Deontic Epistemic *stit* Logic Distinguishing Modes of Mens Rea

Jan Broersen

March 4, 2010

Abstract

Most juridical systems contain the principle that an act is only unlawful if the agent conducting the act has a 'guilty mind' ('mens rea'). Different law systems distinguish different modes of mens rea. For instance, American law distinguishes between 'knowingly' performing a criminal act, 'recklessness', 'strict liability', etc. I will show we can formalize several of these categories. The formalism I use is a complete *stit*-logic featuring operators for *stit*-actions taking effect in 'next' states, S5-knowledge operators and SDL-type obligation operators. The different modes of 'mens rea' correspond to the violation conditions of different types of obligation definable in the logic.

1 Introduction

An important distinction in law is the one between 'actus reus', which translates to 'guilty act', and 'mens rea' for 'guilty mind'. It is a general principle of law that both these conditions should be met for an act to qualify as criminal, that is, guilt not only presupposes a forbidden act as such, also, the performing agent must have committed the act knowingly, intentionally, purposely, etc.[1]. The task of showing that *both* necessary conditions 'actus reus' and 'mens rea' apply to an alleged criminal act, is in law referred to as 'showing concurrence'.

There are different levels of mens rea, each corresponding to different levels of culpability. And, of course, different law systems have different categories. The current North American system works with the following modes, in decreasing order of culpability (as taken from [20]):

- **Purposefully** - the actor has the "conscious object" of engaging in conduct and believes and hopes that the attendant circumstances exist.
- **Knowingly** - the actor is certain that his conduct will lead to the result.

[1] The general principle was already formulated back in 1797, by the English jurist Edward Coke: "actus non facit reum nisi mens sit rea", which is Latin for "an act does not make somebody guilty unless his/her mind is also guilty"

- **Recklessly** - the actor is aware that the attendant circumstances exist, but nevertheless engages in the conduct that a "law-abiding person" would have refrained from.
- **Negligently** - the actor is unaware of the attendant circumstances and the consequences of his conduct, but a "reasonable person" would have been aware.
- **Strict liability** - the actor engaged in conduct and his mental state is irrelevant.

The first class, the one of acts committed *purposefully*, is about acts that are instrumental in reaching an agent's malicious *goal*. The second class is not directly about an agent's intentions, aims or goals, but only about the condition whether or not an agent knows what it is doing. The third class is a little less clear. I think it is defendable to interpret it as the category of acts where an agent knowingly risks an unlawful outcome. For the fourth category, not knowing the (possible, or necessary - that is not made explicit) outcomes is not an excuse: if the agent did not know, it simply should have known. The final category concerns the complete absence of 'mens rea'. This is the category where agents can be culpable without having a 'guilty mind' whatsoever.

I claim the levels of culpability correspond to (1) levels of *excusability* and (2) levels of *deontic strength*[2]. For the first class, the deontic strength is lowest of all and several excuses apply. In particular, for this class an 'actus reus' can be accompanied by the valid excuses: "I did not have bad intentions", "I did not know what I was doing", and so on. For the second category, deontic strength is higher, and fewer excuses apply. In particular, the excuse that there were no bad intentions is no longer acceptable. What counts is that the agent knew what it was doing, irrespective of the goal the act was aimed at. For the third category, where the deontic strength is yet higher, it is not even an excuse that the agent was not sure about the outcome: the agent is liable simply because it took a *risk* that led to an unlawful outcome. In the fourth category, the excuse that the agent simply did not realize the consequences of his act, is no longer valid: for violations of any prohibition in this category it is still liable, because any 'reasonable' agent would have foreseen the consequences. And finally, for the strict liability category, deontic strength is highest of all, and no excuses referring to the mental state of an agent apply at all[3].

In philosophy, the idea that excuses play an important role in distinguishing different modes of acting was put forward by Austin [5]. And many other kinds of excuses than the ones above are thinkable. For instance, among the most

[2] I am not aware of any law or philosophical literature where this triple correspondence has been observed before, but I do not doubt there is.

[3] An additional observation is that for more serious crimes the distinction between the mens rea modes is more relevant than for less serious crimes. If you walk through a red traffic light, the police officer will not take you seriously when you claim you are excused because you did not do it knowingly (you are strictly liable). But, if we consider a case where your way of conduct resulted in some person's death such an excuse is certainly going to be considered.

well-known excuses for violating an obligation are: "I was not *able* to", "I do not agree my act *counts-as* a violation", "I obeyed a stronger, *conflicting* obligation" and "I did not *know* I had to". Of these, in this paper, I will only consider the first and the last one. The first one, about not being able to comply to the obligation, is only a valid excuse if the principle of "ought implies can" applies. The last one, concerning knowledge of the condition that the act is obliged, refers directly to the juridical principle "ignorantia juris non excusat", which translates to "ignorance of or mistake about the law is no defence". So, here the (absence of) excuse is not so much about the mode of acting, as in the modes of mens rea above, but about whether or not the agent knows about the 'deontic status' of the act. This maybe a subtle different with the described modes of mens rea and is not made very clear in the juridical literature. But, in our formalizations it will be.

We will also look at how we can formally define what counts as an 'actus reus'. Also for this, the juridical literature gives exact definitions. In particular, an actus reus cannot be an *involuntary* act. For instance, a person being thrown off a high building, surviving his fall by crashing into another person, who gets killed as the result of functioning as a cushion, has not committed an actus reus, even though the falling person knew that it actually was crashing into the person. The current American Model Penal Code [20] lists what acts count as involuntary acts for which no agent can be liable.

- a reflex or convulsion
- a bodily movement during unconsciousness or sleep
- conduct during hypnosis or resulting from hypnotic suggestion
- a bodily movement that otherwise is not a product of the effort or the determination of the actor, either conscious or habitual

The goal of this paper is to analyze the concepts of actus reus, and the levels of mens rea, culpability, excusability, and deontic strength by formal means. To that end, we define a formal *stit*-logic. The acronym *stit* stands for 'seeing to it that', referring to the central modality of the logic that expresses that groups of agent are responsible for a certain action effect occurring. The main goal of this paper is not to present the formal logic. However, of course, we want the formal basis to be sound, which is why we give a formal semantics and a completeness result.

We will formalize (1) the different modes of mens rea with the exception of the first category concerning purposeful acts, (2) different modes of actus reus, that is, voluntary acts (3) the condition of "ignorantia juris non excusat". The mens rea class of purposeful acts is not considered because I do not consider goals and intentions; I leave this for future research. Almost all the other categories concern conditions referring to an agent's *knowledge* about his actions. And knowledge operators will be a central concern of this paper. More specifically, we will come up with many different notions of obligation (as is common in deontic logic, we will treat obligations and prohibitions on a par, and see prohibitions as obligations to act oppositely), many of which can be associated with one of the classes of mens rea. The formal framework is also very well suited to refine and disambiguate the classes from the juridical literature.

The plan of this paper is as follows. First, in section 2 we define a *stit*-logic that forms the action logic fundament of our investigations. Then, in section 3, we show how to add an epistemic dimension to the base logic, to enable modeling of the notion of 'knowingly doing' that will be central in formalizing the modes of mens rea. Then, in section 4, we will first concentrate on how to represent an 'actus reus', without a deontic connotation. Finally, in section 5, the deontic operators are introduced. In this section we define the different types of obligation that correspond to different modes of mens rea. The final section contains a conclusion and discusses related work, future research, and some strong opinions on the implications of this work.

2 A *stit*-logic affecting 'next' states: XSTIT

In this section we define a complete *stit*-logic where actions take effect in 'next' states: XSTIT. For those unfamiliar with the *stit*-framework: the characters 'stit' are an acronym for 'seeing to it that'. *stit*-logics [8, 9] originate in philosophy, and can be described as endogenous logics of agency, that is, logics of agentive action where actions are not made explicit in the object language. To be more precise, expressions $[A \textit{ stit} : \varphi]$ of *stit*-logic stand for 'agents A see to it that φ', where φ is a (possibly) temporal formula. However, where philosophers write '$[A \textit{ stit} : \varphi]$', we prefer to write '$[A\ \mathsf{stit}]\varphi$' to denote the same notion, to be more in line with standard modal notation. The main virtue of *stit*-logics is that, unlike most (if not all) other logical formalisms, they can express that a choice or action is actually performed / taken / executed by an agent.

The logic XSTIT was first investigated in [13] and used as the basis for deontic operators in the workshop version of the present article [12]. Here we change the logic on several points. In one respect, we make it weaker by no longer defining the next operator as an abbreviation of the agency operator. But, in two respects we make it stronger: by adding an new notion of maximality, and by equating settledness in the next state with Ags-effectivity.[4]

In [15] we used the almost identical name 'X-STIT' for a quite different *stit*-logic. Still, the difference between that logic and the present one is well symbolized by the separation of the 'X' and the acronym 'STIT'. This refers to the fact that that paper's classical *instantaneous stit* logic is extended with a next operator, while in the present *stit*-variant effectivity of *stit*-operators itself refers to next states. In [15], action and time are not 'coupled': next states are

[4] There is an issue with naming logics here. A logic is the subset of valid formulas of a language. So, strictly speaking, by weakening and strengthening earlier definitions, we get another logic, and thus we should use another name. However, the earlier definition was not the *intended* one, and can, in that sense, be said to be mistaken. The present logic is the *intended* XSTIT.

not necessarily the ones brought about by agents in the system[5]. This leads to many differences with the *stit*-logic(s) in [15]. In particular, the present logic drops the axioms in [15] that are due to the instantaneous character of that paper's *stit*-operators, adds axioms that are specific for ensuring effects occur in next states, couples actions and time, and is complete. Also we use a two dimensional semantics, closer to the *stit*-semantics in the philosophical literature.

The fact that in our *stit*-logic we adopt the ontological commitment that actions only take effect in 'next' states, where 'next' refers to immediate successors of the present state, distinguishes the logic from any *stit*-logic in the (philosophical) literature. This choice has as a positive side effect that the logic is axiomatizable (and decidable). The logics of the multi-agent versions of the standard 'instantaneous' *stit*, are undecidable and not finitely axiomatizable [6, 22]. A motivation for only looking at next states comes from computer science, where this is the standard view in formal models of computation. But the main motivation is that this choice fits naturally with the example scenarios we will discuss. These scenarios are all suitably modeled using sets of subsequent choice points where the effects of choices take effect in the next choice point. Actually, I think that it is quite hard to come up with a scenario that really requires we adopt the ontological commitment that effects are instantaneous[6]. Note that we do not assume anything about how distant subsequent choice points should be; they can be arbitrarily close.

Besides the usual propositional connectives, the syntax of XSTIT comprises three modal operators. The operator $\Box\varphi$ expresses 'historical necessity', and plays the same role as the well-known path quantifiers in logics such as CTL and CTL* [21]. Another way of talking about this operator is to say that it expresses that φ is 'settled'. However, settledness does *not* necessarily mean that a property is *always* true in the future (as often thought). Settledness may, in general, apply to the condition that φ occurs 'some' time in the future, or to some other temporal property. This is reflected by the fact that settledness is interpreted as a universal quantification over the *branching* dimension of time, and *not* over the dimension of duration. The operator $[A \text{ xstit}]\varphi$ stands for 'agents A jointly see to it that φ in the next state'. The third modality is the next operator $X\varphi$. It has a standard interpretation as the transition to a next system state. Given a countable set of propositions P and a finite set Ags of agent names, formally the language can be described as:

Definition 2.1 *Given a countable set of propositions P and $p \in P$, and given a finite set Ags of agent names, and $A \subseteq Ags$, the formal language $\mathcal{L}_{XSTIT}$ is:*

$$\varphi \quad := \quad p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid [A \text{ } \mathit{xstit}]\varphi \mid X\varphi$$

In the two earlier accounts of XSTIT [13, 12] we defined the next operator

[5] we discuss the issue of this lack of 'success preservation' in the final section of [15]. In the present logic, the coupling of time and action is guaranteed by the NCUH condition/axiom.

[6] Maybe we should think in the direction of 'mental effects' of choices.

through the abbreviation $X\varphi \equiv_{def} [Ags \ \mathsf{xstit}]\varphi$. However, this has undesirable consequences[7].

Our *stit*-operator concerns, what game-theorists call, 'one-shot' actions. We can also imagine to have a *strategic stit*-operator (see [17]) where it is assumed that groups of agents have multiple subsequent choice points to ensure a certain condition (game-theorists call these 'extensive games'). Such a setting only makes sense if we increase expressivity of the temporal sub-language, and go beyond what can be expressed by the next operator alone. For instance, ensuring a condition 'some time in the future' may in general involve several choices in a row, and is not necessarily accomplished by a one-shot action. But, of course, it cannot be *excluded* that a one shot action determines a long term effect, which justifies why in the one-shot stit-logics in the philosophical literature one studies the stronger temporal operators. However, I think it is somewhat surprising that the philosophical literature does not also study the next operator.

In the description of the structures, below, we will use terminology inspired by similar terminology from Coalition Logic, and call the relations interpreting the *stit*-operator 'effectivity' relations. However, our effectivity relations are *not* just the relational equivalent of the effectivity functions of CL. Our effectivity relations are relative to histories and determine the possible outcomes modulo the history. Effectivity functions in CL are relative to a state, and yield *sets* of possible outcomes.

Before giving the formal definition of the frames, let me point briefly to the differences with 'classical' *stit*-frames, like the ones in the book of Horty [24]. In classical *stit*, as said, effects are instantaneous. To give semantics to that, in the frames the present static state in partitioned into choice sets. In the *stit* logic in this paper effects occur in next states, and thus, the choice partitioning is also with respect to next states (as should be clear from the frame visualizations in fig. 1 and fig. 2). In *stit*-logics, acting, by a group A, is identified with ensuring a condition holds on all dynamic states that may result after execution of the action (all the worlds the act is effective for). In terms of the visualization of fig. 1, the actions, for the single agent whose view on the frame is pictured, appear as ellipses grouping different possible sets of next states. In terms of the visualization of fig. 2, the actions of $Ag1$ appear as columns of the game forms, the actions of $Ag2$ appear as rows, the actions of the empty set of agents appear as the outer rectangles of the game forms, and the actions of Ags appear as the small squares inside the game forms.

After the definition of the frames, we explain the elements they are build from using the two visualizations of XSTIT-frames in fig 1. and fig 2.

Definition 2.2 *An XSTIT-frame is a tuple* $\langle S, H, R_{\Box}, \{R_A \mid A \subseteq Ags\}, R_X \rangle$ *such that:*

- *S is a non-empty set of system states. Elements of S are denoted s, s',*

[7] As a consequence $X\varphi \rightarrow X\Box\varphi$ is derivable, which, with determinism for the X, gives that the frames can only be such that the interpretation of the $\Box$ reduces to the identity relation in next states.

etc[8].

- *H is a non-empty set of system histories. System histories are sets of system states with an ordering derived from the next state relation R_X (defined below). Elements of H are denoted h, h', etc.*
- *Dynamic states are tuples $\langle s, h\rangle$, with $s \in S$ and $h \in H$ and $s \in h$.*
- *$R_\Box$ is a 'historical necessity' relation over dynamic states such that $\langle s, h\rangle R_\Box \langle s', h'\rangle$ if and only if $s = s'$*
- *R_X is a 'next state' relation such that if $\langle s, h\rangle R_X \langle s', h'\rangle$ then $h = h'$, and R_X is serial and deterministic*
- *The R_A are 'effectivity' relations over dynamic states $\langle s, h\rangle$ such that:*
  - *$R_\emptyset = R_\Box \circ R_X$
(empty-group effectivity is system unavoidability / settledness)*
  - *$R_{Ags} = R_X \circ R_\Box$
(Ags effectivity is next system state unavoidability / settledness)*
  - *if $\langle s, h\rangle R_\emptyset \langle s', h'\rangle$ then $\exists s'', h''$ such that $\langle s, h\rangle R_\Box \langle s'', h''\rangle$, and if $\langle s'', h''\rangle R_{Ags} \langle s''', h'''\rangle$ then $\langle s', h'\rangle R_\Box \langle s''', h'''\rangle$ (Ags choice maximality)*[9]
  - *$R_A \subseteq R_B$ for $B \subset A$ (super-groups are at least as effective; in particular, effectivity for the empty 'group' and possibility for the complete group are inherited by all groups)*
  - *For $A \cap B = \emptyset$, if $\langle s, h\rangle R_\Box \langle s', h'\rangle$ and $\langle s, h\rangle R_\Box \langle s'', h''\rangle$ then $\exists s''', h'''$ such that $\langle s, h\rangle R_\Box \langle s''', h'''\rangle$, and if $\langle s''', h'''\rangle R_A \langle s'''', h''''\rangle$ then $\langle s', h'\rangle R_A \langle s'''', h''''\rangle$, and if $\langle s''', h'''\rangle R_B \langle s''''', h'''''\rangle$ then $\langle s'', h''\rangle R_B \langle s''''', h'''''\rangle$ (independence of group agency)*

Fig.1 gives a visualization of an XSTIT-frame-part from the perspective of a single agent. We see the set of static states S pictured as little circles. The set H of histories are pictured as lines through the static states. Roughly, the dynamic states can be associated with the separate branching histories inside the circles representing the static state. However, actually every little branch inside a circle is possibly a *set* of dynamic states, because when histories come together in the past direction we simply do not picture them separately anymore. Furthermore, since the next time relation is serial, meaning there are always next states (in fig. 1 pictured using dotted lines), there are likely to be many more choices ahead when viewing the system from the standpoint of one of the states. Each

[8] In the meta-language we use these symbols both as constant names and as variable names. The same holds for the symbols $h, h', \ldots$ used to refer to histories.

[9] To keep the conditions as readable as possible we tacitly assume universal quantification of unbounded meta-variables over states and histories.

choice point gives extra histories. And this is the reason why the four lines in fig.1 are called 'Hb', for 'history *bundle*'. Note that it is not excluded that there are infinitely many choice points when following histories into the future. This means the number of histories running through a static state can be infinite. This, in turn, means that the number of dynamic states associated with a static state can be infinite. Then, for such a state, the historical necessity equivalence relation ranges over an infinite number of histories. The choices for the agent, as given by the relation R_a are visualized as ellipses in fig.1. To be precise, from any dynamic state built from static state s_1 and any of the histories in the bundles $Hb2, Hb3$ and $Hb4$, through R_a we reach all the dynamic states built from static state s_2 and the bundles $Hb3$ and $Hb4$, plus the dynamic states built from static state s_3 and bundle $Hb2$. And for this agent, from s_1, the choice (action) s_1-choice 2 is effective for φ, if φ is true in all these dynamic states.

We see that the agent does not have much choice in this (partial) example frame. Only in state s_1 the agent has two alternatives (s_1-choice 1 and s_1-choice 2); in all other states only one. Also in state s_2 the agent has only one alternative: which state will result (s_7 or s_8) is decided upon by another agent whose possible choices are not pictured in this figure.

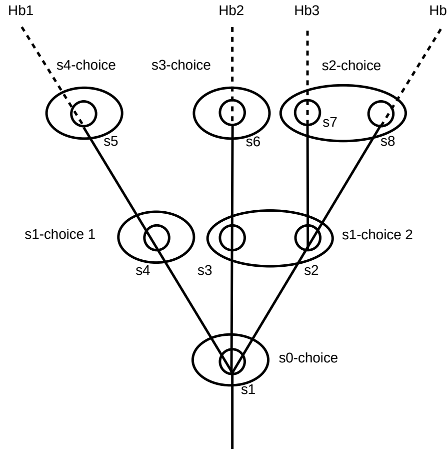

Fig 1. Visualization of a partial XSTIT frame, from the perspective of one agent

To explain the properties concerning the interaction of the effectivity relations for different agents, the visualization of fig. 1 does not suffice. Therefore, in fig. 2, we also visualize a two agent XSTIT frame-part. This picture is less suited to explain the detailed structure of histories and dynamic states (which is why we also give fig. 1), but is better suited for explaining the multi-agent choice structure. The ellipses of fig. 1 are now replaced by rectangles. For each state, the choice structure for reaching a next state is visualized as a two player game form. Before explaining the defined frame conditions in terms of this example frame, we want to emphasize that in this visualization, historical necessity relative to a dynamic state only ranges over all histories through the *small* square determined by that dynamic state. I emphasize this, because in the visualizations of *stit* models in the philosophical literature, that also use squares, historical necessity ranges over all histories within the outer rectangle. The difference is due to the fact that here a game form represents possible next states, while in the philosophical *stit* model visualizations, the rectangles represent a partition of the current state.

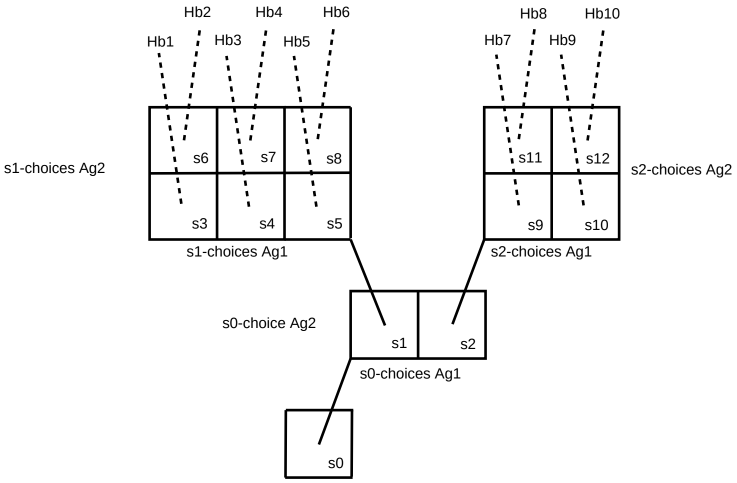

Fig 2. Visualization of a partial two agent XSTIT frame

In terms of the visualization of fig. 2 the condition $R_\emptyset = R_\Box \circ R_X$ says that in each dynamic state (but also each static state) the empty group of agents has exactly one choice, pictured as the big outer rectangle of the game form for the possible next states. More in particular, the inclusion $R_\Box \circ R_X \subseteq R_\emptyset$ says that the empty group of agents has only one choice and has no power; it is not effective to decide between any pair of histories whatsoever. The inclusion $R_\emptyset \subseteq R_\Box \circ R_X$ says in addition that only the outcomes allowed by the empty group of agents are possible as such.

The condition $R_{Ags} = R_X \circ R_\Box$ says that in each dynamic state the complete group of agents has exactly one choice, pictured in fig. 2 as the small square of the game form for the possible next states containing the actual history. The inclusion $R_X \circ R_\Box \subseteq R_{Ags}$ expresses that no agent or group can make a choice between histories that through the next state still run together. That is, even the combined choice power of all agents combined (Ags) cannot separate the histories through the next state. So, what is achieved by Ags, is settled for the next state. This corresponds to what in the philosophical literature is called the principle of 'no choice between undivided histories'. However, in the languages of these logics we cannot express an axiom that corresponds to the principle. Here we get the principle as one of the central axioms. The inclusion $R_{Ags} \subseteq R_X \circ R_\Box$ says that if something is settled for the next state, than that is due to the current choices of the complete group of agents. Note that the next *dynamic* state is *not* determined by the choices of Ags. But we might say that the next *static* state *is*. This is the XSTIT equivalent of the semantic choice in formalisms like ATL [1, 2] and CL [29] that defines that the complete set of agents uniquely determines the next (static) state.

Now we have come the property saying that if $\langle s, h\rangle R_{\emptyset}\langle s', h'\rangle$ then $\exists s'', h''$ such that $\langle s, h\rangle R_{\Box}\langle s'', h''\rangle$, and if $\langle s'', h''\rangle R_{Ags}\langle s''', h'''\rangle$ then $\langle s', h'\rangle R_{\Box}\langle s''', h'''\rangle$. This says that if the empty group of agents allows for the possibility that something will be settled next, than actually the complete group of agents can ensure that something. This is a dynamic version of what in CL and ATL is called the Ags-maximality property. Note however that in the present logic, the choices of Ags are note singleton states, like in CL and ATL. Therefore, we will not refer to the property as Ags-maximality, but as Ags-choice maximality, alluding to the fact that choices are not in general singleton sets. There are more differences between both formalizations of the idea of maximality. We come back to this briefly when we discuss whether or not CL can be seen as a fragment of XSTIT.

The condition $R_A \subseteq R_B$ for $B \subset A$ is known as coalition monotonicity. In terms of the visualization of fig. 2 it says that the smaller squares (choices of the two agents combined) are contained in the larger rectangles that determine the choices of the agents individually.

The independence of agency condition can also be explained in terms of the visualization of the two agent model in fig 2. First we restate the first-order condition in words. Assume we are in a static system state s. Now given two histories h and h' through that state, we can always find a third history h'' such that if group A has an action possibly reaching s' over h'', then the group also can reach s' over h, and if group B has an action possibly reaching s'' over h''', then the group also can reach s'' over h'. This means, in terms of the visualization of the two agent frame in fig. 2 that for any two histories passing through separate smaller boxes within a game form, there is always a history through the unique small box that is part of the choice of *both* agents. This expresses independence of agency, because it says that the intersection of choices of different agents is never empty. If the intersection would be allowed to be empty (little squares falling out of the little game forms in the picture), a choice of one agent would possibly make a choice of another agent impossible.

The independence of agency property is not undisputed. Although Belnap [9] says that "If there are agents whose simultaneous choices are not independent, [...] then we shall need to treat in the theory of agency a phenomenon just as exotic as those discovered in the land of quantum mechanics by Einstein, Podolski and Rosen.", Chellas [18] says that "the correctness of the something happens condition (Chellas' term for independence of agency) must be doubted".

Definition 2.3 *A frame $\mathcal{F} = \langle S, H, R_{\Box}, \{R_A \mid A \subseteq Ags\}, R_X\rangle$ is extended to a model $\mathcal{M} = \langle S, H, R_{\Box}, \{R_A \mid A \subseteq Ags\}, R_X, \pi\rangle$ by adding a valuation π of atomic propositions:*

- *π is a valuation function $\pi : P \longrightarrow 2^{S \times H}$ assigning to each atomic proposition the set of dynamic states in which they are true.*

The truth conditions for the semantics of the operators are standard. The non-standard aspect is the two-dimensionality of the semantics, meaning that we evaluate truth with respect to dynamic states built from a dimension of histories and a dimension of static states.

Definition 2.4 *Truth $\mathcal{M}, \langle s,h\rangle \models \varphi$, of a formula φ in a dynamic state $\langle s,h\rangle$ of a model $\mathcal{M} = \langle S, H, R_\Box, \{R_A \mid A \subseteq Ags\}, R_X, \pi\rangle$ is defined as:*

$$\begin{array}{lll}
\mathcal{M}, \langle s,h\rangle \models p & \Leftrightarrow & \langle s,h\rangle \in \pi(p) \\
\mathcal{M}, \langle s,h\rangle \models \neg\varphi & \Leftrightarrow & \textit{not } \mathcal{M}, \langle s,h\rangle \models \varphi \\
\mathcal{M}, \langle s,h\rangle \models \varphi \wedge \psi & \Leftrightarrow & \mathcal{M}, \langle s,h\rangle \models \varphi \textit{ and } \mathcal{M}, \langle s,h\rangle \models \psi \\
\mathcal{M}, \langle s,h\rangle \models \Box\varphi & \Leftrightarrow & \langle s,h\rangle R_\Box \langle s',h'\rangle \textit{ implies that } \mathcal{M}, \langle s',h'\rangle \models \varphi \\
\mathcal{M}, \langle s,h\rangle \models [A\ \mathsf{xstit}]\varphi & \Leftrightarrow & \langle s,h\rangle R_A \langle s',h'\rangle \textit{ implies that } \mathcal{M}, \langle s',h'\rangle \models \varphi \\
\mathcal{M}, \langle s,h\rangle \models X\varphi & \Leftrightarrow & \langle s,h\rangle R_X \langle s',h'\rangle \textit{ implies that } \mathcal{M}, \langle s',h'\rangle \models \varphi
\end{array}$$

Satisfiability, validity on a frame and general validity are defined as usual.

Definition 2.3 says that, like in standard *stit*-semantics, dynamic states for the same state can have different valuations of atomic propositions. In standard *stit*-formalisms this is actually needed to give semantics to the instantaneous effects of actions. But here, as said, the effects are not instantaneous. Therefore, in the present logic, the fact that different histories through the same state can have different valuations of non-temporal propositions, does not carry much meaning. The reason that alternative histories through the present state are there in the first place is that each future branch point most have 'witnesses' in the form of at least two histories separating. All these histories lead back to the present static state to form different dynamic states in combination with it. And thus, temporal formulas evaluated on these dynamic states might evaluate to different truth values (note that we can nest the X operator *any* finite number of times). That is the reason for having these alternative histories. Now one might have the opinion that *modality-free* formulas should evaluate to the same truth value for all dynamic states based on a static state. That would induce the property $\varphi \rightarrow \Box\varphi$ for φ any '*stit*-operator-free' formula[10] (in [15] we gave a system involving such an axiom[11]). However, this would complicate establishing a completeness result, and does not strengthen the logic in any essential or interesting way. We think that for the present logic in particular, there is no need to impose such a condition. Since actions only take effect in next states, alternative valuations for *atomic* propositions on other histories through the same state are just not relevant for the semantics of our *stit* logic.

Now we go on to the axiomatization of the logic. Actually, axiomatization is fairly easy. The approach we have taken for constructing this logic is to build

[10] In the current set-up of the logic, the only modality-based substitution for φ for which this schema is valid is the one resulting in $\langle\emptyset\ \mathsf{xstit}\rangle\varphi \rightarrow \Box\langle\emptyset\ \mathsf{xstit}\rangle\varphi$. Completeness says we can derive this in the Hilbert system of definition 2.5, which is easy to verify.

[11] In instantaneous *stit* there is a similar concern with the alternative histories through the present instantaneous *choice*. Belnap mentions the problem in [9], pp 31, footnote 4, but does not express any preference regarding introduction of such a property.

up the semantic conditions on frames and the corresponding axiom schemes simultaneously, while staying within the Sahlqvist class. This ensures that the semantics cannot give rise to more logical principles than can be proven from the axiomatization.

Definition 2.5 *The following axiom schemas, in combination with a standard axiomatization for propositional logic, and the standard rules (like necessitation) for the normal modal operators, define a Hilbert system for* XSTIT*:*

	S5 for $\Box$
	KD for each $[A\ \mathsf{xstit}]$
(Det)	$\neg X \neg \varphi \rightarrow X\varphi$
($\emptyset$ = Sett)	$[\emptyset\ \mathsf{xstit}]\varphi \leftrightarrow \Box X\varphi$
(Ags = XSett)	$[Ags\ \mathsf{xstit}]\varphi \leftrightarrow X\Box\varphi$
(Ags-Ch-Max)	$\langle\emptyset\ \mathsf{xstit}\rangle\Box\varphi \rightarrow \Diamond[Ags\ \mathsf{xstit}]\varphi$
(C-Mon)	$[A\ \mathsf{xstit}]\varphi \rightarrow [A \cup B\ \mathsf{xstit}]\varphi$
(Indep-G)	$\Diamond[A\ \mathsf{xstit}]\varphi \wedge \Diamond[B\ \mathsf{xstit}]\psi \rightarrow \Diamond([A\ \mathsf{xstit}]\varphi \wedge [B\ \mathsf{xstit}]\psi)$ *for* $A \cap B = \emptyset$

Theorem 2.1 *The Hilbert system of definition 2.5 is complete with respect to the semantics of definition 2.4.*

Sketch of a proof All axioms are in the Sahlqvist class. This means that all the axioms are expressible as first-order conditions on frames and that together they are complete with respect to the frame classes thus defined, cf. [10, Th.2.42]. Now it is easy to find the first-order conditions corresponding to the axioms. All correspondences are straightforward (mostly inclusions of relations and concatenations of relations), except maybe the one for independence of agency (Indep-G). But for that axiom we can find the corresponding frame condition using the on-line SQEMA system [19].

So, now we know that all axioms correspond to first-order conditions on abstract frames. In particular we know that every formula consistent in the Hilbert system has a model based on an abstract frame. Left to show is that we can associate such an abstract model to a concrete model based on an XSTIT frame as given in definition 2.2. We sketch how to do that. We associate each world of the abstract model to a dynamic world of an XSTIT model: valuations of atoms are directly copied. Then we associate the relation interpreting the X modality in the abstract model to a relation R_X in the XSTIT model: any maximal R_X-connected set of abstract model worlds we define to be a history in the XSTIT model. Now we have to construct the static states for the XSTIT model. We do that by looking at the relation interpreting the modality $[\emptyset\ \mathsf{xstit}]$ in the abstract model. For a given world, we look at all the worlds reachable through $R_\emptyset$. For the worlds thus obtained, we look at all histories through them (because of determinism and seriality, for each world in the abstract model there is a unique history). On all these histories, we go one step back over the R_X-relation (if possible). Each world in the set thus obtained, corresponds to a dynamic state in the XSTIT model, and all together, we take these dynamic states to form a static state. We now have transformed the abstract model into a model in terms of histories, states and dynamic states. Note that the construction is nothing more than a renaming of the one dimensional world structure of an abstract model into the special two dimensional dynamic state structure of an XSTIT model. This means that if the abstract model exists, the corresponding XSTIT model exists. Also, all relational interaction properties stay intact (including independence of agency). So, the formula true on the abstract model must also be true on the XSTIT model.

■

The independence of agency axiom also features in Ming Xu's axiomatization for multi-agent *stit*-logics (see the article in [9]). The present *stit*-logic is different from Xu's in two respects: (1) in the present logic, actions take effect in next states, and (2) the present logic is about groups of agents, while Xu's *stit* only considers individual agents. This shows that the issue of independence of choices of different agents does not depend on the condition that effects are instantaneous or occur in next states.

Pauly's Coalition logic [29] is a logic of ability that is very closely related to *stit*-formalisms. In particular, in [16] it is shown that Coalition Logic can be embedded in instantaneous *stit*-logic. For the present logic, at this point it is still unclear whether or not we can embed Coalition Logic. The tempting translation of Coalition Logic's central modality $[A]\varphi$ as $[A]\varphi := \Diamond[A \text{ xstit}]\varphi$ does not work, because the resulting fragment is not strong enough to validate Coalition Logic's Ags-maximality axiom. The mentioned translation would translate Coalition Logic's maximality axiom into $\neg\Diamond[\emptyset \text{ xstit}]\neg\varphi \rightarrow \Diamond[Ags \text{ xstit}]\varphi$. We can also write this as $\Box\langle\emptyset \text{ xstit}\rangle\varphi \rightarrow \Diamond[Ags \text{ xstit}]\varphi$, where we recognize a variant on the well-known McKinsey property that is not first-order definable. That is not a problem in itself; it is very well possible that non-Sahlqvist axioms are derivable as theorems in a Sahlqvist logic. However, the property is not valid in XSTIT[12]. A counter example in terms of the visualization of fig. 2 is to take a dynamic state built from a history in bundle $Hb3$ and static state s_4 and declare atomic proposition p to be true in it. Now, in the dynamic state one step back along the same history, that is, in the dynamic state built from the same history and static state s_1, we have that $\Box\langle\emptyset \text{ xstit}\rangle p$ is true, while $\Diamond[Ags \text{ xstit}]p$ is false. Note that this does not say that translation of Coalition Logic is not possible. Actually, XSTIT *does* incorporate a notion of Ags-choice maximality (the 'Ags-Ch-Max' axiom). However, the mentioned translation does not translate Coalition's Logics version of maximality to it.

[12] In [13] we claimed embedding of Coalition Logic for that papers version of XSTIT. Although maximality is derivable in that stronger version, we are no longer sure about soundness of the other direction of the mapping. As said, that paper's version of XSTIT is not the intended one.

3 The concept of 'knowingly doing'

In this section we extend XSTIT with epistemic operators $K_a\varphi$ for knowledge of individual agents a. This will enable us to express the concept of 'knowingly doing'. Herzig and Troquard were the first to consider the addition of knowledge operators to a *stit*-logic [23]. Later on the framework was adapted and extended by Broersen, Herzig and Troquard [15, 17]. This section extends earlier work in several ways. In particular, three axioms for the interaction of knowledge and action are proposed. Also the semantics, being two-dimensional, is different from the one in [15]. Finally, the modeled concept is 'knowingly doing', whereas in e.g. [23] the aim is to model 'knowing how'. In my opinion these concepts are different. I think 'knowing how' should be about whether an agent has a plan it knows to be effective. This to me seems an intrinsically strategic issue, one that cannot be approached in a non-strategic *stit*-setting. Also, 'knowing how' is an epistemic qualification concerning an *ability*, while 'knowingly doing' is an epistemic qualification concerning an *action*.

Definition 3.1 *Given a countable set of propositions P and $p \in P$, and given a finite set Ags of agent names, and $a \in Ags$ and $A \subseteq Ags$, we extend the formal language to:*

$$\varphi \quad := \quad p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_a\varphi \mid \Box\varphi \mid [A\ \mathsf{xstit}]\varphi \mid X\varphi$$

We will not fix an epistemic extension of the base XSTIT logic of section 2. Instead we show how to extend the XSTIT frames with an epistemic indistinguishability relation, and than suggest several logical properties for the notion of 'knowingly doing' that *could* be incorporated in an epistemic extension of the XSTIT logic[13]. All the suggested properties are again in the Sahlqvist class, which means that in combination with the definition is section 2 they yield a complete logic. First we extend the frames with the indistinguishability relation and define the semantics.

Definition 3.2 *An* epistemic *XSTIT frame is a tuple $\langle S, H, R_\Box, \{R_A \mid A \subseteq Ags\}, R_X, \{\sim_a \mid a \in Ags\}\rangle$ such that:*

- *$\langle S, H, R_\Box, \{R_A \mid A \subseteq Ags\}, R_X\rangle$ is an XSTIT-frame*
- *The $\sim_a$ are epistemic equivalence relations over dynamic states*

Definition 3.3 *Truth $\mathcal{M}, \langle s, h\rangle \models \varphi$, of a formula φ in a dynamic state $\langle s, h\rangle$ of a model $\mathcal{M} = \langle S, H, R_\Box, \{R_A \mid A \subseteq Ags\}, R_X, \{\sim_a \mid a \in Ags\}, \pi\rangle$ is defined as:*

All relevant clauses from definition 2.4, plus:

$$\mathcal{M}, \langle s, h\rangle \models K_a\varphi \quad \Leftrightarrow \quad \langle s, h\rangle \sim_a \langle s', h'\rangle \text{ implies that } \mathcal{M}, \langle s', h'\rangle \models \varphi$$

Satisfiability, validity on a frame and general validity are defined as usual.

[13] Of course, a danger of this approach to building a logic is that we make it too strong. In particular we always have to make sure the logic is not inconsistent.

With the above definitions we can express that agent a *knowingly* sees to it that φ as $K_a[a \ \mathsf{xstit}]\varphi$, where we slightly abuse notation by denoting $[\{a\} \ \mathsf{xstit}]\varphi$ as $[a \ \mathsf{xstit}]\varphi$. The semantics is in terms of models with epistemic equivalence sets (information sets) containing dynamic states. An agent knowingly does something if its action 'holds' for all the dynamic states in the epistemic equivalence set containing the *actual* dynamic state.

It is important to emphasize that the notion of 'knowingly doing' is entirely different from other notions combining knowledge and action or time in the literature. For instance, if we add epistemic uncertainty relations to temporal logic or dynamic logics, the choice is usually to define them over static *states* [TBD cite]. In that case uncertainty, and thus knowledge, cannot concern actions or choices themselves, but only state-determinate conditions. Only if we let uncertainty range over dynamic states, as for the present logic, we can talk about knowledge of what agents are actually doing.

I will briefly go through the different notions expressible. As said above, 'knowingly doing' is modeled by $K_a[a \ \mathsf{xstit}]\varphi$. Then, 'having the ability to do something', where we assume that 'real' ability involves that the agent knows what it is doing when it 'exercises' the ability, is expressed as $\Diamond K_a[a \ \mathsf{xstit}]\varphi$. With a strategic notion of *stit*, as in [17] or [14] the strategic notion of 'knowing how' can be expressed as $\Diamond K_a[a \ \mathsf{sstit}]\varphi$. However, we will not consider the strategic setting, and thus the 'knowing how' setting here. The notion of 'knowing to have the capacity to cause a certain effect, without knowing what to do to cause that effect', is expressed as $K_a \Diamond[a \ \mathsf{xstit}]\varphi$. An agent seeing to it that it knows something, or, learns, is expressed by $[a \ \mathsf{xstit}]K_a\varphi$. Other variations speak for them selves.

We will now discuss three possible properties for knowingly doing. We will present them as axioms in the language of definition 3.1 and give the corresponding first-order conditions on the frames of definition 3.2. The first property says that what an agent can know about the next state is never more than what it can knowingly do. The axiom is $K_aX\varphi \rightarrow K_a[a \ \mathsf{xstit}]\varphi$ (this property does not hold if the *stit*-operator is replaced by a *deliberative stit*-oparator as defined in section 4).

Proposition 3.1 *The 'ignorance about concurrent choice' (ICC) property, axiomatically expressed as $K_aX\varphi \rightarrow K_a[a \ \mathsf{xstit}]\varphi$, is in the Sahlqvist class and corresponds to the first-order condition $\sim_a \circ R_a \subseteq \sim_a \circ R_X$ on the frames of definition 3.2.*

In terms of the frames, the property says that epistemic equivalence sets are closed under choices[14]. The property ensures that an agent cannot know that two histories belonging to the same choice are different, or, in other words, for any agent the histories within its own choices are indistinguishable. This means that agents cannot know *more* about next states than what is affected by the choices they have. Formulated differently, the property says that agents

[14] An extreme case is where the information sets are exactly the choices in each state. In that case an agent knows all the consequences of his actions.

can only know things about the (immediate) future if they are the result of an action they themselves knowingly perform. Then, an agent *unknowingly* does everything that is (1) true for all the dynamic states belonging to the actual *choice* it makes in the actual *state*, but (2) not true for all the dynamic states it considers possible. In general the things an agent does unknowingly vastly outnumber the things an agent *knows* it does. For instance, by sending an email, we may enforce many, many things we are not aware of, which are nevertheless the result of me sending the email. All these things we do *unknowingly* by knowingly sending the email.

A slightly different way of explaining the property $K_aX\varphi \rightarrow K_a[a \ \mathsf{xstit}]\varphi$ is to say that it is a consequence of the assumption that agents cannot know what actions other agents perform concurrently. The independence property (Indep-G) guarantees that choices of other agents always refine the choices of the agent we consider. Thus, knowing about choices of other agents would mean that the agent would be able to know more about the future state of affairs then is guaranteed by his own action.

The second property we discuss, concerns the idea that the effects of an action that is knowingly performed are known in the next state. We can call this the dynamic version of the well-known 'perfect recall' or 'no forgetting' axiom from the literature on the interaction between epistemic and temporal modalities.

Proposition 3.2 *The 'effect recollection' (ER) property, axiomatically expressed as $K_a[a \ \mathsf{xstit}]\varphi \rightarrow XK_a\varphi$, is in the Sahlqvist class and corresponds to the first-order condition $R_X \circ \sim_a \subseteq \sim_a \circ R_a$ on the frames of definition 3.2.*

According to the property, if agents knowingly see to it that a condition holds in the next state, in that same next state they will recall that the condition holds. Like for the previous property, of course, I do not want to claim that this is a property that is necessarily true for all systems of agents. Yet it is a property that we can impose for idealized agents that are not forgetful.

Finally, we discuss the interaction property giving the relation between a de-dicto and de-re interpretation of knowingly doing: $\Diamond K_a[a \ \mathsf{xstit}]\varphi \rightarrow K_a\Diamond[a \ \mathsf{xstit}]\varphi$.

Proposition 3.3 *The 'uniformity of conformant action' (Unif-Str) property, axiomatically expressed as $\Diamond K_a[a \ \mathsf{xstit}]\varphi \rightarrow K_a\Diamond[a \ \mathsf{xstit}]\varphi$, is in the Sahlqvist class and on the frames of definition 3.2 corresponds to the following first-order condition:*
if $\langle s, h\rangle R_\Box \langle s', h'\rangle$ and $\langle s, h\rangle \sim_a \langle s'', h''\rangle$ then
$\exists s''', h'''$ such that $\langle s', h'\rangle R_\Box \langle s''', h'''\rangle$, and
if $\langle s''', h'''\rangle R_a \langle s'''', h''''\rangle$ then $\langle s', h'\rangle (\sim_a \circ R_a) \langle s'''', h''''\rangle$

This property says that if an agent can knowingly see to it that φ, then it knows that among its repertoire of choices there is one ensuring φ. This property is the *stit*-version of the constraint concerning 'uniform strategies' game theorists talk about. In game theory, *uniform* strategies require that agents have the same choices in all states within information sets. Since in game theory the choices are given names, a constraint is formulated saying that each state within the information set should have choices of the same type (that is, choices with the same name). In the present *stit*-setting, we do not have names. But the intuition that the same choices should be possible in different states of an information set, still applies. The property $\Diamond K_a[a\ \mathsf{xstit}]\varphi \rightarrow K_a\Diamond[a\ \mathsf{xstit}]\varphi$ exactly captures this intuition. It says that if an agent has the possibility to knowingly see to it that φ, then at least one of its choices in the states it considers possible actually ensures φ (that is, a φ-action is possible in all states of the information set). Maybe it is easier to see that the negation of the property, that is $\Diamond K_a[a\ \mathsf{xstit}]\varphi \wedge \widehat{K}_a\Box\langle a\ \mathsf{xstit}\rangle\neg\varphi$ (with $\widehat{K}_a$ the dual of K_a), is contradictory: it would be absurd if an agent has the possibility to knowingly see to it that φ and at the same time would consider it an epistemic possibility that it is settled that whatever it does, it allows for $\neg\varphi$ as a possible outcome. Yet another way of phrasing the property is to say that 'true ability' obeys the property of uniformity of strategies.

4 Modeling the act involved in an actus reus

Now, using only the base logic XSTIT, we can start formalizing the concepts defined in the introduction. First we will consider the notion of 'actus reus'. As explained in the introduction, an actus reus must be a voluntary act. Some aspects of the concept 'voluntary' are captured by the *stit*-notion of 'deliberative action'. A deliberative *stit*-operator adds an extra condition to the standard XSTIT-operator, to avoid the property $[A\ \mathsf{xstit}]\top$. The idea is that agents should not be able to bring about things that will be true inevitably, but only things that without their intervention might not become true. We can easily define a deliberative version of the *stit*-operator.

Definition 4.1 *The deliberative* stit-*operator* $[A\ \mathit{dxstit}]\varphi$ *is defined by:*

$$[A\ \mathit{dxstit}]\varphi \equiv_{def} [A\ \mathit{xstit}]\varphi \wedge \neg\Box X\varphi$$

Proposition 4.1 *The operator* $[A\ \mathit{dxstit}]\varphi$*, is a minimal (i.e., weak) modal operator, not obeying weakening* $[A\ \mathit{dxstit}]\varphi \rightarrow [A\ \mathit{dxstit}](\varphi \vee \psi)$*, or agglomeration* $[A\ \mathit{dxstit}]\varphi \wedge [A\ \mathit{dxstit}]\psi \rightarrow [A\ \mathit{dxstit}](\varphi \wedge \psi)$*, but obeying seriality (D)* $[A\ \mathit{dxstit}]\varphi \rightarrow \langle A\ \mathit{dxstit}\rangle\varphi$*.*

Sketch of a proof The first part of the conjunction is KD and thus satisfies weakening, but the second part not, because of the negation. Because of the negation, the second part satisfies strengthening $\neg\Box X\varphi \rightarrow \neg\Box X(\varphi \wedge \psi)$, but the first part not. The first part satisfies agglomeration, but the second part not. Both parts satisfy the D-axiom.

■

So, deliberateness, as defined in the operator above, seems to capture at least part of what it means to act voluntarily: one could also have acted otherwise, and thus one acts voluntarily. For instance, in the introduction, the crashing into the person breaking the fall of the man thrown off the building is not a voluntary act of the falling man, because the man had no choice but to fall, with the drastic consequence as a result.

However, this is not the only thing we can say about voluntary / deliberate acts. Voluntariness seems to involve more than just having had the possibility to do otherwise. Consider the following example. You carry a very dangerous contagious disease. But you do not know it. You travel by train and choose to sit next to some person and thereby unknowingly see to it that he is fatally infected. Now has an actus reus been committed (assuming spreading fatal diseases is forbidden by law)? The answer must be no. Even though it is true that you did spread the disease, and even though you could have done otherwise, what you did will not count as voluntarily or deliberately spreading the disease, simply because, to a certain extent, you did not know what you were doing.

So deliberateness or voluntariness entails both the possibility to do otherwise and having knowledge of what it is one is doing. Even more, an agent should have knowledge about the side-condition also: if an agent does not know that it could have done otherwise, we would not call the action deliberate. For the epistemic position on the side-condition, we then have two possibilities, motivating two new definitions for deliberate action.

Definition 4.2 *The deliberative* stit *alternatives* $[a \ \mathsf{dxstit}]'\varphi$ *and* $[a \ \mathsf{dxstit}]''\varphi$ *are defined by:*

$$[a \ \mathsf{dxstit}]'\varphi \equiv_{def} K_a[a \ \mathsf{xstit}]\varphi \wedge K_a\neg\Box X\varphi$$

$$[a \ \mathsf{dxstit}]''\varphi \equiv_{def} K_a[a \ \mathsf{xstit}]\varphi \wedge \neg K_a\Box X\varphi$$

The first notion says that deliberativeness requires that the agent not only knowingly performs the action, but also that the agent knows that the result is not settled, and thus that his action is needed to guarantee the result. The second notion has a different side-condition: the agent only considers it possible that the result is not settled.

Proposition 4.2 *The operators* $[a \ \mathsf{dxstit}]'\varphi$ *and* $[a \ \mathsf{dxstit}]''\varphi$ *are minimal (i.e., weak) modal operators, not obeying weakening, or agglomeration, but obeying D.*

Sketch of a proof Considerations similar to those for theorem 4.1 apply. ■

By having suggested some definitions for capturing the voluntariness aspect of an actus reus, we have actually already touched upon the notion of mens rea. This is because talking about *epistemic* aspects of action clearly already introduces 'the mind' as a relevant concept in describing action. But we have not modeled any deontic aspects yet, and thus at this point we still cannot talk about the 'guilt' aspect of mens rea. Deontic aspects will be the subject of the next section.

5 Deontic modalities and modes of mens rea

For the extension of our framework with an operator for 'ought-to-do', we adapt the approach taken by Bartha [7] who introduces Anderson style ([3]) violation constants in *stit*-theory. The approach with violation constants is very well suited for theories of ought-to-do, witness the many logics based on adding violation constants to dynamic logic [26, 11]. However, we believe that the *stit*-setting is even more amenable to this approach. Some evidence for this is found in Bartha's article ([7]), that shows that many deontic logic puzzles (paradoxes) are representable in an intuitive way. And for the present paper a clear advantage of defining obligation as a reduction using violation constants, is that the completeness established for the logics in the previous sections is preserved after addition of the obligation operator. For the violation constant we will use the special proposition $V \in P$.

Bartha [7] defines his reduction for 'obligation to do' within the classical instantaneous *stit*-setting. Here we adapt that to the present situation where actions only take effect in next states. The intuition behind the definition is straightforward: an agent is obliged to do something if and only if by not performing the obliged action, it performs a violation. Since the effect of the obliged action can only be felt in next states, violations also have to be properties of next states. Formally, our definition is given by:

Definition 5.1 *The operator $O[a\ \mathsf{xstit}]\varphi$ expressing obligation of agent a to see to it that φ, under strict liability, is defined by:*

$$O[a\ \mathsf{xstit}]\varphi \equiv_{def} \Box(\neg[a\ \mathsf{xstit}]\varphi \rightarrow [a\ \mathsf{xstit}]V)$$

Proposition 5.1 *The operator $O[a\ \mathsf{xstit}]\varphi$ is KD, that is, it has the same properties as Standard Deontic Logic [31].*

Sketch of a proof Rewrite $\Box(\neg[a\ \mathsf{xstit}]\varphi \rightarrow [a\ \mathsf{xstit}]V)$ as $\Box([a\ \mathsf{xstit}]\varphi \vee [a\ \mathsf{xstit}]V)$. Now the part $[a\ \mathsf{xstit}]V$ does not contain meta-variables (like φ) ranging over arbitrary formulas. This means that the part $[a\ \mathsf{xstit}]V$ is constant as a whole, and does not affect the logical properties of the defined modal operator $O[a\ \mathsf{xstit}]\varphi$. The necessity operator $\Box$ is S5, and $[a\ \mathsf{xstit}]$ is KD. Using standard normal modal logic correspondence theory we conclude that the combined operator $\Box[a\ \mathsf{xstit}]\varphi$ is also KD.

■

The $\Box$ operator in the definition ensures that obligations are 'moment determinate'. This means that their truth only depends on the state, and not on the history (see [24] for a further explanation of this concept). We think that this is correct. But see [30] for an opposite opinion.

In this section we will not consider the 'side conditions' as in the previous sections. But these could, of course, easily be added to model the 'could have done otherwise' aspect of 'deliberateness'. Considering side-conditions would result in yet other categories.

Note that $\neg[a\ \textsf{xstit}]\varphi$ expresses that a does not see to it that φ, which is the same as saying that a 'allows' a choice for which $\neg\varphi$ is a possible outcome. The definition then says that all such choices *do* guarantee that a violation occurs. So the agent is liable, because its action bore the risk of a bad outcome. The above defined obligation is thus a 'personal' one. If, by 'coincidence', φ occurs, apparently due the action of other agents, while the agent bearing the obligation did not make a choice that *ensured* that φ would occur, a violation is guaranteed. So agents do not escape an obligation by having other agents do the work for them.

We can also make the definition a little weaker and say that the agent is only liable if the agent actually guarantees the bad outcome:

Definition 5.2 *The operator $O'[a\ \textsf{xstit}]\varphi$ expressing obligation of agent a to see to it that φ, under strict liability, is defined by:*

$$O'[a\ \textsf{xstit}]\varphi \equiv_{def} \Box([a\ \textsf{xstit}]\neg\varphi \to [a\ \textsf{xstit}]V)$$

Proposition 5.2 *The operator $O'[a\ \textsf{xstit}]\varphi$ is a monotonic (i.e., weak) modal logic obeying the D axiom.*

Sketch of a proof We have to check the properties of the combination $\Box\langle a\ \textsf{xstit}\rangle\varphi$. We recognize a normal simulation of monotonic modal logic. Since S5 obeys D, the monotonic simulation inherits D.

■

Because the above two definitions do not at all refer to an agent's beliefs or other mental state, they both capture variants of the mens rea mode of 'strict liability'. For both definitions it is the case that if there is a violation, the agent is liable whatsoever, independent of whether or not the agent knows what it is doing. But, in my opinion this also includes the mens rea mode of 'negligently'. As described in the introduction, this class concerns those cases where 'a normal person' would have realized the consequences of his action. So, again, it does not matter what that agent knows about what it is doing, it is liable whatsoever. The only difference with the 'strict liability' class is that there can be discussion about what a normal person can foresee, and thus, about whether something should be strictly liable or not.

Now we turn our attention to the mens rea classes 'knowingly' and 'recklessly'. It is clear that to define these, we can use the concept of 'knowingly doing' as defined in the previous section. We have several options, corresponding to different modes of mens rea. We discuss the following three modes:

Definition 5.3 *The operators $OK[a\ \textsf{xstit}]\varphi$, $OK'[a\ \textsf{xstit}]\varphi$ and $OK''[a\ \textsf{xstit}]\varphi$ expressing obligation of agent a to see to it that φ, under respectively the mens rea classes* recklessly*,* knowingly recklessly *and* knowingly*, are defined by:*

$$OK[a\ \textsf{xstit}]\varphi \equiv_{def} \Box(\neg K_a[a\ \textsf{xstit}]\varphi \to [a\ \textsf{xstit}]V)$$

$$OK'[a\ \textsf{xstit}]\varphi \equiv_{def} \Box(K_a\neg[a\ \textsf{xstit}]\varphi \to [a\ \textsf{xstit}]V)$$

$$OK''[a\ \textsf{xstit}]\varphi \equiv_{def} \Box(K_a[a\ \textsf{xstit}]\neg\varphi \to [a\ \textsf{xstit}]V)$$

The first operator, that is $OK[a \text{ xstit}]\varphi$, captures the mens rea mode of 'recklessly'. Here the agent has to knowingly see to it that φ obtains, since otherwise there will be a violation. In other words, if the agent is *reckless*, and does an action that it knows does not exclude an unlawful outcome, it is liable.

The third operator, that is $OK''[a \text{ xstit}]\varphi$, captures the mens rea mode of 'knowingly'. Here there is only a violation if the agent knowingly sees to it that the *opposite* of the lawful outcome φ obtains.

Finally, the second operator, that is $OK'[a \text{ xstit}]\varphi$ defines a mode of mens rea in between 'recklessly' and 'knowingly'. It says that the agent is liable if it knowingly refrains from obtaining φ. So, on the one hand, there is an aspect of recklessness: if the agent knowingly omits to do something, a violation occurs, because omitting may risk an undesirable consequence. On the other hand, if omitting is seen as a form of doing, we can also say that this expresses that there is a violation if the agent knowingly 'does' the for this level of mens rea inexcusable omission.

Proposition 5.3 *The operator $OK[a \text{ xstit}]\varphi$ is KD, that is, it has the same properties as Standard Deontic Logic [31]. The operators $OK'[a \text{ xstit}]\varphi$ and $OK''[a \text{ xstit}]\varphi$ are monotonic (weak) modal operator obeying the D axiom. In particular, the operators do not obey agglomeration.*

Sketch of a proof For $OK[a \text{ xstit}]\varphi$ the proof is similar to the one for theorem 5.1. Here the knowledge modality is extra, which means that we have to investigate the logical behavior of the combination $\Box K_a[a \text{ xstit}]\varphi$, that is, a combination of S5, S5 and KD. This yields KD. For $OK'[a \text{ xstit}]\varphi$ and $OK''[a \text{ xstit}]\varphi$ the proofs are similar to the one for theorem 5.2

■

6 Being excused not knowing the law

In the definitions of the previous section, the focus was on the actus reus itself, and whether or not the actus reus was a knowingly performed act, a reckless act, an omission, etc. That, in itself, has nothing to do with whether or not the agent involved knows about whether or not the act it is conducting is actually an actus reus. So, what the definitions 5.1, 5.2 and 5.3 say, is that obligations *cannot* be escaped by not knowing the law; in whatever way the actus reus is conducted (knowingly, recklessly, etc.) the obligation defines that as an effect there will be a violation. So, for these definitions, the agent cannot come with the excuse that he did not know that he brought about a violation. The definitions say that it does not matter whether or not the bringing about of the violation is knowingly performed. So, the definitions of the previous section actually incorporate the juridical principle of "ignorantia juris non excusat".

However, we might want to define that not knowing about the law is actually an excuse. In that case we have to adapt the definitions.

Definition 6.1 *The operators $KOK[a\ \mathsf{xstit}]\varphi$, $KOK'[a\ \mathsf{xstit}]\varphi$ and $KOK''[a\ \mathsf{xstit}]\varphi$ expressing obligation of agent a to see to it that φ, under respectively the mens rea classes* recklessly, knowingly recklessly *and* knowingly*, avoiding the principle "ignorantia juris non excusat", are defined by:*

$$KOK[a\ \mathsf{xstit}]\varphi \equiv_{def} \Box(\neg K_a[a\ \mathsf{xstit}]\varphi \rightarrow K_a[a\ \mathsf{xstit}]V)$$

$$KOK'[a\ \mathsf{xstit}]\varphi \equiv_{def} \Box(K_a\neg[a\ \mathsf{xstit}]\varphi \rightarrow K_a[a\ \mathsf{xstit}]V)$$

$$KOK''[a\ \mathsf{xstit}]\varphi \equiv_{def} \Box(K_a[a\ \mathsf{xstit}]\neg\varphi \rightarrow K_a[a\ \mathsf{xstit}]V)$$

These definitions require that being obliged to see to something implies one knowingly brings about a violation in case of non-compliance. This means an agent is excused when it does not know it brings about an obligation in case of non-compliance.

Proposition 6.1 *The operator $KOK[a\ \mathsf{xstit}]\varphi$ is KD, that is, it has the same properties as Standard Deontic Logic [31]. The operators $KOK'[a\ \mathsf{xstit}]\varphi$ and $KOK''[a\ \mathsf{xstit}]\varphi$ are monotonic (weak) modal operator obeying the D axiom. In particular, the operators do not obey agglomeration.*

Sketch of a proof No difference with the properties for theorem 5.3 because the difference is only in the constant part of the operator definitions.

■

Note that the definitions of this section take nothing away from the rationale behind the definitions of the previous section. If we want to allow not knowing about the law as an excuse, the definitions of the present section apply, and if we do not want that, we should use the definitions of the previous section.

Of course, looking at the formal structure of the definitions of this section and the previous section, a fourth definition suggests itself: one where it is not necessary to perform the obliged action knowingly, while at the same time, in case of non-compliance, the violation *is* brought about knowingly. But it seems clear right away that this combination is absurd. We cannot knowingly bring about a violation by unknowingly failing to comply with an obligation.

7 Discussion and Conclusions

This paper presents an epistemic temporal *stit*-formalism that is complete with respect to a two-dimensional Kripke semantics. It introduces the new notion of ‘knowingly doing’ and discusses some of its possible properties. Using this notion, new ‘epistemic’ variants of operators for ‘ought-to-do’ are defined. In particular, several modes of ‘mens rea’ and characteristics of what counts as an ‘actus reus’, as defined in the juridical literature, can be analyzed and defined in the framework.

7.1 Implications and general conclusions

The first conclusion to be drawn from this work is that the logic XSTIT and its possible epistemic extensions can function as a sound and complete basis for studying and characterizing the notion of mens rea by characterizing the associated levels of deontic strength as deontic operators. Since the suggested epistemic extensions are based on Sahlqvist properties, and the suggested deontic extensions are based on the introduction of a violation constant, we have a complete logic for all the defined deontic (and non-deontic) operators.

The second general conclusion to be drawn is that our logic framework is very useful for disambiguating and precisely defining action classes from the juridical literature. This is exemplified by the fact that in our definitions a new 'natural' level of mens rea in between 'knowingly' and 'recklessly' popped up. Furthermore, it is clear that I showed quite some restraint in defining different classes; many more subtle combinations are possible, for instance by demanding 'ought implies can', 'side conditions', etc. This suggests that the classification from the juridical literature could be much more subtle and fine-grained than it is, and the present framework could be of help in defining such a classification.

A third conclusion I want to draw is one about deontic logic in general. Sometimes, in discussions with other logicians, I have to defend deontic logic against the claim that there is not a single principle of deontic logic that is non-disputed. To a certain extent that is true. If one aims at designing a 'core' logic of deontic reasoning, one is likely to end up with a very weak system, since for every suggested principle, some deontic logician will raise his hand and come with a concrete scenario and the claim that this is a counter-example. However, my claim would be that such counter-examples often introduce context that interferes with the pure deontic reasoning. For instance, the present paper makes clear that the concept of action itself and the concept of knowledge may interact with the concept of obligation in many different subtle ways, giving rise to a whole plethora of definitions for ought-to-do. And then, action and knowledge are not even the only concepts interfering; there is also time, intention, etc. Then, what the present paper is also a clear example of is the phenomenon that if we want to account for all the modalities that interfere with the pure deontic modalities, and define deontic modalities acknowledging the interactions, we get weaker logics. And this mimics closely the complaint of logicians that there is not a single principle that is not disputed. My impression is thus, that the lack of logical properties is *not* inherent to deontic logic. It is only that deontic modalities often *appear* to be rather weak because they are contaminated with other, non-deontic modalities. And one of the tasks of deontic logicians, as I see it, is to expose the contamination, and bring all interfering modalities to the foreground. In particular, we can view the present work as part of a greater project in search for the 'building blocks' of deontic modalities. And, the building blocks investigated in this paper are 'action' and 'knowledge'.

7.2 Related work

In [28] a logic is presented whose semantics shares several features with ours. In particular, the logic has epistemic indistinguishability relations ranging over dynamic states. However, actions are omitted. In [27] actions are added to this framework by using action names in the models and the object language. So, the authors take a, what we might call 'dynamic logic view' on action. The work focusses on so called 'knowledge based obligations'. The central idea is that when agents get to know more, there are less histories they consider possible, which in turn may induce that the subset of deontically optimal histories, may give rise to new obligations. So the phenomenon being studied is that new knowledge may induce new obligations.

In our setting the phenomenon of getting more obligations by an increase in knowledge can occur in different ways. One way is simply by becoming aware of an obligation, that is, getting to know that one knowingly performs a violation by not performing some obliged action. Another route to enabling that obligations arise as the result of new knowledge, is by adopting the 'ought implies can' principle for the stronger variants of our obligation operator. If agents get to know how to do something knowingly, they might incur an obligation that previously did not apply due to 'ought implies can'. This demonstrates that there seems to be more sides to the problem of 'knowledge based obligation'.

Another well-known interaction between epistemic and deontic modalities is Åqvist's puzzle of 'the knower' [4]. If knowledge is modeled using S5 and obligation using KD (SDL [31]), from $OK\varphi$ we derive $O\varphi$, which is clearly undesirable in an ought-to-be reading. However, this problem does not arise in the present logic, because obligation is strictly limited to apply to *actions*. In particular, if in Åqvist's example, for φ we substitute a *stit*-action $[\alpha\ \mathsf{xstit}]\varphi$, then we can read the derivation as 'the obligation to knowingly see to something implies the obligation to see to that same something'. In the present framework, that is not an undesirably property, but a desirable property obeyed by our definitions, because it is valid that $OK[a\ \mathsf{xstit}]\varphi \rightarrow O[a\ \mathsf{xstit}]\varphi$.

7.3 Future research

The framework we presented asks for extension in several ways. Note first that while the operators for agency are group operators, the operators for knowledge and obligation only refer to single agents. Actually, there are many open questions about how to generalize these operators to group operators. As is well-known, there are several notions of group-knowledge, such as 'shared knowledge', 'common knowledge' and 'distributed knowledge'. Which ones combine with which interaction properties for knowledge and group-action is yet unclear. Likewise we can consider generalizing the obligation operator to a group operator. Given the definitions of section 5 this actually hinges on providing group operators for the knowledge modalities.

Another issue concerns the violation constants. According to the present definitions, they are not relativized to agents or sets of agents. This corresponds to a 'consequentialist's' view on obligation, as in [24], where deontic optimality is determined according to an ordering of all possible histories. We could also take the view, like in [25], that deontic optimality orderings should be relative to agents or groups of agents. For our setting, using violation constants, that would mean that we introduce a violation constant for each agent or each group.

Acknowledgements

This paper is based on a paper presented at the DEON workshop in Luxembourg [12]. Thanks to Leon van der Torre and Ron van der Meijden for organizing this workshop. Thanks to John Horty for forcing me to think of a better term for epistemic agency, which prompted me to come up with 'knowingly doing'. Thanks to Google and Wikipedia for providing swift and fairly accurate initial information on 'knowingly doing' in relation to juridical issues. Thanks to Henry Prakken for suggesting that the theory on causality and responsibility in the juridical domain might form an interesting application domain for *stit*-logics. Finally, thanks to the three anonymous referees, whose comments were very valuable and helped me to recognize some earlier mistakes and omissions.

References

[1] R. Alur, T.A. Henzinger, and O. Kupferman. Alternating-time temporal logic. In *Proceedings of the 38th IEEE Symposium on Foundations of Computer Science*, Florida, October 1997.

[2] R. Alur, T.A. Henzinger, and O. Kupferman. Alternating-time temporal logic. *Journal of the ACM*, 49(5):672–713, 2002.

[3] A.R. Anderson. A reduction of deontic logic to alethic modal logic. *Mind*, 67:100–103, 1958.

[4] L. Åqvist. Good samaritans, contrary-to-duty imperatives, and epistemic obligations. *NOUS*, 1:361–379, 1967.

[5] J. Austin. A plea for excuses. *Proceedings of the Aristotelian Society*, (7), 1956.

[6] Philippe Balbiani, Olivier Gasquet, Andreas Herzig, François Schwarzentruber, and Nicolas Troquard. Coalition games over Kripke semantics: expressiveness and complexity. In Cédric Dègremont, Laurent Keiff, and Helge Rückert, editors, *Festschrift in Honour of Shahid Rahman*. College Publications, 2008. to appear.

[7] Paul Bartha. Conditional obligation, deontic paradoxes, and the logic of agency. *Annals of Mathematics and Artificial Intelligence*, 9(1-2):1–23, 1993.

[8] N. Belnap and M. Perloff. Seeing to it that: A canonical form for agentives. *Theoria*, 54:175–199, 1988.

[9] N. Belnap, M. Perloff, and M. Xu. *Facing the future: agents and choices in our indeterminist world.* Oxford, 2001.

[10] P. Blackburn, M. de Rijke, and Y. Venema. *Modal Logic*, volume 53 of *Cambridge Tracts in Theoretical Computer Science.* Cambridge University Press, 2001.

[11] J.M. Broersen. *Modal Action Logics for Reasoning about Reactive Systems.* PhD thesis, Faculteit der Exacte Wetenschappen, Vrije Universiteit Amsterdam, februari 2003.

[12] J.M. Broersen. A logical analysis of the interaction between 'obligation-to-do' and 'knowingly doing'. In L.W.N. van der Torre and R. van der Meyden, editors, *Proceedings 9th International Workshop on Deontic Logic in Computer Science (DEON'08)*, volume 5076 of *Lecture Notes in Computer Science*, pages 140–154. Springer, 2008.

[13] J.M. Broersen. A complete stit logic for knowledge and action, and some of its applications. In M. Baldoni, T. Cao Son, M.B. van Riemsdijk, and M. Winikoff, editors, *Declarative Agent Languages and Technologies VI (DALT 2008)*, volume 5397 of *Lecture Notes in Computer Science*, pages 47–59, 2009.

[14] J.M. Broersen. A stit-logic for extensive form group strategies. In *WI-IAT '09: Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, pages 484–487, Washington, DC, USA, 2009. IEEE Computer Society.

[15] J.M. Broersen, A. Herzig, and N. Troquard. A normal simulation of coalition logic and an epistemic extension. In *Proceedings Theoretical Aspects Rationality and Knowledge (TARK XI), Brussels.*

[16] J.M. Broersen, A. Herzig, and N. Troquard. From coalition logic to STIT. 157(4):23–35, 2006. Proceedings LCMAS 2005.

[17] J.M. Broersen, A. Herzig, and N. Troquard. A STIT-extension of ATL. In Michael Fisher, editor, *Proceedings Tenth European Conference on Logics in Artificial Intelligence (JELIA'06)*, volume 4160 of *Lecture Notes in Artificial Intelligence*, pages 69–81. Springer, 2006.

[18] Brian F. Chellas. Time and modality in the logic of agency. *Studia Logica*, 51(3/4):485–518, 1992.

[19] W. Conradie, V. Goranko, and D. Vakarelov. Algorithmic correspondence and completeness in modal logic I: The core algorithm SQEMA. *Logical Methods in Computer Science*, 2(1):1–26, 2006.

[20] Markus D. Dubber. *Criminal Law: Model Penal Code.* Foundation Press, 2002.

[21] E.A. Emerson. Temporal and modal logic. In J. van Leeuwen, editor, *Handbook of Theoretical Computer Science, volume B: Formal Models and Semantics*, chapter 14, pages 996–1072. Elsevier Science, 1990.

[22] Andreas Herzig and Francois Schwarzentruber. Properties of logics of individual and group agency. In Carlos Areces and Rob Goldblatt, editors, *Advances in Modal Logic*, volume 7, pages 133–149. College Publications, 2008.

[23] Andreas Herzig and Nicolas Troquard. Knowing How to Play: Uniform Choices in Logics of Agency. In Gerhard Weiss and Peter Stone, editors, *5th International Joint Conference on Autonomous Agents & Multi Agent Systems (AAMAS-06), Hakodate, Japan*, pages 209–216. ACM Press, 8-12 May 2006.

[24] J.F. Horty. *Agency and Deontic Logic.* Oxford University Press, 2001.

[25] Barteld Kooi and Allard Tamminga. Moral conflicts between groups of agents. *Journal of Philosophical Logic*, 37(1):1–21, 2008.

[26] J.-J.Ch. Meyer. A different approach to deontic logic: Deontic logic viewed as a variant of dynamic logic. *Notre Dame Journal of Formal Logic*, 29:109–136, 1988.

[27] E. Pacuit, R. Parikh, and E. Cogan. The logic of knowledge based obligation. *Knowledge, Rationality and Action a subjournal of Synthese*, 149(2):311–341, 2006.

[28] Rohit Parikh and Ramaswamy Ramanujam. A knowledge based semantics of messages. *Journal of Logic, Language and Information*, 12(4):453–467, 2003.

[29] Marc Pauly. A modal logic for coalitional power in games. *Journal of Logic and Computation*, 12(1):149–166, 2002.

[30] H. Wansing. Obligations, authorities, and history dependence. In H. Wansing, editor, *Essays on Non-classical Logic*, pages 247–258. World Scientific, 2001.

[31] G.H. von Wright. Deontic logic. *Mind*, 60:1–15, 1951.