided to swap them for inference questions more closely related to factors which have been previously found to relate to intent, namely foresight, freedom to make decisions, causation and desire. Experiment 2 found a difference in people's judgement of causation between human and AI, so we thought it would be interesting to study this on a per scenario basis.

Experiment 3: Design

The study used a mixed design with two experimental groups and four within-subject factors. Participants were divided into two groups, one judging AI pilots and the other group judging human pilots. The aim of the third experiment was to investigate the within-subject effect of giving participants a definition of intent (or Cthofrjk) and measuring the differences elicited compared with their judgment of pilot intent according to their own understanding of the term.

There were four within-subject factors, three the same as in Experiment 1 - legality, benefit and wind - and the final one being whether the participant was asked to use their own definition or the provided one. This final within-subject factor was counterbalanced - participants were randomly divided as to whether they were asked to give their judgments of Cthofrjk in the first eight evidence sets or in the second eight. This aspect of the design allowed for controlling the potential confounding effects of participants being asked to make judgments of intent first which could make participants more likely to think that the Cthofrjk definition that they subsequently saw was related to intent. The experiment consisted of the same three phases as in Experiment 1: training, testing, and survey.

Experiment 3: Procedure

Apart from the differences stated in this section, the procedure was the same as the procedure used in Experiment 1. Training proceeded as in Experiment 1. Participants were introduced to the pilot they would be evaluating - human or AI - and, when appropriate, the definition of Cthofrjk they would be using for subsequent intent inferences. After responding to 8 randomly selected experimental items, they were told to either switch from using their own definition intent to using a formal definition of Cthofrjk, or vice versa. To control against any ordering effect, participants in each group were split further between those that were given the definition for the first set of 8 questions, or the second

set. In addition to intent, participants made judgments on the pilot's driving skills, desire, willingness to break the law, foresight, causality and autonomy to make decisions freely. The survey phase was mostly the same as for the Cthofrjk group in Experiment two. The two causal attributions questions were removed from this section. Participants were additionally asked how likely they thought that the drone entering no-fly zones was either due to a pilot mistake or mechanical fault in the drone.

Experiment 3: Measures and Materials

In this section there is a description of new measures and materials that were unique to experiment 3.

- *Inferences*
  On a 10-point analogue scale, participants made judgments on a pilot's skills, willingness to break the law, foresight, whether the pilot made their decision freely, whether the pilot caused the drone to reach its final destination in the way that it did, and whether the drone flew how the pilot desired it to.
- *Mistakes*
  On a 10-point analogue scale, participants were asked "How likely was it that a drone entering a no-fly zone was caused by pilot mistake?" and "How likely was it that a drone entering a no-fly zone was caused by mechanical fault in the drone?"

Experiment 3: Participants

We performed a power analysis for the within-between interaction ANOVA. The 74 participants, divided into two groups and measured over two levels, were sufficient to detect an effect size of 0.6 with significance level of 1% and power of 99%. This was sufficient for the smallest significant contrast shown in Experiment 2. Participants had to be above the age of 18 and were recruited with the same language and residency criteria as the previous experiments. The participants were recruited via Prolific. There were 38 in the human pilot group and 36 in the AI pilot group.

4.1 Experiment 3: Results

Mean intent responses in Experiment 3 are shown in Figure 5.

The within subjects repeated measures ANOVA shown in Table A8 revealed significant main effects of Wind ($F(1, 70) = 98.5$, $p < 0.001$, $\eta_p^2 = 0.584$, $\omega^2 = 0.481$), Legality ($F(1, 70) = 28.4$, $p < 0.001$, $\eta_p^2 = 0.288$, $\omega^2 = 0.153$) and Benefit ($F(1, 70) = 11.6$, $p = 0.001$, $\eta_p^2 = 0.142$, $\omega^2 = 0.045$). The size of the main effects is shown in Table 7.

Table A9 shows the between group effects of the repeated measures ANOVA. The Pilot grouping is significant ($F(1, 70) = 6.075$, $p = 0.016$, $\eta_p^2 = 0.080$, $\omega^2 = 0.035$). The average difference in intent between Human and AI

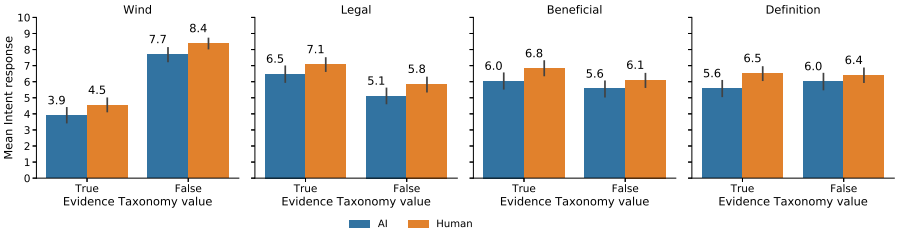


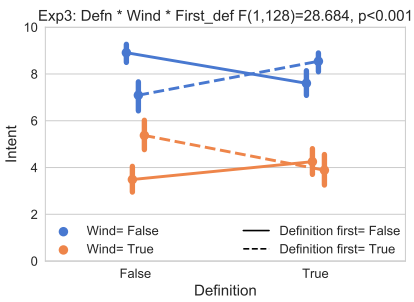
Fig. 6: Experiment 3 mean intent responses by group.

pilots is 0.65 as shown in Table 7. This is a reverse from Experiment 2, where at least within the formal definition group, AI was 0.48 points more intentional.

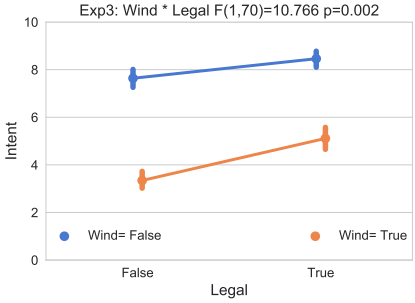
Compared to previous experiments a greater number of interactions were found (6) by the ANOVA. These are shown in Figures 7 and 8. This might be a function of the smaller sample size (74). The most significant interaction in terms of eta and omega, and the only one involving definition order was between Definition, Wind and First Definition type ($F(1, 70) = 28.7$, $p < 0.001$, $\eta_p^2 = 0.291$, $\omega^2 = 0.147$), This effect is seen in Figure 7a. It seems that the effect of wind is lessened the second time that participants see the evidence set – they gave an intent score closer to 5 - regardless of which definition type they used first. Whilst not symmetric, this seems more likely to be an artefact of the experiment rather than a function of intent definitions. The other significant interactions were Wind and Legality ($F(1, 70) = 10.766$, $p = 0.002$, $\eta_p^2 = 0.133$, $\omega^2 = 0.031$), Definition and Legality ($F(1, 70) = 10.898$, $p = 0.002$, $\eta_p^2 = 0.135$, $\omega^2 = 0.044$), Wind, Legality and Benefit ($F(1, 70) = 6.949$, , $p = 0.01$, $\eta_p^2 = 0.090$, $\omega^2 = 0.020$), and Wind, Legality and Pilot ($F(1, 70) = 5.339$, , $p = 0.024$, $\eta_p^2 = 0.071$, $\omega^2 = 0.014$). Finally a five way interaction was shown to be significant between Definition, Wind, Legality, Benefit and Pilot ($F(1, 70) = 4.312$, $p = 0.042$, $\eta_p^2 = 0.058$, $\omega^2 = 0.007$). This is displayed in Figure 8, note the larger size of the error bars in this figure due to the smaller group averages.

			95% CI for Mean Difference					
Variable	Comparison	Estimate	Lower	Upper	SE	df	t	p
Definition	formal - your	-0.152	-0.513	0.209	0.181	70	-0.841	0.405
Wind	True - False	-3.787	-4.548	-3.026	0.382	70	-9.923	< .001
Legal	True - False	1.319	0.825	1.813	0.248	70	5.327	< .001
Benefit	True - False	0.607	0.252	0.962	0.178	70	3.410	0.001
Pilot	AI - Human	-0.651	-1.179	0.124	0.264	70	-2.465	0.016
First_D	Your - Formal	0.165	-0.362	0.692	0.264	70	0.624	0.535

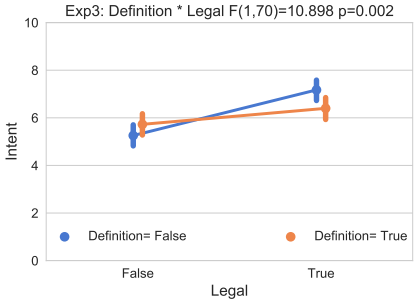
Table 7: Experiment 3: Intent Contrasts, Intent scores are averaged across the other levels and groups not being contrasted. The t-test variant used does not assume equal variances for the within group contrasts.



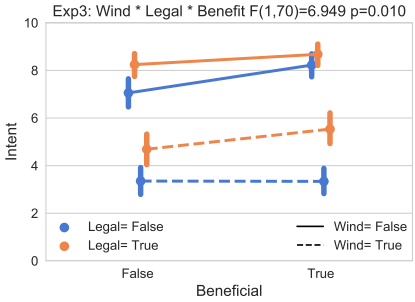
(a) Exp3: Defn*Wind*FirstDef



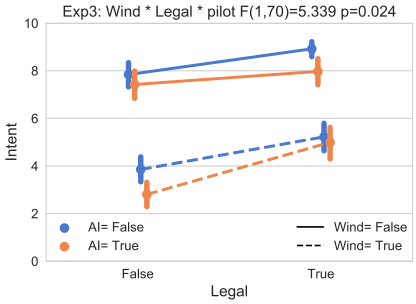
(b) Exp3: Wind*Legal



(c) Exp3: Defn*Legal



(d) Exp3: Wind*Legal*Benefit



(e) Exp3:Wind*Legal*Pilot

Fig. 7: Experiment 3 Significant interactions at a 5% level. 5% Error bars in charts calculated by resampling.

After the main body of questions participants were asked to assess in the event of a drone flying into a no-fly zone, how likely that was caused by a pilot error or a mechanical or hardware error. The summary results are shown in Table 8. A simple repeated measures ANOVA, confirmed that the Difference between Error types was significant ($F(1, 72) = 7.471$, $p = 0.008$, $\eta_p^2 = 0.094$, $\omega^2 = 0.046$) but the effect of pilot type was not (F(1,72)=0.030,

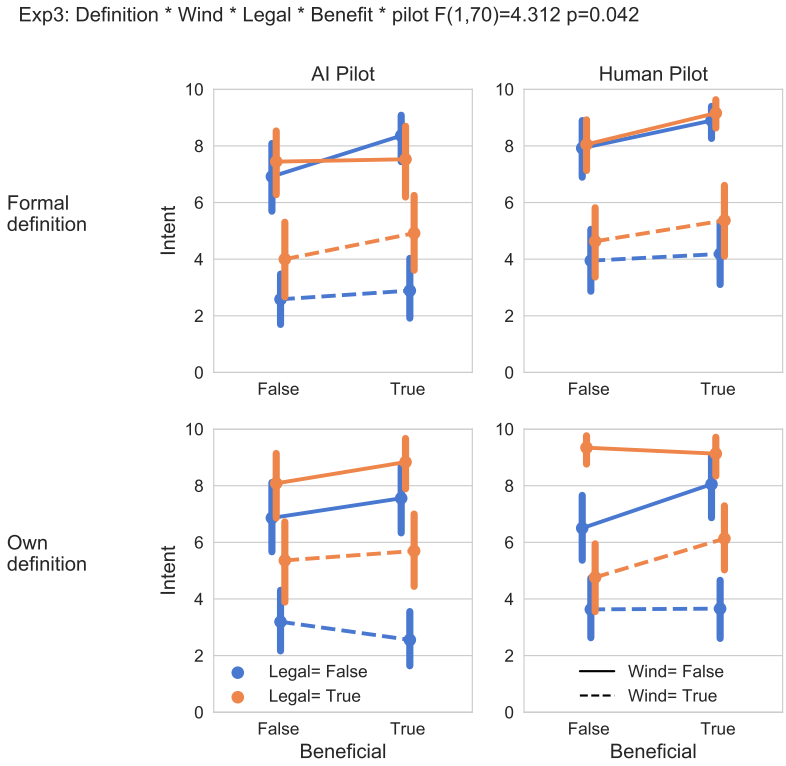


Fig. 8: Experiment 3 Significant interactions at a 5% level. 5% Error bars in charts calculated by resampling.

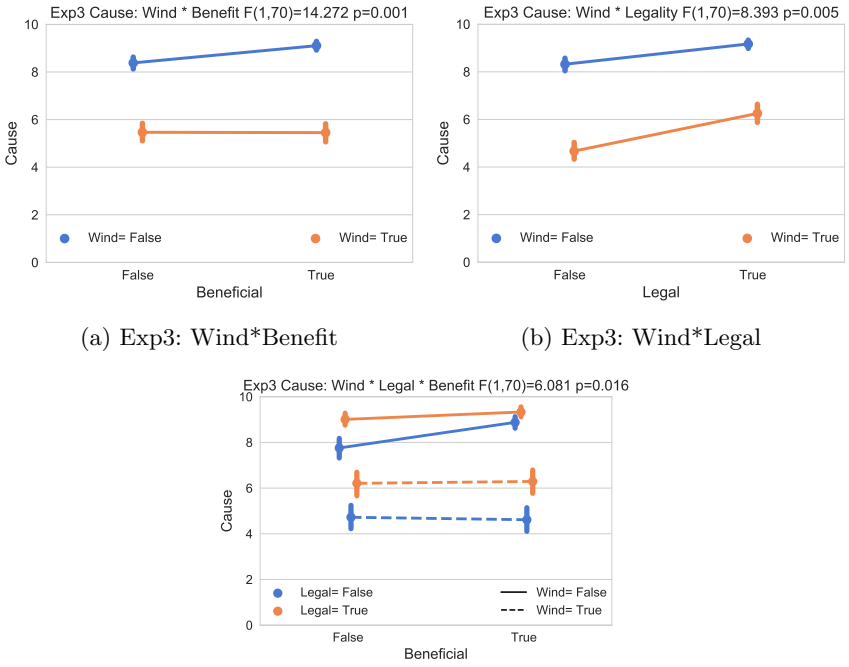
Pilot Group	pilot mistake		drone error	
	mean	std	mean	std
AI	4.94	2.41	4.28	2.01
Human	5.37	1.94	3.97	2.39

Table 8: Participants were asked to assess the chance of errors in the pilot and the drone causing movement into no-fly zones.

p=0.862). A t-test ($t(72) = -2.733$, $p = 0.008$) indicated a significant average effect of 1.0 and confidence interval [0.3-1.7] – Participants judged that the pilots were more likely to have caused an error than the drone when moving into a no-fly zone regardless of whether the pilot was human or AI.

Participants were also asked how much they agreed with a series of statements after giving their intent judgment. One of the statements was "The pilot caused the drone to reach its final destination in the way that it did". We performed a separate repeated measures ANOVA on this variable shown in Tables

A13 and A14. Three within subject main effects were significant according to the ANOVA: Wind ($F(1,72) = 111, p < 0.001, \eta_p^2 = 0.614, \omega^2 = 0.426$), Legality ($F(1,72) = 71.6, p < 0.001, \eta_p^2 = 0.506, \omega^2 = 0.142$) and Benefit ($F(1,72) = 9.306, p = 0.003, \eta_p^2 = 0.117, \omega^2 = 0.013$). Three interactions were significant Wind and Benefit ($F(1,72) = 14.272, p < 0.001, \eta_p^2 = 0.169, \omega^2 = 0.016$) and Wind and Legality ($F(1,72) = 8.393, p = 0.005, \eta_p^2 = 0.107, \omega^2 = 0.013$) and Wind, Legality and Benefit ($F(1,72) = 6.081, p = 0.016, \eta_p^2 = 0.080, \omega^2 = 0.006$). These are shown in Figure 10.



(a) Exp3: Wind*Benefit

(b) Exp3: Wind*Legal

(c) Exp3: Wind*Benefit*Definition

Fig. 9: Experiment 3: Causality significant interactions

No between subject effects were significant - neither the type of pilot or whether participants were given the definition first or second (Table A14). Contrasts are shown in Table 9. A visual comparison between intent and causal responses is shown in Figure 10.

4.2 Experiment 3 Discussion

The signs, ordering and approximate magnitude of main effects of the taxonomy were repeated (Table 7) from Experiments 1 and 2. However the difference within participants when using their own definition of intent and when using

			95% CI for Mean Difference					
Variable	Comparison	Estimate	Lower	Upper	SE	df	t	p
Wind	True - False	-3.284	-3.899	-2.660	0.310	73	-10.593	< .001
Legal	True - False	1.216	0.938	1.516	0.145	73	8.413	< .001
Benefit	True - False	0.355	0.121	0.577	0.112	73	3.155	0.002
Definition	own - Formal	-0.166	-0.424	0.111	0.134	73	-1.231	0.222
Pilot	Human - AI	0.236	-0.412	0.885	0.325	70	0.727	0.470

Table 9: Experiment 3: Causality Contrasts

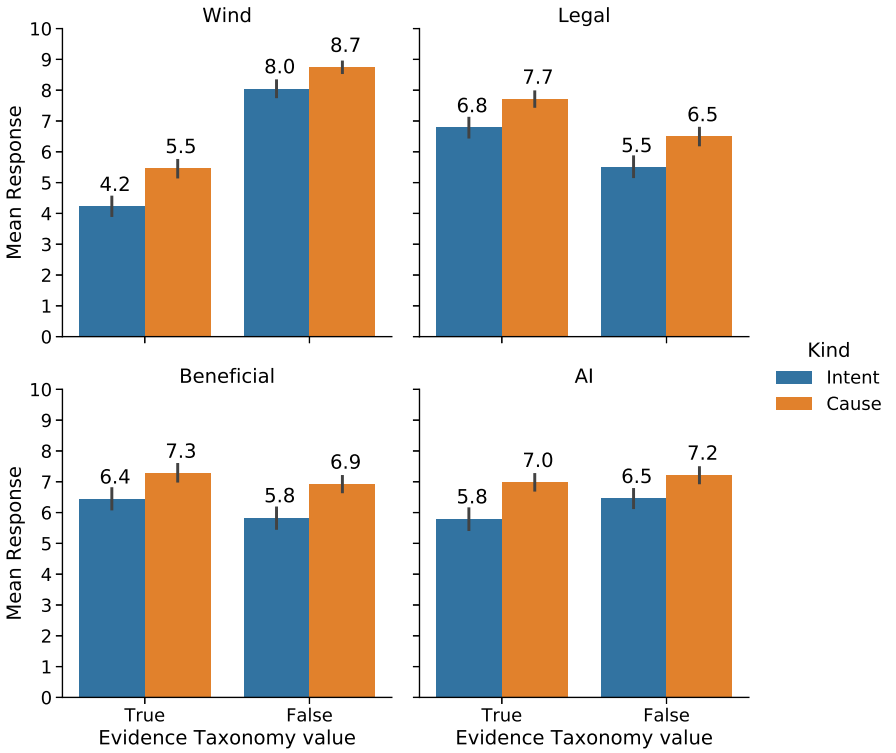


Fig. 10: Experiment 3 comparison between mean elicited values of intent and causality: Main effects are mirrored between two variables. All differences are significant except between AI (pilot) groups, where there is no significant difference in responses.

the formal one is not statistically significant. The interactions seen in Experiment 2 between the definition and wind/legality were not repeated. This is shown in Figure 7.

The within participant design for this experiment demonstrated that providing a definition of intent generally did a decent job of recreating participants' intent judgments since the main contrast between the questions answered with and without definition was not statistically significant.

As with Experiment 2 a small but statistically significant difference was found between participants who were judging AI or human pilots, with AI pilots being judged as less intentional. This was found to be the case in Hidalgo et al. (2021). Unlike Experiment 2, participants did not judge both human and AI pilots.

In Experiment 2 we found a difference in causal judgements between AI and Human, however this was not repeated in Experiment 3 with no significant difference in causal ratings according to the ANOVA in Table A9. In this setting, participants were asked about causality per scenario whilst in Experiment 2, it was a general question. Additionally in Experiment 2, participants considered both human and AI pilots, whilst in Experiment 3 participants only considered one pilot type. However participants did differentiate between AI and human in their intent judgments so the no difference result cannot be easily explained by suggesting participants were judging AI pilots as humans.

By asking about the causal effect of the pilot on the drone's movement, we were able to see that the effect of the evidence taxonomy was mirrored for Wind, Legality and Benefit. We also note that the causality results were cleaner in the sense that fewer interactions were detected by the ANOVA and none that included Definition or Pilot variables. An advantage of also eliciting causal judgments is the larger research body surrounding the concept. An interesting question to consider is how the concepts of intent and causality judgments influence other. As we saw in the introduction, a classical theory of intent requires causality as an input. Leaving the narrow, physical view of causality aside momentarily, judgments of intent are likely to influence judgments of causality. This is particularly relevant in legal contexts, where courts have to distinguish between different causes of harm to determine whether a relevant one exists. Lagnado and Channon (2008) find intent does increase judgments of causality.

We hypothesised that the lower intent score for illegal moves was due to an error hypothesis being formed in the respondent's head. If this were true, we would expect to see the causality rating of illegal behaviour to be lower than legal behaviour. This was indeed the case, with an average effect of 1.2 ($t(72) = 8.4, p < 0.001$) from Table 9 which is similar to the effect on intent of 1.3 ($t(72) = 5.3, p < 0.001$) from Table7.

The question of whether the pilot caused the drone's flight path should not depend on whether that path was beneficial to the pilot (as short as possible) in cases where causality is known with confidence. The effect of a move being beneficial is significant for causality ($t(72) = 3.2p = 0.002$), but small at 0.355. This is similar to the effect on intent which was 0.6 ($t(72) = 3.4, p = 0.001$).

5 Experiments 1,2,3 Combined Results

In addition to the questions which we have so far discussed, participants were also asked at the end of each experiment about how responsible they felt The Pilot's Programmer/Instructor and Employer were for any harms caused be the pilot's actions. In Experiment 1 participants were only asked about the pair corresponding to the pilot type that they had been answering questions. In the other two experiments participants were asked about both types. We excluded the data from Experiment 3 where participants were additionally asked about the pilot type they had not previously been considering. We performed a standard ANOVA on the data (Table B15 in supporting materials) and found that participants and found one significant effect - Responsibility scores were significantly higher for the AI pilot's employer and programmer ($F(1, 910) = 14.897$, $p < 0.001$, $\eta_p^2 = 0.0016$, $\omega^2 = 0.015$). Responsibility scores did not significantly differ between the instructor/employer role or differ between groups that had been given the formal definition of intent, used their own or used both. The simple effects are shown in Table 10.

		95% CI for Mean Difference					
Comparison	Estimate	Lower	Upper	SE	df	t	p
Human - AI	-0.895	-1.350	-0.440	0.232	910	-3.860	< .001
Instructor - Employer	-0.210	-0.665	0.245	0.232	910	-0.907	0.365
The Formal - Both	-0.025	-0.630	0.580	0.308	910	-0.082	0.935
Your - Both	0.187	-0.416	0.790	0.307	910	0.607	0.544

Table 10: Responsibility Contrasts combined across Exps 1,2 & 3

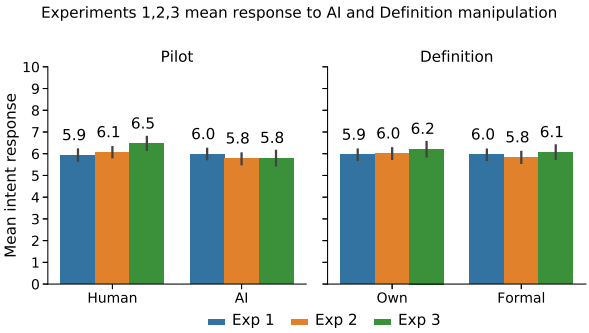
At the end of each experiment, we asked those participants which had been considering AI pilots, in a free text question, whether they thought AI could have intent. Encoding this in a binary way by reading each response and encoding it as yes if the response was predominantly affirmative, the results are shown in Table B16. We see that approximately participants were 50% likely to say yes, versus 33% for No, which indicates participants are not overwhelmingly negative to the idea. Within experiment 3, half of the participants who had not been asked to consider AI pilots in the survey yet their response to the question (23 Yes, 11 No, 4 undecided) was not different from the group asked to only consider AI pilots (20 Yes, 13 No, 3 undecided). The Chi squared statistic to test homogeneity between Experiment and group was 12.4 which has a p-value of 0.26. Thus, the hypothesis that the experimental treatment did not affect responses could not be rejected.

6 General Discussion

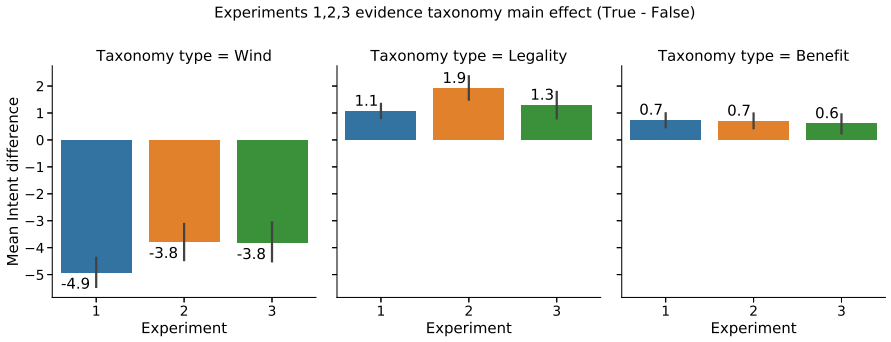
We found evidence of a small difference but statistically significant difference in the judgments of intent between AI and Human Pilots in Experiments 2 and

3 but not 1. In Experiment 3 Human Pilots were judged to be 0.65 points (on a 0-10 inclusive scale) more intentional; in Experiment 2 this was 0.3 points.

Across the three experiments, providing a definition or not did not elicit dramatic differences in intent inferences. Experiment 2 onwards did not use the word 'intent' in the provided definition preventing participants from using their own definition. Several significant interactions were found between the definition and elements of the evidence taxonomy in Experiment 2 however these were not repeated in Experiment 3. We conclude that, in our experimental setting, the definition did a decent job of recreating people's natural concept of intent.



(a) Exp1,2,3: Definition and AI mean intent responses



(b) Exp 1,2,3: Taxonomy main effects

Fig. 11: Experiment 1,2,3: Main effects

The main effects of the taxonomy were repeated across the three experiments in sign, with similar magnitude. The presence of wind (a proxy for causality) lowers intent judgments by between 3.8 and 4.9 points. Legal actions are thought to be more intentional by between 1.1 and 1.9 points and actions which are beneficial (and thereby desirable) to the pilot are between 0.6 and 0.7 points more intentional. The main effects for the three experiments are

aggregated in Figure 11. The negative effect of Wind on intent is likely related to the strong association between causality and intent.

Whilst our experiments did find a difference in attributions of intent between AI and human, it was not large. This is consistent with other recent studies surrounding lay persons treatment of intent in a Robot similarly but not identically to Humans De Graaf and Malle (2018, 2019); Kneer (2020). The phenomena of people treating even abstract objects such as moving geometric shapes as if they had intent has been observed in research since Heider and Simmel (1944). Thellman, Silvervarg, and Ziemke (2017) also find no difference when participants were asked to judge the described behaviour of humans or humanoid robots in conjunction with visual depiction of the actor. They adopt the terminology of Dennett (1987) and term the human phenomena of attributing intent to the behaviour of other actors as adopting "the intentional stance". People might not actually believe that the other actor has intent, but they respond as if it has. They distinguish Dennett-type intent inferences from people actually believing the other actors have intent illustrating the difference with a cartoon example. People's ability to understand the character's mental states is not the same as believing that those cartoon character have any genuine agency and mental states which they term Searle-type intentionality (After Searle (1999)). It seems to us that in a legal context, courts would require this type of intentionality, since intent is something that should be established as a factual beyond all reasonable doubt. This gives weight to the approach of providing a formal definition of intent which jurors can test against evidence they are presented about the AI actor.

The consistent negative effect of illegality on intent is of interest. We expected, given previous studies (Knobe, 2003a; Pettit & Knobe, 2009), that illegal moves would be seen as being more intentional. The sensitivity of intent to 'Moral valence' has even been shown to be present amongst Judges in Kneer and Bourgeois-Gironde (2017). Related is the finding that norm-violations are found to be more causal regardless of outcome Alicke, Rose, and Bloom (2011). That they weren't here, suggests a tendency for participants, in this setting, to seek alternative explanations when norm-breaking behaviour is observed. Molden (2009) finds that people do use what he terms a positive-intention heuristic; outcomes which are positive are seen as more intentional and to a lesser extent, actions which are positive are also seen as more intentional. Similarly, Thellman et al. (2017) find that positive behaviours are more intentional than negative behaviours in humans and humanoid robots. Intuitively, as an observer, if you were to observe a car doing something strange like drive across the centre of a roundabout, your conclusion might not be that the driver (or autopilot) is choosing to do that to speed up their journey. Instead, you might conclude that something had either gone wrong with the car or the driver. Given its novelty and generality across pilot type, we think this error-assumption effect is worth studying in other contexts to see whether it replicates. The effect could be because this study differs from many previous studies on intent because the stimuli are not vignette-based; participants are asked instead to make inferences from evidence of behaviour.

That non-beneficial moves lower intent is intuitive, though the effect was small in our study. The stimuli were not an ideal design to unambiguously separate beneficial from non-beneficial moves, and we think further experiments could investigate this effect. Whether a move was beneficial or not would correspond to the motive of a drone pilot's actions. From a legal perspective, motive can provide circumstantial evidence as to intent, but the lack of it should not disprove its presence. As previously mentioned, it is established in Law that something need not be desired for it to be intended so perhaps participants didn't feel that they needed to understand the precise motivations of the pilot's actions to judge that it had intended something. The weak effect of this variable in our taxonomy could show that our participants do not disagree.

The finding that participants placed higher responsibility on the instructors and employers of AI than those of human pilots indicates some reluctance to place responsibility on AI in the event of harm (only legal persons can commit crimes for example). Given the lack of ability to sanction AI, this could be seen as evidence of the existence of a responsibility gap.

A criticism of specificity is valid. The results of this experiment pertain to flying unmanned drones through a city and may only be limited to that situation. Follow up experiments should try to test across different domains, as De Graaf and Malle (2018) show, whether the intent of an AI is judged the same as a human, varies with setting. One very important thing that differs between the judgment of human and AI actors, is that humans are recognised to be, at some level, the same. Any individual is mostly given the same rights and obligations as any other. AI actors on the other hand come in any number of different designs and capabilities. It might be that the judgment of intent of an AI actor is very dependent on the specific AI, in a way that does not occur when humans judge other humans. Thus any study contrasting judgments of AI versus human behaviour has a specificity limitation.

7 Conclusion

The results of our three experiments generally agree on the following:

1. AI pilots are judged to have less intent than human pilots but the effect is small. This was the case when participants judged the pilots in isolation or whether they judged both types.
2. Providing a folk definition of direct intent does not change judgments of intent in a large way either in Humans or AI. We tested this finding to see whether respondents were simply ignoring the definition in favour of using their own definition when the definition was labelled as intent and found no large difference in responses. We also tested this finding in a within participant setting, again finding no large difference.

People only differentiating slightly between human and AI agents when judging mental states agrees with a large body of Human Robot Interaction (HRI) research which has often found that people are willing to ascribe mental states to robots as if they were humans Malle et al. (2016); Thellman et al. (2017) but

subtle differences exist. Where our research differs, is that we have provided a minimum set of evidence sufficient to identify intent in human and AI agent according to a legal folk definition of intent. The emphasis of this work is not exploring the phenomena of humans ascribing mental states to AI, it is instead exploring what might happen if the law settles upon a definition of intent in AI and lay-people are asked to detect its presence given some evidence.

The second finding indicates that lay-people's attribution of intent is robust to being given a definition and that the definition used in this study is not drastically different from people's instinctive definition. Both indications should be of comfort to courts which often worry about how to define intent and more recently doubt that intent can exist and can be judged in AI.

With regards to the evidence taxonomy the following three main effects were reproduced in each experiment:

1. When the agent is deemed not to have controlled a movement (through the presence of wind), their movement is judged to be much less intentional.
2. Legal actions (movements) are judged to be more intentional than illegal ones.
3. Beneficial actions are judged to be more intentional, but the effect is generally small.

The effect of wind, which reduces pilot control, on attributions of intent is predictable. Illegal actions receiving a lower intention score is somewhat surprising given existing literature on norm transgression and intent. We hypothesise that participants have a resistance to labelling norm-transgression as deliberate, if they are not given unambiguous evidence. This is something to be further explored. The weak effect of beneficial movements is consistent with the legal position that intended results do not need to be desired, a position which is much debated within the psychology community. The result could also be just an artefact of the experimental setting. This caveat applies across any research concerning attitudes towards Autonomous AI given the almost infinite variety of forms they can take. Further work is needed to validate the findings of this research across different AI forms and functions.

In Experiment 3 judgments of causality are shown to be affected in the same way by the evidence taxonomy as judgments of intent. Causality is long considered to be a requirement of intent, but more recently, the intent of the actor has also been shown to affect judgments about the causality of their actions. Whilst Experiment 2 indicated people judged AI pilots as slightly less causal after the main survey, they did not differentiate on a per scenario basis in Experiment 3.

Across all experiments, when asked whether an AI could have intent, roughly 50% of responses were positive, which indicates that lay-people are not universally hostile to the idea. This indicates that respondents were not all *imagining* that the AI Pilot had intent, some at least believed that it could have intent. When asked about the responsibility of the Pilot's instructors and employer, those participants who had considered AI pilots gave a higher

response by 0.9 on a 1-10 scale. This indicates that people attribute less responsibility to an AI agent. The difference is not large given the legal (non) status of AI and indicates that people attribute some responsibility to an AI agent. This is at odds with the legal position which denies any legal personality to AI agents.

Acknowledgments.

Declarations

Funding

- Author 1 is supported by xxxx
- Authors 1 and 2 were supported in this research by xxxxx

Competing interests

The authors have no competing interests.

Ethics approval

Not Applicable

Consent to participate

Consent for publication

Availability of data and materials

All data is available at https://github.com/intentExperiment/flyingdrones1

Code availability

Not Applicable

Authors' contributions

Authors contributed equally.

Appendix A Supporting information

Cases	Sum of Squares	df	Mean Square	F	p	η^2	η_p^2	ω^2
Wind	6173.603	1	6173.603	270.929	< .001	0.458	0.688	0.583
Wind * AI	4.934	1	4.934	0.217	0.643	3.659e-4	0.002	0.000
Wind * Definition	2.312	1	2.312	0.101	0.751	1.714e-4	8.241e-4	0.000
Wind * AI * Definition	18.323	1	18.323	0.804	0.372	0.001	0.006	0.000
Residuals	2802.777	123	22.787					
Legal	292.129	1	292.129	63.497	< .001	0.022	0.340	0.118
Legal * AI	0.599	1	0.599	0.130	0.719	4.444e-5	0.001	0.000
Legal * Definition	5.805	1	5.805	1.262	0.263	4.305e-4	0.010	5.623e-4
Legal * AI * Definition	0.255	1	0.255	0.055	0.814	1.889e-5	4.498e-4	0.000
Residuals	565.885	123	4.601					
Benefit	140.232	1	140.232	34.362	< .001	0.010	0.218	0.062
Benefit * AI	0.042	1	0.042	0.010	0.919	3.106e-6	8.343e-5	0.000
Benefit * Definition	2.579	1	2.579	0.632	0.428	1.913e-4	0.005	0.000
Benefit * AI * Definition	1.331	1	1.331	0.326	0.569	9.872e-5	0.003	0.000
Residuals	501.968	123	4.081					
Wind * Legal	2.034	1	2.034	0.631	0.429	1.509e-4	0.005	0.000
Wind * Legal * AI	1.636	1	1.636	0.508	0.477	1.214e-4	0.004	0.000
Wind * Legal * Definition	0.405	1	0.405	0.126	0.723	3.007e-5	0.001	0.000
Wind * Legal * AI * Definition	3.367	1	3.367	1.045	0.309	2.497e-4	0.008	7.303e-5
Residuals	396.483	123	3.223					
Wind * Benefit	23.514	1	23.514	10.464	0.002	0.002	0.078	0.011
Wind * Benefit * AI	14.187	1	14.187	6.313	0.013	0.001	0.049	0.006
Wind * Benefit * Definition	1.432	1	1.432	0.637	0.426	1.062e-4	0.005	0.000
Wind * Benefit * AI * Definition	0.275	1	0.275	0.122	0.727	2.037e-5	9.929e-4	0.000
Residuals	276.410	123	2.247					
Legal * Benefit	7.181	1	7.181	2.415	0.123	5.326e-4	0.019	0.002
Legal * Benefit * AI	8.411	1	8.411	2.829	0.095	6.238e-4	0.022	0.003
Legal * Benefit * Definition	1.600	1	1.600	0.538	0.465	1.186e-4	0.004	0.000
Legal * Benefit * AI * Definition	5.983	1	5.983	2.012	0.159	4.437e-4	0.016	0.002
Residuals	365.712	123	2.973					
Wind * Legal * Benefit	19.894	1	19.894	8.913	0.003	0.001	0.068	0.009
Wind * Legal * Benefit * AI	0.325	1	0.325	0.146	0.703	2.410e-5	0.001	0.000
Wind * Legal * Benefit * Definition	4.730	1	4.730	2.119	0.148	3.508e-4	0.017	0.001
Wind * Legal * Benefit * AI * Definition	4.091	1	4.091	1.833	0.178	3.034e-4	0.015	0.001
Residuals	274.556	123	2.232					

Table A1: Experiment 1 Within subject effects Anova. Significant effects highlighted.

Cases	Sum of Squares	df	Mean Square	F	p	η^2	η_p^2	ω^2
AI	0.562	1	0.562	0.044	0.834	4.167e-5	3.605e-4	0.000
Definition	0.050	1	0.050	0.004	0.950	3.718e-6	3.218e-5	0.000
AI * Definition	0.087	1	0.087	0.007	0.934	6.476e-6	5.605e-5	0.000
Residuals	1557.713	123	12.664					

Table A2: Experiment 1: Between Subjects Effects

Appendix B Responsibility Information

	F	df1	df2	p
Wind Legal Benefit	1.071	3	123	0.364
Wind Legal No-Benefit	0.429	3	123	0.732
Wind Not-Legal Benefit	1.261	3	123	0.291
Wind Not-Legal No-Benefit	4.335	3	123	0.006
No-Wind Legal Benefit	0.959	3	123	0.415
No-Wind Legal No-Benefit	1.505	3	123	0.217
No-Wind Not-Legal Benefit	3.334	3	123	0.022
No-Wind Not-Legal No-Benefit	0.408	3	123	0.747

Table A3: Experiment 1 Levene's test for Equality of Variances within groups

Cases	Sum of Squares	df	Mean Square	F	p	η^2	η_p^2	ω^2
AI	45.264	1	45.264	7.441	0.007	0.001	0.055	0.010
AI * Definition	18.160	1	18.160	2.986	0.086	5.203e-4	0.023	0.003
Residuals	778.594	128	6.083					
Wind	7249.090	1	7249.090	143.559	< .001	0.208	0.529	0.430
Wind * Definition	1174.090	1	1174.090	23.251	< .001	0.034	0.154	0.105
Residuals	6463.449	128	50.496					
Legal	1944.762	1	1944.762	82.906	< .001	0.056	0.393	0.241
Legal * Definition	227.958	1	227.958	9.718	0.002	0.007	0.071	0.033
Residuals	3002.546	128	23.457					
Benefit	257.733	1	257.733	28.077	< .001	0.007	0.180	0.056
Benefit * Definition	1.179	1	1.179	0.128	0.721	3.377e-5	0.001	0.000
Residuals	1174.979	128	9.180					
AI * Wind	3.522	1	3.522	0.543	0.463	1.009e-4	0.004	0.000
AI * Wind * Definition	3.522	1	3.522	0.543	0.463	1.009e-4	0.004	0.000
Residuals	830.728	128	6.490					
AI * Legal	2.516e-4	1	2.516e-4	4.799e-5	0.994	7.208e-9	3.750e-7	0.000
AI * Legal * Definition	7.035	1	7.035	1.342	0.249	2.015e-4	0.010	4.826e-4
Residuals	671.023	128	5.242					
Wind * Legal	26.813	1	26.813	2.447	0.120	7.682e-4	0.019	0.004
Wind * Legal * Definition	38.667	1	38.667	3.529	0.063	0.001	0.027	0.006
Residuals	1402.538	128	10.957					
AI * Benefit	0.624	1	0.624	0.173	0.678	1.788e-5	0.001	0.000
AI * Benefit * Definition	4.417	1	4.417	1.226	0.270	1.265e-4	0.009	2.324e-4
Residuals	461.165	128	3.603					
Wind * Benefit	53.008	1	53.008	5.702	0.018	0.002	0.043	0.010
Wind * Benefit * Definition	3.958	1	3.958	0.426	0.515	1.134e-4	0.003	0.000
Residuals	1189.900	128	9.296					
Legal * Benefit	9.571	1	9.571	1.842	0.177	2.742e-4	0.014	0.001
Legal * Benefit * Definition	33.818	1	33.818	6.509	0.012	9.688e-4	0.048	0.008
Residuals	664.991	128	5.195					
AI * Wind * Legal	25.379	1	25.379	4.507	0.036	7.271e-4	0.034	0.005
AI * Wind * Legal * Definition	0.379	1	0.379	0.067	0.796	1.084e-5	5.249e-4	0.000
Residuals	720.722	128	5.631					
AI * Wind * Benefit	5.472	1	5.472	0.944	0.333	1.568e-4	0.007	0.000
AI * Wind * Benefit * Definition	1.522	1	1.522	0.263	0.609	4.359e-5	0.002	0.000
Residuals	741.740	128	5.795					
AI * Legal * Benefit	18.391	1	18.391	2.799	0.097	5.269e-4	0.021	0.003
AI * Legal * Benefit * Definition	3.637	1	3.637	0.554	0.458	1.042e-4	0.004	0.000
Residuals	841.017	128	6.570					
Wind * Legal * Benefit	2.178	1	2.178	0.255	0.614	6.240e-5	0.002	0.000
Wind * Legal * Benefit * Definition	9.559	1	9.559	1.120	0.292	2.739e-4	0.009	2.472e-4
Residuals	1092.595	128	8.536					
AI * Wind * Legal * Benefit	8.591	1	8.591	1.663	0.200	2.461e-4	0.013	9.243e-4
AI * Wind * Legal * Benefit * Definition	0.803	1	0.803	0.155	0.694	2.300e-5	0.001	0.000
Residuals	661.248	128	5.166					

Table A4: Experiment 2 Repeated measures ANOVA: Within Subjects Effects. Significant effects highlighted.

Cases	Sum of Squares	df	Mean Square	F	p	η^2	η_p^2	ω^2
Definition	16.361	1	16.361	0.695	0.406	4.687e-4	0.005	0.000
Residuals	3012.478	128	23.535					

Table A5: Experiment 2 Repeated measures ANOVA: between Subjects Effects

	F	df1	df2	p
AI Wind Legal Benefit	0.925	1	128	0.338
AI Wind Legal No-Benefit	3.924	1	128	0.050
AI Wind Not-legal Benefit	6.009	1	128	0.016
AI Wind Not-legal Not Benefit	6.967	1	128	0.009
AI No-Wind Legal Benefit	10.954	1	128	0.001
AI No-Wind Legal NotBenefit	11.119	1	128	0.001
AI No-Wind Not Illegal Benefit	58.103	1	128	< .001
AI No-Wind Not Illegal NoBenefit	6.130	1	128	0.015
Hum Wind Legal Benefit	0.026	1	128	0.873
Hum Wind Legal No-Benefit	2.223	1	128	0.138
Hum Wind Not-legal Benefit	0.724	1	128	0.396
Hum Wind Not-legal No-Benefit	7.763	1	128	0.006
Hum No-Wind Legal Benefit	10.904	1	128	0.001
Hum No-Wind Legal No-Benefit	3.874	1	128	0.051
Hum No-Wind Not-legal Benefit	34.390	1	128	< .001
Hum No-Wind Not-legal No-Benefit	14.782	1	128	< .001

Table A6: Experiment 2 Levene's test for Equality of Variances

		95% CI for Mean Difference						
Variable	Comparison	Estimate	Lower	Upper	SE	df	t	p
Wind	False - True	3.782	3.119	4.352	0.337	129	11.208	< .001
Legal	True - False	1.914	1.514	2.355	0.219	129	8.724	< .001
Benefit	True - False	0.703	0.441	0.967	0.132	129	5.308	< .001
Pilot	AI - Human	-0.289	-0.509	-0.081	0.109	129	-2.656	0.009

Table A7: Experiment 2 Contrasts. Intent scores are averaged across groups not being contrasted. The t-test variant does not assume equal variances.

Cases	Sum of Squares	df	Mean Square	F	p	η^2	η_p^2	ω^2
Defn	6.826	1	6.826	0.708	0.403	3.672e-4	0.010	0.000
Defn * pilot	21.144	1	21.144	2.192	0.143	0.001	0.030	0.005
Defn * F_D	5.132	1	5.132	0.532	0.468	2.761e-4	0.008	0.000
Defn * pilot * F_D	10.098	1	10.098	1.047	0.310	5.433e-4	0.015	2.103e-4
Residuals	675.207	70	9.646					
Wind	4229.268	1	4229.268	98.457	< .001	0.228	0.584	0.481
Wind * pilot	0.518	1	0.518	0.012	0.913	2.789e-5	1.724e-4	0.000
Wind * F_D	109.377	1	109.377	2.546	0.115	0.006	0.035	0.015
Wind * pilot * F_D	37.495	1	37.495	0.873	0.353	0.002	0.012	0.000
Residuals	3006.874	70	42.955					
Legal	513.019	1	513.019	28.377	< .001	0.028	0.288	0.153
Legal * pilot	1.713	1	1.713	0.095	0.759	9.215e-5	0.001	0.000
Legal * F_D	53.105	1	53.105	2.937	0.091	0.003	0.040	0.013
Legal * pilot * F_D	1.166	1	1.166	0.065	0.800	6.275e-5	9.207e-4	0.000
Residuals	1265.505	70	18.079					
Benefit	108.760	1	108.760	11.627	0.001	0.006	0.142	0.045
Benefit * pilot	5.082	1	5.082	0.543	0.464	2.734e-4	0.008	0.000
Benefit * F_D	0.743	1	0.743	0.079	0.779	3.999e-5	0.001	0.000
Benefit * pilot * F_D	15.886	1	15.886	1.698	0.197	8.547e-4	0.024	0.003
Residuals	654.789	70	9.354					
Defn * Wind	14.032	1	14.032	0.855	0.358	7.550e-4	0.012	0.000
Defn * Wind * pilot	0.558	1	0.558	0.034	0.854	2.999e-5	4.848e-4	0.000
Defn * Wind * F_D	470.956	1	470.956	28.684	< .001	0.025	0.291	0.147
Defn * Wind * pilot * F_D	58.476	1	58.476	3.562	0.063	0.003	0.048	0.016
Residuals	1149.324	70	16.419					
Defn * Legal	112.422	1	112.422	10.898	0.002	0.006	0.135	0.044
Defn * Legal * pilot	0.609	1	0.609	0.059	0.809	3.275e-5	8.423e-4	0.000
Defn * Legal * F_D	0.896	1	0.896	0.087	0.769	4.821e-5	0.001	0.000
Defn * Legal * pilot * F_D	6.566	1	6.566	0.637	0.428	3.533e-4	0.009	0.000
Residuals	722.125	70	10.316					
Wind * Legal	67.322	1	67.322	10.766	0.002	0.004	0.133	0.031
Wind * Legal * pilot	33.386	1	33.386	5.339	0.024	0.002	0.071	0.014
Wind * Legal * F_D	1.854	1	1.854	0.297	0.588	9.976e-5	0.004	0.000
Wind * Legal * pilot * F_D	0.061	1	0.061	0.010	0.922	3.273e-6	1.389e-4	0.000
Residuals	437.723	70	6.253					
Defn * Benefit	3.550	1	3.550	0.521	0.473	1.910e-4	0.007	0.000
Defn * Benefit * pilot	1.553	1	1.553	0.228	0.635	8.354e-5	0.003	0.000
Defn * Benefit * F_D	11.107	1	11.107	1.630	0.206	5.976e-4	0.023	0.002
Defn * Benefit * pilot * F_D	6.633	1	6.633	0.973	0.327	3.569e-4	0.014	0.000
Residuals	477.136	70	6.816					
Wind * Benefit	9.609	1	9.609	1.337	0.251	5.170e-4	0.019	0.001
Wind * Benefit * pilot	0.685	1	0.685	0.095	0.758	3.688e-5	0.001	0.000
Wind * Benefit * F_D	18.035	1	18.035	2.510	0.118	9.703e-4	0.035	0.005
Wind * Benefit * pilot * F_D	20.064	1	20.064	2.792	0.099	0.001	0.038	0.006
Residuals	503.019	70	7.186					
Legal * Benefit	0.158	1	0.158	0.031	0.861	8.474e-6	4.440e-4	0.000
Legal * Benefit * pilot	0.061	1	0.061	0.012	0.913	3.303e-6	1.731e-4	0.000
Legal * Benefit * F_D	5.962	1	5.962	1.177	0.282	3.208e-4	0.017	4.919e-4
Legal * Benefit * pilot * F_D	10.972	1	10.972	2.166	0.146	5.903e-4	0.030	0.003
Residuals	354.584	70	5.065					
Defn * Wind * Legal	9.737	1	9.737	1.892	0.173	5.239e-4	0.026	0.003
Defn * Wind * Legal * pilot	0.776	1	0.776	0.151	0.699	4.173e-5	0.002	0.000
Defn * Wind * Legal * F_D	11.753	1	11.753	2.283	0.135	6.323e-4	0.032	0.004
Defn * Wind * Legal * pilot * F_D	0.155	1	0.155	0.030	0.863	8.316e-6	4.288e-4	0.000
Residuals	360.295	70	5.147					
Defn * Wind * Benefit	0.040	1	0.040	0.018	0.894	2.129e-6	2.540e-4	0.000
Defn * Wind * Benefit * pilot	7.212	1	7.212	3.241	0.076	3.880e-4	0.044	0.003
Defn * Wind * Benefit * F_D	3.827	1	3.827	1.720	0.194	2.059e-4	0.024	9.874e-4
Defn * Wind * Benefit * pilot * F_D	3.757	1	3.757	1.689	0.198	2.021e-4	0.024	9.445e-4
Residuals	155.738	70	2.225					
Defn * Legal * Benefit	0.496	1	0.496	0.090	0.765	2.667e-5	0.001	0.000
Defn * Legal * Benefit * pilot	10.179	1	10.179	1.852	0.178	5.477e-4	0.026	0.003
Defn * Legal * Benefit * F_D	2.106	1	2.106	0.383	0.538	1.133e-4	0.005	0.000
Defn * Legal * Benefit * pilot * F_D	8.180	1	8.180	1.488	0.227	4.401e-4	0.021	0.001
Residuals	384.738	70	5.496					
Wind * Legal * Benefit	46.703	1	46.703	6.949	0.010	0.003	0.090	0.020
Wind * Legal * Benefit * pilot	0.537	1	0.537	0.080	0.778	2.890e-5	0.001	0.000
Wind * Legal * Benefit * F_D	0.020	1	0.020	0.003	0.956	1.095e-6	4.324e-5	0.000
Wind * Legal * Benefit * pilot * F_D	3.804	1	3.804	0.566	0.454	2.047e-4	0.008	0.000
Residuals	470.472	70	6.721					
Defn * Wind * Legal * Benefit	3.716	1	3.716	0.953	0.332	1.999e-4	0.013	0.000
Defn * Wind * Legal * Benefit * pilot	16.817	1	16.817	4.312	0.042	9.048e-4	0.058	0.007
Defn * Wind * Legal * Benefit * F_D	5.253	1	5.253	1.347	0.250	2.826e-4	0.019	7.775e-4
Defn * Wind * Legal * Benefit * pilot * F_D	0.389	1	0.389	0.100	0.753	2.094e-5	0.001	0.000
Residuals	272.979	70	3.900					

Table A8: Experiment 3: Intent within Subjects Effects. F_D group refers to whether participants saw formal definition for first set of 8 questions, or were asked to use their own definition of intent.

Cases	Sum of Squares	df	Mean Square	F	p	η^2	η_p^2	ω^2
pilot	125.162	1	125.162	6.075	0.016	0.007	0.080	0.035
F_D	8.017	1	8.017	0.389	0.535	4.313e-4	0.006	0.000
pilot * F_D	10.517	1	10.517	0.510	0.477	5.658e-4	0.007	0.000
Residuals	1442.144	70	20.602					

Table A9: Experiment 3: Intent between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η^2	η_p^2	ω^2
Error type	39.278	1	39.278	7.471	0.008	0.053	0.094	0.046
Error type * Pilot ID	4.900	1	4.900	0.932	0.338	0.007	0.013	0.000
Residuals	378.539	72	5.257					

Table A10: Experiment 3: Error Attribution Repeated Measures ANOVA: Within Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η^2	η_p^2	ω^2
Pilot ID	0.133	1	0.133	0.030	0.862	1.797e-4	4.197e-4	0.000
Residuals	316.387	72	4.394					

Table A11: Experiment 3: Error Attribution Repeated Measures ANOVA: Between Subjects Effects

		95% CI for Mean Difference					
Comparison	Estimate	Lower	Upper	SE	df	t	p
Drone err - Pilot err	-1.031	-1.782	-0.279	0.377	72	-2.733	0.008
Human grp - AI grp	0.060	-0.627	0.747	0.345	72	0.174	0.862

Table A12: Experiment 3 Error Attribution

Cases	Sum of Squares	df	Mean Square	F	p	η^2	η_p^2	ω^2
Defn	7.241	1	7.241	1.367	0.246	6.107e-4	0.019	7.485e-4
Defn * F_D	11.437	1	11.437	2.158	0.146	9.646e-4	0.030	0.002
Defn * pilot	1.301	1	1.301	0.245	0.622	1.097e-4	0.003	0.000
Defn * F_D * pilot	6.662	1	6.662	1.257	0.266	5.618e-4	0.018	5.254e-4
Residuals	370.912	70	5.299					
Wind	3171.921	1	3171.921	111.312	< .001	0.268	0.614	0.426
Wind * F_D	1.179	1	1.179	0.041	0.839	9.945e-5	5.908e-4	0.000
Wind * pilot	0.281	1	0.281	0.010	0.921	2.373e-5	1.410e-4	0.000
Wind * F_D * pilot	80.868	1	80.868	2.838	0.097	0.007	0.039	0.012
Residuals	1994.708	70	28.496					
Legal	444.139	1	444.139	71.620	< .001	0.037	0.506	0.142
Legal * F_D	2.390	1	2.390	0.385	0.537	2.015e-4	0.005	0.000
Legal * pilot	15.036	1	15.036	2.425	0.124	0.001	0.033	0.003
Legal * F_D * pilot	3.729e-4	1	3.729e-4	6.013e-5	0.994	3.145e-8	8.590e-7	0.000
Residuals	434.093	70	6.201					
Benefit	35.840	1	35.840	9.306	0.003	0.003	0.117	0.013
Benefit * F_D	1.966	1	1.966	0.511	0.477	1.658e-4	0.007	0.000
Benefit * pilot	1.178	1	1.178	0.306	0.582	9.934e-5	0.004	0.000
Benefit * F_D * pilot	0.293	1	0.293	0.076	0.783	2.474e-5	0.001	0.000
Residuals	269.584	70	3.851					
Defn * Wind	10.852	1	10.852	2.224	0.140	9.152e-4	0.031	0.002
Defn * Wind * F_D	14.837	1	14.837	3.040	0.086	0.001	0.042	0.004
Defn * Wind * pilot	1.841	1	1.841	0.377	0.541	1.553e-4	0.005	0.000
Defn * Wind * F_D * pilot	1.339	1	1.339	0.274	0.602	1.129e-4	0.004	0.000
Residuals	341.619	70	4.880					
Defn * Legal	9.091	1	9.091	3.090	0.083	7.667e-4	0.042	0.003
Defn * Legal * F_D	0.272	1	0.272	0.092	0.762	2.291e-5	0.001	0.000
Defn * Legal * pilot	1.371	1	1.371	0.466	0.497	1.156e-4	0.007	0.000
Defn * Legal * F_D * pilot	0.198	1	0.198	0.067	0.796	1.668e-5	9.595e-4	0.000
Residuals	205.973	70	2.942					
Wind * Legal	39.218	1	39.218	8.393	0.005	0.003	0.107	0.013
Wind * Legal * F_D	0.249	1	0.249	0.053	0.818	2.104e-5	7.621e-4	0.000
Wind * Legal * pilot	0.760	1	0.760	0.163	0.688	6.413e-5	0.002	0.000
Wind * Legal * F_D * pilot	9.606	1	9.606	2.056	0.156	8.101e-4	0.029	0.002
Residuals	327.106	70	4.673					
Defn * Benefit	0.809	1	0.809	0.253	0.616	6.822e-5	0.004	0.000
Defn * Benefit * F_D	0.086	1	0.086	0.027	0.870	7.219e-6	3.824e-4	0.000
Defn * Benefit * pilot	0.234	1	0.234	0.073	0.787	1.976e-5	0.001	0.000
Defn * Benefit * F_D * pilot	5.773	1	5.773	1.806	0.183	4.869e-4	0.025	0.001
Residuals	223.732	70	3.196					
Wind * Benefit	41.050	1	41.050	14.272	< .001	0.003	0.169	0.016
Wind * Benefit * F_D	0.003	1	0.003	0.001	0.973	2.785e-7	1.640e-5	0.000
Wind * Benefit * pilot	8.764	1	8.764	3.047	0.085	7.392e-4	0.042	0.002
Wind * Benefit * F_D * pilot	4.415	1	4.415	1.535	0.220	3.723e-4	0.021	6.351e-4
Residuals	201.331	70	2.876					
Legal * Benefit	6.472	1	6.472	1.998	0.162	5.458e-4	0.028	0.001
Legal * Benefit * F_D	3.835	1	3.835	1.184	0.280	3.234e-4	0.017	2.433e-4
Legal * Benefit * pilot	2.093	1	2.093	0.646	0.424	1.766e-4	0.009	0.000
Legal * Benefit * F_D * pilot	0.070	1	0.070	0.022	0.883	5.914e-6	3.091e-4	0.000
Residuals	226.754	70	3.239					
Defn * Wind * Legal	0.887	1	0.887	0.337	0.563	7.484e-5	0.005	0.000
Defn * Wind * Legal * F_D	2.422e-4	1	2.422e-4	9.208e-5	0.992	2.043e-8	1.315e-6	0.000
Defn * Wind * Legal * pilot	2.788	1	2.788	1.060	0.307	2.351e-4	0.015	6.563e-5
Defn * Wind * Legal * F_D * pilot	1.711	1	1.711	0.651	0.423	1.443e-4	0.009	0.000
Residuals	184.120	70	2.630					
Defn * Wind * Benefit	0.548	1	0.548	0.174	0.677	4.621e-5	0.002	0.000
Defn * Wind * Benefit * F_D	1.276	1	1.276	0.406	0.526	1.076e-4	0.006	0.000
Defn * Wind * Benefit * pilot	0.164	1	0.164	0.052	0.820	1.381e-5	7.443e-4	0.000
Defn * Wind * Benefit * F_D * pilot	6.196	1	6.196	1.973	0.165	5.225e-4	0.027	0.001
Residuals	219.816	70	3.140					
Defn * Legal * Benefit	0.892	1	0.892	0.259	0.612	7.523e-5	0.004	0.000
Defn * Legal * Benefit * F_D	0.583	1	0.583	0.169	0.682	4.919e-5	0.002	0.000
Defn * Legal * Benefit * pilot	0.861	1	0.861	0.250	0.619	7.261e-5	0.004	0.000
Defn * Legal * Benefit * F_D * pilot	2.365	1	2.365	0.687	0.410	1.995e-4	0.010	0.000
Residuals	241.028	70	3.443					
Wind * Legal * Benefit	17.803	1	17.803	6.081	0.016	0.002	0.080	0.006
Wind * Legal * Benefit * F_D	0.013	1	0.013	0.004	0.948	1.067e-6	6.175e-5	0.000
Wind * Legal * Benefit * pilot	0.661	1	0.661	0.226	0.636	5.579e-5	0.003	0.000
Wind * Legal * Benefit * F_D * pilot	2.922e-4	1	2.922e-4	9.981e-5	0.992	2.465e-8	1.426e-6	0.000
Residuals	204.944	70	2.928					
Defn * Wind * Legal * Benefit	5.926	1	5.926	1.974	0.164	4.998e-4	0.027	0.001
Defn * Wind * Legal * Benefit * F_D	0.634	1	0.634	0.211	0.647	5.345e-5	0.003	0.000
Defn * Wind * Legal * Benefit * pilot	2.823	1	2.823	0.940	0.336	2.380e-4	0.013	0.000
Defn * Wind * Legal * Benefit * F_D * pilot	0.951	1	0.951	0.317	0.575	8.018e-5	0.005	0.000
Residuals	210.107	70	3.002					

Table A13: Experiment 3: Causal ratings, within Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η^2	η_p^2	ω^2
F_D	7.196	1	7.196	0.231	0.633	6.069e-4	0.003	0.000
pilot	16.487	1	16.487	0.528	0.470	0.001	0.007	0.000
F_D * pilot	0.275	1	0.275	0.009	0.926	2.316e-5	1.257e-4	0.000
Residuals	2185.289	70	31.218					

Table A14: Experiment 3: Causal ratings, between Subjects Effects

Cases	Sum of Squares	df	Mean Square	F	p	η^2	η_p^2	ω^2
employee	150.979	1	150.979	14.897	< .001	0.016	0.016	0.015
role	8.332	1	8.332	0.822	0.365	8.821e-4	9.027e-4	0.000
Defn	9.509	2	4.754	0.469	0.626	0.001	0.001	0.000
employee * role	5.141	1	5.141	0.507	0.477	5.442e-4	5.571e-4	0.000
employee * Defn	32.538	2	16.269	1.605	0.201	0.003	0.004	0.001
role * Defn	10.274	2	5.137	0.507	0.603	0.001	0.001	0.000
employee * role * Defn	6.693	2	3.346	0.330	0.719	7.085e-4	7.251e-4	0.000
Residuals	9222.558	910	10.135					

Table B15: ANOVA - Responsibility for harm caused by a pilot, ratings taken over Experiments 1,2 and 3

Experiment	Pilot	Definition	No	Unsure	Yes	**Total**	%No	%Unsure	%Yes
1	AI	The Formal	9	8	14	**31**	29	26	45
1	AI	Your	9	8	17	**34**	26	24	50
2	Both	The Formal	26	14	23	**63**	41	22	37
2	Both	your	21	9	37	**67**	31	13	55
3	AI	Both	13	3	20	**36**	36	8	56
3	Human	Both	11	4	23	**38**	29	11	61
Total			**89**	**46**	**134**	269			

Table B16: Experiments 1,2,3: Response to question: Do you think AI can have intent?

References

Abbott, R., & Sarch, A. (2020). Punishing artificial intelligence: Legal fiction or science fiction. *Is Law Computable?*, 323–384. Retrieved from https://doi.org/10.5040/9781509937097.ch-008

Adams, F., & Steadman, A. (2004). Intentional Action in Ordinary Language : Core Concept or Pragmatic Understanding? *Analysis*, *64*(2), 173–181. Retrieved from http://www.jstor.org/stable/3329124

Alicke, M.D., Rose, D., Bloom, D. (2011). Causation, Norm Violation and Culpable Control. *Journal of Philosophy*, *108*(12), 670–696. Retrieved from https://www.jstor.org/stable/23142912

Alldredge, J. (2015). The " CSI Effect " and Its Potential Impact on Juror Decisions. *Themis: Research Journal of Justice Studies and Forensic Science*, *3*. Retrieved from https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=1027&context=themis

Banks, J. (2019). A perceived moral agency scale: Development and validation of a metric for humans and social machines. *Computers in Human Behavior*, *90*(November 2017), 363–371. Retrieved from https://doi.org/10.1016/j.chb.2018.08.028

Bigman, Y.E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*(March), 21–34. Retrieved from https://doi.org/10.1016/j.cognition.2018.08.003

Bratman, M.E. (1990). What is Intention? P.R. Cohen, J. Morgan, & M.E. Pollock (Eds.), *Intentions in communication* (chap. 2). MIT Press.

Coffey, G. (2009). Codifying the Meaning of 'Intention' in the Criminal Law. *The Journal of Criminal Law*, *73*(5), 394–413. Retrieved from https://doi.org/10.1350/jcla.2009.73.5.590

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380. Retrieved from https://doi.org/10.1016/j.cognition.2008.03.006

De Graaf, M.M., & Malle, B.F. (2018). People's Judgments of Human and Robot Behaviors: A Robust Set of Behaviors and Some Discrepancies. *ACM/IEEE International Conference on Human-Robot Interaction*(March), 97–98. Retrieved from https://doi.org/10.1145/3173386.3177051

De Graaf, M.M., & Malle, B.F. (2019). People's Explanations of Robot Behavior Subtly Reveal Mental State Inferences. *ACM/IEEE International Conference on Human-Robot Interaction*, *2019-March*, 239–248. Retrieved from https://doi.org/10.1109/HRI.2019.8673308

Dennett, D. (1987). *The Intentional Stance*. MIT Press.

Dietvorst, B.J., & Bartels, D.M. (2021). Consumers Object to Algorithms Making Morally Relevant Tradeoffs Because of Algorithms' Consequentialist Decision Strategies. *Journal of Consumer Psychology*. Retrieved from https://doi.org/10.2139/ssrn.3753670

Furlough, C., Stokes, T., Gillan, D.J. (2021). Attributing Blame to Robots: I. The Influence of Robot Autonomy. *Human Factors*, *63*(4), 592–602. Retrieved from https://doi.org/10.1177/0018720819880641

Ginther, M.R., Shen, F.X., Bonnie, R.J., Hoffman, M.B., Jones, O.D., Marois, R., Simons, K.W. (2014). The language of Mens Rea. *Vanderbilt Law Review*, *67*(5), 1327–1372.

Heider, F., & Simmel, M. (1944). An Experimental Study of Apparent Behavior. *The American Journal of Psychology*, *57*(2), 243–259.

Hidalgo, C.A., Orghian, D., Canals, J.A., de Almeida, F., Martin, N. (2021). *How Humans Judge Machines*. MIT Press.

Ishowo-Oloko, F., Bonnefon, J.-F., Soroye, Z., Crandall, J., Rahwan, I., Rahwan, T. (2019). Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nature Machine Intelligence*, *1*(11), 517–521. Retrieved from http://dx.doi.org/10.1038/s42256-019-0113-5

Johnson, D.G., & Verdicchio, M. (2019). AI, agency and responsibility: the VW fraud case and beyond. *AI and Society*, *34*(3), 639–647. Retrieved from http://dx.doi.org/10.1007/s00146-017-0781-9

Jörling, M., Böhm, R., Paluch, S. (2019). Service Robots: Drivers of Perceived Responsibility for Service Outcomes. *Journal of Service Research*, *22*(4), 404–420. Retrieved from https://doi.org/10.1177/1094670519842334

Kenny, A. (2013). Intention and Side Effects: the Mens Rea for Murder. J. Keown & R.P. George (Eds.), *Reason, morality, and law: The philosophy of john finnis* (pp. 109–117). Oxford Scholarship Online. Retrieved from https://doi.org/10.1093/acprof:oso/9780199675500.001.0001

Klass, A.B. (2007). Punitive damages and valuing harm. *Minnesota Law Review*, *92*(1), 83–160.

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., Tenenbaum, J.B. (2015). Inference of intention and permissibility in moral decision making. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, *1*(1987), 1123–1128.

Kneer, M. (2020). Can a robot lie ? *Cognitive Science*, *45*.

Kneer, M., & Bourgeois-Gironde, S. (2017). Mens rea ascription, expertise and outcome effects: Professional judges surveyed. *Cognition*, *169*(August), 139–146. Retrieved from https://doi.org/10.1016/j.cognition.2017.08.008

Knobe, J. (2003a). Intentional action and side effects in ordinary language. *Analysis*, *63*, 190–194. Retrieved from https://doi.org/10.1111/1467-8284.00419

Knobe, J. (2003b). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology*, *16*(2), 309–324. Retrieved from https://doi.org/10.1080/09515080307771

Knobe, J. (2004). Intention, Intentional Action and Moral Considerations. *Analysis*, *64*(2), 181–187.

Knobe, J. (2006). The Concept of Intentional Action: A Case Study in the Uses of Folk Psychology. *Philosophical Studies*, *130*(2), 203–231. Retrieved

from https://www.jstor.org/stable/4321796

Knobe, J., & Malle, B. (1997). The Folk Concept of Intentionality. *Journal of Experimental Social Psychology*, *33*(33), 101–121.

Lagnado, D.A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, *108*(3), 754–770. Retrieved from https://doi.org/10.1016/j.cognition.2008.06.009

Leike, J., Martic, M., Krakovna, V., Ortega, P.A., Everitt, T., Lefrancq, A., ... Legg, S. (2017). *AI Safety Gridworlds.* Retrieved from http://arxiv.org/abs/1711.09883

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

List, C., & Pettit, P. (2011). *Group Agency: The Possibility, Design and Status of Corporate Agents.* Oxford Scholarship Online. Retrieved from https://doi.org/10.1093/acprof:oso/9780199591565.001.0001

Malle, B.F., & Nelson, S.E. (2003). Judging mens rea: The tension between folk concepts and legal concepts of intentionality. *Behavioral Sciences and the Law*, *21*(5), 563–580. Retrieved from https://doi.org/10.1002/bsl.554

Malle, B.F., Scheutz, M., Forlizzi, J., Voiklis, J. (2016). Which robot am I thinking about? The impact of action and appearance on people's evaluations of a moral robot. *ACM/IEEE International Conference on Human-Robot Interaction*, *2016-April*(October 2017), 125–132. Retrieved from https://doi.org/10.1109/HRI.2016.7451743

McManus, R.M., & Rutchick, A.M. (2019). Autonomous Vehicles and the Attribution of Moral Responsibility. *Social Psychological and Personality Science*, *10*(3), 345–352. Retrieved from https://doi.org/10.1177/1948550618755875

Mele, A.R., & Cushman, F. (2007). Intentional Action, Folk Judgments, and Stories: Sorting Things Out. *Midwest Studies in Philosophy*, *31*(1), 184–201. Retrieved from https://doi.org/10.1111/j.1475-4975.2007.00147.x

Molden, D.C. (2009). Finding meaning in others' intentions: The process of judging intentional behaviors and intentionality itself. *Psychological Inquiry*, *20*(1), 37–43. Retrieved from https://doi.org/10.1080/10478400902744295

Mueller, P.A., Solan, L.M., Darley, J.M. (2012). When Does Knowledge Become Intent? Perceiving the Minds of Wrongdoers. *Journal of Empirical Legal Studies*, *9*(4), 859–892. Retrieved from https://doi.org/10.1111/j.1740-1461.2012.01269.x

Parsons, S. (2000). Intention in Criminal Law: why is it so difficult to find? *Mountbatten Journal of Legal Studies*, *4*(1 & 2), 5–19. Retrieved from https://doi.org/10.1017/s0841820900001375

Pettit, D., & Knobe, J. (2009). The pervasive impact of moral judgment. *Mind and Language*, *24*(5), 586–604. Retrieved from https://doi.org/10.1111/j.1468-0017.2009.01375.x

Quillien, T., & German, T.C. (2021). A simple definition of 'intentionally'. *Cognition*, *214*(June), 104806. Retrieved from https://doi.org/10.1016/j.cognition.2021.104806

Robinson, P.H., & Darley, J.M. (1995). *Justice, Liability, and Blame : Community Views and the Criminal Law* (1634th ed.). Faculty Scholarship at Penn Law. Retrieved from https://scholarship.law.upenn.edu/faculty_scholarship/1634

Searle, J.R. (1999). *Mind, Language and Society: Philosophy in the real world.* Basic Books.

Simester, A.P., Spencer, J.R., Stark, F., Sullivan, G.R., Virgo, G.J. (2019). Mens Rea. *Simester and sullivan's criminal law* (7th ed., pp. 137–190). Hart.

Smith, J.C. (1990). A note on " intention ". *Criminal Law review*, *85*.

Smith, V.L. (1993). When prior knowledge and law collide - Helping jurors use the law. *Law and Human Behavior*, *17*(5), 507–536. Retrieved from https://doi.org/10.1007/BF01045071

Sommers, R. (2021). Experimental Jurisprudence. *Science*, *373*(6553). Retrieved from https://doi.org/10.2139/ssrn.3680107

The American Law Insitute (2017). *General Requirements of Culpability.* Retrieved from https://archive.org/details/ModelPenalCode_ALI/page/n31/mode/2up

The Sentencing Council (2019). *General guideline: overarching principles.*

Thellman, S., Silvervarg, A., Ziemke, T. (2017). Folk-psychological interpretation of human vs. humanoid robot behavior: Exploring the intentional stance toward robots. *Frontiers in Psychology*, *8*(NOV), 1–14. Retrieved from https://doi.org/10.3389/fpsyg.2017.01962

Tobia, K. (2021). Law and the Cognitive Science of Ordinary Concepts. *Law and mind* (pp. 86–96). Cambridge University Press. Retrieved from https://doi.org/10.1017/9781108623056.005

Tobia, K. (2022). Experimental Jurisprudence. *University of Chicago Law Review*, *89*.

Williams, G. (1987). Oblique intention. *The Cambridge Law Journal*, *46*(3), 417–438. Retrieved from https://doi.org/10.1017/S0008197300117453

Yeo, N. (2020). Mistakes and knowledge in algorithmic trading : the Singapore Court of Appeal case of Quoine v B2C2. *Journal of International Banking and Financial Law*, *35*(5), 300–305.

Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition*, *120*(2), 202–214. Retrieved from http://dx.doi.org/10.1016/j.cognition.2011.04.005