Action Planning for Autonomous Systems with respect to Safety Aspects

Philipp Ertle*, Dennis Gamrad†, Holger Voos* and Dirk Söffker†
†Chair of Dynamics and Control,
University of Duisburg-Essen,
47057 Duisburg, Germany
Email: {gamrad, soeffker}@uni-due.de
*ZAFH Autonome mobile Serivceroboter
University of Applied Sciences Ravensburg-Weingarten,
88241 Weingarten, Germany
Email: {ertle, voos}@hs-weingarten.de

***Abstract*—Autonomous systems are often needed to perform tasks in complex and dynamic environments. For this class of systems the traditional safety assuring methods are not satisfying because these systems cannot be analyzed completely during development phase. In order to realize a more flexible safety analysis the internal representation of the outside world that is learned by an autonomous Cognitive Technical System is used to identify hazardous situations. The so called safety principles are the hazard knowledge. These can be added to the system prior to operating time without losing the possibility of adjusting or expanding this hazard knowledge during operating time. How these safety principles are generally designed and implemented is explained in this contribution. Furthermore, as underlying Cognitive Technical System provides anticipation capabilities it is possible to expand the planning process in order to take hazard information into account. Finally, in a simulation example is shown how the autonomous system determines possible future actions and evaluates them with regard to hazards in order to provide a plan with acceptable risk.**

***Index Terms*—autonomous system, safe, cognitive technical system, risk assessment, planning**

I. Introduction

The demand for autonomous systems is rising because of an increasing need for autonomous technical systems that are able to execute challenging and complex tasks in unstructured dynamic environments. In many cases, they have to collaborate with humans in a natural and intuitive way and adapt themselves to varying conditions [21]. Typical examples are autonomous service robots. In order to realize a high level of autonomy, the imitation of human cognition including the representation of knowledge, learning, anticipation, planning etc. can be helpful. If an autonomous system comprises a representational level and has capabilities, like learning and planning, it can be considered as cognitive (see [19]). However, these cognitive capabilities and the highly flexible interaction with human users and the environment also increase the risk that any action of the autonomous system could lead to injury of humans or damages to the environment. Therefore, autonomous systems are safety-critical systems that require a safety strategy in order to assure that no unacceptable risks exist [6],[2]. However, it seems to be impossible to foresee all complex interactions between autonomous technical system and the environment and therefore the derivation of a safety strategy and the realization of a satisfying risk assessment during the development phase also seems to be impossible [21]. Hence, if safety analysis cannot be established during the development phase sufficiently, it has to be ensured that the safety aspects are observed during operating time [21]. In order to keep the system's autonomy these supervision mechanisms have to be realized by the system itself. There are only few contributions (see for instance [13],[12],[3]) dealing with safety aspects of autonomous systems. All of them present either partial solutions or they are outlining the basic problem without mentioning a satisfying solution.

In this contribution, a novel approach including a cognitive architecture for the realization of safe autonomous systems is presented. The core idea is that the autonomous systems should have an internal representation of the outside world. This generated internal representation—a formalized description of the environment and systemic interaction—enables application of numerous methods. It is shown how the internal representation can be used to evaluate the system's environment with respect to safety aspects. This again is used for safe action planning. In this work, the mental model of the cognitive architecture representing the structure of the real world is extended by a safety/risk model containing general principles for the evaluation of situations. Hence, it is possible to consider safety aspects during planning and interaction. Furthermore, the approach also allows the learning of the mental model.

The following sections describe the used cognitive architecture as the underlying method, its formalization of interaction and its representational level. Then, the general method for recognizing hazards in the system's environment are outlined. Finally, the implementation of the proposed approach is described and illustrated using a simulation example.

II. Realization of cognitive technical systems

In this section, an approach for the realization of Cognitive Technical Systems interacting autonomously with its environment is presented. The representation of knowledge which is the key feature of cognitive systems is based on Situation-Operator-Modeling and implemented by high-level Petri Nets. As framework for the representational level and the cognitive functions a cognitive architecture is used. Within the architecture a mental action space depending on the system's knowledge and the current situation is generated and used for planning. The following subsections describe the methodical background—the Situation-Operator-Modeling approach—and the cognitive architecture in detail.

A. Modeling and Analysis of Interactions

For the modeling of interaction, the Situation-Operator-Modeling (SOM) approach (see [18],[19]) can be applied. Within the SOM approach the processes in the real world are considered as sequences of scenes and actions, which are modeled as situations (time-fixed description of the considered system or problem) and operators (changes within the considered system) respectively. A situation s_i consists of characteristics c_i and relations r_i. In technical systems, the characteristics can be physical values measured by the sensors. The relations represent an inner structure of the situation, which extends the classical situation calculus (see [11]) by linking the characteristics to each other through arbitrary functions. The operators o_i have the same quality as the relations of the situation. An operator transfers a situation to another ($o_i : s_x \rightarrow s_y$). Depending on its functionality, the characteristics, the relations or both can be changed. The condition whether an operator can be applied is described by the operator's assumptions. An operator on a higher hierarchical level can be build by the combination of several operators termed as meta operator ($o_{i \rightarrow n} : s_i \rightarrow s_n$). Furthermore, SOM realizes a graphical representation, in which situations are illustrated by gray ellipses with black dots denoting characteristics and white circles denoting relations. Such as the relations (or passive operators), also the active operators are represented by white circles. A detailed description about this underlying approach is given in [18],[19]. Due to structural similarities, SOM-based models can be represented by high-level Petri Nets (HPNs), which may be modeled graphically, simulated, and analyzed by established software tools. The SOM situation can be represented by a place with one or several tokens and the SOM operator can be related to the transition with guard functions and current bindings. In this contribution, the software Renew (Reference net workshop) for Reference nets (see [10]), a special HPN formalism, is applied. As a special feature, Renew allows Java objects and nets as tokens. However, it does not provide automated state space generation or own analysis functions, like other common HPN tools. Hence, these functionalities have to be realized by the model itself.

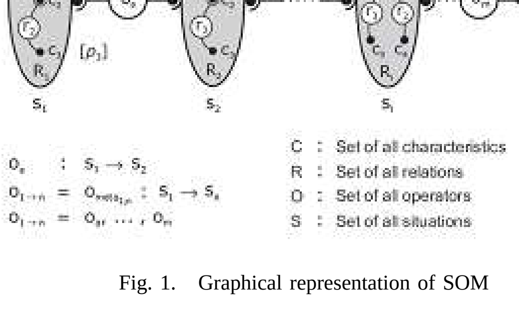

Fig. 1. Graphical representation of SOM

B. Representational level for Cognitive Technical System

The described approach for modeling, simulation, and analysis of Human-Machine-Interaction can also be used to realize a representational level for knowledge, which is considered here as the key feature of cognitive systems. The representational level is realized within a cognitive architecture, which simulates human cognition from a phenomenological engineering-oriented point of view and builds a unique framework (see [14]) for the realization of cognitive functions. Represented knowledge and the whole information processing are characterized by an open and variable structure, a symbolic representation, and the ability to reduce complexity. Hence, the approach is not restricted to a few problems or application fields, the learned facts can be traced by humans, and difficult tasks/problems can be simplified.

The proposed cognitive architecture (see Fig. 1) comes with the known three levels for skill-based, rule-based, and knowledge-based input/output behavior (see [17]. The connections to the sensors and actuators of the technical system to be controlled (ego system) are realized by sensing and execution modules. The input from the sensing module (represented as situation) is further interpreted by a perception module applying relations stored in a perception model. The actions of the ego systems as well as the dynamic of the outside world (other agents etc.) are represented in an action model in detail. The action model is used as a basis for the calculation of a mental action space which can also be extended from interaction directly. Action and perception model are realized by sequences of so called operator nets consisting of characteristic lists as assumptions and function nets which describe the effects of actions and interpretation principles for situations, respectively. The combination of action model, mental action space, and perception model builds the mental model of the architecture. The planning module takes a goal situation and the current interpreted situation as inputs and uses the mental action space to generate a plan containing a sequence of operators from the current situation to the desired situation. The execution of plans is organized by the execution module. During the execution of a plan or during random interaction as well, it is checked whether the mental model corresponds to the reality. If differences between the mental model and the reality are detected, the mental model can be modified accordingly by two modules for the learning of operators and the learning of the situation. Furthermore the learning modules are used to interpret the previously learned knowledge for generalization and complexity reduction. A detailed description regarding the modeling of interaction, knowledge representation, and cognitive functions can be found in [9].

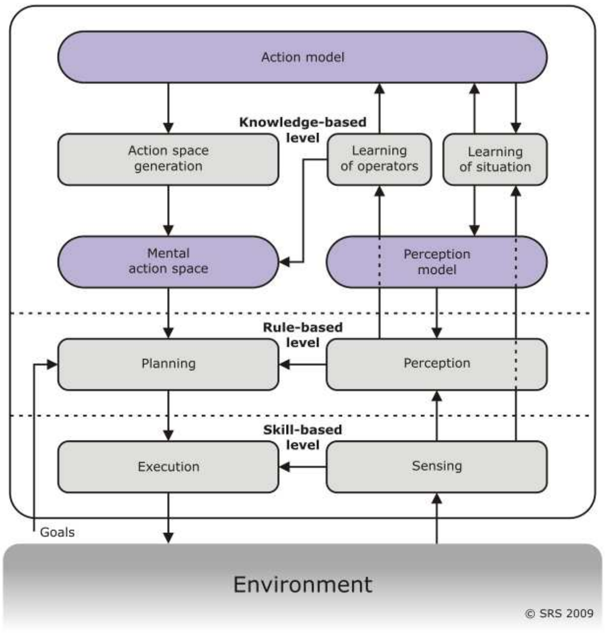

Fig. 2. Structure of the proposed cognitive architecture

III. Safety Aspects of Autonomous Systems

The general safety strategy includes detection of hazards, evaluation and implementing of countermeasures by detailed examination of the system and its operating environment [6]. When this strategy is applied to autonomous systems a fundamental problem appears: In accordance to [16] autonomy is needed when it is difficult to foresee all future situations the system will get into during normal interaction with its environment. For complex environments there is need to design and built autonomous systems without first explicitly identify the full range and scope of the interaction dynamics. Moreover, in [8] is stated that there is need for achieving emergence effects by explicitly omitting inflexible a priori plans or strategies and fixed internal functional structures. From that follows finally and strictly spoken that there is available for safety analysis neither a definable environment nor functional predefined system. The few contributions that are addressed to the topic 'safety of autonomous systems' have in common that they are no overcoming this problem.

In [13] it is introduced a method, how to link outcomes of safety surveillance components to an extended fault tree analysis, in order improve fault detection. The system is assumed as safe when there are no faults. In other words the system is safe when it is conform to its specifications. This approach therefore requires for a complete specification of the system's component dependencies beforehand.

In [12] and [15] is described a method how a probabilistic model can be defined to enable decision-making under considering of safety aspects. The hazard analysis—realized by a fault tree analysis—is the basis of the quantitative and probabilistic description of accident or hazardous events. Indeed, it is proposed to map relations between accident events and causal factors but it is suggested a 'hand-crafted' fault tree analysis observing static event combinations and static accident severity information.

In [3] is outlined how a certification process of autonomous systems could be put into practice. One central aspect of this approach is to consider possible hazards of future scenarios. The effort to generate (a complete) hazard analysis therefore depends very strongly on the complexity of the operating environment. The question arises, why is there need for autonomous capabilities if the system environment is analyzed completely?

The mentioned approaches are important to increase safety aspects of autonomous systems, albeit they cannot raise a claim to make autonomous systems safe. Decision risks or action risks that are very closely related to interaction dynamic (see [20]) are from central interest for autonomous system. These are focused in this contribution. In accordance to [21] it is proposed to supplement the traditional safety methods by adding so called dynamic risk assessment. With the help of generalized cause-consequence dependencies the risk assessment dynamically generates limits. In the following it is shown how dynamic online risk analysis can be realized methodically and can be put into practice. If, namely, a method is made available that is able to automatically identify hazards, it could be allowed to safely modify the internal representation of the outside world or the internal structure of the system (e.g. by learning methods) during operation.

The core idea is that first of all future situations are predicted based on a representation of the real world and then the deviated mental action model is automatically assessed with the help of 'general safety principles'. Hence, the information about dangerous states or situations can be used to identify actions or whole action sequences which results in these dangerous situations. This again can be used within the planning process.

In this approach the safety assessment is applied mental action space which represents the anticipation capabilities of the underlying Cognitive Technical System (see Fig. 2). The mental action space is evaluated, supplemented with safety information and then provided to the planning algorithm again. Herein, a first step comprises a check of the provided situations and a search for applicable safety principle in a database. If an suitable safety principle is found then an associated risk value can be computed with the help of the connected instruction for risk determination. The processing of these instructions generates risk information that added to the respective situation. The risk information describe the risk of any particular situation. This additional information allows the planning of actions that also take safety aspects into account. In this contribution a probabilistic based planning method is proposed for this purpose.

The risk information is encoded with risk values. These combine both, accident severity and probability ($Risk = S_{Acc} \cdot P_{Acc}$). Therefore, accidental events ($P_{Acc} = 1$) or hazards ($0 \geq P_{Acc} > 1$) can be described. The severity is assumed to be $0 \geq S_{Acc} \geq 1$ whereby $S_{Acc} = 1$ denotes worst case accident. In this contribution the risk values are figured out as percentage values. For instance, a risk value of $risk = 100\%$ is the incidence of the worst case accident.

For the realization of on-line risk analysis some requirements are to mention. Besides the representational level for knowledge,the system should comprise

1) a risk function based on a general (dynamic) risk model in a knowledge base (safety principles),
2) a measure of the distance from actual state to any hazardous states,
3) a planning algorithm considering risk information and
4) a method and measure to describe the quality of the overall risk evaluation.

These topics are explained in detail in the following sections.

A. Risk Function and General Risk Models

In this approach the safety assessment is applied on the anticipation capabilities of the underlying Cognitive Technical System—on the mental action space (see Fig. 2). The overall process is to generate the mental action space (anticipations), evaluate it, supplement it with risk information and then provide it to the planning algorithm again (see Fig. 3). The kinds of hazards that occur during operation time differ with respect to the acting system level. Every hierarchical level needs for level specific risk factors [20]. This contribution focuses the knowledge-based level. On this knowledge-based level there are more abstract hazards and hazards based on decisions. Often, these hazards are side effects of actions. For example, they are caused by the infringing of conventions, by the manipulation of dangerous objects or by the interacting of objects in the environment. For instance, an automatic park system could park on the fire line or a domestic robot could place any object on a hot cooking plate.

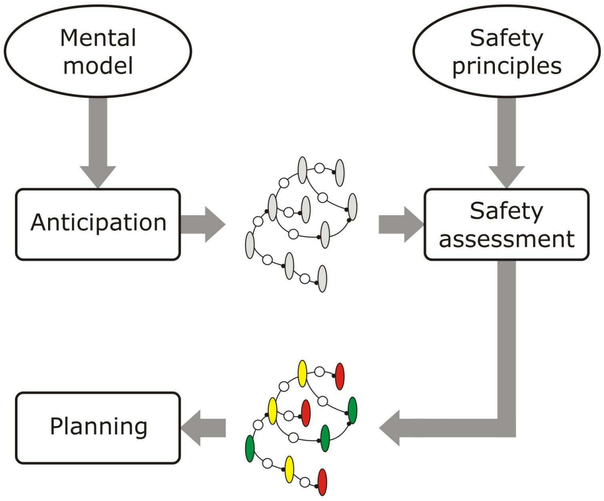

Fig. 3. Safety assessment of a mental representation

A set of fundamental safety principles should constrain the system's degrees of freedom strongly by universal quantifiers—a conservative set of safety principles. This conservative core contains principles that are principally valid, e.g. never kill a human. The principles are constrains that can be refined during operation. The refinement of the dynamic safety knowledge is from central interest especially with regard to environment's multiplicity. The basic requirements of dynamic safety principles are considered in principle, but they are not detailed in this contribution.

Basically, it is important to formulate these as universal principles. Due to the problem that future situations are not known in detail, it seems to be reasonable to implement principles that are valid in all future situations. For example, heated plastic materials generate toxic vapors or start to burn. So, the principle that the combination of intense heat and plastic materials is dangerous is not true in all cases, but not wrong in 'conservative' safety concept. If an exception is known it could be added to the knowledge, in all other cases the system would avoid to approximate 'plastic' to 'heat sources' generally. The safety principle could also be cause-and-effect chains that are derived, for instance, from natural laws (interaction of objects) or from social conventions. The main difference to an conventional safety strategy is that the limitations are realized in a dynamic knowledge representation. This knowledge can be extended or adapted in principle.

Basically, the principles are valid premises which can be used with (valid) observations for deductive conclusion. If the rules exists '*plastic object close to a heat source is hazardous*' and there would be the observation '*plastic object is close to heat source*' then the conclusion '*hazard occurs*' would be correct. Key problems will be the complexity and the completeness of this so called 'safety knowledge'. The completeness is linked to 'assessment quality discussion' below (see III-C). The complexity problem is linked to the 'safety principle learning topic' and will be discussed in future research work. Basically, the problem can be circumvented by applying a two step approach: 1) basic and important safety principles must be included beforehand 2) refinement of the principles is allowed under specific circumstances (training phase, learning by supervision etc.). In this contribution is shown how this can be realized with rule based knowledge systems. Basically, rule based systems are not sufficient for integrating of new knowledge. Therefore, knowledge management methods, for example 'Truth-Maintenance-Systems', have to be applied in future to ensure the consistence of knowledge.

A further problem that also appears in the mentioned example above is to describe the term 'hazard' numerically. If the condition is fulfilled, that plastic material is too close to the heat source, how dangerous is it in numbers? Most often the risk seems to depend on additional parameters, for instance, on temperature and distance. If the condition of a safety principle is fulfilled, then the determination of the respective risk value takes place by computation of the related risk function. The risk function basically is generated with known risk values. Most often extreme values or uncritical values are known. These can be used as basic information to approximate a risk function (see also [20]). During operating time future experiences could be considered by updating the respective risk function.

B. Measure of distance to hazardous states

In the proposed approach is described how hazardous situations can be detected with the help of safety principles. The Cognitive Technical System in this approach enables to generate possible future situations, thus, future hazardous situations are predictable as well. To remain in a safe situation, situations with unacceptable risks must be avoided [2] and therefore transitions into situations with unacceptable risk are prohibited. Furthermore, the risk could be reduced additionally, if a maximum 'distance' to hazardous situations is kept. From that follows, that a 'distance' to a hazardous situation or better be called 'safety margin' is from central interest, regardless of whether they are represented by time, steps or actions. With regard to decisions on a deliberative level the consideration of the safety margin enables to be fault tolerant with respect to decisions. For instance, if the safety evaluation with safety margin of n actions detects no unacceptable risks then it is ensured that despite the executing of n arbitrary actions (of situational allowed actions) the system remains in safe state. In case of systems with learning capabilities, safety in exploration activity can be enabled.

The Markov-Transition-Matrix is used to test for the reachability of future states or situations. The action space which is provided by the Technical Cognitive System (see II) contains the actual situation and future possible situations. Future situations are reachable by applying allowed actions (transitions). Each situation has a set of allowed actions. There are no preferred actions to enable a experience independent evaluation. Therefore, the actions space can be seen as a stochastic, homogeneous Markov-Chain of first order and the Markov-Transition-Matrix can be applied to compute the probability of transitions.

To transform the action space into the transition matrix the operators (transition between two situations) are transferred into a Markov transition matrix with the dimension of $j \times j$ with j situations in the action space. Each allowed operator increases the transition probability between its initial and final situation. The sum of each matrix row is normalized to 1. Final states such as goal state or 'death' states (used in the example later on) are absorbing states (no allowed operators available). For instance, if there are two allowed operators in (initial) situation S_1 then the transition probability is '0.5' in each case: to reach final situations S_x or S_y. Thus, the state space in form of transitions from S_i to S_j via operator $o_{i,j}$ can be translated to the transition matrix M:

$$M = \begin{Bmatrix} P(o_{1,1}) & \dots & P(o_{i,1}) \\ \vdots & \ddots & \vdots \\ P(o_{1,j}) & \dots & P(o_{i,j}) \end{Bmatrix}$$

With the help of rising M to the power of n the transition probability from S_i to S_j in n steps can be computed. Furthermore, each situation $S_1 \dots S_i$ contains k risk values (from evaluation of k safety principles), therefore a risk matrix containing the risk values $1 \dots k$ could be generated:

$$R = \begin{Bmatrix} R_1(S_1) & \dots & R_1(S_i) \\ \vdots & \ddots & \vdots \\ R_k(S_1) & \dots & R_k(S_i) \end{Bmatrix}$$

The product of transition probability matrix M^n and the transposed risk matrix generates the risk matrix RP_n for taking n random actions into account:

$$RP_n = M^n \times R^T$$

With regard to planned action the risk matrix can be used as look-up table to find out the corresponding risk. Depending on which RP_n is computed (with regard to n) different safety margins can be taken into account. The safety margin can be seen as a combined measure how far hazardous situations are away and how likely their incidence is. For instance, if a safety margin of '1' is considered and there are three allowed actions in situation S_2 and one of these leads to a 'mortal' situation S_3 (severity $S_{Acc} = 1$), then changing from situation S_1 to S_2 will result in a future random action risk (decision error) of $risk_1 = 33\%$ (see Fig. 4 left). If a safety margin of '2' is considered then the change to situation S_1 (assumed there are three allowed actions in starting situation S_1 and in intermediate situation S_2) will result to $risk_1 = 0\%$ (no dangers are 'in sight' with regard to safety margin '1') and $risk_2 = 0.33 \cdot 0.33 \approx 11\%$ (see Fig.4 right). From this results that in situations with no risk ($risk1_{1/2} = 0\%$) one/two random actions could be applied without reaching a dangerous/'mortal' situation. In the case of $risk_2 = 11\%$ follows that there is a risk of 11% to 'die' ($S_{Acc} = 1$) when two descion errors (random actions) occur.

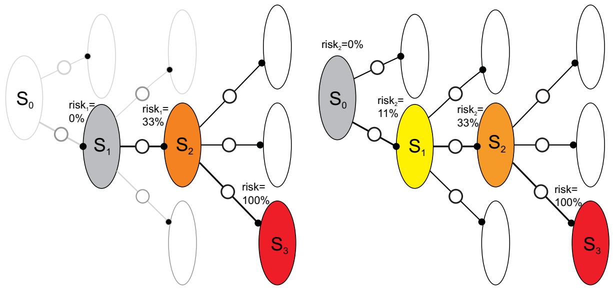

Fig. 4. Safety margin for one or two actions

Thus, information describing the distance to dangerous situations allow 1) to reduce the set of situations which should be reachable by the system and 2) to plan with consideration of risk optima. It seems to be reasonable to optimize risks also under consideration of respective task benefit. Higher risks seem to be accepted when the advantage to risk something is counterbalanced.

C. Expanded subject-matter of safety in case of autonomous systems

There seems to be a lack of definitions for danger causes in the traditional safety terminology. For example, if an autonomous mobile service robot comes to close to a fireplace and starts burning or the service robot provokes fire by putting a plastic object on a cooking plate in accordance to its task 'bring dishes to kitchen', how could these problems be called? Should it be called a failure, an error, fault, disturbance, defect or malfunction (definitions in accordance to [2] [1])? In case of the fire-catching robot it is 1) a defect or disturbance when the temperature sensor is out of order and the system therefore is not able to detect the increase of surrounding temperature or 2) a malfunction when there are requirements in the specifications but no implementation on this issue. A fire-catching robot without temperature sensor has a failure when it is damaged by the fire—the robot loses its capabilities despite being in full accordance to the specifications. Indeed, these specifications are incomplete in the described case because the knowledge of accurate environment description is missing. In the traditional sense, this problem could be treated as incomplete specification or specification mistake.

To prevent hazardous actions as mentioned above, 'safety principles' are enabling the system to recognize hazards and thus they enable to keep away from them. If these 'safety principles' are assumed to be a kind of knowledge (due to their degree of abstraction), then hazards can only occur (assuming a perfect reliable and fault-free system) due to lack of 'safety knowledge'. From this follows, safety can be improved only by 'completing' the knowledge with regard to dangers causes. If it is assumed that the completeness of the safety knowledge is impossible, then any providing of safety evidence can only take place empirically. However, on that point, it is from central interest to consider the quality of the safety assessment. For example, an indicator for 'level of familiarity' could be useful to recognize entering in a new environment or executing a complete unknown task. An indicator for 'routine' allows for increasing or decreasing of the autonomous system's execution speed. These topics are not part of this contribution but will be addressed in future contributions.

IV. Implementation

A. Safety assessment of situations and actions

Taking the proposed approach for modeling complex interactions between an autonomous system and its environment into account, the question arises whether situations or operators should be examined for safety assessment. Intuitively, people tend to focus on dangerous actions (operators), because actions are are changing things. This contribution however, safety aspects are always related to the actual situation in the SOM approach and dangerous actions are those actions that lead to hazardous situations. Therefore, it is assumed that the definition of hazardous situations is sufficient.

B. Connection of safety principles and action space

The action space is a set of possible future situations that will result from applying respective operators. Each situation is a representation of the actual perceived situations of the real world which contains all relevant information in form of characteristics. The characteristics are derived from the system's inputs applying an interpretation process. The goal of any risk assessment is to add one additional characteristic as one further characteristics [20] to any situation. These risk characteristics are generated by interpreting the current situation with the help of safety principles. Of course this method is suffering from the same restrictions as the overall system: phenomena that are not included in any sensory or input information cannot be considered at all. In order to be able to evaluate safety aspects at an upper deliberative level, also more abstract information must be processed. To give an example, a service robot should fulfill the task 'put the plastic bowl beside the kitchen sink'. The fulfillment of this task requires the capability to identify objects. On basis of these high level abstraction mechanisms, also high level safety principles must be generated. The reliability of the underlying process has to be considered as well. Additionally, the availability of all relevant information has to be considered. In order to simplify this complexity, the following conditions are assumed to be fulfilled (while knowing well that they are not trivial):

- All underlying processes are free of faults and failures.
- The representation of the relevant system's environment is complete.

If the representation is asumed to be complete, the service robot for instance is able to recognize all required objects like the plastic bowl. However, the robot might also detect the flat surface of the stove as an ideal place to deposit the bowl. Placing the bowl on a hot cooking plate may lead to the emission of toxic vapors or even fire an hence possible descision exists that lead to hazardous situations. A general formulation of a safety principle could be as follows. All detectable objects are included in a knowledge base, where objects are associated to attributes. For example, the object O_1 '*bowl*' is contained in the knowledge base with the corresponding attributes like '*liquid container*', '*plastic*', '*open on topside*', etc. This approach follows the 'object and attribute concept' in [4]. The object O_2 '*stove*' may be related to the attributes '*extreme heat*', '*in kitchen*', etc. Using this approach safety principles can be formulated as shown in Fig. 5 (according to [5]). If the

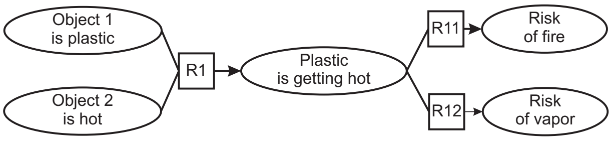

Fig. 5. Safety principles for interaction of plastic and heat

matching safety principles are found in a current situation, the corresponding risk value must be calculated. Each principle defines how the respective risk value has to be computed. For example, if the risk strongly depends on physical facts like distance, a suitable risk function would be defined that takes these parameters (distance of two objects, e.g.) into account, respectively. In many cases it would be helpful to use a quantitative approach like fuzzy methods instead of a qualitative functions, see also [20].

C. Planning under consideration of safety aspects

After assessing the risk of possible future situations, a the planning algorithm now can also take the possible risks of any planned action into account. Therefore, the determined risk values are added as transition (operator) costs. Hence, a weighted graph results. A transition from an arbitrary situation to a situation containing a risk value of 5% will cost 5, e.g. In order to determine whether the goal situation is reachable and what are the respective costs, several methods can be applied. In this contribution a Dijkstra algorithm is applied to calculate the path from an initial to a desired situation regarding the minimum costs. If large graphs have to be analyzed, also heuristic methods could be applied to reduce the processing time [5].

Finally, a system would be called safe if all occurring risks are lower than the tolerable risk threshold. So the planning algorithm should minimize the cost (i.e. overall risk) on the one side but also has to consider the risk-benefit relation with respect to maximum allowable risk.

V. Experimental results

A. Experimental scenario

As an example application, an arcade game [7] is chosen, where an autonomous agent interacts in a grid-based environment. This example is perfect in order to illustrate the derived methodologies and can be arbitrarily extended to achieve higher degrees of complexity. The environment consist of different kinds of fields and the agent can perform the actions 'up', 'down', 'left', and 'right'. In general, the task consists of first picking up a certain number of 'emeralds', by entering related fields, and then finishing the level by leaving the scenario through an exit door (see Fig. 6). Here, the agent is performing actions with help of the proposed cognitive architecture. Any situation S_i, as the input of the architecture, consists of the characteristics 'x-position' (integer), 'y-position' (integer), 'type of the current field' (string), and 'collected points' (integer). This example is now used to illustrate the action planning including safety aspects based on mental action space. The agent has to leave the level by reaching the exit door in the lower right corner, which can be reached by using a lot of different paths. Furthermore, an emerald can be picked up to increase the amount of collected points. This is no necessary condition to finish the scenario, however it can be used as an additional factor influencing the planning process. The considered scenario also contains three acid fields and a hostile monster agent performing horizontal movements. A collision with either the monster or the acid fields would lead to the undesired 'death' of the agent. In the considered application example a safety principle P_1 is defined.

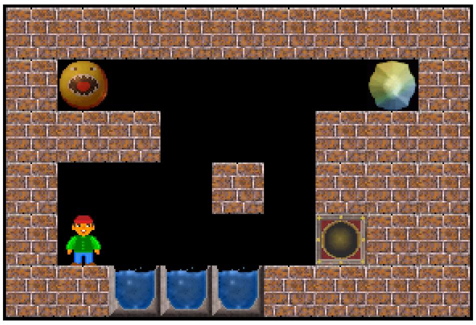

Fig. 6. Example scenario of the arcade game

$$P_1 : Pos\left(player\right) = Pos\left(mortal\_field\right) \rightarrow risk = 1$$
$$A : mortal\_field = \{acid\_field, monster\}$$

The conditional parts become true, if the position of the player is equal to the position of a '$mortal\_field$' (which again could contain 'attributes A' '$acid\_field$' or '$monster$'). The related risk is defined as 1 (accident severity $S_{Acc} = 1$ means 'death' of agent, accident probability $P_{Acc} = 1$ when the condition of safety principle is fulfilled and $risk = S_{Acc} \cdot P_{Acc}$).

B. Experimental results

The mental action space is generated based on the initial situation of the agent (lower left corner) and contains all possible future situations and actions. In Fig. 1, a part of the mental action space as it results in in this example is depicted using the SOM symbolic and colors in accordance to their present risk. Here, the hostile monster and the emerald in the upper right corner are not considered for a simplification. After the generation of the mental action space, it is evaluated with two different safety margins. These can be identified by the index: $risk_1, risk_2$. The risk value means in accordance to the above mentioned principle that agent's 'death' is described with severity of '1'. Therefore, the risk in respective situations is assumed as 100% ('mortal') because the event 'agent death' is certain and multiplied with severity '1'.In case of $risk_1 = 33\%$ every third random future action (starting from respective situation) will lead to agent's 'death'. In case of $risk_2 = 33\%$ the performing of two random actions will lead in every third case to agent's 'death'. From this results that in situations with no risk ($risk_{1/2} = 0\%$) one/two random actions could be applied without reaching a dangerous/'mortal' situation.

The complete actions space and evaluation of the simplified example is illustrated in Fig. 7. The shortest route will lead over situation $S_{3,4,5,8}$ to the goal situation S_{21}. This path is dangerous because it is close to the acid fields ($S_{9,10,11}$: colored red). Any decision error or exploration step will lead with high probability to the agent's 'death'.

For planning the Dijkstra algorithm is used. It generates paths under consideration of its costs. Thus, each transition from $S_i \rightarrow S_j$ costs: $cost_{ij} = 1 + risk\left(S_j\right)$, whereby the risk value is assumed to be expressed in percent $(0 \dots 100)$. The overall costs of the path $\vec{S}$ with the elements $S^0 \dots S^n$ or n actions (with $n \in \mathbb{N}^*$) are calculated $cost_{0,n} = n + \sum_{x=1}^{n} risk\left(S^x\right)$. This example cost formula is a very safety oriented variant. The plan with lowest costs is: $costs(\vec{S} : S_{2,6,7,12,13,18,16,8,21})$. The goal is reachable without any risks when considering a safety margin of one action. The plan does not change by considering safety margin of two actions but the maximum risk rises to $risk_2 = 11.1\%$ and the costs amount 37 ($\approx 11.1(\%){+}8.3(\%){+}8.3(\%){+}9(steps)$).

Path 1: $maxrisk_2(8) = 11,1\%$, $n = 9$ steps, $costs = 37$, $emeralds = 0$:
$o_{a:up}$,$o_{a:right}(8\%)$,$o_{a:right}(8\%)$,$o_{a:up}$,$o_{a:right}$,
$o_{a:right}$,$o_{a:down}$,$o_{a:down}(11\%)$,$o_{a:right}$

By examination of the level picture (Fig. 6), it can be seen that there is a bottleneck caused by the ledge directly above the player's starting position. It is not possible to pass the acid containers with a greater safety margin.

When the dynamic elements (monster, emerald) are taken into account, too, the complexity of the evaluation is increasing. The action space contains 816 situations (instead of 23), 1662 operators, 22 different goal constellations and therefore it is too big to be illustrated. If the scenario is evaluated again with safety margin '1' then the goal can be reached without any risks. For shortest route there are used 18 steps and 30 steps for the case that the emerald is collected. The planning process under consideration of safety margin '2' and collecting the emerald, results in:

Path 11: $maxrisk_2(383) = 37,5\%$, $n = 42$ steps, $costs = 111$, $emeralds = 1$:
$o_{m:right}$,$o_{m:right}$,$o_{m:right}$,$o_{m:right}$,$o_{m:right}$,$o_{m:left}$,
$o_{a:up}$,$o_{m:left}$,$o_{a:right}$,$o_{m:left}$,$o_{a:right}$,$o_{m:left}$,
$o_{a:up}$,$o_{m:left}$,$o_{a:right}$,$o_{m:right}$,$o_{a:right}$,$o_{m:right}$,
$o_{a:down}$,$o_{m:right}$,$o_{m:right}$,$o_{m:right}$,$o_{m:left}$,$o_{a:up}$,
$o_{m:left}$,$o_{a:up}$,$o_{m:left}$,$o_{a:right}$,$o_{m:left}$,$o_{a:right}$,
$o_{m:left}$,$o_{a:left}$,$o_{m:right}$,$o_{a:left}$,$o_{m:right}(38\%)$,$o_{a:down}$,
$o_{m:right}(31\%)$,$o_{a:down}$,$o_{m:right}$,$o_{a:down}$,$o_{m:right}$,$o_{a:right}$

When the emerald should be collected, then costs of 42 actions and maximum risk of 37.5% or overall costs of 111 have to be accepted. It can be seen by evaluating the plan, how the movement of the monster is considered for avoiding of collisions. When the emerald should be collected, then costs of 42 actions and maximum risk of 37.5% or overall costs of 111 have to be accepted. When the actions sequence is regarded it can be seen how the movement of the monster is taken into account to avoid collisions. The autonomous system waits until the monster is in the left corner. This moment is used to pass the rock in the center of the scenario. If the monster again is in the left corner, the moment is used to collect the emerald.

VI. Summary and Outlook

A new safety assurance approach is needed because it is difficult to realize a safety analysis for autonomous systems with traditional methods. With so called safety principles a solution is presented how hazards can be recognized dynamically. The generating of plans with respect to hazard information can be realized with the help of the underlying Technical Cognitive System's learning and anticipation capabilities.

The presented approach enables to combine general safety knowledge during the design phase with mechanisms for knowledge refinement and completion during the operating time. Hence, maximum efforts assuring safety take place during the development phase. The application, refinement, and completion of the safety knowledge can be performed by the autonomous system itself during operating time. This is described in principle and additionally it is demonstrated in a simulation example.

First of all, the Cognitive Technical Systems generates the 'action space' consisting of future reachable situations. These situations are evaluated with regard to hazards and supplemented with this risk information. After that further risk values are added to situations that are related to dangerous situations in order to additionally take nearby dangerous situations into account. Furthermore, all risk information is used, amongst other things, to calculate the costs of possible plans.

This is demonstrated in a small simulation example. The results of the simulation example are plans for fulfilling given task and respective costs with regard to risks and benefits. It can be seen clearly how the autonomous system generates plans with maximum 'distance' to hazardous situations. Additionally, the behavior of dynamic (moving) elements in the environment is taken into account. Finally, this contribution shows that it fundamentally is possible to enable autonomous Technical Cognitive Systems to perform a risk assessment online and, furthermore, to generate plans taking this hazard information into account. Therefore, the result is an optimal action sequence with regard to perceived hazards what can be reproduced and checked with the simulation example. The remaining risks, if available, can either be limited to an acceptable risk or counterbalanced with task fulfillment benefits.

Initially, the applied safety principles are kept very simple, in order to show primarily the functionality of the presented method. Surely, there will be various problems when extending this method to real world problems. This and the learning of new and consistent safety knowledge are not part of this contribution but will be addressed in future research work.

References

[1] "VDI/VDE 3542-1 safety terms for automation systems qualitative terms and definitions," 2000.

[2] "EN 61508-4 functional safety of electrical/electronic/programmable electronic safety-related systems - part 4: Definitions and abbreviations," 2002.

[3] R. D. Alexander, N. J. Herbert, and T. P. Kelly, "Structuring safety cases for autonomous systems," in *System Safety, 2008 3rd IET International Conference on*, 2008, pp. 1–6.

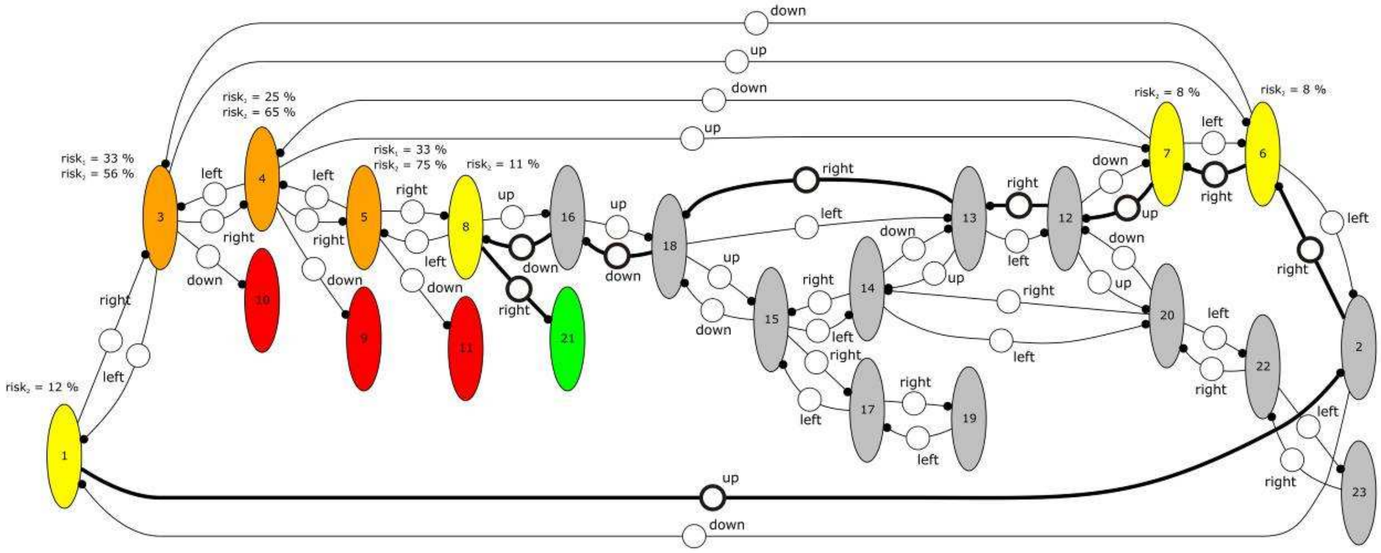

Fig. 7. Graphical representation of the mental action space

[4] Baumgartner, “A survey of upper ontologies for situation awareness,” in *International Conference on Knowledge Sharing and Collaborative Engineering (KSCE ’06)*, St. Thomas, USA, 2006.

[5] I. Boersch, *Wissensverarbeitung : eine Einführung in die Künstliche Intelligenz für Informatiker und Ingenieure*, 2nd ed. Heidelberg: Elsevier, Spektrum Akademischer Verl., 2007.

[6] J. Börcsök, *Functional Safety: Basic Principles of Safety-related Systems*. Heidelberg: Hüthig, 2007.

[7] A. Entertainment, “Rocks’n’Diamonds,” 2004. [Online]. Available: http://www.artsoft.org/rocksndiamonds/

[8] J. Gowdy, “Emergent architectures: A case study for outdoor mobile robots,” Ph.D. dissertation, Robotics Institute, Carnegie Mellon University, 2000.

[9] O. Kummer, *Referenznetze*. Berlin: Logos Wissenschaftsverlag, 2002.

[10] J. McCarthy, “Situations, actions and causal laws,” Stanford University, Technical Report, 1963.

[11] A. Newell, “You can’t play 20 questions with nature and win: Projective comments on the papers of this symposium,” in *Visual information processing*. New York: Academic Press, 1973, p. 283–308.

[12] C. Pace and D. W. Seward, “A model for autonomous safety management in a mobile robot,” vol. 1, 2005, pp. 1128–1133.

[13] J. Åslund, J. Biteus, E. Frisk, M. Krysander, and L. Nielsen, “Safety analysis of autonomous systems by extended fault tree analysis,” *International Journal of Adaptive Control and Signal Processing*, vol. 21, no. 2-3, pp. 287–298, 2007.

[14] J. Rasmussen, “Skills, principles, knowledge: Signals, signs, and symbols, and other distinctions in human performance models,” vol. 13, 1983, p. 257–267.

[15] D. Seward, C. Pace, and R. Agate, “Safe and effective navigation of autonomous robots in hazardous environments,” *Autonomous Robots*, vol. 22, no. 3, pp. 223–242, 2007.

[16] T. Smithers, “Autonomy in robots and other agents,” *Brain and Cognition*, vol. 34, pp. 88–106, 1997.

[17] D. Söffker, *Systemtheoretic Modeling of the knowledge-guided Human-Machine-Interaction (In German)*, ser. Habilitation Thesis, University of Wuppertal, 2001. Berlin: Logos Wissenschaftsverlag, 2003.

[18] ——, “Interaction of intelligent and autonomous systems - part i: Qualitative structuring of interactions.” ser. 14, vol. 4, 2008, p. 303–339.

[19] G. Strube, C. Habel, L. Konieczny, and B. Hemforth, “Kognition,” in *Handbuch der Künstlichen Intelligenz, G. Görz, C.-R. Rollinger und J. Schneeberger (ed.)*. Oldenbourg Wissenschaftsverlag, 2004, vol. 4, pp. 19–72.

[20] H. Voos and P. Ertle, “Online risk assessment for safe autonomous mobile robots - a perspektive,” vol. 7th Workshop on Advanced Control and Diagnosis, Zielona Góra, PL, 2009.

[21] A. Wardziński, “Safety assurance strategies for autonomous vehicles,” in *SAFECOMP ’08: Proceedings of the 27th international conference on Computer Safety, Reliability, and Security*. Berlin, Heidelberg: Springer-Verlag, 2008, p. 277–290.