Intents in Actions

Meir Friedenberg
meir@cs.cornell.edu
Cornell University
Ithaca, NY, USA

Joseph Y. Halpern
halpern@cs.cornell.edu
Cornell University
Ithaca, NY, USA

ABSTRACT

What outcomes does an agent intend in performing an action? We develop a principled approach to this question and provide a new formal definition of intent in a causal framework. Our definition is modular, draws on ideas from philosophy of law, and works in many natural cases where earlier proposed definitions did not.

CCS CONCEPTS

• **Applied computing → Law**.

KEYWORDS

intent, causality, causal models

ACM Reference Format:
Meir Friedenberg and Joseph Y. Halpern. 2025. Intents in Actions. In *Proceedings of The 20th International Conference on Artificial Intelligence and Law (ICAIL '25).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

G.E.M. Anscombe, in her groundbreaking work on intention [3], distinguished three different kinds of intentions and explored the relationship between them. The first of these, "expression of intention for the future", has to do with plans to either do or bring something about in the future. The second, "intentional action", is what we might think of as having done something on purpose. The third, "intention in acting" (also sometimes known as "intention with which"), is about the reasons for which an agent performed an action.

Our focus is on providing a formal definition of the third type of intent, intention in acting. While interest in developing planning systems led to a long and influential line of work formalizing the first type of intent (see e.g. [9, 20, 22, 24]), relatively little has been done on formalizing intention in acting. Two significant exceptions are the work of Kleiman-Weiner et al. [19] and Halpern and Kleiman-Weiner [16] (HK from now on), which provide definitions similar in spirit to ours. (We discuss how our work relates to theirs in Section 5.) The third type of intent seems just as relevant for both legal reasoning and AI as the first type is, for a different set of problems. In common law, some form of criminal intent (or *mens rea*) is generally necessary in order to be held criminally liable. Glanville Williams, an influential scholar of criminal law, wrote that "With one exception, an act is intentional as to a consequence if it is done with (motivated by) the wish, desire, purpose or aim (all synonyms in this context) of producing the result in question" [29].[1]

Moving to AI, consider an autonomous agent such as a robot or autonomous vehicle trying to understand the intent of a human principal so as to either be able to assist the principal or be able to perform a task that the principal desires, what is important for the autonomous system to understand is not what future plans the principal has, but rather what outcome the principal is trying to achieve by acting in that manner. Similarly, *legible* or *intent-expressive* robot motion was defined in the human-robot interaction literature as "motion that enables an observer to quickly and confidently infer the correct goal" [10], which seems connected to the type of intent that we consider here rather than the notion prevalent in the planning literature.

With this in mind, we propose a definition of what it means for an action to be motivated by a conjunctive outcome, and then show how to go from motivation to intent. While this is arguably the most common case (and is the only case that has been considered in previous work), we show by example that cases of disjunctive intent and motivation also arise quite naturally. Our definitions of motivation and intent in the conjunctive case can be extended easily to get corresponding definitions in the disjunctive case. We then compare our approach to work done on legal definitions of intent (specifically, that of Duff [11]), work on intent in the planning community (with a focus on that of Cohen and Levesque [9]), work done in the stit ("seeing to it that") framework [1, 7, 21], and the definition given by HK.

Like the work of HK and Kleiman-Weiner et al. [19], our definition of intent uses causality in a significant way (in fact, we define intent in a causal framework). The use of causality is perhaps the main feature distinguishing our work from previous work on intention like that of CL and work in the stit framework. We believe that our use of causality is critical. As we point out in Section 5, causality plays a key role in the legal definition of intent. Moreover, we believe that the inability to express a causal connection leads to problems in the CL approach (again, see Section 5) and will lead to analogous problems in other approaches.

The rest of this paper is organized as follows: In the next section, we review causal models and how we model an agent's epistemic state. In Section 3 we provide our definition of intent when intentions

ICAIL '25, June 16–20, 2025, Chicago, IL, USA
© 2025 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

[1] The one exception is what he terms "oblique intention", where one doesn't desire an outcome but does believe it is certain to occur. It also bears mentioning that this is not the only notion of intention considered in the legal literature; for example, in R. v. Mohan the Criminal Division of the Ontario Court of Appeal held that "specific intent" could be defined as "a decision to bring about a certain consequence" or as the "aim." (The Supreme Court later reversed this decision, though not on this basis.) Though the second of these definitions ("aim") seems to accord with the one we consider here, the first seems to accord more closely with Anscombe's first notion of intent. A thorough analysis of the different notions of intent considered in the legal literature is beyond the scope of this work, as our primary goal is to draw on the relevant legal theory to help formalize this notion of intent.

are conjunctions; in Section 4, we extend the definition to the disjunctive case. We discuss related work in Section 5, and conclude with some discussion of future work in Section 6.

2 CAUSAL MODELS AND EPISTEMIC STATES

Our framework is based on the causal models framework of Halpern and Pearl [17]. We briefly review it here. A causal scenario is described by a set of variables and their values. We distinguish between *exogenous* variables, those whose values are determined by factors outside of the model, and *endogenous* variables, those whose values are determined by factors within the model. The values of the variables are related by *structural equations.*

A causal model $M = (\mathcal{S}, \mathcal{F})$ consists of a *signature* $\mathcal{S}$ and a set $\mathcal{F}$ of structural equations. A signature $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$ consists of a (finite but non-empty) set $\mathcal{U}$ of exogenous variables, a (finite but non-empty) set $\mathcal{V}$ of endogenous variables, and range function $\mathcal{R}$ mapping every variable in $\mathcal{U} \cup \mathcal{V}$ to the finite set of values it can take on. Because we need to consider an agent's action, like HK, we assume that there is an *action variable* A in $\mathcal{V}$, where $\mathcal{R}(A)$ is the set of actions available to the agent. It is straightforward to extend this framework to multiple agents and multiple actions but, for simplicity, we stick to a single agent and a single action.

$\mathcal{F}$ associates with each endogenous variable $X \in \mathcal{V}$ a structural equation, a function denoted F_X such that $F_X : (\times_{U\in\mathcal{U}}\mathcal{R}(U)) \times (\times_{Y\in\mathcal{V}-\{X\}}\mathcal{R}(Y)) \to \mathcal{R}(X)$; that is, F_X determines the value of X, given the values of all the other variables in $\mathcal{U} \cup \mathcal{V}$. Rather than writing something like $F_X(Y, Z, U) = Y + U$, we typically write $X = Y + U$.

We also want to be able to reason about interventions in these models. If $\vec{X}$ is a vector of endogenous variables and $\vec{x}$ is a vector of values such that $x_i \in \mathcal{R}(X_i)$ for all i, the model $M_{\vec{X}\leftarrow\vec{x}}$ is the same as M except that, for each i, the structural equation for X_i is replaced by $X_i = x_i$. Intuitively, this captures intervening to override the causal structure and instead set the value of X_i to be x_i.

An assignment of values to all of the exogenous variables is called a *context.* A *causal setting* is a tuple $(M, \vec{u})$ consisting of a model M and a context $\vec{u}$. A *world* w is a complete assignment to the endogenous variables. As is standard in the literature, in this paper we restrict to *acyclic* models, where there are no cycles in the causal dependencies among the variables (where a variable X depends on Y if there is some setting of the variables other than X and Y such that the value of X given by F_X can change when the value of Y changes). This ensures that the values of all variables are uniquely determined by the structural equations, given a causal setting. We let $w_{M,\vec{u}}$ denote the (unique) world determined by the causal setting $(M, \vec{u})$. Thus, $w_{M_{\vec{X}\leftarrow\vec{x}},\vec{u}}$ is the world that results from intervening to set $\vec{X}$ to $\vec{x}$ in $(M, \vec{u})$. We sometimes write $w_{M,\vec{X}\leftarrow\vec{x},\vec{u}}$ rather than $w_{M_{\vec{X}\leftarrow\vec{x}},\vec{u}}$. For a Boolean combination φ of *primitive events*, formulas of the form $X = x$ for $X \in \mathcal{V}$ and $x \in \mathcal{R}(X)$, we can define what it means for φ to be true at a world w, written $w \models \varphi$, in the obvious way.

Finally, in order to talk about the beliefs and desires of an agent, we follow HK in defining the agent's *epistemic state* to be a tuple $(\Pr, \mathcal{K}, \mathbf{u})$. The set $\mathcal{K}$ consists of the causal settings that the agent considers possible, all of which have the same signature for their respective models; Pr is a probability distribution over $\mathcal{K}$ describing the agent's (subjective) beliefs about the likelihood of each causal setting in $\mathcal{K}$; and $\mathbf{u}$ is a utility function over worlds that describes the agents preferences by mapping each world to a real-valued utility. The assumption that we can model an agent's preferences using a real-valued utility function (which is also made by Kleiman-Weiner et al. [19]) is clearly quite nontrivial, although it is standard in AI and decision theory. That said, while this use of utility makes the exposition easier, we do not believe that it is a critical component of our approach. We could consider alternatives to utility, such as just assuming a partial order on outcomes. Moreover, even if we use a real-valued utility, we can use different decision rules, ranging from expected utility maximization to prospect theory [18] to regret minimization [23, 25]. These issues should be largely orthogonal to the issues we focus on here. We briefly return to these issues in Section 6.

3 INTENTION: THE CONJUNCTIVE CASE

In this section we consider the common case where intentions are taken to be conjunctive formulas of the form $\bigwedge_{1\le i\le N} Y_i = y_i$. Every formula in this section should be taken to have this form, sometimes written $\vec{Y} = \vec{y}$. For simplicity, we require here and throughout this paper that no Y_i is the action variable A. As we noted in the introduction, the notion of intent we are trying to capture seems to be closely tied to the notion of motivation, so we begin by providing conditions that we take to constitute motivation, and then add one more condition to define intent.

3.1 Motivation

Before we begin describing the conditions themselves, let us be clear about where we are headed:

Definition 3.1. The conjunctive formula φ *motivated action a in epistemic state* $\mathcal{E}$ if the four conditions PB, CA, Rel, and Des (defined below) are satisfied.

We next define the four conditions carefully. As we said, we assume throughout this section that φ is a conjunctive formula of the form $\vec{Y} = \vec{y}$.

Possibility of Bringing About: For an agent to have been motivated by φ to take action a, she must have considered it possible that taking action a might bring about φ.[2] Let

$$[[[A \leftarrow a]\varphi]]_{\mathcal{K}} = \{(M, \vec{u}) \in \mathcal{K} : w_{M,A\leftarrow a,\vec{u}} \models \varphi\},$$

the set of causal settings in $\mathcal{K}$ where φ is true after doing a. A natural first attempt at capturing this condition is to require

$$\Pr([[[A \leftarrow a]\varphi]]_{\mathcal{K}}) > 0. \qquad (1)$$

This is not quite right, though. The inequality above captures that "she believes the result may occur, given her action." A stronger condition of "believing that she can bring the result about" seems necessary; an agent cannot be motivated by a desire to bring about φ if she believes she cannot affect φ. Thus, we instead require that there be some causal setting and some alternative action a' such

[2] As the English suggests, we mainly expect this condition to apply if the agent does not currently believe that φ holds. But all of what we say makes sense even if the agent does currently believe that φ holds. For example, the agent might believe that φ will stop holding if nothing is done, or if some alternative to action a is performed.

that φ occurs in the setting if action a is taken but not if a' is taken. Formally, we say that $(M, \vec{u}, a')$ *PB-certifies* φ for a if

$$w_{M,A\leftarrow a,\vec{u}} \models \varphi \text{ and } w_{M,A\leftarrow a',\vec{u}} \not\models \varphi.$$

The condition we want is then:

PB. There exists a setting $(M, \vec{u})$ with $\Pr((M, \vec{u})) > 0$ and an action $a' \in \mathcal{R}(A)$ such that $(M, \vec{u}, a')$ PB-certifies φ for a.

Note that PB implies the simpler requirement (1) above. We can think of the first half of PB-certification (essentially, (1)) as a sufficiency condition; it says that performing a suffices to bring about φ (in some setting). The second half can be viewed as a necessity condition; performing a is necessary to bring about φ (again, in some setting), in the sense that, but for a, φ would not have occurred (in particular, if a' had been performed).

Cause of Action: For this condition, what we intuitively want to capture is the idea that, no matter what action was taken, obtaining outcome φ would suffice to get the agent at least as much expected utility as she'd expect to get under any alternative. Let

$$\mathrm{EU}_{\mathcal{E}}[\vec{Y} \leftarrow \vec{y}] = \sum_{(M,\vec{u})\in\mathcal{K}} \Pr((M, \vec{u}))\mathbf{u}(w_{M,\vec{Y}\leftarrow\vec{y},\vec{u}});$$

thus, $\mathrm{EU}_{\mathcal{E}}[\vec{Y} \leftarrow \vec{y}]$ is the agent's expected utility under epistemic state $\mathcal{E}$ given intervention $\vec{Y} \leftarrow \vec{y}$. We then require that for φ (i.e., $\vec{Y} = \vec{y}$) to motivate a, the following holds:

CA. For all actions $a' \in \mathcal{R}(A)$,

$$\mathrm{EU}_{\mathcal{E}}[\vec{Y} \leftarrow \vec{y}; A \leftarrow a'] \geq \max_{a''\in\mathcal{R}(A)} \mathrm{EU}_{\mathcal{E}}[A \leftarrow a''].$$

Thus, getting φ for sure, no matter what action is performed, is at least as good an outcome as we'd expect from just performing an action, without necessarily getting φ. (Note that performing an action may result in φ with some probability, but will not in general guarantee φ.)

CA lets us deal with a well-known set of examples that cause problems for other definitions of intention. For example, Chisholm [8] considers Tom, who wanted to poison his uncle, and accidentally killed a pedestrian in a car crash on his way to buy the poison. It then turn out that the pedestrian was in fact Tom's uncle. But we do not want to say that Tom intended to kill his uncle by running over the pedestrian (although some definitions of intention do). CA blocks this. If Tom considers it quite unlikely that the pedestrian is his uncle, then the expected utility of killing the pedestrian and performing some other act will be lower than that of Tom just killing his uncle.

Desirability: Unfortunately, a φ that doesn't really matter to the agent could still satisfy PB and CA, suggesting that these conditions do not sufficie to capture motivation. To see this, consider the following example:

Example 3.2. Consider a scenario with binary variables B, C, and D (taking on values 0 and 1), as well as an action variable A with (exactly) two possible values, a and a'. If action a is taken then B gets value 0 and C and D both get values 1. If action a' is taken then B gets value 100 and C and D get values 0. The agent's utility is 100 if $C + D = 1$, and is $100 - B$ if $C + D \neq 1$. Thus, with action a, she gets utility 100 (because $B = 0$), while with a', she gets 0. But note that the formula $C = 1$ in fact satisfies conditions PB for action a and CA. $C = 1$ holds with a but not a', so PB holds, and if we intervened to ensure $C = 1$, then action a' would also give utility 100. However, it seems strange to say that the agent was motivated by $C = 1$, as we can see from the equations that the value of C really isn't contributing to her utility here. Intuitively, what's going wrong is that $C = 1$ would actually be helpful under the other actions, and CA doesn't do anything to test whether $C = 0$ is really desirable as an outcome of a.

To remedy this, we require φ to be a cause of the expected utility being high with action a. Formally, we require

Des. φ (i.e. $\vec{Y} = \vec{y}$) is a cause of the high expected utility of a; that is, there exists a value $\vec{y}'$ such that

$$\mathrm{EU}_{\mathcal{E}}[\vec{Y} \leftarrow \vec{y}'; A \leftarrow a] < \mathrm{EU}_{\mathcal{E}}[\vec{Y} \leftarrow \vec{y}; A \leftarrow a].$$

Intuitively, $\vec{Y} \leftarrow \vec{y}$ can be thought of as a "but-for cause" of the (high) expected utility of a (and thus is what makes a desirable), because some other value of $\vec{Y}$ would have decreased the utility. (See [15] for a more detailed discussion of actual causality.) Of course, we can say even more if we have a more quantitative notion of desirability; see Section 6 for more discussion of this issue.

Relevance: Finally, we need a minimality condition to ensure that all of the conjuncts of φ are actually relevant. It is quite possible to add intuitively irrelevant conjuncts to φ and still have conditions PB, CA, and Des hold. For example, consider a simple scenario where there are three Boolean variables, B, C, and D, as well as action variable A. The agent gets a utility of 1 for B being true, 2 for C being true, 0 for D being true, and 0 for any of B, C, or D that are false. Her total utility is the sum of the utilities for each of the variables. Action a_0 (deterministically) causes B to be true and C and D to be false, whereas action a_1 (deterministically) causes B to be false and C and D to be true. The agent takes action a_1, as it gives her a utility of 2, whereas a_0 would give her only 1. Intuitively, we would want $C = \mathit{true}$ to have motivated her action and, indeed, it satisfies PB, CA, and Des. Unfortunately, though, it is not hard to check that $C = \mathit{true} \wedge D = \mathit{true}$ also satisfies these conditions. Yet we do not want $D = \mathit{true}$ to be part of the intent. So we require that φ be *conjunct-minimal*, that is, there does not a exist a formula φ' that results from removing some conjuncts from φ and also satisfies the first three conditions. Formally,

Rel. There does not exist a formula φ' such that φ extends φ' with extra conjuncts (i.e., φ is logically equivalent to a formula of the form $\varphi' \wedge \varphi''$, where φ'' has at least one conjunct not in φ') and φ' satisfies PB, CA, and Des.

3.2 Intent

So what distinguishes motivation from intention? We propose an additional condition that we believe is necessary for intention beyond PB, CA, Des, and Rel.

Likely Occurrence: As HK already noted, it seems unusual to say that an agent intended an outcome if that outcome was unlikely to occur. For example, someone who buys a lottery ticket may be motivated by the possibility of winning the jackpot. While it seems reasonable to say "they hope to win the jackpot", it seems odd to say "they intended to win the jackpot". Similarly, when a surgeon performs an operation that they believe has low likelihood of success, it seems more natural (at least in our dialects of English) to say that they acted on the hope of saving the patient rather than with the intent of doing so. This additional condition says that there is some threshold δ (that we would expect to be context-dependent, and might depend on, say, the situation, the action, and/or the outcome) such that the probability of φ happening if a is performed is at least δ:

LO. $\Pr([[[A \leftarrow a](\varphi)]]_{\mathcal{K}}) \geq \delta$.

Our definition of intent is the same as that given earlier for motivation (Definition 3.1), except that we add the extra condition LO to PR, CA, Des, and Rel.

3.3 Parts of Intents

In natural language, it often seems acceptable to say that an agent intended a part of what we have defined as intent. For example, if an agent plants a bomb because she wants to kill both Alice and Bob, we are often comfortable saying that she intended to kill Alice (or intended to kill Bob), not just that she intended to kill both of them. But with our definition, it is *not* necessarily the case that she intended to just kill Alice (or Bob): just killing one might not give enough expected utility to exceed that of the alternative actions. To capture this natural-language usage, we define a *part of an intent*:

Definition 3.3. Given an epistemic state $\mathcal{E}$, action a, and formula φ, we say that φ *was part of an intent in taking action a in epistemic state* $\mathcal{E}$ if there exists a φ' that extends φ such that (i) φ' is an intent in taking action a and (ii) there exists an $(M, \vec{u}, a')$ that PB-certifies both φ and φ' for a.

Part (ii) of this definition ensures that the agent's action has the possibility of bringing about φ, and does so in the same circumstances as it might affect the "full intent."

As we noted, it seems common to refer to these partial intents simply as intents in natural language. A corresponding notion can be defined for partial motivations. (As Halpern [15] noted, there is also an analogous phenomenon with causality, where we refer to parts of conjunctive causes as simply "causes" in natural language.)

3.4 Example

We now consider how our definition can be applied to a legal example. The example highlights the role of modeling.

Suppose that Defendant Smith ran over a pedestrian, Jones, crossing a crosswalk. The question of whether this is involuntary manslaughter, second-degree murder, or first-degree murder depends on Smith's intentions. First-degree murder requires premeditation and a deliberate intent to kill; second-degree murder involves an intentional killing without premeditation; finally, involuntary manslaughter is an unintentional killing that results from negligence or recklessness, not an intent to kill. In our framework, these distinctions can be captured by the choice of causal model and utility function.

Consider three versions of the story. In version 1, suppose that Smith knew that Jones crossed that particular crosswalk every Monday at noon. He left home at 11:45 and carefully planned his route so that he could be at the crosswalk at noon; in particular, he slowed down before reaching the crosswalk so that he wouldn't get there too early. He then deliberately ran over Jones. In version 2, Smith left home at 11:50 happened to arrive at the crosswalk just before noon. He noticed that Jones, his sworn enemy, was crossing. In a fit of rage, he ran over Jones. Finally, in version 3, Smith left home at 11:50, got distracted while driving, and accidentally ran over Jones. Clearly, in these three versions of the story, we have first-degree murder, second-degree murder, and involuntary manslaughter, respectively.

To model these versions of the story, we consider a causal model with variables DT (for Smith's departure time, which is either 11:45 or 11:50), S (for driving speed), which is either slow or normal, AI (or arrival time at intersection), which is either 11:55 or 12:00, D (for distracted), which is either 0 or 1, SDJ (for Smith detests Jones), which is either 0 or 1, RO (for Smith runs over Jones), which is either 0 or 1, and JD (for Jones dies), which is either 0 or 1. The equations should be obvious. For example, if Smith leaves at 11:45 and drives normally, he arrives at the intersection at 11:55; if he leaves at 11:45 and drives slowly or leave at 11:50 and drives normally, he arrives at noon. If he is at the intersection at noon and either detests Jones or is distracted, then he runs over Jones; otherwise he does not. Jones dies if Smith runs over Jones. The utility function for versions 1 and 2 is such that the utility is high if Jones dies and low otherwise; for version 3, it is just the opposite. The action variable (i.e., the variable A in the definitions) is RO; the outcome of interest is $JD = 1$.

Now it is easy to check that Smith ran over Jones with the intent for Jones to die in versions 1 and 2 (i.e., using the utility function for versions 1 and 2), but not in version 3:

- PB clearly holds for all version of the story;
- CA holds for (the utility function of) versions 1 and 2, but not for version 3;
- similarly, Des holds for versions 1 and 2, but not for version 3;
- Rel holds trivially (for all versions of the story);
- since we have taken the outcome to be (deterministically) determined by the action, LO trivially holds (for all three versions).

Of course, nothing in our formalism tells us which is the appropriate utility function to take; nor does the formalism tell us whether there was premeditation. While we can capture premeditation using the values of variables in the model, investigative work will be required to determine these values. What our formalism does do is provide a common framework to discuss issues relevant to intent; while people may disagree as to whether there was intent, by using our framework, they can at least agree on what it would take to determine intent, rather than talking past each other.

4 INTENTION: THE DISJUNCTIVE CASE

In the previous section we considered what are perhaps the most common form of intents, intents that are conjunctions of primitives. Not all intents are of this form though; there are many examples where what is intended is most naturally thought of as a disjunction. For example, consider a scenario where a domestic terrorist wants to make a statement and so intends to kill any one (or more) senator(s); this seems most naturally captured by saying that what the terrorist intends is the disjunction over each senator dying. Or consider a cyber-criminal who connects to a device in the hopes of either being able to find private information on it or being able to hold access to the device for ransom (and would have been willing to illegally connect for either outcome). We expect disjunctive intents to be particularly prevalent when combined with beliefs. For example, a human agent might believe that the robot Robbie intends to either get coffee for Alice or help Ann with her project. (Note that this is different from intending to get coffee for Alice or intending to help Ann; intention does not necessarily distribute over disjunction, because of condition LO.)

In this subsection we consider formulas of the form

$$\varphi = \bigvee_{1\leq j\leq M} \left(\bigwedge_{1\leq i\leq N_j} Y_{j,i} = y_{j,i} \right)$$

as objects of intent. Our basic approach is to require that all of our earlier conditions apply to each of the disjuncts of φ; each disjunct on its own ought to be good enough to get the agent better utility than the alternatives, ought to be desirable, etc. We can define corresponding conditions to those of the previous section for these cases.

PB$_\vee$. For each disjunct of φ, PB holds.

CA$_\vee$. For each disjunct of φ, CA holds.

Des$_\vee$. For each disjunct of φ, Des holds.

The relevance condition requires a little bit more work. It is helpful to first think about the purely disjunctive case, that is, formulas φ of the form $\bigvee_{1\leq j\leq M} Y_j = y_j$. Whereas in the conjunctive case we had to worry about irrelevant primitives being included, in the disjunctive case we have to worry about relevant primitives being left out. For example, consider a scenario where there are three Boolean variables, F, G, and H, in addition to the action variable A. The agent gets utility 5 from F being true, and 10 from either G or H being true; there is no additional utility from both G and H being true. Action a_0 deterministically causes F to be true and both G and H to be false; a_1, on the other hand, causes F to be false and $G \land \neg H$ and $\neg G \land H$ each to be true with probability .5. The agent takes action a_1, getting a utility of 10, since the alternative would give only a utility of 5. Intuitively, we want $G \lor H$ to have motivated the action, and indeed it will satisfy PB$_\vee$, CA$_\vee$, and Des$_\vee$. Unfortunately, so does G by itself.

For purely disjunctive formulas, we want φ to be *disjunct maximal*, that is, there is no primitive event $Y' = y'$ that is not the same as any disjunct of φ such that $\varphi \lor Y' = y'$ satisfies PB$_\vee$, CA$_\vee$, and Des$_\vee$. This ensures that every relevant possibility that motivated the agent is included. While the minimality condition for the conjunctive case and the maximality condition for the disjunctive case may seem different, they are really enforcing the same requirement; they both essentially ensure that we select a formula that is maximal in terms of the causal settings where it is true, that is, the least restrictive formula. For our more general case, then, the requirement is as follows:

Rel$_\vee$. For each disjunct of φ, Rel holds. Moreover, every conjunctive formula that satisfies PB, CA, Des, and Rel is equivalent to some disjunct of φ.

We go from motivation to intent using the exact same condition LO as in the previous section. It is worth noting that LO is the only condition that we apply to the full formula rather than the individual disjuncts separately. A simple example illustrates why. Consider a raffle where if the agent buys a ticket she is nearly guaranteed to win a prize, but because there are so many possible prizes there is no one prize that she is likely to win; we still want to say she intended the disjunction of winning each of those possible prizes, which was a very likely outcome. Note that this is different from condition PB, where we do want to apply it disjunct by disjunct; if the agent cannot effect a particular disjunct then it really is not part of the story of what motivated her action.

Finally, it's worth considering whether a similar "part of an intent" notion as in the previous section applies to these disjunctive cases. Unfortunately, here the story seems quite complicated. It might be tempting to apply our previous notion in a disjunct-by-disjunct manner, and so say that if she intended $(W = w \land X = x) \lor (Y = y \land Z = z)$ then $W = w \lor Y = y$ is part of her intent. But this doesn't always correspond well to natural-language usage. Imagine a scenario where a terrorist committed an act with the intent to either kill a specific senator and disable the camera filming him or kill a specific congressman and disable the camera. We might well feel comfortable saying that she intended to kill the senator or the congressman. But we would likely not feel comfortable with the seemingly symmetric case of saying she intended to either kill the senator or disable the congressman's camera. Natural-language usage seems to require that in a disjunctive intent, all disjuncts be of the same "type". We believe that a richer framework, for example, one that can talk about types, would be needed to capture this phenomenon. In light of this, we leave the notion of parts of intents for disjunctive formulas to future work.

So far we have dealt with formulas in what we call DNF$^+$, disjunctive normal form formulas with no negated primitive events. It would seem that dealing with any arbitrary formula φ should then be straightforward: we simply convert φ to DNF$^+$ (which is always possible) and apply the approach above. The problem with this is that our approach is not purely semantic; the syntax of a formula matters. In particular, we can have two logically equivalent formulas in DNF$^+$ where one is intended and the other is not. Consider, for example, a formula $X = 1$ that the agent intends. This is logically equivalent to $(X = 1 \land Y = 0) \lor (X = 1 \land Y = 1)$ if Y is a Boolean variable, but this second formula will not be intended since, for example, $X = 1 \land Y = 1$ will not satisfy Rel (since the agent intends $X = 1$), and therefore the formula will not satisfy Rel$_\vee$. It in fact seems to us that there are times when people would say they intended $X = 1$ but not feel comfortable saying they intended $(X = 1 \land Y = 0) \lor (X = 1 \land Y = 1)$, so the syntactic dependence does not seem so unreasonable. But then conversion to DNF$^+$ does not simply resolve the problem for all formulas. We leave to future work the question of whether there is a better way of dealing with arbitrary formulas and, in particular, syntactic dependence.

5 RELATED WORK

It would be impossible for us to provide a thorough survey and analysis of the extensive literature across computer science, philosophy, and law on the topic of intent. Instead, we will focus on a few points about how the current work relates to three lines of work that seem particularly relevant.

First, we highlight the connection between our work and some influential prior work in the philosophy of law. R.A. Duff, in *Intention, Agency and Criminal Liability* [11], put forth the following definition:

> *We now have an account of what it is to act with the intention of bringing about a specified result, and to succeed in doing so:*
>
> A. *The agent wants (or desires) that result.*
> B. *She believes that what she does might bring that result about.*
> C. *She acts as she does because of that want and that belief.*
> D. *What she does causes that result.*

Our formal definition is similar in spirit (by design!) to that of Duff. Indeed, Duff's conditions inspired the naming of our conditions; we wanted to emphasize what seem to us like conceptual similarities. Roughly, our condition PB corresponds to Duff's B (it is arguably slightly stronger, in that Duff's B can be viewed as just saying that the agent considers it possible that what she does is sufficient to bring about the result while, as we observed, PB involves a necessity condition along with sufficiency), our CA to his C (with the caveat that, as we observed, while CA says that φ is desirable, it does not quite say that φ is what motivated the agent to do a), and our Des to his A. Our condition LO is our replacement of Duff's condition D: condition D says that that a causes the outcome φ, while LO says that performing a brings about φ with high probability. If we conflate "causes" with "brings about" (which is not quite right, but is not a bad gloss), then LO is a weakening of (4). We would argue that LO is actually closer to the way people (and the law) think of intent than Duff's condition D: We would say that Alice intended Bob to die if she planted a bomb that she believed would kill Bob, but it failed to go off. But Alice's action certainly did not cause Bob to die since, in fact, he did not die. The law does care about actual causality: Alice would be charged with murder if her bomb caused Bob's death; if the bomb fails to go off, she can be charged only with attempted murder. But Alice can still intend an outcome even if her action does not in fact cause it (or if the agent attempts the action but fails).

Note that Duff's definition does not have an analogue of our condition Rel. As we argued above, something along the lines of Rel seems necessary to avoid incorrect inferences in even some fairly simple cases. Despite that difference, the similarity between the definitions makes us optimistic that our definition is applicable to the legal domain; in future work, we hope to explore connections between our approach and various legal approaches to intent.

Arguably the most influential work on intent in the CS community is the groundbreaking work of Cohen and Levesque [9] (CL from now on) and Rao and Georgeff [24]. Drawing on ideas from Bratman's highly influential philosophical work on intention (see, e.g., [5]), each developed modal logics for reasoning about intent. These works were interested in how agents can use intents to make plans and drew on ideas from the philosophy of action; by way of contrast, our work is focused on retrospectively determining intents and connects to ideas from philosophy of law.

For example, CL's definition centers around *commitment*; roughly speaking, an agent intends an outcome φ if she chooses to commit to bringing it about. CL's notion of intent differs in two significant ways from ours. The first is that their notion describes a property of an agent ("she intends to bring about φ"), whereas ours describes a property of an agent and her action ("her intent in doing a was to bring about φ"). The second is that, following Bratman, CL take the philosophical stance that intention is irreducible to beliefs and desires, but rather is a cognitive state that guides resource-bounded agents in planning for the future. Clearly, we are not taking that stance; our goal is to determine intent using the desires and beliefs of an agent.

Importantly, for us, an agent's beliefs include beliefs about the causal structure of the world, which is not the case for CL. As we have emphasized, the causal structure plays a particularly significant role in our approach. It may also be useful for fixing a flaw in CL's definition. They "require the agent to think he is about to do something *bringing about* [φ]" (emphasis changed). This statement is causal in nature ("bringing about φ"), but CL require only that the agent believes he will take an action a, this action will be immediately followed by φ being true, and that the agent has agency over a occurring.[3] Indeed, CL's formalism cannot express counterfactual or causal statements. But consider, for example, an extremely punctual baker. In his village, the bell tower clock chimes at exactly 7:00 every morning; he is committed to opening the doors of his bakery right before it does. It seems quite odd to say that the baker intended for the clock to chime, given that he had no control over it chiming. But it is the case that he believes he will take an action (opening his door), this action will be immediately followed by the clock chiming, and that he has agency over this occurring (if he does nothing then the event "door opens followed by clock chiming" will not happen, but if he does act, it will). Moreover, before he opens his door the bell does not chime, but after he opens his door, the bell chimes. Thus, according to the CL definition, opening the door brings about the bell chiming. By the same token, just before opening the door, according to the CL definition, the baker thinks he is about to do something (namely, opening the door) that will bring about the bell chiming.

What is missing in the CL definition is the causal connection: opening the door does not bring about the clock chiming. Of course, we could try to deal with this by extending CL to allow counterfactual reasoning. However, even with the ability to capture counterfactual reasoning, it is still highly nontrivial to capture causality (see [15] for extensive discussion of this issue); as we hope this paper shows, it is also quite nontrivial to capture intention.

Intention has also been formalized using a STIT ("seeing to it that") formalism. For example, Broersen [7] also tries to capture the notion of intention as used in the law. The American Model Penal Code [2] divides criminal intent into four states of mind listed in decreasing order of culpability: purposely, knowingly, recklessly, and negligently. Broersen focuses on formalizing the first two (and, more generally, intention), using the logic XSTIT [6], extended with modal operators for knowledge and intentional action. The knowledge operators are quite standard (and, in particular, satisfy the standard axioms for knowledge in the logic S5; see [12]). But the intention operators do not capture what (to us, at least) seems like a key property of intention: causality.

[3] More precisely, CL [9, p. 228] define what it means for a to bring about p (in their notation) as $(\text{HAPPENS} \neg p?; a; p?)$. Thus, a brings about p at a certain point if p doesn't hold but, after performing a, it does.

The law certainly recognizes the role of causality. For example, according to the Model Penal Code, a "person acts purposely with respect to a material element of an offense when: (i) if the element involves the nature of his conduct or a result thereof, it is his conscious object to engage in conduct of that nature or to cause such a result ... " [2, Section 2.02(2)(a)]. Our use of causal models to define intent lets us capture this causal connection. The knowingly condition also involves causality; indeed, the Model Penal Code describes knowingly as follows: "A person acts knowingly with respect to a material element of an offense when ... he is aware that it is practically certain that his conduct will cause such a result" [2, Section 2.02(2)(b)].[4] The lack of a direct way to capture causality is perhaps the key difference between our approach and the stit approach more generally (see, e.g., [1, 21]), as well as what distinguishes our approach from that of CL and from formalizations of intention from what has been called the database perspective [26, 27]. As we said above in the context of CL, while it is certainly possible to add counterfactual reasoning to all these approaches, we believe that it would still be nontrivial to use them to capture causality and intention.

Finally, the work most similar in spirit to ours is that of Kleiman-Weiner et al. [19] and the subsequent work of HK. Both papers define intent using a counterfactual condition somewhat in the spirit of CA, with HK using causal models and epistemic states of the type we used here. We highlight the key differences between our work and the earlier work, with a focus on HK (although essentially the same problems arise in [19]).

HK consider only conjunctive intents. As we have argued, there seem to be natural cases where intents are best described as disjunctions of particular outcomes. There is no obvious way to extend the HK definition to the disjunctive case. Our approach, on the other hand, allows us to go from the conjunctive case to the disjunctive case in a natural way.

A second major difference is perhaps most easily demonstrated using an example introduced by HK, where a surgeon performs an operation that she believes has only a 20% chance of saving her patient, but is also the only hope. The outcome of the patient living satisfies HK's main counterfactual condition, but unfortunately so does the outcome of the patient dying; roughly speaking, this is because if the patient was guaranteed to die, then it wouldn't matter which action the surgeon took. To deal with this problem, HK use their counterfactual condition just to identify which variables the agent intended to affect, and then declare that the values the agent intended for those variables were whichever ones maximized utility and occurred with positive probability. But this approach does not always give reasonable results. Consider a scenario where an agent can choose between two actions, a_0 and a_1. If she takes a_0 she deterministically gets $1,000. If she takes a_1, though, she is entered into a lottery where she'll get $1,500 with probability .9, $500,000 with probability .001, and nothing with probability .099. The agent takes action a_1, which gives higher expected utility. According to HK's definition, the agent's intent in taking a_1 was to win $500,000, which doesn't seem right; clearly the $1,500 outcome is also playing a major role in motivating her decision.

Our definition gets the surgery example right for arguably the right reason: the surgeon was motivated by the possibility of the patient living because that was better than the expected outcome of her not performing surgery, which is that the patient dies. The outcome of the patient dying after surgery, on the other hand, would not satisfy this condition, and so did not motivate the surgery. And in our previous example with the lottery, since we allow disjunctive intents, our definition can capture the role of the possibility of winning $1,500 in characterizing the agent's motivation.

A third difference between the definitions is most easily seen by considering Example 3.2. In that case, HK's definition will in fact give that $C = 1$ was intended. More generally, it doesn't seem to properly handle outcomes that would be helpful under an alternative action but are not under the action actually taken. Our clause Des is designed to make sure we handle those cases properly.

Finally, to get their definition to give reasonable answers in some examples, HK need to assume that there is a special *reference set* of actions (a subset of the set of actions), which are the only ones considered. The problem (as HK themselves acknowledge) is that there is no obvious principled approach to choosing the reference set. Our definition can handle these examples without the need to appeal to a reference set.

Ward et al. [28] give a definition of intent based on that of HK, and show, among other things, that their definition can be used to infer the intentions of reinforcement learning agents and language model from their behavior. While the definition of Ward et al. is, perhaps not surprisingly, similar in spirit to ours, like that of HK, it appeals to reference sets and does not handle the disjunctive case.

It is also worth briefly mentioning the work of Ashton [4], which was done independently of our work and is also motivated by a desire to formalize intent. One strength of that paper is that care is taken to delineate different types of intent considered in British criminal law and to try to determine what distinguishes each of them. That said, the definitions proposed are only what Ashton calls "semi-formal", in ways that seem to leave open considerable ambiguity as to what they really mean and how they should be applied. Because of these ambiguities, it is difficult to compare Ashton's definitions to ours.

6 CONCLUSION AND FUTURE WORK

In this work we have provided a new formalization of what we might call *intent as motivation for action.* We believe that our formalization should prove relevant for areas like human-robot interaction, as well as to legal scholars. One advantage of our definition is that it is highly modular. While we believe that we have done a reasonable job of formalizing each of our five conditions for intent, if further research suggests a better way of defining any of them in certain domains, it should be easy to swap any subset of them out.

We have focused on "all-or-nothing" notions of motivation and intent; either the agent was motivated by/intended φ in performing action a, or she wasn't/didn't. We believe that an important direction for future work involves getting more quantitative notions of motivation and intent. After all, people do say in natural language "φ motivated

[4] While we have not directly attempted to capture the notions of "purposely" and "knowingly", our models are rich enough to do so. For example, we could capture the "practically certain" requirement of knowingly using probability, in the spirit of our condition LO.

me to some extent". We sketch some preliminary attempts at doing this that we hope to expand on in future work, to give a sense of what can be done and why the problem is nontrivial.

We start with *degrees of intention*. Consider a scenario where an agent can choose between two lotteries, the first of which will give her a utility of 1000 for sure, and the second of which will give her 900 with probability 0.5 and 1200 with probability 0.5. While it seems like the 900 outcome actually plays a role in the agent's decision to choose the second lottery (although perhaps to a lesser extent than the 1200 outcome), our definition in Section 3 will only give 1200 as motivating the agent's action. There is a straightforward modification to our definitions that lets us capture the intuition that the 900 outcome motivated the agent but to a lesser extent than the 1200 outcome: we simply extend CA to a condition β-CA, where $\beta \in [0,1]$, and require that for all actions $a' \in \mathcal{R}(A)$,

$$\mathrm{EU}_{\mathcal{E}}[\vec{Y} \leftarrow \vec{y}; A \leftarrow a'] \geq \beta \max_{a'' \neq a \in \mathcal{R}(A)} \mathrm{EU}_{\mathcal{E}}[A \leftarrow a''].$$

Clearly, 900 satisfies .9-CA (so it might make sense to say that 900 motivated the agent to degree .9), but not 1-CA. But this definition gets at only one aspect of motivating to a certain degree. For example, if in the second lottery the agent received 900 with probability .01 and 1200 with probability .99, it doesn't seem reasonable to say that 900 motivated the agent to degree .9. There seem to be many different factors that affect how we judge the relative extent to which two possible outcomes motivate an agent to take an action. For example, the probability of the outcome, the utility of the outcome, and how the outcome compares to alternative outcomes all play a role. We need to take all of these into account when defining degree of intention.

Turning to probabilistic intent, consider another scenario where an agent can choose between two lotteries. This time, in the first one (which corresponds to taking action a_0) she gets \$1,000 for sure, while in the second one (which corresponds to taking action a_1) she gets \$1,000 with probability .9, \$500,000 with probability .001, and nothing with probability .099. The agent takes action a_1. Intuitively, what motivated her was that, with high probability, she did at least as well as with a_0, but there was, in addition, a small probability of getting \$500,000. Just saying that she was motivated by getting \$1,000 or \$500,000 (which is the case, according to our definitions above) misses out on the probabilistic aspects.

This example suggests that we might want to allow objects of intent that have the form $\varphi = \alpha_1 : \psi_1, \ldots, \alpha_M : \psi_M$, where each ψ_j is a conjunctive formula of the type we considered in Section 3 and the ψs are disjoint events (more precisely, $[[\psi_i]]$, $i = 1, \ldots, M$, are disjoint sets). We can think of these as formulas in a richer language. Intuitively, an agent intends such a formula if she intends ψ_j with probability α_j. So in the previous example, the relevant formula is $0.9 : D = 1,000; 0.001 : D = 500,000$, where D is the variable that describes how much the agent gets.

While we can extend our definitions to deal with probabilistic intents in a number of ways, we have yet to find one that is problem-free. Combining our extensions with ideas about degrees of intension seems to help to some extent. We hope to explore these issues in more detail.

A related issue is that, in our model, we assume that the agent has a single utility function. But agents are often driven by a number of (possibly conflicting) utility functions they want to maximize. Arguably, at the time an action is taken, there is one primary utility function that is driving the action; we can take our definition of intent as considering that utility function. But we could take a somewhat broader view. When it comes to motivation, we can imagine a more nuanced balancing act between utility functions; it would be of interest to try to extend our framework to handle multiple utility functions. We could similarly imagine that an agent's uncertainty is characterized, not by a single probability measure, but by a set of them. There have been a number of proposals for making decisions when uncertainty is represented by a set of probability measures (see [13] for an overview); these could be extended to deal with multiple utility functions as well. We believe that we should be able to modify our definitions in a natural way to deal with these approaches to decision making, although we have not checked details.

Besides exploring more quantitative notions of intent, we would also like to investigate connections between the definitions of this paper and legal theories of intent, with the goal of clarifying points that are of legal relevance.

Finally, while in this paper we have focused on how to define intent, we think it is critical to explore algorithmic questions as to whether intent can be determined or at least approximated under reasonable conditions, as well as whether there are useful axiomatic characterizations of intent.

While computing causality is D_1^P-complete in general [14, 15], here (in PB and Des) we consider only but-for causality ($A = a$ is a cause of φ if φ would not have occurred had A had a different value), which is much simpler to verify. But checking Rel seems to be computationally difficult, although we have not verified this. The difficulty of checking CA, Des, and LO depends in part on how the causal model is presented. For example, if there are n binary variables in the model, there may be exponentially many possible outcomes. If we describe their utility explicitly as part of the description of the model, then the size of the model is exponential in the number of variables, and checking CA, Des, and LO (and Rel, for that matter) should be polynomial in the szie of the model. On the other hand, if there is a more compact description of the utility function, then the complexity of checking CA, Des, and LO could also be high in the number of variables. That said, this complexity should not be a problem in cases where there are few variables, which come up often in practice (e.g., in the law). It would also be worth investigating whether there are important special cases for which we can get good complexity results even for settings with a large number of variables.

Turning to logical aspects, one advantage of both CL and the stit approach is that they start with logical languages and consider axioms, in some cases providing complete axiomatizations. It should certainly be possible to do something analogous in the case of intent. It would be of particular interest to get an axiomatic characterization of our notion of intent.

ACKNOWLEDGMENTS

The work of Meir Friedenberg and Joe Halpern was supported in part by NSF grants IIS-1703846 and IIS-1718108, ARO grant W911NF-22-1-0061, MURI grant W911NF-19-1-0217, and a grant from the Cooperative AI Foundation.

REFERENCES

[1] A. I. R. Abarca and J. M. Broersen. 2023. A stit logic of intentionality. In *4th International Workshop on Dynamic Logic: New Trends and Applications (DaLi22)*. 125–153.

[2] American Law Institute. 1984. *Model Penal Code, annotated.* American Law Institute Publishers.

[3] G. E. M. Anscombe. 1957. *Intention.* Basil Blackwell, Oxford, U.K.

[4] H. Ashton. 2022. Definitions of intent suitable for algorithms. *Artificial Intelligence and Law* (2022), 1–32.

[5] M. Bratman. 1987. *Intention, Plans, and Practical Reason.* Harvard University Press, Cambridge, MA.

[6] J. M. Broersen. 2011. Deontic epistemic stit logic distinguishing modes of *mens rea. Journal of Applied Logic* 9, 2 (2011), 127–152.

[7] J. M. Broersen. 2011. Making a start with the stit logic analysis of intentional action. *Journal of Philosophical Logic* 40, 4 (2011), 499–530.

[8] R. M. Chisholm. 1966. Freedom and action. In *Freedom and Determinism*, K. Lehrer (Ed.). Random House, New York, NY.

[9] P. R. Cohen and H. J. Levesque. 1990. Intention is choice with commitment. *Artificial Intelligence* 42, 2–3 (1990), 213–261.

[10] A. D. Dragan, K. C. Lee, and S. S. Srinivasa. 2013. Legibility and predictability of robot motion. In *8th ACM/IEEE International Conference on Human-Robot Interaction.* 301–308.

[11] R. A. Duff. 1990. *Intention, Agency and Criminal Liability: Philosophy of Action and the Criminal Law.* Blackwell.

[12] R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. 1995. *Reasoning About Knowledge.* MIT Press, Cambridge, MA. A slightly revised paperback version was published in 2003.

[13] J. Y. Halpern. 2003. *Reasoning About Uncertainty.* MIT Press, Cambridge, MA. A second edition was published in 2017.

[14] J. Y. Halpern. 2015. A modification of the Halpern-Pearl definition of causality. In *Proc. 24th International Joint Conference on Artificial Intelligence (IJCAI 2015).* 3022–3033.

[15] J. Y. Halpern. 2016. *Actual Causality.* MIT Press, Cambridge, MA.

[16] J. Y. Halpern and M. Kleiman-Weiner. 2018. Towards formal definitions of blameworthiness, intention, and moral responsibility. In *Proc. Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18).* 1853–1860.

[17] J. Y. Halpern and J. Pearl. 2005. Causes and explanations: a structural-model approach. Part I: causes. *British Journal for Philosophy of Science* 56, 4 (2005), 843–887.

[18] D. Kahneman and A. Tversky. 1979. Prospect theory, an analysis of decision under risk. *Econometrica* 47, 2 (1979), 263–292. https://doi.org/10.2307/1914185

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

[20] K. Konolige and M.E. Pollack. 1993. A representationalist theory of intention. In *Proc. Thirteenth International Joint Conference on Artificial Intelligence (IJCAI '93).* 390–395.

[21] E. Lorini and A. Herzig. 2008. A logic of intention and attempt. *Synthese* 163, 1 (2008), 45–77.

[22] J.-J.Ch. Meyer, W. van der Hoek, and B. van Linder. 1999. A logical approach to the dynamics of commitments. *Artificial Intelligence* 113, 1-2 (1999), 1–40.

[23] J. Niehans. 1948. Zur preisbildung bei ungewissen erwartungen. *Schweizerische Zeitschrift für Volkswirtschaft und Statistik* 84, 5 (1948), 433–456.

[24] A.S. Rao and M.P. Georgeff. 1991. Modeling rational agents within a BDI-architecture.. In *Principles of Knowledge Representation and Reasoning: Proc. Second International Conference (KR '91).* 473–484.

[25] L. J. Savage. 1951. The theory of statistical decision. *J. Amer. Statist. Assoc.* 46 (1951), 55–67.

[26] Y. Shoham. 2009. Logical theories of intention and the database perspective. *Journal of Philosophical Logic* 38, 6 (2009), 633–648.

[27] M. van Zee, D. Doder, L. van der Torre, M. Dastani, T. Icard, and E. Pacuit. 2020. Intention as commitment toward time. *Artificial Intelligence* 283 (2020).

[28] F. R. Ward, M. MacDermott, F. Belarinelli, F. Toni, and T. Everitt. 2024. The reasons that agents act: intention and instrumental goals. In *Proc. Twenty-Third International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024).* 1901–1909.

[29] G. Williams. 1987. Oblique intention. *The Cambridge Law Journal* 46, 3 (1987), 417–438.