J Philos Logic (2011) 40:499–530
DOI 10.1007/s10992-011-9190-6

Making a Start with the *stit* Logic Analysis of Intentional Action

Jan M. Broersen

Received: 16 November 2009 / Accepted: 17 January 2010 / Published online: 2 May 2011

Abstract This paper studies intentional action in *stit* logic. The formal logic study of intentional action appears to be new, since most logical studies of intention concern intention as a static mental state. In the formalization we distinguish three modes of acting: the objective level concerning the choices an agent objectively exercises, the subjective level concerning the choices an agent knows or believes to be exercising, and finally, the intentional level concerning the choices an agent intentionally exercises. Several axioms constraining the relations between these different modes of acting will be considered and discussed. The side effect problem will be analyzed as an interaction between knowingly doing and intentionally doing. Non-successful action will be analyzed as a weakening of the epistemic attitude towards action. Finally, the notion of 'attempt' will be briefly considered as a further weakening in this direction.

Keywords Agency · Indeterminism · Action theory · Modal logic · Formal epistemology

1 Introduction

This paper studies intention as a mode of acting. This is quite different from studying intention as one of the elements of mental states of agents. Within the computer science community working on logical approaches to AI, the best-known paper on the latter subject is the one by Cohen and Levesque [16]. The

J. M. Broersen (✉)
Intelligent Systems Group, Department of Information and Computing Sciences,
Faculty of Science, Universiteit Utrecht, PO Box 80.089, 3508 TB Utrecht, The Netherlands
e-mail: broersen@cs.uu.nl

difference in subject between the present work and that work is best explained by Cohen and Levesque themselves [16, p 216]:

> Most philosophical analysis has examined the relationship between an agent's doing something intentionally and that agent's having a present-directed intention. Recently, Bratman [7] has argued that intending to do something (or having an intention) and doing something intentionally are not the same phenomenon, and that the former is more concerned with the coordination of an agent's plans. We agree, and in this paper we concentrate primarily on future-directed intentions. Hereafter, the term "intention" will be used in that sense only.

In this paper we study the interpretation of intention explicitly excluded by Cohen and Levesque: "intentionally doing". The difference is parallelled by differences in the formal apparatuses to study the notions. Cohen and Levesque use a first-order logic where action types are represented in the same way as in Dynamic Logic [24, 40]. More precisely: to talk about action types they use a translation of propositional Dynamic Logic into their first-order language. Although this approach enables them to reason about several important properties of action, like pre and post condition reasoning in the context of action (type) composition, it does not enable them to reason about acting as such. In the formalism they put forward there is no object level construct for expressing, for instance, that it is currently true that "agent *agt* writes a paper". As in Dynamic Logic, expressivity is limited to conditional assertions like "if action *a* were to be executed, it would have as an effect that a paper is written" and non-conditional assertions like "action *a* is executable". The reason that Cohen and Levesque do not need an operator for action is that they do not study intentional action, but intention as a mental state. For our study of intention as a mode of acting, we use *stit* logic. *Stit* logic does enable us to talk about action unconditionally. For the present study, another advantage of using *stit* logic rather than Dynamic Logic is that it is still unclear how to express properties and aspects of agency in Dynamic Logic (examples are: refraining, deliberate choice, independence of agency, regularity, etc.).

Having pointed out the difference with the work of Cohen and Levesque, we want to stress that there are also issues that arise under both interpretations of 'intention'. In particular, one of the central issues in the work of Cohen and Levesque is that intention is not closed under side effects of action (the well-known denstist's example). In our framework we will analyze the same problem in the context of intentional action.

In philosophy, the understanding of the nature of intentional action is a central theme, with contributions from Anscombe [3], Davidson [17], Chisholm [15], Searle [42], Mele and Moser [37], and, more recently Knobe [31]. But our main motivation for the present work comes from the literature on law and deontic logic. As is well known, for a judge deciding on a verdict, there is a lot

of difference between murder, manslaughter, homicide, killing in self-defense, etc. Yet, all these acts concern one objective *physical* event: that of causing someone's death. The difference is in the mode of acting, that is, in the mental state by which an agent's act is accompanied at the time of conduct (the legal literature speaks of 'showing concurrence').

In criminal law, the different modes of acting correspond with different categories of culpability. And it is the judge's task to assess to which category a case belongs. 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 [18]):

- *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.
- *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

In this paper we will be only concerned with the first two categories. We formalize the distinctions between the other categories in [12] (which extends and corrects [9]).

The first category, the one of acts committed *purposefully*, is about acts that are instrumental in reaching an agent's *goal*. So, this is the category of intentional action. From a legal perspective it is important to asses whether or not the intention in the action is malicious. The second category 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 plan of this paper is as follows. First, in Section 2 we introduce the logic XSTIT from [12] as the base formalism in which we perform our analysis. Section 3 introduces the combined operators we use to express the epistemic and intentional attitudes towards action. In Section 4 we concentrate on the notion of 'knowingly doing', mostly taken from [12]. Then, in Section 5 we present our view on the notion of intentionally doing, and discuss the relation and difference with knowingly doing. The well-known side-effect problem will be cast in terms of the difference between intentionally doing and knowingly doing. In Section 7 we observe that intentionally doing as defined in Section 5 does not leave room for intentional action being non-successful. We show how to adapt the properties to allow for non-successful action. In particular, we will weaken the notion of 'knowingly doing' to its belief equivalent. In Section 8 we consider, as a suggestion for future work, how to weaken the epistemic attitude

towards action performance even further, and discuss the concept of 'attempt'. Finally Section 9 discusses more future work and conclusions.

2 A Group *stit* Logic Affecting 'Next' States: XSTIT

As the basis for the investigation we use the the logic XSTIT first presented in [10] and [9], and corrected in [12]. Since XSTIT is more extensively introduced in [12], here we will give a briefer exposition of XSTIT and its extension with epistemic attitudes.

XSTIT is a complete *stit* logic where actions take effect in 'next' states. For those unfamiliar with the *stit* framework: the characters 'stit' are an acronym for 'seeing to it that'. *Stit* logics [5, 6] originate in philosophy, and can be described as endogenous logics of agency, that is, logics of agentive choice where action types are not made explicit in the object language. To be more precise, expressions $[A\ stit : \varphi]$ of *stit* logic stand for 'agents A see to it that φ', where φ is a (possibly) temporal formula. However, where the founding fathers of *stit* theory write '$[A\ stit : \varphi]$', we prefer to write '$[A\ \mathsf{stit}]\varphi$', to be more in line with standard modal logic notation. The main virtue of *stit* logics is that, unlike most (if not all) other logical formalisms relating to action, they can express that a choice is actually exercised by an agent. This relates directly to the notion of truth in the semantics; truth is always evaluated against what we call 'dynamic states' which consist of history-state pairs. This reflects that *truth* of a formula says something about the dynamics of the agents in the system, that is, about which choices they exercise and about what is true as the result of that.

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. A motivation for interpreting *stit* modalities in terms of effects in next states comes from computer science, where this is the more common view in formal models of computation (transition systems). 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 [4, 26].

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* [19]. 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 those not familiar with *stit* theory often think). 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 $[A\ \mathsf{xstit}]\varphi$ stands for 'agents A jointly see to it that φ in the next state'.

The modality $X\varphi$ is the next operator. It has a standard interpretation as the transition to a next static 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 $\mathcal{L}_{\mathsf{XSTIT}}$ is:

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

Our *stit* operator concerns, what game-theorists call, 'one-shot' actions. We can also imagine to have a *strategic stit* operator (see [13]) where it is assumed that groups of agents have multiple subsequent choice points to ensure a certain condition (game-theorists call settings like these 'extensive games').

In the description of the structures, below, we will use terminology inspired by similar terminology from Coalition Logic (CL) [39], 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.

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

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

- S is an infinite set of static states. Elements of S are denoted s, s', etc.[1]
- $H \subseteq 2^{2^S \setminus \emptyset} \setminus \emptyset$ is a non-empty set of histories, which are ordered infinite sub-sets of S. 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 $s \in h$. Histories receive their order from the next state relation R_X over dynamic states: s' is next of s on h if and only if $\langle s, h\rangle R_X \langle s', h\rangle$.[2]
- R_X is a 'next state' relation that is serial and deterministic, and if $\langle s, h\rangle R_X \langle s', h'\rangle$ then $h = h'$
- $R_\Box$ is a 'historical necessity' relation over dynamic states such that $\langle s, h\rangle R_\Box \langle s', h'\rangle$ if and only if $s = s'$
- 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 static state unavoidability / settledness)

[1] 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.

[2] To keep the conditions listed here as readable as possible we tacitly assume universal quantification of unbounded meta-variables over static states, histories and groups.

- $R_A \subseteq R_B$ 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 $\langle s_1, h_1\rangle R_\Box \langle s_2, h_2\rangle$ and $\langle s_1, h_1\rangle R_\Box \langle s_3, h_3\rangle$
  then $\exists s_4, h_4$ such that $\langle s_1, h_1\rangle R_\Box \langle s_4, h_4\rangle$,
  and if $\langle s_4, h_4\rangle R_A \langle s_5, h_5\rangle$ then $\langle s_2, h_2\rangle R_A \langle s_5, h_5\rangle$,
  and if $\langle s_4, h_4\rangle R_B \langle s_6, h_6\rangle$ then $\langle s_3, h_3\rangle R_B \langle s_6, h_6\rangle$
  (independence of group agency)

In Fig. 1, we visualize a two agent XSTIT frame-part. For each state, the choice structure for reaching next states as determined by effectivity relations and history bundles is visualized as a two player game form. In *stit* logics, acting, by a group A, is identified with ensuring that a condition holds on all (dynamic) states that may result from exercising a choice (all the worlds the choice is effective for). In terms of the visualization of Fig. 1, the choices of $Ag1$ appear as columns of the game forms, the choices of $Ag2$ appear as rows, the choice of the empty set of agents (which does not depend on the actual history, and is thus moment determinate) appears as the outmost rectangle of a game form, and the choices of Ags appear as the smallest squares inside the game forms.

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 *smallest* square determined by that dynamic state. I emphasize this, because in the visualizations of *stit* models in the philosophical literature, that also use game

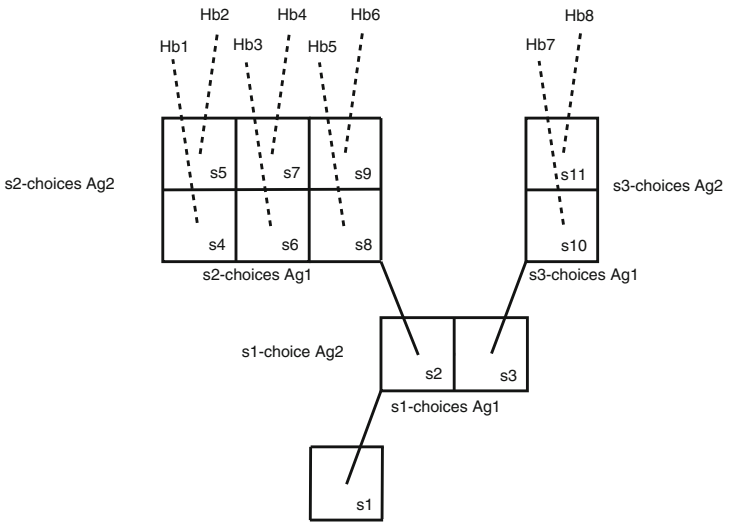

Fig. 1 Visualization of a partial two agent XSTIT frame

forms, historical necessity ranges over all histories within the *outmost* rectangle. The difference is due to the fact that here a game form represents possible next states, while in the traditional *stit* model visualizations, the rectangles represent a partition of the current moment.

In terms of the visualization of Fig. 1 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 option,[3] pictured as the outmost 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 option 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. 1 as the smallest 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 exercise a choice that separates histories which in the next state run together again. That is, even the 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 *stit* literature is called the principle of 'no choice between undivided histories'. The inclusion $R_{Ags} \subseteq R_X \circ R_{\Box}$ says that if something is settled for the next state, then 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 [39] that defines that the complete set of agents uniquely determines the next state.[4]

The condition $R_A \subseteq R_B$ for $B \subset A$ is known as coalition (anti) monotonicity. In terms of the visualization of Fig. 1 it says that the smallest squares (choices of the two agents combined) are contained in the larger rectangles that determine the choices of the agents individually. The reason that we do not have the condition $R_A \subset R_B$ for $B \subset A$ is that it is always possible to add an agent to the system that has no power at all; an agent with the same powers as the empty set of agents.[5] Note that it cannot be the case that genuine choices (that is, choices for which objective alternatives exist) of different agents are identical (that is, correspond to exactly the same bundle of histories), because this conflicts with independence of agency.

[3] We avoid the word 'choice' here, since some insist that it is intrinsic to the meaning of 'choice' that there is an alternative. One could argue however, that the same is true for the meaning of 'option' or even for the meaning of 'possibility'. Later on we will distinguish between 'choices' and 'genuine choices' to make the distinction.

[4] In the semantics of CL and ATL there is no distinction between static states and dynamic states; the states featuring in the semantics of these formalisms are best thought of as our 'static' states.

[5] However, one may argue that such agents are no agents at all. But, here we do not make the distinction between agents and nature more precise.

The independence of agency condition is a modal confluency property. In terms of the visualization of the two agent frame in Fig. 1 it says that for any history through a column choice of agent 1 and any history through a row choice of agent 2, there is always a history through the unique smallest box that is in the choice of *both* agents (in terms of the figure: it is in the intersection of a row and a column). 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 (smallest squares falling out of the little game forms in the picture), choice exertion of one agent would possibly make a choice of another agent impossible.

Contrary to what is suggested by Fig. 1 we do not have that different choices of the same agent cannot overlap and that the combined choice of agents is always exactly the intersection of the choices of the agents participating in the collective choice. In modal logic we cannot characterize these conditions as properties of the frames, since we cannot characterize that intersections are empty. But since these conditions result in much tidier visualizations, we assume them in the figures.

Definition 2.3 A frame $\mathcal{F} = \langle S, H, R_X, R_\Box, \{R_A \mid A \subseteq Ags\}\rangle$ is extended to a model $\mathcal{M} = \langle S, H, R_X, R_\Box, \{R_A \mid A \subseteq Ags\}, \pi\rangle$ 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 $\mathcal{M} = \langle S, H, R_X, R_\Box, \{R_A \mid A \subseteq Ags\}, \pi\rangle$ is defined as:

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

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

Definition 2.3 says that, like in standard *stit* semantics, dynamic states based on the same static state can have different valuations of atomic propositions. This might be considered counter-intuitive, since it would make sense to assume that such propositions do not express truths about the dynamics of the agent system, which implies that their valuation should be uniform over

the histories based on the same static state (that is, their valuation should be 'moment determinate' [28]). It is not problematic to adapt the semantics with this extra condition. To preserve completeness for the axiomatization we give below, we then would have to add the axiom $p \rightarrow \Box p$ for p any 'modal-operator-free' formula.

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 $[A \text{ xstit}]$

$\neg X \neg \varphi \rightarrow X\varphi$ (Det)

$\Box X\varphi \leftrightarrow [\emptyset \text{ xstit}]\varphi$ ($\emptyset$-SettX)

$[Ags \text{ xstit}]\varphi \leftrightarrow X\Box\varphi$ (Ags-XSett)

$[A \text{ xstit}]\varphi \rightarrow [A \cup B \text{ xstit}]\varphi$ (C-Mon)

$\Diamond[A \text{ xstit}]\varphi \wedge \Diamond[B \text{ xstit}]\psi \rightarrow \Diamond([A \text{ xstit}]\varphi \wedge [B \text{ xstit}]\psi)$ for $A \cap B = \emptyset$ (Indep-G)

Theorem 2.1 (From [12]) *The Hilbert system of Definition* 2.5 *is complete with respect to the semantics of Definition* 2.4.

In the rest of the paper, we discuss logical properties not in terms of the multi-agent frames of the type pictured in Fig. 1, but in terms of single agent 'views' on such frames. To explain this, in Fig. 2 we first give agent 1's view on the frame of Fig. 1. In this visualization, the choices for agent 1, as given by the relation $R_{\{1\}}$, appear as ellipses grouping different possible sets of next states. We see the set of static states S pictured as little circles. The choices of the other agent appear here as non-determinism of an unspecified source. Strictly speaking elements from the set H of histories are not pictured. The lines through the static states in the picture represent 'history bundles' (which explains the names 'Hb' in the picture). In this figure (but also in Fig. 1) branching of time is then represented as branching of bundles of histories. Since this is only a partial frame, from the viewpoint of any static state there may still be infinitely many choices ahead, which means that the number of histories in a bundle through any pictured static state can also be infinite.

Figure 2 shows how agency is about exercising control over non-determinism. All the choices of agent 1 in the picture represent different possibilities for the agent's potential to control non-determinism. For instance, in static state s_1 it has a choice between two deterministic alternatives. In static state s_2 it has nothing to choose from since there is only one non-deterministic alternative; what state results depends on the choice of the other agent. In s_3 the agent has a choice between three alternatives that are in themselves

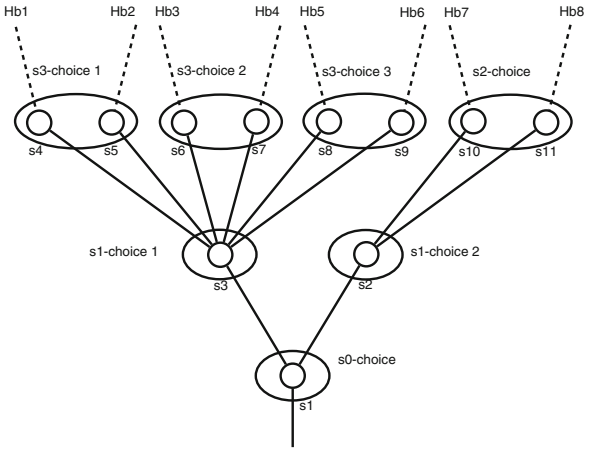

Fig. 2 Visualization of the partial XSTIT frame of Fig. 1, from the perspective of agent 1

non-deterministic. Note that we talk about the choices as concerning 'the potential' for controlling non-determinism. This is because the choices in the picture only represent an agent's *objective* possibilities to control non-determinism. In Section 4 we will add an agent's epistemic attitude towards these objective choices. This means that we will be able to express to what extent an agent *knows* about its abilities to control non-determinism. And in Section 5 we will add the intentional attitude, enabling us to investigate logical properties governing *intentional* control over non-determinism.

3 Operators for Knowledge and Intention

In his Philosophical Investigations §621 Wittgenstein famously asked: "What is left over if I subtract the fact that my arm goes up from the fact that I raise my arm?". The question is rhetorical, emphasizing Wittgenstein's point that the subtraction yields no residue.[6] In conflict with this view, here we start from the position that "my arm goes up" and "I raise my arm" are essentially different: the first is the objective action while the second is the simultaneous intentional action. We will explore the idea that any agentive act can be considered at three different levels: (1) the objective level, i.e., the choice actually exercised, (2) the subjective level, which is about what an agent knows or believes to be choosing, and (3) the intentional level, which is about the intentional choice exercised. The objective level is accounted for by the choices modeled in the base XSTIT logic of the previous section. To account for the other levels of consideration, we extend XSTIT with epistemic operators $K_a\varphi$ for knowledge of individual agents a, and operators $I_a\varphi$ for

[6]Thanks to Menno Lievers for pointing this out.

'agent a intends φ'. Subjective action is then modeled with the combined modality[7] $K_a[a\ \mathsf{xstit}]\varphi$, and intentional action with $I_a[a\ \mathsf{xstit}]\varphi$. We will not discuss present (or immediate next state) directed intentions for φ of the form $\Box I_a[a\ \mathsf{xstit}]\varphi$. Also we will not discuss future directed intentions for φ of the form $\Box I_a F[a\ \mathsf{xstit}]\varphi$, where the F-operator is read as 'some time in the future'.[8] This temporal operator is not in the object language of the systems we consider here.

Definition 3.1 We extend the syntax of Definition 2.1 with an operator for knowledge and intentional action, resulting in:

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

Note that the stit-operators concern groups of agents, while the knowledge and intention operators concern individual agents. In this paper we do not want to consider the intricacies of the action versions of group knowledge and group intention. We extend XSTIT's semantic basis by the following definitions.

Definition 3.2 The class of general KI-extended XSTIT frames consists of frames $\mathcal{F} = \langle S, H, R_X, R_\Box, \{R_A \mid A \subseteq Ags\}, \{\sim_a \mid a \in Ags\}, \{i_a \mid a \in Ags\}\rangle$ such that:

- $\langle S, H, R_X, R_\Box, \{R_A \mid A \subseteq Ags\}\rangle$ is an XSTIT-frame
- The $\sim_a$ are epistemic equivalence relations over dynamic states (corresponding to the modal frame class S5).
- The i_a are intention equivalence relations over dynamic states (corresponding to the modal frame class S5).

In what follows, we will consider subclasses of the above general frame class, where the relations $R_\Box$, R_A, R_X, $\sim_a$ and i_a obey first-order Sahlqvist conditions modeling the interactions between the associated modalities. We choose this set-up, because we want to have the freedom to discuss different, but sometimes equally defendable axioms for the interactions. Each set of axioms gives a different logic and a different frame class as a subset of the above general class. The whole point of this exercise is thus not so much to put forward the logic of intentional action, but to show that the present framework enables us to study the different possibilities and considerations for designing such a logic.

[7]Note that we abuse syntax of the object language by denoting singleton sets of agents by the name of the single agent in the set.

[8]The future directed intention for 'seeing to it that φ some time in the future', i.e., $\Box I_a F[a\ \mathsf{xstit}]\varphi$ should not be confused with the present directed (immediate) intention to 'see to it that φ at some time in the future': $\Box I_a[a\ \mathsf{xstit}]F\varphi$. An example of the first is to intend to kill a certain agent some time in the future, while an example of the second is to intend to now kill the agent at some time in the future, for instance by placing a booby-trap in his car.

When we consider KI-extended XSTIT frames without the intention relations, we will speak of K-extended (general) XSTIT frames. We can now extend Definition 2.4 with clauses for truth conditions for the knowledge operator and the intention operator.

Definition 3.3 The truth conditions for the knowledge operator K_a and the intention operator I_a are defined as:

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

Herzig and Troquard were the first to consider the addition of knowledge operators to a *stit* logic [27]. Later on the framework was adapted and extended by Broersen, Herzig and Troquard [13, 14]. The epistemic fragment of the present logic extends our earlier work on epistemic *stit* in several ways. In particular, new properties for the interaction of knowledge and action are proposed. Also the semantics, being two-dimensional, is different from the one in [14]. Finally, the modeled concept is 'knowingly doing', whereas in e.g. [27] the aim is to model 'knowing how'.

Intention operators have been considered in the *stit* framework by Lorini and Herzig [33, 34] and by Semmling and Wansing [43]. However, in both these works, like in the work of Cohen and Levesque, the emphasis is on intention as a mental state, and not on intention as a mode of acting.

4 Knowingly Doing

With the above definitions we can express that agent *a knowingly* sees to it that φ as $K_a[a\ \mathsf{xstit}]\varphi$ [9]. Semantically: an agent knowingly does φ if φ holds for all the dynamic states in the epistemic equivalence set containing the *actual* dynamic state. In [14] we also called this 'conformantly' doing, in analogy with the notion of conformant planning [22], which looks at plans that are successful under incomplete knowledge about the current state.

We will go briefly through some notions that are expressible. As said above, 'knowingly doing' which is short for 'knowingly seeing to it that φ' is modeled by $K_a[a\ \mathsf{xstit}]\varphi$. Then, 'having the ability to do something', where we assume that 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'[9] notion of *stit*, as in [13] or [11] the strategic notion of 'knowing how' can be expressed as $\Diamond K_a[a\ \mathsf{sstit}]\varphi$. However, here we will not consider the strategic or 'knowing how' setting. The notion of 'knowing to have the capacity to cause a certain

[9] What is meant by 'strategic' here is that an action possibly involves several subsequent choices. In game theory one refers to such settings as 'extensive games'.

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 themselves.

Let us now consider the concept of knowingly doing in terms of the defined frames and the models based on them. Figure 3 visualizes a possible way to add agent 1's epistemic indistinguishability relation to the frame of Fig. 2. We need some background knowledge to interpret this visualization in the right way. We know that the epistemic indistinguishability (or, equivalence) relation $\sim_a$ partitions the dynamic states of a frame. However, we will even assume here that for every static state, $\sim_a$ partitions the dynamic states based on it (which corresponds to the axiom K-S we discuss below). Now, equivalence classes of dynamic states based on a given static state s are hard to visualize by picturing them directly as *separate* bundles at point s in the figure. Therefore, in Fig. 3, such equivalence classes are visualized indirectly as dotted rectangles grouping all possible states next of s. Then, for any specific dynamic state, by application of the combined operator $K_a[a\ \mathsf{xstit}]\varphi$ (that is, by following elements of the concatenated relations $\sim_a \circ R_{\{a\}}$) we reach all the dynamic states within a specific dotted rectangle. From the picture it is clear that these dynamic states are always a subset of all possible next states.

In Fig. 3 we see that from static state s_3, there are three objective choices for the agent (s3-choice 1, s3-choice 2 and s3-choice 3), while there are two choices the agent can knowingly exercise (the two dotted rectangles grouping choices together). The dotted rectangles represent the two sets of states reachable through $\sim_a \circ R_{\{a\}}$ from different dynamic states based on static state s_3. In this particular frame, in s_3 the agent cannot distinguish between s3-choice 2 and s3-choice 3; as far as it knows, these choices are identical, which is visualized by the dotted rectangle surrounding them.

We now briefly discuss possible properties for the interaction between knowledge and action.

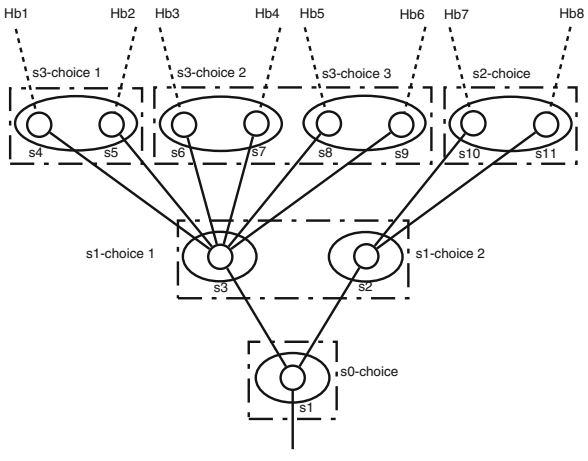

Fig. 3 Knowingly doing in a K-extended XSTIT frame

Definition 4.1 The 'knowledge of next states' (KX) property, the 'recollection of effects' (ER) property, the 'uniformity of strategies' (Unif-Str) property, and the 'static state knowledge' (K-S) property are defined as the axioms:

$$K_a X\varphi \rightarrow K_a[a\ \mathsf{xstit}]\varphi \tag{KX}$$

$$K_a[a\ \mathsf{xstit}]\varphi \rightarrow X K_a\varphi \tag{ER}$$

$$\Diamond K_a[a\ \mathsf{xstit}]\varphi \rightarrow K_a\Diamond[a\ \mathsf{xstit}]\varphi \tag{Unif-Str}$$

$$K_a\Box\varphi \leftrightarrow \Box K_a\varphi \tag{K-S}$$

The axioms express intuitive properties for the interactions of knowledge and action. Since they are discussed extensively in [12], here they will only be briefly explained.

The 'knowledge of next states' (KX) property expresses that the only way in which an agent can be certain about what holds next is by seeing to it itself.[10] In terms of the K-extended frame of Fig. 3 the axiom says that the ellipses visualizing the objective choices are always contained inside the dotted rectangles, that is, knowingly doing is closed under objective choices. Axiom (ER) expresses that if agents knowingly see to something, then they know that something is the case in the resulting state. Axiom (Unif-Str) expresses that if an agent can knowingly see to something, it knows seeing to that something is one of its causal capacities.[11] For instance: the fact that I can knowingly break the cup by throwing it on the floor implies that I know to have the causal power to break the cup. For an example concerning the absence of the implication in the opposite direction, consider the case of a blind person in a room with a light switch (see [14]): the blind person knows it has the causal power to ensure the room is sufficiently lighted, but it has no means to knowingly see to it. Finally, (K-S) says that settledness and knowledge commute. This says that agents can be uncertain about the choices of other agents and their own objective choices, but never about the static state they are in. So, if we also want to reason about uncertainty of the static states agents are in, this property is too strong. However, for the purposes of the present paper that only considers uncertainty as related to the mode of acting, we can safely assume it.

Proposition 4.1 *The axioms given in Definition* 4.1 *are all in the Sahlqvist class. Therefore, they all correspond to first order conditions on the frames of Definition* 3.2 *and can be added to the Hilbert system of Definition* 2.5 *to obtain a complete system.*

[10] This property can be refined, resulting in the two properties $K_a[A\ \mathsf{xstit}]\varphi \rightarrow K_a[a\ \mathsf{xstit}]\varphi$ for $a \in A$ and $K_a[A\ \mathsf{xstit}]\varphi \rightarrow K_a\Box[A\ \mathsf{xstit}]\varphi$ for $a \notin A$. Since these refinements are not relevant here, we do not discuss them.

[11] The stronger property $\Diamond K_a\varphi \rightarrow K_a\Diamond\varphi$ is also appropriate, but we will not discuss it here.

Also for subjective action we can establish an independence property. If objective choices of agents are independent (the Indep-G axiom of Section 2) then also knowingly exercised choices are independent.

Proposition 4.2 *Given the axioms of* XSTIT *and the axioms for knowingly doing of Definition* 4.1 *we can derive* independence of subjective choice*, which is defined as:*

$$\Diamond K_a[a\ \mathit{xstit}]\varphi \wedge \Diamond K_b[b\ \mathit{xstit}]\psi \rightarrow \Diamond(K_a[a\ \mathit{xstit}]\varphi \wedge K_b[b\ \mathit{xstit}]\psi) \quad \text{(Indep-K)}$$

Using correspondence theory we conclude the property follows. The independence of agency property from XSTIT says that intersections of choices of different agents are never empty. Now since a knowingly exercised (i.e., subjective) choice always contains at least an objective choice (axiom KX), intersections of subjective choices of different agents are also never empty.

That knowledge has an entirely different character here than in most other systems with epistemic operators, is maybe best explained through the notion of 'moment determinacy' [28]. Semantically, moment determinacy of an operator M is defined by the condition that the truth value of M is independent of the history h in dynamic states $\langle s, h\rangle$. Syntactically, moment determinacy can be defined as follows: M is moment determinate if $M\varphi \rightarrow \Box M\varphi$ is valid. An example of a moment determinate modality is 'unconditional obligation', where what an agent is obliged does not depend on what it does, or on what some other agent does.[12]

Now, in the present framework, knowledge is not moment determinate. We cannot conclude to $K_a\varphi \rightarrow \Box K_a\varphi$, because that does not hold for the substitution $[[a\ \mathsf{xstit}]\psi/\varphi]$. And this seems right: an agent's knowledge should not only depend on the static context. If we can assume that an agent knows what it does when it chooses something, what it knows depends on what choice it exercises, and not only on the static state it is in.

5 Intentionally Doing

We will discuss the concept of intentionally doing in terms of the frame of Fig. 4. The dotted ellipses represent intentional actions. We follow the same visualization pattern as before: from state s_1 the smaller closed ellipses represent the objective choices interpreting the modality $[a\ \mathsf{xstit}]\varphi$, the dotted rectangles represent the choices the agent can knowingly exercise interpreting the combined modality $K_a[a\ \mathsf{xstit}]\varphi$, and, finally, the dotted ellipses represent the choices the agent may exercise intentionally interpreting the combined modality $I_a[a\ \mathsf{xstit}]\varphi$. Now, assume the actual dynamic state is one based on

[12]Note however that, many interesting examples of obligation in deontic logic are of the kind that does depend on what agents do (if you drive a car, you need to carry your license; if you kill, you have to kill gently [21]). See [44] for a discussion on the moment determinateness of obligation.

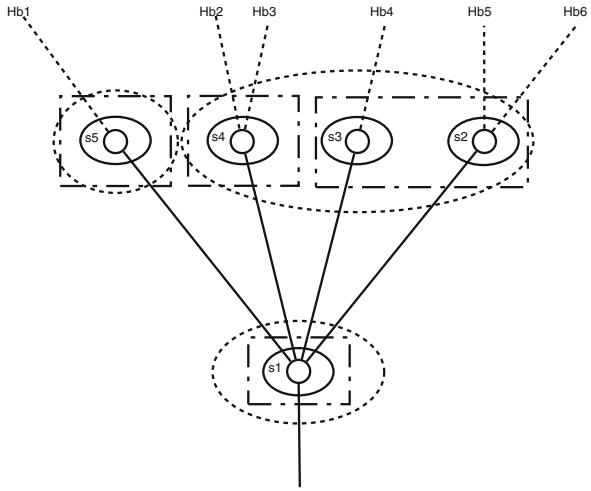

Fig. 4 Knowingly doing and intentionally doing in a KI-extended XSTIT frame

static state s_1 and a history from the bundle $Hb4$. Then, the objective choice exercised by the agent in s_1 is the one visualized by the small ellipse around s_3. However, the intentional choice exercised is the one visualized by the dotted ellipse around s_4, s_3 and s_2. So, the agent intends to be doing what holds in *all* the dynamic states based on the static states s_4, s_3 and s_2 and the histories running through them. At the same time the agent knows to be doing the action visualized by the dotted rectangle around the states s_3 and s_2. So, it knows to be doing what holds in *all* the dynamic states based on the static states s_3 and s_2. So, what it intends to be doing is also what it knows to be doing. And what it knows to be doing is also what it does (the small ellipse around s_3). But not the other way around. What it actually does is possibly *more*[13] than what it knows to be doing (on the dynamic states formed with s_3, things may hold that are not true on *all* the dynamic states formed with s_3 and s_2). And, what it knows to be doing is possibly more[14] than it intends to be doing (on the dynamic states formed with s_3 and s_2, things may hold that are not true on *all* the dynamic states formed with s_4, s_3 and s_2).

Note that in the frame of Fig. 4 from static state s_1 the agent has *two* possibilities for exercising an intentional choice. This means that intentional action in not moment determinate. Relative to the history under consideration, one of the two intentional choices is exercised. Once more this emphasizes that we do not model intention as a static mental state.[15]

[13]The exception is if s_3 and s_2 are bisimular. However, bisimularity concerns models, while here we are concerned with logic properties at the level of frames.

[14]The exception is if s_4, s_3 and s_2 are bisimular.

[15]If, in this example, we would have to associate the two possible intentional actions to a static intention holding for the state s_1, maybe we can think of this static intention here as a disjunction of the two intentions in both intentional acts.

Let us now investigate possible logics corresponding to the type of frames exemplified by Fig. 4. In particular we are interested in logical properties for the (combined) intentional action modality $I_a[a\ \mathsf{xstit}]\varphi$ and in logical properties of interactions of other modalities with this combined modality.

Since we model present directed intention in action as an S5 modality, and since the logic of the central XSTIT modality $[a\ \mathsf{xstit}]\varphi$ is KD, through correspondence theory and composition of the relations interpreting the modalities we get that the logic of the combined modality $I_a[a\ \mathsf{xstit}]\varphi$ is also KD. This mirrors that intentional action is consistent, that is, it cannot be consistent that an agent intentionally sees to it that φ and at the same time intentionally sees to it that $\neg\varphi$. The corresponding D-axiom is $\neg(I_a[a\ \mathsf{xstit}]\varphi \wedge I_a[a\ \mathsf{xstit}]\neg\varphi)$. In terms of the frame pictured in Fig. 4 this says that from any dynamic state, we can reach all the states within the dotted ellipse representing the current intentional choice by following (elements of) the composed relation $i_a \circ R_{\{a\}}$.

Now we turn to the interaction of intentional action with knowledge. First, it seems correct to assume that intentional action has its result among the states the agent *knows* to be possible next states. In XSTIT (Section 2) we can easily derive an axiom expressing directly that objective action takes effect in next states: $\Box X\varphi \rightarrow [a\ \mathsf{xstit}]\varphi$. Here, for intentional action, we then need a variant of this axiom, involving subjectively (i.e., epistemically) possible next states.

Definition 5.1 The 'intentional actions take effect in K-subjectively possible next states' (X-Eff-I) property is defined as the axiom:

$$\Box K_a X\varphi \rightarrow I_a[a\ \mathsf{xstit}]\varphi \qquad \text{(X-Eff-I)}$$

In terms of the frame visualized in Fig. 4 the property (X-Eff-I) says that dotted ellipses only contain states that are also contained in some dotted rectangle. This enforces that if it is settled that the agent knows that φ in the next state, which means the agent cannot do anything about it, the agent cannot but intend that φ holds next. In a next section we come back to this property when we discuss to what extent intentional action should imply that an intentional or subjective alternative should be possible.

Now we go to the second interaction with knowingly doing. A crucial property of intentional action seems to be that an agent only performs an intended action if that same agent performs that action knowingly.[16] If I send an email, and by doing that I do not *knowingly* cause a server to break down,

[16]However, see the book of Mele [36] for a thorough discussion on the possibly conflicting position that unconscious intentional choices are possible. Mele uses this position to argue against the claim that it follows from experimental results in neuroscience (Benjamin Libet [32]) that agents have no free will.

I clearly do not *intentionally* bring down the server by sending the email.[17] In the literature on intentional action this position was defended in [23], and discussed in e.g. [20]. Within the present framework, we can capture this property of intentionally doing by the following axiom.

Definition 5.2 The 'intentionally doing implies knowingly doing' (I ⇒ K) property is defined as the axiom:

$$I_a[a\ \mathsf{xstit}]\varphi \rightarrow K_a[a\ \mathsf{xstit}]\varphi \qquad \text{(I} \Rightarrow \text{K)}$$

In terms of the frame visualized in Fig. 4, together with property (X-Eff-I), the property (I ⇒ K) says that any dotted rectangle visualizing what the agent knows to be doing is contained entirely within the dotted ellipse visualizing what the agent intentionally does.

All the constraints considered so far can be added to the system we had so far to obtain a complete system. Henceforth, we will refer to the resulting system as 'the base system' for intentional action.

Proposition 5.1 *The axioms defined in Definitions* 5.2 *and* 5.1 *are all in the Sahlqvist class. Therefore, they all correspond to first order conditions on the frames of Definition* 3.2 *and can be added to the Hilbert system of Definition* 2.5 *to obtain a complete system.*

As for objective choices and for subjective choices, for intentional choices we can formulate and prove an independence property. And this is how it should be. If objective choices of agents are independent (the Indep-G axiom of Section 2), and if knowingly exercised choices are independent, then certainly intentional choices are independent.

Proposition 5.2 *Given the axioms of XSTIT, the axioms for knowingly doing of Definition* 4.1*, KD for intentional action, and the interaction axioms* (X-Eff-I) *and* (I ⇒ K)*, we can derive* independence of intentional choice*, which is defined as:*

$$\Diamond I_a[a\ \mathit{xstit}]\varphi \wedge \Diamond I_b[b\ \mathit{xstit}]\psi \rightarrow \Diamond(I_a[a\ \mathit{xstit}]\varphi \wedge I_b[b\ \mathit{xstit}]\psi) \qquad \text{(Indep-I)}$$

[17] One reviewer replies that maybe an agent might intentionally see to it that a flipped coin lands heads up (the coin is manipulated and has two heads, and the agent also intends heads) without knowingly seeing to it that the coin lands head up (the agent is not aware of the manipulation and believes tails is a possible outcome). However, since the agent in this scenario believes that failing to bring up heads is a possible outcome of its choice, we do not regard this as an example of intentional action, but as an example of 'attempt', for which indeed, as we discuss in Section 8, the epistemic attitude towards action is weakened. Independent of this, it is an interesting aspect of this example that we can read a 'Frankfurt style manipulation' of the coin in it (even though in Frankfurt style arguments the manipulation concerns a device interfering with an agent's neural activity). We plan to discuss such cases in a separate paper discussing and criticizing manipulation arguments against compatibilism by representing the associated scenarios in epistemic extensions of XSTIT.

We do not give the formal derivation of the property in the modal axiomatization. However, from correspondence theory, we can see that the property must follow. The independence of agency property from XSTIT says that intersections of choices of different agents are never empty. Now since an intentional choice always contains at least a knowingly exercised choice (the axioms KD for intentional action, (X-Eff-I) and (I $\Rightarrow$ K)) which in turn always contains at least an objective choice (axiom KX for knowingly doing), it follows that intersections of intentional choices of different agents are also never empty.

6 Deliberateness, Side Conditions and Side Effects

We call a choice 'deliberate' if it is the result of some form of deliberation on the side of the agent exercising the choice. It is obvious that intentional action falls in this category. In *stit* theory, deliberateness has been modeled through so called 'side conditions' [29]. These side conditions are used to interdefine deliberate and normal variants of *stit* operators. In our *xstit*-setting this interdefinability takes the following form: $[a\ \mathsf{dxstit}]\varphi \equiv_{def} [a\ \mathsf{xstit}]\varphi \wedge \Diamond X\neg\varphi$ and $[a\ \mathsf{xstit}]\varphi \equiv_{def} [a\ \mathsf{dxstit}]\varphi \vee \Box X\varphi$. The idea behind the side condition $\Diamond X\neg\varphi$ in the definition of deliberate versions of *stit* operators is that a choice can only be deliberate if there is an alternative choice that would not have guaranteed the same outcome. Or, in other words, a choice cannot be deliberate if the agent did not have an alternative, that is, if the outcome of the agent's choice was already settled. Now, of course, a similar intuition applies to intentional action. However, in the previous section side conditions did not play a role at all. Therefore, in this section we consider deliberateness and side conditions as related to intentional action. And we show that the introduction of side conditions in the definition of intentional action avoids the side effect problem for intentional action. To our knowledge, this relation between side conditions and side effects has not been suggested before.

6.1 Can we Infer Intentionality from Epistemic Conditions?

An agent deliberates with the information it has, that is, with the conditions it knows and beliefs to be true about its environment and its capacities to change its environment. This means that deliberateness of an intentional action is about the interaction with the agent's epistemic attitude towards action. First we briefly discuss what we think is a wrong way to model this interaction.

Let us go back to what we already said about the interaction in the previous section. We said that intentionally doing implies knowingly doing (axiom (I $\Rightarrow$ K)). But what about the other side of the coin? To what extent can we conclude that an agent intentionally does something on the basis of what it knows to be doing? Prima facie there is something to say for the position that a logical inference in this direction makes sense. One might even go as far as

to say that there is no distinction between knowingly doing and intentionally doing: if an agent knowingly does something then it does it intentionally, since given the circumstance that it knows what it is doing, if what it does is not what it intends to do, it should have chosen to do something else. However, this is too simple a view for several reasons. First of all, it is for good reasons that the legal literature distinguishes between purposeful acts and knowingly performed acts, and attaches different levels of culpability to them, as discussed in the introduction. Second, if we argue for a logical inference from knowingly doing to intentionally doing following the prima facie viewpoint just mentioned, we should somehow account for the side condition that "it could have chosen to do something else". This means then that we would have to consider to add one of the following principles as axioms to the system:

$$K_a[a\ \mathsf{xstit}]\varphi \wedge \Diamond K_a[a\ \mathsf{xstit}]\neg\varphi \rightarrow I_a[a\ \mathsf{xstit}]\varphi$$

$$K_a[a\ \mathsf{xstit}]\varphi \wedge \Diamond K_a\neg[a\ \mathsf{xstit}]\varphi \rightarrow I_a[a\ \mathsf{xstit}]\varphi$$

$$K_a[a\ \mathsf{xstit}]\varphi \wedge \Diamond\neg K_a[a\ \mathsf{xstit}]\varphi \rightarrow I_a[a\ \mathsf{xstit}]\varphi$$

The first property models the side condition as "the agent can knowingly ensure $\neg\varphi$", the second property models it as "the agent can knowingly refrain from φ", and the third models it as "the agent can perform an action different from knowingly ensuring φ". These are all different possibilities (ascending in strength) for adding to the base system that from an agent's *epistemic* attitude towards its action and action possibilities we can derive the *intentionality* of its action.

However, we should be sceptic about the appropriateness of these principles. First, in several respects, the operator $I_a[a\ \mathsf{xstit}]\varphi$ is already very strong. In the next subsection we will argue that we should weaken it, by which we avoid unwanted derivations concerning side effects. Second, it appears to me that by adding principles like these we confuse the logic of an observer with the logic of the agent acting. It is as if an abductive reasoning principle of an observer agent ("it must be that that agent does this intentionally, otherwise it would have done something else") is turned into a deductive reasoning principle for the acting agent itself.

Let us now come back to the direction of inference leading from an intentional towards an epistemic attitude towards action. In particular, let us consider the converse directions of the three properties above. In our opinion this direction of inference is very intuitive. Part of it is, of course, already incorporated by the axiom ($\mathrm{I} \Rightarrow \mathrm{K}$). But also, the circumstance of the agent knowing that an alternative action is possible (the side condition) seems a necessary condition for genuine intentional action: it reflects that intentional

action presupposes a deliberation effort on the part of the agent in the sense that it at least has considered to refrain from the intentional act. In the next section we define a version of the operator for intentional action that satisfies the converse direction of the third property above, thereby also avoiding the well-known 'side effect problem'.

6.2 Deliberate Intentional Action

Based on the notion of deliberative *stit* from standard *stit* theory [29], and based on the observation in the previous section that in the interaction between the intentional and epistemic attitude towards actions, the direction of inference should be from the former to the latter, in this section we come to a definition of deliberate intentional action. The side condition will be an epistemic condition, reflecting that the deliberation leading to the intentional act is based on what an agent knows about its capacity to bring about changes in its environment. The definition then says that an agent's act qualifies as intentional if what the agent does is what it intends to do and if it also had the possibility to knowingly refrain from what it does. Formally, we get the following definition.

Definition 6.1 The modality 'agent a deliberately intentionally sees to it that next φ', denoted $[a\ \mathsf{xint}]\varphi$ is defined as:

$$[a\ \mathsf{xint}]\varphi \equiv_{def} I_a[a\ \mathsf{xstit}]\varphi \wedge \Diamond\neg K_a[a\ \mathsf{xstit}]\varphi$$

6.3 Side Effects and Double Effects

In terms of static intentions and beliefs, the side effect problem is the problem of whether or not intentions should be closed under knowledge (or belief). For instance, if I intend to go to the dentist, and I know (or belief) going implies having pain, it should not follow that I intend to have pain [16]. Or, from an agent's intention to bomb a terrorist's home, and its believe that the house is next to a school it should not follow that the agent intends to bomb the school. Of course it is possible to argue that the agent does intend the pain and that it does intend to bomb the school, since the agent maybe could have known how to avoid these situations. But then we again make an inference in what we regarded to be the wrong direction: from the epistemic attitude towards an intentional attitude.

In the setting of this paper, concerning modes of acting, the side effect problem gets a different flavor. Here the side effect problem concerns the inappropriateness of, for instance, the inference from the premisses that an agent intentionally visits the dentist, and all ways of knowingly visiting the dentist are also ways to knowingly get pain, to the conclusion that the agent

intentionally gets pain.[18] Formally this means that we do not want the following closure axiom to be derivable:

$$I_a[a\ \mathsf{xstit}]\varphi \wedge \Box(K_a[a\ \mathsf{xstit}]\varphi \rightarrow K_a[a\ \mathsf{xstit}]\psi) \rightarrow I_a[a\ \mathsf{xstit}]\psi \qquad \text{(SE)}$$

However, using correspondence theory, it is not difficult to see that the axiom (SE) is derivable in the base system of Section 5 (dotted boxes are contained in dotted ellipses, so if φ is true in all points of a dotted ellipse, and if for all dotted boxes inside the ellipse it holds that if φ holds for all points in the dotted box, then ψ holds for all points in the box, then ψ holds at all points inside the ellipse).

A first possible reaction to this problem can be that even though the property (SE) has a counter intuitive aspect, it is too strong a requirement for the logic to demand that such a weak property should not hold. The property is rather weak, since it says that *only* if *any* possible way to knowingly do φ is also a way of doing ψ, we derive the intentionality of the side effect. In many situations where side effects play a role, this requirement is not met. Let us take a closer look at the dentist's example. Assume that $K_a[a\ \mathsf{xstit}](d \wedge p)$ = "knowingly visiting the dentist and have pain" and $K_a[a\ \mathsf{xstit}]d$ = "knowingly visiting the dentist". Clearly $K_a[a\ \mathsf{xstit}](d \wedge p)$ is a way of doing $K_a[a\ \mathsf{xstit}]d$ and we have as a logical fact that $\Box(K_a[a\ \mathsf{xstit}](d \wedge p) \rightarrow K_a[a\ \mathsf{xstit}]d)$. Now assume the agent intentionally visits the dentist but has no other way of doing that than by going to the dentist and *risking* pain. Formalizing the situation of the agent, we come to the set of formulas $Th = \{I_a[a\ \mathsf{xstit}]d,\ K_a[a\ \mathsf{xstit}](d \wedge p),\ \neg\Diamond K_a[a\ \mathsf{xstit}](d \wedge \neg p)\}$, that is, (1) the agent intentionally visits the dentist, (2) knowingly visits the dentist in a way that causes him pain, and (3) does not know a way of visiting the dentist ensuring that there is no pain. Clearly we derive that the agent knowingly sees to it that it has pain ($Th \vdash K_a[a\ \mathsf{xstit}]p$). But, we cannot derive that the agent intentionally sees to it that it has pain ($Th \nvdash I_a[a\ \mathsf{xstit}]p$). The axiom (SE) does not apply here. This is because in this situation all ways of knowingly visiting the dentist are ways of knowingly *risking* pain. For the axiom (SE) to apply it should have been the case that all ways of knowingly visiting the dentist are ways of knowingly *ensuring* that there will be pain. Of course, one can object that it is strange that in this modeling example the agent actually knowingly 'ensures' it has pain while apparently it could have made the choice for only 'risking' the pain. But then we are again back in the situation of the previous section where we consider the question whether it makes sense to derive intentional attitudes from epistemic attitudes.

[18] See also Bratman's example of the 'marathon man' where the side effect is wearing down his sneakers [8], and Harman's example of the sniper where the side effect is alarming the enemy with the blast of the firing rifle [25].

A second, more appropriate reaction to the derivability of (SE) in the base system of Section 5 is to admit that we have to weaken the properties of its intentional action operator $I_a[a\ \mathsf{xstit}]\varphi$. Now, the kind of weakening we need to avoid (SE) is the one provided by the definition of 'deliberate intentional action' in Definition 6.1. We might say that just like the defined notion of deliberate *stit* avoids logic properties like $[a\ \mathsf{xstit}]\top$, the defined notion of 'deliberate intentional action' avoids the property (SE) (property (SE) does not hold with the $I_a[a\ \mathsf{xstit}]$ operators replaced by $[a\ \mathsf{xint}]$ operators).

We believe, the solution to the side effect problem provided by Definition 6.1 is intuitive. The core of the solution is that the side effect problem is no problem if we look at an intentional act as the outcome of a process of deliberation. For instance, imagine that an intention is the result of a process of estimating maximal expected utility (MEU). In the weighing process the side effects of actions of course have been accounted for. The positive utility of the main effect outweighs the negative utility of the side effect. In such a situation it makes sense to say that the agent intends the effect, and thus the action associated with the main utility, but does not intend the effect associated with the side effect. By considering side conditions, here we just introduce enough of this weighing process into the logic to solve the problem and not derive unwanted conclusions.

There is however one possible caveat. One might claim that even in case the agent knows of no way to refrain from seeing to it that φ, it is possible that the agent intentionally see to it that φ, for the reason that if it *would have had* the possibility to refrain from its action it *would have done so*. Note that in this argumentation for the intentionality of an action, the side condition is entirely hypothetical, while in the operator for deliberate intentional action in Definition 6.1 the side condition is a concretely existing possibility the agent knows about.

If we accept this view on intentional action, Definition 6.1 no longer applies, since it defines deliberateness relative to concrete alternative possibilities. It seems then that if we want to take this suggestion seriously, we would have to introduce a theory of model update in this setting, to model hypothetical or counter factual situations (see [41] for the relation between counter factuals and updates). However, in the present setting this step is not necessary. We can distinguish between alternative actions and alternative choices. If we make this distinction, then we can claim that alternative choices are always possible, even though they can be unsuccessful and thus do not lead to an alternative action or outcome. The issue of success of choices will be discussed in Sections 7 and 8.

As a final note for this section we observe that there is an interesting relation between the side effect problem and the ethical doctrine of the double effect[19]

[19]Thanks to Thomas Müller for pointing to this connection.

(going back to Thomas d'Aquino). Very roughly, the doctrine says that side effects are excusable if the primary intention of an act was aimed at something that is considered to be 'good' (think about trying to justify collateral damage in warfare). This concept has a strong deontological aspect,[20] and we leave its discussion to another paper where we plan to consider the operators presented here as building blocks of a deontic action logic.

7 Non-successful Choice

Axiom (I $\Rightarrow$ K) ensures that an intended action is also a knowingly performed action. Knowingly performed actions are successful actions in the sense that the actual dynamic state is among the dynamic states in the epistemic equivalence class (game theorists would say: 'the information set'). Axiomatically, we have that from (I $\Rightarrow$ K) $I_a[a\ \mathsf{xstit}]\varphi \rightarrow K_a[a\ \mathsf{xstit}]\varphi$ and from the veridicality of knowledge we derive that $I_a[a\ \mathsf{xstit}]\varphi \rightarrow [a\ \mathsf{xstit}]\varphi$. Then with axioms (*Ags*-XSett) and (C-Mon) we derive that $I_a[a\ \mathsf{xstit}]\varphi \rightarrow X\Box\varphi$. Finally, with standard normal modal reasoning, we arrive at $I_a[a\ \mathsf{xstit}]\varphi \rightarrow X\varphi$. This derived theorem says that intentional action is successful: what an agent intentionally does is also what happens.

But, for intentional action this is often simply not the case. What we intentionally do, is not necessarily what happens. For instance, the environment (including other agents) may behave unexpectedly, causing the actual action to be completely different than the intended action. It can even be the case that we intentionally perform an action and achieve the opposite. For instance, we perform the intentional action of securing a precious vase that is too close to the edge of a table, and by doing so, we cause it to fall to the ground.

The system built so far can be adapted to allow for the fact that intentional action is not successful, in an elegant way. What we need to do is to allow for a possible discrepancy between what an agent believes to be doing and what objectively happens. So, what we need to do, is to weaken the notion of knowingly doing to its *belief* analog. We do not have a good word for the notion thus resulting; maybe 'believing to do' is the phrase that comes closest.

Let us explain the concept of believing to do and the way it allows intentional action to be non-successful in terms of an example frame. In Fig. 5 we see a situation where in s_1 an agent has two objective choices which are non-deterministic due to a second agent -not pictured- that has the possibility to choose simultaneously. Assume that the actual dynamic state is one based on s_1 and one of the histories of the bundles $Hb5$ or $Hb6$. Now, also assume that the agent's intended choice is the one visualized by the dotted ellipse around s_3 and s_4. Finally, assume that the agent implements this intended choice by believing to exercise the choice visualized as the dotted rectangle around s_3 (we assume it does not implement its intended choice by believing to exercise the

[20] Going against a purely consequentionalist view.

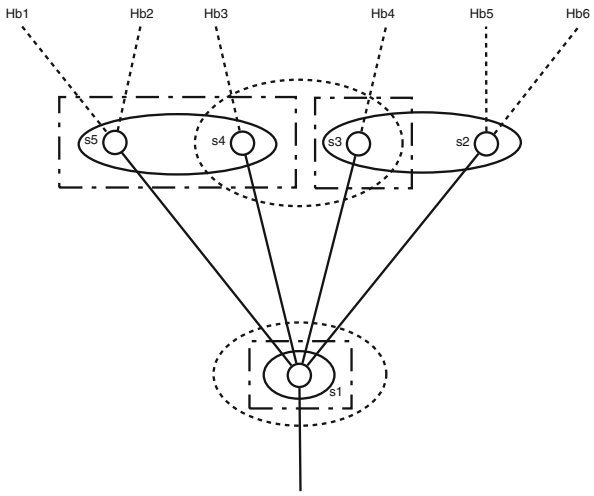

Fig. 5 Unsuccessful action in a BI-extended XSTIT frame

choice represented by the dotted rectangle around s_5 and s_4, since it considers it possible this results in s_5, which is not according to the intention in its choice). Now this agent is in for a surprise. The action it believes to do, is not the action it really performs. The agent believes to end up in s_3, but it ends up in s_2 due to unexpected choice interference of the other agent.[21] So, its intended action is unsuccessful.

The general semantic picture is thus that we want to allow for the situation where the actual dynamic state is not among the dynamic states that are epistemically accessible. Let us now very briefly present the resulting logic. We change the knowledge operator in a belief operator, resulting in the following syntax.

Definition 7.1 We extend the syntax of Definition 2.1 with an operator for belief and intentional action, resulting in:

$$\varphi \ldots := p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid [A \text{ xstit}]\varphi \mid X\varphi \mid B_a\varphi \mid I_a\varphi$$

Definition 7.2 The class of general BI-extended XSTIT frames consists of frames $\mathcal{F} = \langle S, H, R_X, R_\Box, \{R_A \mid A \subseteq Ags\}, \{b_a \mid a \in Ags\}, \{i_a \mid a \in Ags\}\rangle$ such that:

– $\langle S, H, R_X, R_\Box, \{R_A \mid A \subseteq Ags\}\rangle$ is an XSTIT-frame

[21] We consider the situation where a choice is unsuccessful due to unexpected simultaneous choice interference of other agents (or nature) to be the typical one. We take the viewpoint that agents can never be mistaken about the choice they exercise themselves. They can however be mistaken about which *action* they do, because they can be mistaken about simultaneous choice exertion of other agents (or nature).

- The b_a are epistemic accessibility relations over dynamic states obeying seriality and positive and negative introspection (corresponding to the modal frame class KD45).
- The i_a are intentional accessibility relations over dynamic states obeying seriality, transitivity and euclidicity (corresponding to the modal frame class KD45).

Definition 7.3 The clause for the truth condition of belief is:

$$\mathcal{M}, \langle s, h\rangle \models B_a\varphi \Leftrightarrow \langle s, h\rangle b_a \langle s', h'\rangle \text{ implies that } \mathcal{M}, \langle s', h'\rangle \models \varphi$$

In the logic we now have KD45 instead of S5 for the individual epistemic operators. For the interaction axioms corresponding to appropriate conditions on the relations $R_\Box$, R_A, R_X, b_a and i_a (corresponding to particular subclasses of the general frames of Definition 7.2), it is more difficult to find appropriate candidates. We cannot simply turn all the axioms for knowingly doing in Section 4 into belief equivalents. However, the axioms (Rec-eff), (Unif-Strat) and (B-S) do have belief analogs.

Definition 7.4 The 'B-recollection of effects' (B-ER) property, the 'B-uniformity of strategies' (B-Unif-Str) property, and the 'static state belief' (B-S) property are defined as the axioms:

$$B_a[a\ \mathsf{xstit}]\varphi \rightarrow X B_a\varphi \qquad \text{(B-ER)}$$

$$\Diamond B_a[a\ \mathsf{xstit}]\varphi \rightarrow B_a\Diamond[a\ \mathsf{xstit}]\varphi \qquad \text{(B-Unif-Str)}$$

$$\Box B_a\varphi \rightarrow B_a\Box\varphi \qquad \text{(B-S)}$$

For knowingly doing in terms of knowledge, in Definition 4.1 we had the KX property. We explained in Section 4 that in terms of the frames of Fig. 3 this says that the ellipses visualizing the objective choices are always contained inside the dotted rectangles visualizing the subjective choices, that is, knowingly doing is closed under objective choices. But that is exactly the property we do not want here, to allow for a discrepancy between what an agent believes to be doing and what it actually does. So here, what one believes to be doing is not closed under the objective causal capabilities one has. In Fig. 5 this is visualized by the ellipse around s_2 not being contained in the dotted rectangle around s_3. So, in the example pictured by the frame (we assumed the actual history, the one that we evaluate truth of formulas against, is one in the bundle Hb6 and that relative this history the agent believes to be exercising the choice represented by the dotted rectangle around s_3), in s_1 the agent believes it has the power to ensure the conditions of the dynamic states based on s_3, but, in reality, it ends up in s_2, satisfying the possibly different conditions in this state.

Finally, and most importantly, we also need new versions of the axioms concerning the interaction of intention and the epistemic operator. We get:

Definition 7.5 The 'intentional actions take effect in B-subjectively possible next states' (BX-Eff-I) property, and the 'intentionally doing implies believing to do' (I ⇒ B) property are defined as the axioms:

$$\Box B_a X\varphi \rightarrow I_a[a\ \mathsf{xstit}]\varphi \qquad \text{(BX-Eff-I)}$$

$$I_a[a\ \mathsf{xstit}]\varphi \rightarrow B_a[a\ \mathsf{xstit}]\varphi \qquad \text{(I} \Rightarrow \text{B)}$$

Proposition 7.1 *The axioms defined in Definitions* 7.4 *and* 7.5 *are all in the Sahlqvist class. Therefore, they all correspond to first order conditions on the frames of Definition* 7.2 *and can be added to the Hilbert system of Definition* 2.5 *to obtain a complete system.*

In terms of the frame visualized in Fig. 5 the property (BX-Eff-I) says that dotted ellipses only contain states that are also contained in some dotted rectangle. This captures that the agent must belief that the effects of intentional actions are actually possible for the agent to achieve in the next state. And in terms of the same frame, the property (I ⇒ B) says that any dotted rectangle visualizing what the agent believes to be doing is contained entirely within the dotted ellipse visualizing what the agent intentionally does. Together the properties capture that agents can only perform intentional actions that they can belief to be doing.

Mele [35] gives the example of a person absolutely believing to win a lottery, apparently having completely misplaced confidence in her ability to predict the outcome. The question posed is whether or not winning the lottery is an intentional action in case the person wins, despite the odds. In our view, the answer to the question whether or not this is an intentional action should not depend on the outcome of the lottery. In the formalization given in the present section, the winning of the lottery is an intentional action that is successful. If this agent would not have won, which would have been far more likely, she would still have performed an intentional action. However, to her surprise, and her surprise only, the action was unsuccessful.

We conclude this section with the claim that in the logic with knowledge replaced by belief, we indeed no longer derive that intentional action is necessarily successful. We do not have that from *(I-B)* $I_a[a\ \mathsf{xstit}]\varphi \rightarrow B_a[a\ \mathsf{xstit}]\varphi$ we derive that $I_a[a\ \mathsf{xstit}]\varphi \rightarrow [a\ \mathsf{xstit}]\varphi$, because belief is not like knowledge veridical.

8 Attempt and Future Research

In Section 5 we started with the axiom (I ⇒ K) saying that intentionally doing implies knowingly doing. In Section 7 we weakened the property to

(I ⇒ B) saying that intentional action implies that one *believes* to do an action 'implementing' the intention in the action. Here we want to briefly consider what happens if we weaken the relation between the two attitudes towards action even further. This means we come in the territory of the notion of 'attempt'.[22] One can argue that an important characteristic of attempt is that an agent does not fully know or fully belief that the outcome will be as it intends. In particular, besides its belief in the possibility of success, it believes in the possibility of failure. If we accept this as part of the definition of attempt,[23] we cannot model it by the intentional action operators defined in Sections 5 and 7. In particular, it follows that if we want to model attempt by an operator $[a\ \mathsf{xatt}]\varphi$, the following minimal requirements apply.

Definition 8.1 The 'attempt implies knowing the possibility to fail' (A ⇒ K-Pos-Fail) property and the 'attempt implies knowing the possibility to succeed' (A ⇒ K-Pos-Succ) property are defined as the axioms:

$$[a\ \mathsf{xatt}]\varphi \rightarrow K_a\neg[a\ \mathsf{xstit}]\varphi \qquad \text{(A ⇒ K-Pos-Fail)}$$

$$[a\ \mathsf{xatt}]\varphi \rightarrow K_a\neg[a\ \mathsf{xstit}]\neg\varphi \qquad \text{(A ⇒ K-Pos-Succ)}$$

However, at this point it is not clear yet if on the basis of these axioms we can build a logic of attempt just as we did for intentional action in the previous sections. Indeed the axioms of Definition 8.1 can be read as weakening the epistemic attitude in the relation with intention as expressed by (I ⇒ K) and (I ⇒ B). However, it seems clear that these minimal requirements are not sufficient for a sensible notion of attempt. The point is that there is a wide range of actions that fall under the definition of attempt if we stick to only the above two properties to characterize the relation between the epistemic attitude and the intentional attitude in action. One extreme is exemplified by the nuclear explosion example of Mele and Moser [37]. Some agent can prevent a nuclear explosion by typing a 10 digit code it does not know. It tries anyway. The above requirements are met, because this agent knows there is a possibility it succeeds and there is a possibility that it fails. So this would count as an attempt. But then, under this definition, practically any intentional action would qualify as an attempt, because the properties (A ⇒ K-Pos-Fail) and (A ⇒ K-Pos-Succ) are very weak.

What seems to be missing from this view on attempt is that if an agent can do things is several ways, an attempt is never a way of performing the intended action that implements the agent's intention less *likely* than some

[22]The notion of attempt studied here is different from the one studied by Lorini and Herzig [33]. Lorini and Herzig consider attempts relative to action types α, and see them as the mental counterparts of a potentially performed bodily movement α of the agent.

[23]Note that we have a subjective notion of attempt in mind. In an objective notion of attempt an agent does not necessarily belief that an attempt can go wrong; the fact that it can go wrong can be objectively true and at the same time not believed by the agent.

other way to implement it. It seems simply absurd to qualify any action that counts as a way to perform an intended action while satisfying the conditions of Definition 8.1 as an attempt; only the 'best' ways to perform it, that is, the ways most likely yielding the right result, should count as attempts. However, this brings us in the territory of probabilistic reasoning. How to combine *stit* logics with probabilistic reasoning is an entirely new subject of study that we leave to future research.

9 Conclusion

We have presented an *xstit* logic analysis of intentional action. We have discussed how by considering side conditions in its formalization we can avoid that intentional action is closed under knowledge about side effects. Also we have shown how to represent intentional action that is possibly not successful. We argued that the distinction between successful and non-successful action only makes sense if there can be a distinction between what agents belief to do and what they actually do. If these coincide there is success. If these do not coincide, there is failure. Finally we discussed the possibility of weakening the relation between the epistemic attitude and intentional attitude in action even further to account for a model of 'attempt'.

On no grounds we can pretend to have shown that certain axioms are definitely valid for reasoning about intentional action. And even less we can pretend to have given all the axioms that make sense for reasoning about intentional action. But what we do hope to have convinced the reader of is that the semantic framework put forward here is suited to study the logical properties of intentional action.

In the introduction it was explained that the main motivation for this work comes from the legal literature. And in Section 4 it was mentioned that we plan to consider the operators studied here as the building blocks of a deontic logic, in the same way as in [12] the concept of knowingly doing is taken as the basis for studying the legal concept of 'mens rea'.

Several examples from the philosophical literature had to be left unexplained. For instance there is Davidson's example of the wild pigs [17] and Chisholm's example of the murderous nephew [15]. Although most aspects of these examples we can already model in the present framework, it seems that these examples call for an intrinsically extensive game view on the matter where action possibly involves an arbitrary number of basic actions to be performed subsequently (for instance, Davidson's example concerns the intentionality of a *complex* action composed of an unsuccessful shooting action followed by an unlikely environmental event compensating it). This means the present theory should be generalized in the direction of [11]. See also [30] for an ATL-based approach to defining intention in an extensive game setting.

Finally there is the formalization of the notion of 'moral luck' [38, 45]. One way in which an agent can be said to be morally lucky is when the intention of his action is bad, but circumstances cause that the action does not work out

as badly as intended. It is clear we can represent aspects of this in the present framework.

We are aware of the fact (not only because it was explicitly pointed out by a reviewer) that the formal analysis of intentional action proposed in this paper does not address many of the subtle and fundamental issues reported in the very extensive philosophical literature on the subject. Indeed the whole point of a formal analysis is to shed new light on these issues. However, it is not that this paper tries to hide or ignore the many points raised in the philosophical literature. The main reason that many of them are not addressed here is that this paper does not pretend to be making anything more than a start with such a formalization.

Acknowledgements Thanks to Rosja Mastop and Thomas Müller and the anonymous referees for all their valuable suggestions and comments.

References

1. Alur, R., Henzinger, T., & Kupferman, O. (1997). Alternating-time temporal logic. In *Proceedings of the 38th IEEE symposium on foundations of computer science* (pp. 100–109). Florida.
2. Alur, R., Henzinger, T., & Kupferman, O. (2002). Alternating-time temporal logic. *Journal of the ACM, 49*(5), 672–713.
3. Anscombe, G. (1963). *Intention* (2nd ed.). Ithaca, NY: Cornell University Press.
4. Balbiani, P., Gasquet, O., Herzig, A., Schwarzentruber, F., & Troquard, N. (2008). Coalition games over Kripke semantics: expressiveness and complexity. In C. Dègremont, L. Keiff, & H. Rückert (Eds.), *Dialogues, logics and other strange things, essays in honour of Shahid Rahman* (pp. 5–26). College Publications.
5. Belnap, N., & Perloff, M. (1988). Seeing to it that: A canonical form for agentives. *Theoria, 54*(3), 175–199.
6. Belnap, N., Perloff, M., & Xu, M. (2001). *Facing the future: Agents and choices in our indeterminist world*. Oxford.
7. Bratman, M. (1984). Two faces of intention. *Philosophical Review, 93*, 375–405.
8. Bratman, M. (1987). *Intention, plans, and practical reason*. Cambridge Massachussetts: Harvard University Press.
9. Broersen, J. M. (2008). A logical analysis of the interaction between 'obligation-to-do' and 'knowingly doing'. In L. v. d. Torre & R. v. d. Meyden (Eds.), *Proceedings 9th international workshop on Deontic Logic in Computer Science (DEON'08). Lecture Notes in Computer Science* (Vol. 5076, pp. 140–154). Springer.
10. Broersen, J. M. (2009a). A complete STIT logic for knowledge and action, and some of its applications. In M. Baldoni, T. C. Son, M. van Riemsdijk, & M. Winikoff (Eds.), *Declarative Agent Languages and Technologies VI (DALT 2008). Lecture Notes in Computer Science* (Vol. 5397, pp. 47–59).
11. Broersen, J. M. (2009b). 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* (pp. 484–487). Washington, DC: IEEE Computer Society.
12. Broersen, J. M. (2011). Deontic epistemic *stit* logic distinguishing modes of Mens Rea. *Journal of Applied Logic, 9*(2), 127–152.

13. Broersen, J. M., Herzig, A., & Troquard, N. (2006). A STIT-extension of ATL. In M. Fisher (Ed.), *Proceedings Tenth European Conference on Logics in Artificial Intelligence (JELIA'06)*. *Lecture Notes in Artificial Intelligence* (Vol. 4160, pp. 69–81). Springer.
14. Broersen, J. M., Herzig, A., & Troquard, N. (2007). A normal simulation of coalition logic and an epistemic extension. In D. Samet (Ed.), *Proceedings Theoretical Aspects Rationality and Knowledge (TARK XI), Brussels* (pp. 92–101). ACM Digital Library.
15. Chisholm, R. (1966). Freedom and action. In K. Lehrer (Ed.), *Freedom and determinism*. Random House.
16. Cohen, P., & Levesque, H. (1990). Intention is choice with commitment. *Artificial Intelligence, 42*(3), 213–261.
17. Davidson, D. (1980). *Essays on actions and events*. Oxford: Clarendon Press.
18. Dubber, M. D. (2002). *Criminal law: Model penal code*. Foundation Press.
19. Emerson, E. (1990). Temporal and modal logic. In J. v. Leeuwen (Ed.), *Handbook of theoretical computer science, volume B: Formal models and semantics* (Chapt. 14, pp. 996–1072). Elsevier Science.
20. Falvey, K. (2000). Knowledge in intention. *Philosophical Studies, 99*, 21–44.
21. Forrester, J. (1984). Gentle murder, or the adverbial Samaritan. *Journal of Philosophy, 81*(4), 193–197.
22. Goldman, R. P., & Boddy, M. S. (1996) Expressive planning and explicit knowledge. In *Proceedings of the 3rd international conference on Artificial Intelligence Planning Systems (AIPS-96)* (pp. 110–117). AAAI press.
23. Hampshire, S. (1981). *Thought and action*. University of Notre Dame Press.
24. Harel, D., Kozen, D., & Tiuryn, J. (2000). *Dynamic logic*. The MIT Press.
25. Harman, G. (1976). Practical reasoning. *Review of Metaphysics, 29*, 431–463.
26. Herzig, A., & Schwarzentruber, F. (2008). Properties of logics of individual and group agency. In C. Areces & R. Goldblatt (Eds.), *Advances in modal logic* (Vol. 7, pp. 133–149). College Publications.
27. Herzig, A., & Troquard, N. (2006). Knowing how to play: Uniform choices in logics of agency'. In G. Weiss & P. Stone (Eds.), *5th International joint conference on Autonomous Agents & Multi Agent Systems (AAMAS-06), Hakodate, Japan* (pp. 209–216). ACM Press.
28. Horty, J. (2001). *Agency and deontic logic*. Oxford University Press.
29. Horty, J. F., & Belnap, N. D. (1995). The deliberative stit: A study of action, omission, and obligation. *Journal of Philosophical Logic, 24*(6), 583–644.
30. Jamroga, W., Hoek, W. V. D., & Wooldridge, M. (2005) Intentions and strategies in game-like scenarios. In *Progress in artificial intelligence: Proceedings of EPIA 2005. Lecture Notes in Artificial Intelligence* (Vol. 3808, pp. 512–523). Springer.
31. Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis, 63*, 190–193.
32. Libet, B. (2004). *Mind time: The temporal factor in consciousness*. Cambridge, MA: Harvard University Press.
33. Lorini, E., & Herzig, A. (2008). A logic of intention and attempt. *Synthese, 163*(1), 45–77.
34. Lorini, E., Troquard, N., Herzig, A., & Castelfranchi, C. (2007). Delegation and mental states. In *Proceedings of sixth international joint conference on Autonomous Agents and Multiagent Systems (AAMAS'07)*. New York: ACM Press.
35. Mele, A. (Ed.) (1997). *The philosophy of action*. Oxford: Oxford Univeristy Press.
36. Mele, A. (2009). *Effective intentions: The power of conscious will*. Oxford: Oxford Univeristy Press.
37. Mele, A., & Moser, P. (1994). Intentional action. *Nous, 28*, 39–68.
38. Nagel, T. (1979). Moral luck. In *Mortal questions* (pp. 24–38). Cambridge University Press.
39. Pauly, M. (2002). A modal logic for coalitional power in games. *Journal of Logic and Computation, 12*(1), 149–166.
40. Pratt, V. (1976). Semantical considerations on Floyd-Hoare logic. In *Proceedings 17th IEEE symposium on the foundations of computer science* (pp. 109–121). IEEE Computer Society Press.
41. Ryan, M., & Schobbens, P.-Y. (1997). Counterfactuals and updates as inverse modalities. *Journal of Logic, Language and Information, 6*(2), 123–146.

42. Searle, J. (1983). *Intentionality*. Cambridge: Cambridge University Press.
43. Semmling, C., & Wansing, H. (2008). From BDI and stit to bdi-stit logic. *Logic and Logical Philosophy, 17*, 185–207.
44. Wansing, H. (2001). Obligations, authorities, and history dependence. In H. Wansing (Ed.), *Essays on non-classical logic* (pp. 247–258). World Scientific.
45. Williams, B. (1982) Moral luck. In *Moral luck* (pp. 20–39). Cambridge University Press.