Solving Bratman's video game puzzle in two formalisms

Robert Trypuz[1,2]
trypuz@loa-cnr.it
Emiliano Lorini[3,4]
emiliano.lorini@istc.cnr.it
Laure Vieu[2,3]
vieu@irit.fr

[1] *Università di Trento*
[2] *Laboratorio di Ontologia Applicata, ISTC, CNR, Trento*
[3] *Institut de Recherche en Informatique de Toulouse, CNRS*
[4] *Instituto di Scienze e Tecnologie della Cognizione, CNR, Roma*

Abstract

The aim of this paper is to propose two formal solutions to a well-known problem in the recent philosophical literature called the Video Game puzzle. The Video Game puzzle was introduced by Michael Bratman in the 80's and then became a challenge for any theory of intention and action. A few approaches to this puzzle have been discussed in philosophy, but to the best of our knowledge, no solution has been spelled out in a logical formalism.

The first solution proposed in this paper is based on a rich first-order theory of time, actions and intentions in which individual actions are reified. The second one is based on a multi-modal logic of intention and action in which the concept of attempt is formalized by means of a modal operator for attempt.

Keywords: action, intention, rationality

1 Introduction

In [3, 4], Bratman introduced the Video Game (VG) puzzle which then became a challenge for any theory of intention and action.

The VG scenario is composed of two similar video games. In each of them the player's task is to guide a missile into a certain target. It is assumed that the player is skilled in playing these games and is ambidextrous, so he can play one game with each hand. The player knows that the two games are so linked that it is impossible to hit both targets; if both targets are about to be hit simultaneously the machines shut down. He also knows that there is a reward for hitting either target. Taking into account that it is difficult to hit either target, the player decides to try to guide missile 1 towards target 1 and simultaneously to try to guide missile 2 towards target 2, seeing the risk of shutting down the machines as outweighed by the increase in his chances of hitting a target. Eventually, the player hits target 1 and fails to hit target 2 [3].

Because the player's behavior, i.e., his guiding two missiles towards two targets, is assumed to be rational, any theory of action and intention should describe the intention(s) under which these actions are performed. And here comes the challenge, because the most intuitive and widely accepted thesis linking actions and intentions, called the Simple View, seems to be falsified by the VG scenario, given the strong consistency hypothesis, i.e., that for a rational agent his intentions must be consistent with his beliefs. The Simple View states that *if A is an intentional (rational) action of an agent a, then a intends to do A*. According to this view, if I hit target 1 intentionally, then I intend to hit target 1. But as Bratman argued: "Given the symmetry of the case I must also intend to hit target 2. But given my knowledge that I cannot hit both targets, these two intentions fail to be strongly consistent"[3]. This reasoning led Bratman to conclusion that "Since my relevant intentions in favour of hitting target 1 are the same as those in favour of hitting target 2, I have neither intention. So the Simple View is false" [3].

The VG puzzle then takes the form of a question: *with what intention(s) the successful action of "hitting target 1" and the unsuccessful one of "hitting target 2" have been done*?

The philosophical literature describes different approaches to this question. Here we examine how two different logical formalisms are able to answer and solve the puzzle. The first answer is based on a rich first-order theory of time, actions and intentions in which individual actions are reified [25].

The second one is based on a multi-modal logic of intention and action in which types of attempts are modal operators [15, 14]. The assumptions behind these two theories are quite different, but they agree in considering that actions are non-deterministic, an essential feature for modelling the VG scenario. They also agree in accepting the strong consistency hypothesis, and describe the intentions underlying the player's actions, thus solving the VG puzzle. In the remainder, we successively present the two formal frameworks and how they model the VG scenario. We conclude by briefly comparing the solutions and situating them in the relevant philosophical literature.

2 OntoSTIT+ meets the VG puzzle

2.1 Introduction to OntoSTIT+

We briefly describe here OntoSTIT+, a sorted first-order theory whose language and axiomatics is fully developed in [25]. OntoSTIT+'s domain contains agents, moments, histories, intervals, action tokens, action courses and arbitrary non-temporal entities (including agents). The language includes a set of primitive predicates Δ_+ establishing relations between individuals of the first six sorts, a set of primitive predicates, Π_+ over individuals of the last sort, intended to describe the states of affairs that hold at a moment m and a history h (in short, at an *index* m/h), a set of predicates Θ of basic action types linking action tokens to their participants (of the last sort), and two sets of predicates $\Omega\Theta$ and $\Gamma\Theta$ describing respectively the expected outcomes and the preconditions associated to an action type. Bold, lowercase letters stand for constants.

As OntoSTIT+ is the extension of a first-order equivalent of Chellas's "seeing to it that" logic of agency [5, 2], it is grounded in a branching time structure made of moments and histories (maximal chains of moments), which grasps the indeterminacy of the future. Moments are ordered by means of the precedence relation "Pre" which is a strict partial order. $Pre(m, m')$ stands for "a moment m precedes a moment m'". $PreEq(m, m') \triangleq Pre(m, m') \vee m = m'$. For any moment m, there is some history—a possible evolution of the world—h such that m is in h ($In(m, h)$) and vice versa. Intervals are bounded stretches of histories, so they are linearly ordered and convex. Any interval has a beginning and an end, which are unique. By $InI(m, i)$ we describe that a moment m is in the interval i.

Actions are considered as having a duration and as being non-deterministic. This means that action *tokens* may unfold into different action *courses* on histories that branch during their execution. Obviously, all courses of a same action token start at the same moment, but they may have different durations. $CO(c,t)$ stands for "an action course c is a course of an action token t". Each action course runs through a particular interval i ($RT(c,i)$) and lies on some history h ($LOn(c,h)$). The agent of each action token exists and is unique. $AgO(a,t)$ stands for "a is agent of the action token t". We assume that one and only one basic action type A_i from Θ applies to each action token. An action course is *successful* if and only if the expected-outcomes predicate ($OA_i \in \Omega\Theta$) associated with the basic action type of its action token holds at the end. So, the same action token might have successful courses on certain histories and unsuccessful courses on others.

2.2 Extension to mental attitudes

We extend OntoSTIT+ with modal operators, here limited to Bel and Int for agents' beliefs and outcome intentions. Actually, we introduce three kinds of intention. They result from marrying a logical or AI distinction between *outcome intention that (to be)* and *intention to do* [6, 27, 12, 11] and a philosophical distinction between *prior (future–directed) intention to do* and *intention-in-action (present–directed) to do* [23, 3, 4, 24]. The first two are based on first-order predicates, while the third uses the modal operator Int.

Intention-in-action (to do) is expressed by $IntDo_a^{m/h}(t)$, which stands for "agent a intends at m/h to do the action token t" and it is defined in the following way[1]:

- $IntDo_a^{m/h}(t) \triangleq \exists c, i(CO(c,t) \wedge LOn(c,h) \wedge AgO(a,t) \wedge RT(c,i) \wedge InI(m,i))$

By definition, this intention is contemporaneous with all courses of t.

Prior intention (to do) is expressed by $PIntDo\text{-}A_{ia}^{m/h}(\overrightarrow{x})$, which stands for "agent a at m/h believes that he is going to do some action of type A_i with respect to patient(s) $\overrightarrow{x}$" and is defined[2] in the following way:

- $PIntDo\text{-}A_{ia}^{m/h}(\overrightarrow{x}) \triangleq Bel_a^{m/h} \exists m', t(Pre(m,m') \wedge A_i(t,a,\overrightarrow{x}) \wedge IntDo_a^{m'/h}(t))$

The last type of intention, outcome intention (that), has the form $Int_a^{m/h}\varphi(m',h')$, which stands for "agent a intents at m/h that $\varphi(m',h')$". Formulas of type $\varphi(m',h')$ are Boolean combinations of atoms built with

predicates of Π_+. This intention is axiomatically constrained so that φ is in the power of agent a, and m' is in the future of m on history h'. This intention persists during the execution of the action the agent undertakes to obtain φ (but the last moment), until a believes that φ obtains or is impossible, or a simply stops desiring φ. Indeed, if an agent intends some state of affair φ, he does not believe that φ is impossible, he believes that φ is not the case now and he desires φ. The axiom ensuring that an agent's intention that φ entails that he does not believe that φ is impossible captures the *strong consistency principle*:

(SC$_O$) $Int_a^{m/h}\varphi(m', h') \rightarrow \neg Bel_a^{m/h}(\forall m''(Pre(m, m'') \rightarrow \neg\varphi(m'', h'))$

We assume that an action of a given type is not always done for the same reason. So, in order to relate action tokens with their motivating outcome intention a series of "Intention-Outcome" (IO) operators is finally introduced. $IO\text{-}A_{ia,t}^{m/h}\varphi(m', h')$ implies that at m/h, a course of the token t of the action type A_i is going on together with the corresponding intention in action, a intends φ, and the expected outcome of t (i.e., OA_i) implies φ:

- $IO\text{-}A_{ia,t}^{m/h}\varphi(m', h') \rightarrow (IntDo_a^{m/h}(t) \wedge A_i(t, a, \overrightarrow{x}) \wedge Int_a^{m/h}\varphi(m', h') \wedge (OA_i(m', h', a, \overrightarrow{x}) \rightarrow \varphi(m', h')))$

As just explained, we assume that the action course of the token t *stops* at the index in which the outcome intention that φ is dropped and the corresponding $IO\text{-}A_i\varphi$ stops being true.

2.3 Modelling the VG scenario

We assume that $Hit \in \Pi_+$, and that the propositions $Hit(m, h, \mathbf{mis1}, \mathbf{tar1})$ and $Hit(m, h, \mathbf{mis2}, \mathbf{tar2})$ mean "at index m/h target 1 (resp. 2) has just been hit by missile 1 (2)". Among the actions types which the player **p** is able (skilled) to perform, is the action type *hitting a target with a missile*, or, in other words, *guiding a missile into a target*. We denote it for clarity and concision "*Guide*". We specify that the expected outcomes of this type of action, described by the predicate "*OGuide*", imply that the missile directed by the player hits the target. Then we specify the rules of the game, which say that it is impossible to hit both targets, i.e., if one target is hit the other cannot be hit either simultaneously or later:[3]

(VG1) $OGuide(m, h, \mathbf{p}, \mathbf{mis1}, \mathbf{tar1}) \rightarrow Hit(m, h, \mathbf{mis1}, \mathbf{tar1})$

(VG2) $OGuide(m, h, \mathbf{p}, \mathbf{mis2}, \mathbf{tar2}) \rightarrow Hit(m, h, \mathbf{mis2}, \mathbf{tar2})$.

(VG3) $Hit(m, h, \mathbf{mis1}, \mathbf{tar1}) \rightarrow$
$\forall m', h' \, ((PreEq(m, m') \wedge In(m', h')) \rightarrow \neg Hit(m', h', \mathbf{mis2}, \mathbf{tar2}))$

(VG4) $Hit(m, h, \mathbf{mis2}, \mathbf{tar2}) \rightarrow$
$\forall m', h' \, ((PreEq(m, m') \wedge In(m', h')) \rightarrow \neg Hit(m', h', \mathbf{mis1}, \mathbf{tar1}))$

Obviously (VG3) and (VG4) are in the player's beliefs, because he knows the rules of the game. The player knows that "there is a reward for hitting either target", so he starts desiring having hit either target and forms the intention at index $\mathbf{m}/\mathbf{h}$ that either target is hit some time later:

(VG5) $\exists m', h' \, Int_{\mathbf{p}}^{\mathbf{m}/\mathbf{h}}(Hit(m', h', \mathbf{mis1}, \mathbf{tar1}) \vee Hit(m', h', \mathbf{mis2}, \mathbf{tar2}))$

Clearly, this intention and the fact that the player believes (VG3) and (VG4) are not in contradiction with the strong consistency principle (SC_O).

Now how the player deals with this disjunctive intention to decide on which actions to do is a question of strategy. In this framework, we assume the player starts searching for the actions that may lead to achieving the content of its outcome intention (VG5). From (VG1) and (VG2) he infers that the needed basic action type is "*Guide*" and that he can instantiate this action type with two sets of patients, $\mathbf{mis1}, \mathbf{tar1}$ and $\mathbf{mis2}, \mathbf{tar2}$, since we have as theorems:

- $OGuide(m, h, \mathbf{p}, \mathbf{mis1}, \mathbf{tar1}) \rightarrow$
$(Hit(m, h, \mathbf{mis1}, \mathbf{tar1}) \vee Hit(m, h, \mathbf{mis2}, \mathbf{tar2}))$

- $OGuide(m, h, \mathbf{p}, \mathbf{mis2}, \mathbf{tar2}) \rightarrow$
$(Hit(m, h, \mathbf{mis1}, \mathbf{tar1}) \vee Hit(m, h, \mathbf{mis2}, \mathbf{tar2}))$

In Bratman's scenario, the player decides to do the two actions simultaneously. Thus two prior intentions to do, namely, the prior intention to do *Guide* on $\mathbf{mis1}, \mathbf{tar1}$ and the prior intention to do *Guide* on $\mathbf{mis2}, \mathbf{tar2}$ are generated:

- $PIntDo\text{-}Guide_{\mathbf{p}}^{\mathbf{m}/\mathbf{h}}(\mathbf{mis1}, \mathbf{tar1})$ and $PIntDo\text{-}Guide_{\mathbf{p}}^{\mathbf{m}/\mathbf{h}}(\mathbf{mis2}, \mathbf{tar2})$.

At a later index $\mathbf{m}^*/\mathbf{h}^*$ the player decides to start the two action tokens t and t' of basic action type *Guide* on $\mathbf{p}, \mathbf{mis1}, \mathbf{tar1}$ and $\mathbf{p}, \mathbf{mis2}, \mathbf{tar2}$ respectively. At this index the prior intentions turn into the intentions-in-action

to do t and t' and it is required that the player $\mathbf{p}$ does not believe that the preconditions of *Guide* (*PGuide*) applied to both sets of patients do not hold at $\mathbf{m}^*/\mathbf{h}^*$. The motivating outcome-intention is linked to these actions by the operator *IO-Guide*. We therefore have:

(VG6) $\exists t(Guide(t, \mathbf{p}, \mathbf{mis1}, \mathbf{tar1}) \wedge IntDo_{\mathbf{p}}^{\mathbf{m}^*/\mathbf{h}^*}(t) \wedge$
$\exists m, h\ \textit{IO-Guide}_{\mathbf{p},t}^{\mathbf{m}^*/\mathbf{h}^*}(Hit(m, h, \mathbf{mis1}, \mathbf{tar1}) \vee Hit(m, h, \mathbf{mis2}, \mathbf{tar2})))$

(VG7) $\exists t'(Guide(t', \mathbf{p}, \mathbf{mis2}, \mathbf{tar2}) \wedge IntDo_{\mathbf{p}}^{\mathbf{m}^*/\mathbf{h}^*}(t') \wedge$
$\exists m, h\ \textit{IO-Guide}_{\mathbf{p},t'}^{\mathbf{m}^*/\mathbf{h}^*}(Hit(m, h, \mathbf{mis1}, \mathbf{tar1}) \vee Hit(m, h, \mathbf{mis2}, \mathbf{tar2})))$

As one can see, for starting both action tokens the player must have only the disjunctive outcome intention (VG5), which activates two intentional actions and give rise to two intentions-in-action. While (VG5) and the two *IO-Guide* formulas persist, the two action tokens are executed. When target 1 is hit, the courses of both t and t' end, only that of t being successful.

In OntoSTIT+ all actions are intentional and any particular action token is always done with the corresponding intention-in-action (to do). In this sense this theory assumes the Simple View. In the VG scenario, we have in addition the outcome intention that either target is hit, and this is the motivating intention associated to both actions. As advocated in the philosophical literature about unintentional side effects, in this theory it is possible for an agent to intend that φ and believes that $\varphi \rightarrow \psi$, without intending that ψ. But even if we were to allow for this inference and assume that (VG1) and (VG2) hold, we would *not* have $\exists m', h'\ Int_{\mathbf{p}}^{\mathbf{m/h}} Hit(m', h', \mathbf{mis1}, \mathbf{tar1}) \wedge$ $\exists m', h'\ Int_{\mathbf{p}}^{\mathbf{m/h}} Hit(m', h', \mathbf{mis2}, \mathbf{tar2})$, because atoms formed on the predicate *OGuide* cannot be arguments of the *Int* operator since $OGuide \notin \Pi_+$. OntoSTIT+ dissociates the (prior of present-directed) intention to do an action and the intention that its expected outcomes hold. Nevertheless, it requires that some motivating outcome intention be present. This may of course be the same intention that the standard effects of the expected outcomes hold but not necessarily.

3 $\mathcal{LIA}$ meets the VG puzzle

3.1 Introduction to $\mathcal{LIA}$

In $\mathcal{LIA}$ the standard operators of dynamic logic of the form $[\alpha]$ (where α is an element of a set of action names ACT—which corresponds to action types in OntoSTIT+) are substituded by modal operators for attempt of the form $[\![\alpha]\!]$. The main difference between $\mathcal{LIA}$ and standard dynamic logic is that the dynamic primitives are not atomic actions, but atomic attempts. In standard dynamic logic formula $[\alpha]\,\varphi$ is read "φ holds after the occurrence of action α", whereas in $\mathcal{LIA}$ formula $[\![\alpha]\!]\,\varphi$ is read "φ holds after the agent's attempt to do α". Hence, formula $[\![\alpha]\!]\,\bot$ has to be read "the agent does not try to do α". The dual of $[\![\alpha]\!]$ is $\langle\!\langle\alpha\rangle\!\rangle$ and formula $\langle\!\langle\alpha\rangle\!\rangle\varphi$ is read "the agent tries to do α and φ holds after the agent's attempt". Hence, formula $\langle\!\langle\alpha\rangle\!\rangle\top$ has to be read "the agent tries to do α".

The notion of attempt formalized in $\mathcal{LIA}$ corresponds to the notion of volition proposed in the philosophical literature [8, 9, 13, 21]. Furthemore, an action name α in ACT is supposed to denote a so-called *basic action* [7, 10, 23], that is, a particular kind of bodily movement under the voluntary control of the agent that the agent can intend to perform without thinking how to perform it. For example, α might denote a simple motor pattern such as "raising an arm", "moving a leg", but also a complex motor pattern specialized for the accomplishment of a specific task and the achievement of a specific result such as "grasping an object", "tying one's shoes", "closing a window", etc.

In $\mathcal{LIA}$ *trying* (or *attempting*) *to do* a basic action α is conceived as a mental event which is caused by a proximal intention to do α and which consists in an agent exerting voluntary control over the initiation and the execution of the bodily movement α, that is, an attempt is the mental counterpart of the intentional performance of a bodily movement. In $\mathcal{LIA}$ there are not only operators for attempt of the form $[\![\alpha]\!]$ but also the two operators Bel and $Goal$ for modelling mental attitudes of agents. The original $\mathcal{LIA}$ also includes the operators X ("next"), G ("always") and $UNTIL$ of linear temporal logic. In this paper we abstract from time and we only present the fragment of the logic in which we can reason about mental attitudes and attempts. The operator Bel is a standard $KD45$ doxastic operator expressing what an agent currently believes ($Bel\varphi$ is read "the agent believes that φ"). The operator $Goal$ is a KD operator and refers to chosen goals of the agent,

i.e., goals that the agent decides to pursue ($Goal\varphi$ is read "the agent has the chosen goal that φ"). For the sake of simplicity, we suppose that $Goal\varphi$ corresponds to the particular form of intention called "outcome intention that" described above. That is, $Goal\varphi$ can also be read "the agent intends that φ holds".[4]

Every operator $[\![\alpha]\!]$ is supposed to satisfy the axioms and rules of inference of the basic normal modal logic K. We here focus on some additional principles:

($\mathbf{Alt}_{Att}$) $\quad \langle\!\langle\alpha\rangle\!\rangle\varphi \rightarrow [\![\beta]\!]\,\varphi$

($\mathbf{SC}_L$) $\quad Bel\varphi \rightarrow \neg Goal\neg\varphi$

($\mathbf{PosIntr}_{Goal}$) $\quad Goal\varphi \rightarrow BelGoal\varphi$

($\mathbf{NegIntr}_{Goal}$) $\quad \neg Goal\varphi \rightarrow Bel\neg Goal\varphi$

($\mathbf{IA}$) $\quad \langle\!\langle\alpha\rangle\!\rangle\top \leftrightarrow Goal\langle\!\langle\alpha\rangle\!\rangle\top$

Axioms $\mathbf{PosIntr}_{Goal}$ and $\mathbf{NegIntr}_{Goal}$ are principles of positive and negative introspection for chosen goals. Axiom $\mathbf{SC}_L$ is the strong consistency axiom relating beliefs with chosen goals. According to Axiom $\mathbf{Alt}_{Att}$, all attempts correspond to transitions to the same world, that is, all attempts occur in parallel. Thus, if the agent tries to do two different actions α and β at the same time, all effects of the attempt to do α and all effects of the attempt to do β are effects of the joint occurrence of the two attempts. This explains why formula $\langle\!\langle\alpha\rangle\!\rangle\top$ can be read "the agent tries to do α" instead of "it is possible that the agent tries to do α". Axiom **IA**—for "Intentional Attempt"— establishes that the agent tries to do α if and only if he intends to try to do α. The direction "$\leftarrow$" of **IA** says that if an agent intends to try to do α he goes through the mental effort of doing α (which consists in trying to do α). This direction of the axiom highlights the causal relation between present-directed intention and attempt. According to the direction "$\rightarrow$" of **IA** attempts are by definition intentional. Indeed, an agent cannot really try to do some action α without having (at least) the intention to try to do α (compare **IA** with the definition of $IntDo_a^{m/h}(t)$ in Section 2).

In order to model the relationship between (basic) action occurrences and attempts, further formal constructs have to be introduced in the logic. Given a set $\Delta = \{\langle\!\langle\alpha\rangle\!\rangle\bot | \alpha \in ACT\}$ of all *simple attempt formulas* of the form $\langle\!\langle\alpha\rangle\!\rangle\top$, a set of objective formulas OBJ is defined. OBJ is the smallest

superset of the set Π of propositional atoms and the set Δ of simple attempt formulas such that: if $\varphi, \psi \in OBJ$ then $\neg\varphi, \varphi \vee \psi \in OBJ$. From this, a function Pre is introduced which assigns an objective formula in OBJ to each basic action, that is: $Pre : ACT \longrightarrow OBJ$. $Pre(\alpha)$ is supposed to denote the *execution precondition of action* α. The notion of execution precondition is fundamental for defining the notion of (basic) action occurrence. Formula $\langle\alpha\rangle\varphi$, which has to be read "the agent does α in a successful way and φ is true after α's occurrence", is defined starting from the primitive notions of attempt and execution precondition. That is, for any α, we have:

- $\langle\alpha\rangle\varphi =_{def} \langle\langle\alpha\rangle\rangle\varphi \wedge Pre(\alpha)$.

This means that in $\mathcal{LIA}$ the successful execution of a basic action α is an attempt to do α when the execution precondition of α holds:

- $\langle\alpha\rangle\top =_{def} \langle\langle\alpha\rangle\rangle\top \wedge Pre(\alpha)$.

Formula $\langle\alpha\rangle\top$ has to be read "the agent does α in a successful way". For example, $\langle raiseArm\rangle\top$ has to be read "the agent raises his arm in a successful way". $[\alpha]\,\varphi$ is given as an abbreviation of $\neg\langle\alpha\rangle\neg\varphi$ for any $\alpha \in ACT$. Thus, $[\alpha]\,\varphi$ has to be read "after the agent performs α in a successful way, it is the case that φ" ($[\alpha]\,\bot$ has to be read "the agent does not perform α in a successful way"). The equivalence between $\langle\alpha\rangle\top$ and $\langle\langle\alpha\rangle\rangle\top \wedge Pre(\alpha)$ should be conceived as a non-standard way to express *execution laws*. In fact, *execution laws* have been traditionally expressed by taking actions as primitive elements, without decomposing them into more elementary constituents (viz. attempts). For instance, in Situation Calculus [22] it is supposed that an action α is executable if and only if its execution precondition $Poss(\alpha)$ is true. Thus, in Situation Calculus the concept of attempt only appears implicitly in the concept of execution precondition and there is no clear distinction between the former and the latter. On the contrary, in $\mathcal{LIA}$ attempt and execution precondition are clearly distinguished in the formal specification of execution laws.

The distinction between *trying to do* α and *doing* α *in a successful way* is fundamental for distinguishing two subspecies of *intention to do something*: the *intention to do something in a successful way* and the *intention to try to do something*. The agent's *intention to do* α *in a successful way* is defined in $\mathcal{LIA}$ as the agent's chosen goal to do α in a successful way. Formally:

- $IntDo(\alpha) =_{def} Goal\langle\alpha\rangle\top$.

The agent's *intention to try to do* α is defined as the agent's chosen goal to try to do α. Formally:

- $IntTry(\alpha) =_{def} Goal\langle\langle\alpha\rangle\rangle\top$.

Thus, if an agent intends to do α in a successful way, he also intends to try to do α, i.e., $IntDo(\alpha) \rightarrow IntTry(\alpha)$ is a theorem of the present logic. Moreover, an agent may intend to try to do α without intending to do α in a successful way, i.e., $IntTry(\alpha) \wedge \neg IntDo(\alpha)$ might be true. The plausibility of such a consequence has been defended by several philosophers [3, 4, 19, 20]. In fact, we can imagine plenty of scenarios where an agent intends to try to do something without intending to accomplish it. For example, suppose that i promises to pay j a certain amount of euros if j tries to raise an arm within five seconds. Agent i assures agent j that he need not actually raise an arm in a successful way for getting the amount of euros. It is plausible to say that j intends to try to raise an arm even if he does not intend to succeed in raising an arm. In fact, j does not care whether his trying is going to succeed or to fail.[5]

The following abbreviations are given in order to express that the agent intends to do α and β in parallel and the agent intends to try to do α and β in parallel:

- $IntDo(\alpha \wedge \beta) =_{def} Goal(\langle\alpha\rangle\top \wedge \langle\beta\rangle\top)$;
- $IntTry(\alpha \wedge \beta) =_{def} Goal(\langle\langle\alpha\rangle\rangle\top \wedge \langle\langle\beta\rangle\rangle\top)$.

3.2 Modelling the VG scenario

We suppose that the action of hitting target 1 and the action of hitting target 2 are basic actions of the video game player. Under this assumption, formulas of type $\langle\langle hitT1\rangle\rangle\top$ and $\langle\langle hitT2\rangle\rangle\top$, which respectively express the facts that "the player tries to hit target 1" and that "the player tries to hit target 2", sound acceptable.

The rules of the game say that it is impossible to hit the two targets in the same time, that is, it is impossible that the player successfully hits target 1 and successfully hits target 2. The definition of successful action being based on the preconditions, this means that the execution precondition of the action of hitting target 1 and the execution precondition of the action of hitting target 2 must be inconsistent. In fact, the player can hit target 1

when he tries to do this, only if missile 1 is correctly oriented towards target 1 and either the player does not try to hit target 2 or missile 2 is not correctly oriented towards target 2. Formally:

- $Pre(hitT1) = M1TowardsT1 \wedge ([\![hitT2]\!] \bot \vee \neg M2TowardsT2)$.

In a similar way, the player can hit target 2 when he tries to do this only if, missile 2 is correctly oriented towards target 2 and either the player does not try to hit target 1 or missile 1 is not correctly oriented towards target 1. Assuming that the player tries to hit both targets, it follows that

- $Pre(hitT1) \rightarrow \neg Pre(hitT2)$.

Since $Pre(hitT1)$ and $Pre(hitT2)$ are inconsistent, the player must believe that he cannot successfully hit both target 1 and target 2. In fact, the following formula can be inferred by definitions of $\langle\alpha\rangle$ and $[\alpha]$ and standard modal principles:

(A) $Bel\neg(\langle hitT1\rangle\top \wedge \langle hitT2\rangle\top)$

According to Axiom $\mathbf{SC}_L$, a rational agent cannot intend that φ if he believes that $\neg\varphi$. Therefore, from (A) we can infer $\neg IntDo(hitT1 \wedge hitT2)$. This means that in the VG scenario, if the player is rational, he cannot intend to hit target 1 and to hit target 2 at the same time. Now the point is: how can we model the player's decision to play the two video games simultaneously? The nice aspect of $\mathcal{LIA}$ is that it makes the notion of attempt available and allows us to model the player's intention to try to hit both targets. In our view, this is exactly the player's intention in the VG example when he decides to have a go at both games at once. Formally:

(B) $IntTry(hitT1 \wedge hitT2)$ that is, $Goal(\langle\langle hitT1\rangle\rangle\top \wedge \langle\langle hitT2\rangle\rangle\top)$

Obviously, (B) and (A) are not inconsistent. Thus, the player intends to hit both targets even if he believes that he cannot hit both targets in a successful way. Moreover, $IntTry(hitT1 \wedge hitT2)$ implies $\langle\langle hitT1\rangle\rangle\top \wedge \langle\langle hitT2\rangle\rangle\top$ (by Axiom **IA**). This explains how the player's intention to try to hit both targets brings it about that the player mobilizes his energy and tries to hit both targets.

4 Discussion

4.1 Relation with the philosophical literature

We start the discussion by comparing the two presented solutions of the VG puzzle with the ones already existing in the philosophical literature. Two main approaches towards the VG puzzle have been discussed. The first one is based on intention to try (e.g., Mele [18]) and the second one considers the disjunctive intention to hit target 1 or to hit target 2 (e.g., Tuomela [26]).

The $\mathcal{LIA}$ solution is clearly based on the one proposed by Mele who claims that the player "tries to hit target 1 (with a video missile fired with his right hand) while also trying to hit target 2 (with a video missile fired with his left hand) and he intends to try to hit target 1 while also intending to try to hit target 2" [18, p. 131]. Mele stressed that possessing intention to try to hit target 1 and intention to try to hit target 2 escapes the principle of strong consistency, since "successfully executing an intention to *try* to A entails *trying* to A. It does not entail A-ing" [18, p. 133].

On the other hand, the solution of the VG puzzle in OntoSTIT+ appears to be related to the answer given by Tuomela who says that the player acts under "intention by his shooting behavior to hit 1 or hit 2" [26, p. 65]. Similarly to Tuomela's solution, in OntoSTIT+, the player acts under the disjunctive intention, but OntoSTIT+ refers to intention *that*, whereas Tuomela considers the intention *to do*. In OntoSTIT+, both intentions to do are actually present.

4.2 Comparison of two frameworks

The two frameworks largely differ in expressivity. While the first uses a rich temporal structure, and a rich notion of action with duration, with action types and two sorts of action individuals, the second doesn't even require time to solve the puzzle, and uses only operators based on names for actions and attempts, i.e., action types. They both model several kinds of intentions, but the first focuses on the ontological differences between (prior) intention-to-do, intention-that and intention-in-action, while for the second, intention-that, intention-to-do and intention-to-try are all based on a single intention operator, the difference lying in the use of action or attempt operators.

In the light of this as well as in the light of how they are related to the literature, it would seem that the two solutions are of a different essence.

The puzzle is solved in OntoSTIT+ by exploiting the independence between intention-to-do an action and intention-that its outcomes be true, with the same disjunctive intention-that motivating to two different intentions-to-do. On the other hand, the puzzle is solved in $\mathcal{LIA}$ by distinguishing between attempts and intention-to-try and actions and intention-to-do, the player intending to try both actions but not intending to do both actions.

However, it appears that since actions in OntoSTIT+ are non-deterministic, they actually correspond to attempts in $\mathcal{LIA}$. And successful actions are defined in terms of "action/attempt" in both frameworks.[6] So, intentions-to-do in OntoSTIT+ correspond to intentions-to-try in $\mathcal{LIA}$. The fact that in OntoSTIT+ there is no inference from intention to do an action to intention that all the action's outcomes be true ends up being very similar to the fact that in $\mathcal{LIA}$, intention-to-try doesn't imply intention to do successfully. Thus OntoSTIT+ is also related to Mele's solution, which shows that the two existing approaches towards the VG puzzle are not incompatible.

Notes

[1]This fact restricts also the concept of action to those actions which do no go beyond bodily movement such as *running a marathon* or *signing a document*. In order to allow for actions which do go beyond bodily movement, we need replace this definition by the following axiom:

- $IntDo_a^{m/h}(t) \rightarrow \exists c, i(CO(c,t) \wedge LOn(c,h) \wedge AgO(a,t) \wedge RT(c,i) \wedge InI(m,i))$

and accept the predicate "$IntDo$" as primitive.

[2]This is a definition schema.

[3]Because we are in branching time, we need to quantify over future moments and all alternative histories.

[4]We are aware that this assumption is over-simplistic in two senses. First, an intention-that is not reducible to a chosen goal. As in OntoSTIT+, and as argued in [1, 23, 16], an agent intends that φ if and only if he wants φ to be true and believes that φ is under his control. So, an intention-that is a complex mental attitude which involves both a goal and a particular kind of belief. Second, as taken into account in OntoSTIT+, an intention-that is by definition a future-directed mental state. That is, when an agent intends that φ then he wants φ to be true at some point in the future. Here we abstract from time and we simply do not consider this temporal aspect of intention-that.

[5]Nevertheless, it has to be noted that other authors have opposed to the idea that an agent can intend to try to do something without intending to do it in a successful way. For instance, McCann has committed himself to the view that "intending to try to do α" entails "intending to do α in a successful way" [17]. For McCann, in fact, *trying* is a vague concept and *intending to try* is totally vacuous unless it is supplemented by a more

substantive intention. Differently from McCann's conception of *trying*, in $\mathcal{LIA}$ *trying* has a precise denotation. Therefore, we do not see any problem in having it as a possible content of intention.

[6]The success of an action/attempt is not based on the same grounds, because of a different assumption regarding the duration of actions. In $\mathcal{LIA}$ (with time), attempts are one-step, so the success of an attempt depends only on its preconditions being satisfied as explained above. In OntoSTIT+, actions have any duration, so the success of an action lies in preconditions as well as other factors (for instance, the agent may change its mind and abort the action during its execution).

References

[1] A. Baier. Act and intent. *Journal of Philosophy*, 67:648–658, 1970.

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

[3] M. Bratman. Two faces of intention. *Philosophical Review*, 93:375–405, 1984.

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

[5] B. F. Chellas. *The logical form of imperatives*. Perry Lane Press, Stanford, CA., 1969.

[6] P. Cohen and H. Levesque. Intention is choice with commitment. *Artificial intelligence*, 42:213–261, 1990.

[7] A. Danto. Basic actions. *American Philosophical Quarterly*, pages 141–148, 1965.

[8] L. Davis. *Theory of Action*. Prentice-Hall, Englewood Cliffs, N. J., 1979.

[9] C. Ginet. *On Action*. Cambridge University Press, Cambridge, 1990.

[10] A. Goldman. *A Theory of Human Action*. Prentice-Hall, Englewood Cliffs NJ, 1970.

[11] B. Grosz and S. Kraus. Collaborative plans for complex group action. *Artificial Intelligence*, 86(2):269–357, 1996.

[12] A. Herzig and D. Longin. C&L Intention Revised. In M-A. Williams D. Dubois, Ch. Welty, editor, *Principles of Knowleadge Representation and Reasoning*. Menlo Park, California, AAAI Press, 2004.

[13] J. Hornsby. *Actions*. Routledge & Kegan Paul, London, 1980.

[14] E. Lorini. *Variations on intentional themes: From the generation of an intention to the execution of an intentional action*. PhD thesis, University of Siena, Department of Philosophy, Siena (Italy), 2007.

[15] E. Lorini, A. Herzig, and C. Castelfranchi. Introducing "attempt" in a modal logic of intentional action. In M. Fisher, W. Van der Hoek, and B. Konev, editors, *Proceedings Tenth European Conference on Logics in Artificial Intelligence (JELIA06), LNAI, vol. 4160*. Springer Verlag, Liverpool, 2006.

[16] E. Lorini, N. Troquard, A. Herzig, and C. Castelfranchi. Delegation and mental states. *Proceedings of 6th International Joint Conference on Autonomous Agents in Multi-Agent Systems (AAMAS'07)*, 2007.

[17] H. McCann. Rationality and the range of intention. *Midwest Studies in Philosophy*, 10:191–211, 1986.

[18] A. Mele. Intending and Trying: Tuomela vs. Bratman at the Video Arcade. In M. Sintonen, P. Ylikoski, and K. Miller, editors, *Realism in Action, Essays in the Philosophy of the Social Science*. Kluwer Academic Publishers, 2003.

[19] A. R. Mele. She intends to try. *Philosophical Studies*, 55:101–106, 1989.

[20] A. R. Mele. *Springs of Action: Understanding Intentional Behavior*. Oxford University Press, Oxford, 1992.

[21] B. O'Shaughnessy. Trying (as the mental "pineal gland"). *Journal of Philosophy*, 70:365–86, 1973.

[22] R. Reiter. *Knowledge in action: logical foundations for specifying and implementing dynamical systems*. MIT Press, Cambridge, 2001.

[23] J. Searle. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press, New York, 1983.

[24] J. R. Searle. *Rationality in Action*. The MIT Press, 2001.

[25] N. Troquard, R. Trypuz, and L. Vieu. Towards an ontology of agency and action: From STIT to OntoSTIT+. In *Proceedings of the Fourth International Conference FOIS 2006, Baltimore, Maryland (USA), November 9-11*, pages 179–190, 2006.

[26] Raimo Tuomela. *The Importance of Us*. Stanford University Press, 1995.

[27] Michael Woolridge. *Reasoning about Rational Agents*. The MIT Press, 2000.