Causal Knowledge and the Logic Of Dynamics in the Development Of Autonomous Intelligent Agents

Bogdan Enciu

National Institute for R&D in Informatics
8-10 Averescu Avenue,
71316, Bucharest
ROMANIA
E-mail:enciu@std.ici.ro / enciu@risc.ici.ro

"Vere scire, esse per causas scire."
Francis Bacon

Following the debates hosted by the "Fuzzy Sets" Seminar, the article gives evidence, in a personal manner, from the author's position and his concerns, some ideas of older or more recent source, related with the problem of causality and causal knowledge embedded into an appropriate logic frame with repercussions on the development of intelligent agents. The action is seen as the main factor which determines the change, with its dynamic contradictory character of the alternative actualization and latency of some opposite phenomena, and its possible consequences.

1. The actions of any agent, be a human or an intelligent system, undertaken to pursue the attainment of a goal within a certain context, are preceded by the decisions made on the basis of some acquired knowledge. Accepting the idea from the motto and if considering that the agent has, in order to produce certain changes to reach a certain goal, to act in the environment, the performances of the agent will be influenced by its causal knowledge in that the agent should know not only the results of an action but also its consequences. The agent can also (re)act in a reflex mode (through reactive actions) to some stimuli but only the actions directed towards a goal achievement can be said to bear the agent's intentions.

The action of an agent can induce a change that can be a result of the action or a consequence of that action. The result could be viewed in a logical relation with the action, while the consequence can be viewed in a causal relation with the action [14], the distinction between the result and the consequence being given, mainly, by the agent's intention. Therefore, the common elements bearing on both relations are the agent's intention and the modality of its acting (given that the result and the consequence can derive from the same action).

The knowledge of the causal connections, as relations between generic events/phenomena – meaning that an event/a phenomenon is the cause of the appearance of another event/phenomenon (effect), is reflected, in principle, in the physical, chemical, biological, economic laws expressed by functional relations. The problem can also be considered in terms of conditional relations [15].

2. As references in approaching some causal theories, the deontic logic, the logics of change and action elaborated by von Wright, can be considered. The deontic logic (a logic of the norms) assumes and includes the formal frame and the principles of the propositional classic logic, and of the logic of change and action [14]. The logic of change describes state transformations, transitions from one certain (generic) state of things to another (generic) state of things, looking upon events as arranged pairs of states. The logic of action is a logic of acts that make changes between states of things. In the logic of action, functors that express the action of and the abstinence from determining the appearance of a state are introduced, a logic of action formula describing the attitude of the agent with regard to a circumstance that is not explicitly expressed in that formula. By substituting some functors and by adding the axioms of the elementary teleologic, von Wright's logic of action can be transformed into a logical theory of purposes [9]. The ever done endeavour to model the activities/the actions of the agents has led to the elaboration and development of modal logical systems, mixed modal systems, temporal teleologic systems, modal extensions to predicate calculus, some of them constituting the basis for solving decision problems or being applied to the agent systems that execute elementary actions and take part in communication.

3. In order to solve the problems, not few, that ask for a common sense reasoning associated with action domains, knowledge of the form: 'the phenomenon ϕ causes the phenomenon ψ' (in the sense that ϕ is a sufficient condition for ψ) is not necessary (or maybe it cannot be

established) but only knowledge that describes conditions under which the phenomena are caused (in the form: 'necessarily, if ϕ (true) then the fact that ψ (true) is caused') [7]. In this latter form it can express "dynamic" causal laws (which reminds of the idea of a practical syllogism [15]: in the context of the necessary action taken to reach a goal, the form under which causality can be seen is given by a necessary but insufficient condition for reaching the goal) or "static" causal laws (both phenomena, consequences of the same action, occur simultaneously – the phenomenon ϕ accompanies the phenomenon ψ).

The formalization in [7] considers a causal theory D which contains a full description of the conditions under which phenomena are caused, an interpretation I for a propositional language with the set of literals L, and defines $D^I = \{q \mid \exists p,\ p \Rightarrow q \in D \text{ and } I \models p\}$ as the set of consequences of all causal laws in D whose antecedents are true in interpretation I, $p \Rightarrow q$ expressing a causal law ('p' and 'q' are formulas of the propositional language but '$p \Rightarrow q$' is out of the language and is not the material implication). The interpretation I is defined as causally explained according to D if I is the unique model of D^I.

In a modal approach, using the 'possible world' concept (a possible world 'v' can be seen as a deontic alternative to a world 'w' satisfying the condition wRv, where R is the accessibility relation [5]), introducing a monadic modal operator 'C' (with appropriate properties) and considering that having $p \in P$ and $Cq \in P$ ('Cq' can be read 'q is caused'), knowledge expressed in the previous second form, can be formalized as: $p \rightarrow Cq$. Having a set of possible worlds W and accepting that in the world $w \in W$, the truth value for 'p' is known but not also for 'Cq' and admitting other worlds from W, $v : P \rightarrow \{0,1\}$, then $v(Cq) = 1 \Leftrightarrow w(p) = 1$ for $\forall v \in W \mid wRv$, where R would be a causal relation between worlds.

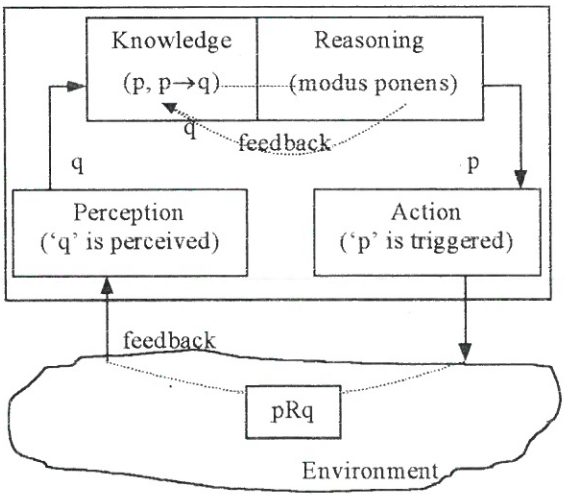
A modal non-monotonic logic (called UCL) for representing causal knowledge (of the second form, previously presented), based on the principle of universal causation, is introduced in [11]. A causal law ($p \rightarrow Cq$) has associated a pair (I,S) where I is an interpretation and S is a set of interpretations which I belongs to ($I \in S$). Based on the S5 modal system, UCL imposes two conditions on the truth of the propositions: $(I,S) \models p \Leftrightarrow I \models p$, and: $(I,S) \models Cp \Leftrightarrow \forall I' \in S$, $(I',S) \models p$ (i.e. a proposition is true in (I,S) if and only if it is true in I and, respectively, a proposition is necessarily true if and only if it is true in any (I,S)). The interpretation I can be seen as a "now" possible world and S as the set of "now and in the future" possible worlds. Given a theory T, the pair (I,S) is a model of T if $(I,S) \models p$, $\forall p \in T$. The interpretation I is defined as being causally explained by T if $(I,\{I\})$ is the unique model of T. This logic embeds some of the causal theories proposed till now (including [7]) for reasoning about actions [11].

4. Establishing the causal character of a relation between two generic events implies the notion of contrafactual condition [15] (provided an event is not present, if it were made present then the other event would also have appeared). Therefore, the causal relation could be determined on the basis of the idea of action.

A model that would describe the causal/determination relations between generic events/phenomena (considering the agent's action as determining the appearance of a certain state or, in other words, triggering the cause of a state appearance) could be that of a dynamic system obtained by modeling the logic of change with the help of the concept of possible world. A dynamic system can be thought of as a set of relations between possible worlds or, formally, as a function $f : L \rightarrow L$, where L is a set of possible worlds [4]. We consider A_0 a set of elements and $R_1, R_2, \ldots, R_n$ as possible relations between these elements. We form the sets $B_1 = \{y \mid xR_1y,\ x \in A_0\}$ and $A_1 = A_0 \cup B_1$. By induction there results $A_{n+1} = f(R_n, A_n)$ which can be seen as a dynamic system, where A_i represents a state (a possible world) and R_i is a command. If we consider a subset A_0 from the set of propositions ($A_0 \subset P$), the set of tautologies T and a function $F : P \times P \rightarrow P$, $F(p,q) = p \rightarrow q$ (material implication), the reverse image $F^{-1}(A_0 \cup T)$ can be seen as R_1, operating as a 'modus ponens' applied to the set $A_0 \cup T$, thereby 'q' is detached. The same as above, we obtain $A_1 = A_0 \cup T \cup B_1$ ($q \in B_1$, in its turn, 'q' could have the form: '$r \rightarrow s$'). By induction here results $A_{n+1} = f(R_n, A_n)$ and also $R_n = g(A_n)$ (the relations R_i are different at each step depending on the modification of the sets A_i).

Given that the agent has the capacity to learn (to "seize" the determining relations between certain phenomena), this can induce changes meant to force the occurrence of other phenomena which can thus be considered as being caused by its action. In other words, the agent can perform experiments as a method of inductive learning. Therefore in the case that the agent does not know a certain causal relation/connection, it can act in the environment (inducing 'p'), and perceive the

effect ('q') of the action, the relation between 'p' and 'q' being in the environment. The interaction leads to the discovery of 'q' (the "extraction" of 'q' happens as it does in the 'modus ponens' reasoning). Actually, the environment itself contains the knowledge that the agent would require in order to reach its goals and this knowledge can be learned just from the interaction with the situations created/turned up in the environment [3]. If the effect of an action is not at variance with the goal, the reaction of the environment is considered to be positive and able to strengthen the initial action of the agent (two situations can be distinguished: in the case when the realized action, in accordance with the context, is further necessary, the action will be given the highest priority; in the case when the realized action is no longer necessary - although following its assessment the action could be declared as 'realized' and its priority might keep the same or even higher - an event assessment could reveal the necessity of the execution of another action, which will be given the highest priority [3]). If the effect of the action does not fit the goal, the reaction of the environment is considered to be negative and the agent's next action will be different (the previous action will have a lower priority). Therefore the effect (the reaction of the environment) acts on the cause that produced it. This fact, which determines the change of the relation R_i, is the 'feedback' [4] to be seen as a characteristic of the causal connections.



Provided the agent's knowledge base contains causal knowledge, the agent may carry out action planning activities (by the internal feedback, a necessary condition for the agent to be "conscious" of its actions), the external feedback making it possible to acknowledge the fulfilment/failure of an action.

5. An interesting aspect to approach by the representation of some of the conditions under which phenomena occur, consists of the infinitesimal variations of some parameters which can make unexpected changes take place in the environment where the agent performs. The importance of such phenomena (singular points) was remarked by Maxwell [10]. Such phenomena/events, somehow of an uncertain/undetermined nature and with a low probability of occurrence, should be considered provided their occurrence considerably influences the agent's activity (although it might seem to uselessly charge a well-"determined" application). Such a situation raises the problem of the impossibility of perceiving the fine variations of some parameters (which would make it impossible to detect an event that would be the cause of the occurrence of another event - this in the instance the causal relation is known) as well as the problem of the knowledge about the cause(s) of the phenomenon which might be missing. Still the occurrence of an event/phenomenon can be accompanied by another detectable event/phenomenon (which could, as previously shown, be expressed as a condition under which the first event occurred). Therefore, the agent has the possibility of acting on the occurrence of an event even though its occurrence was not directly detected.

6. The representation of some causal knowledge under the form of some functions (as physical laws) may not satisfy the condition for an agent to perform certain tasks in the sense of reaching some goals (the agent will modify somehow the environment, its freedom of action being obviously conditioned by the laws of the environment, it will not merely be allowed to know the environment and act upon it only for the survival - this can be seen as a goal in itself). The explanation resides in that that a complex environment, with various restrictions, and with other agents performing in that environment, has no possible functional descriptions, a logic-conditional description of that knowledge being imposed.

As to intelligent agents, notions such as belief, commitment, desire, choice, even emotion are mentioned [13], which can be an expression of the agent's intention (make it to intend the triggering of an action). Such features, which are imposed on an intelligent agent, raise some problems concerning the adoption of and the manifestation of a characteristic "attitude" to the changes in the environment. Changes determined by natural phenomena or other agents, can formally be seen as being generated by the alternative actualization and latency of certain contradictory events/propositions

(somehow in the sense mentioned in [6]) or as a transition from one possible world to another (considering Leibniz's view on the concept of possible world as a non-contradictory world meaning that 'p' and 'non-p' cannot exist simultaneously). One problem is the "moral attitude" which can generate contradiction or contrariety ('to help friends and to harm enemies' is in contradiction with 'to harm friends and to help enemies' but 'to help friends and to help enemies' inspires contrariety [1]). A moral attitude can be given by specifying a value system with (predefined) priorities (we tried to underline the importance of a value system in a conceptual schema of an architecture for intelligent agents proposed in [3]). Another problem (also a facet of contraries but at a different level) arises from the fact that conflicts between the goals (or, better said, between a goal and a sub-goal) of an activity can appear and manifest at the level of the performed actions. This problem (which implies the notion of ability in the sense given in [2]) could be solved by establishing dynamic updated priorities associated with the agent's possible actions. The fact that other agents also act in the environment can be seen as the main factor which can create contrariety for an agent.

7. The contradictory, antagonistic character of the activity of the cerebral hemispheres, which can be considered as significant for human intelligence, is also deducible from the observation made on human subjects who have undergone surgical separation of the cerebral hemispheres at the level of the corpus callosum performed to ameliorate the effects of epileptic seizures. Following surgery, it has been observed that, at the level of perception and of manipulation by the hands, the results offered by the right hand and the left hand were completely different, a phenomenon known as 'alien hand syndrome' [16]. The syndrome was also common with subjects who suffered some brain injuries. If touching with both hands the same object successively, the subject will give different answers about the shape or the structure of the object. When the subject was asked and had to answer with 'yes or no' by indicating the answer with a finger from each hand on a sheet of paper, it was noticed that some subjects indicated a different answer with each hand. Often, the left hand could not help acting contrary to the right hand.

We can assume that the link between both hemispheres is decisive in the process of (re)acting to stimuli. The breaking-down of communication between cerebral hemispheres leads to a contradictory behaviour, as a result of some contradictory "impulses", each tendency being opposed a contrary one. Beside the opinion that these phenomena prove the dual character of the human consciousness, we can consider that the arbitrating mechanism of these "contradictions" (finished with the final option) constitutes a determinant element of the intelligence. On the other side, these tendencies, which are proved as being naturally contradictory, seem to be in the sense of some affirmations with philosophical character as "the positive proposition must imply the existence of the negative proposition and reverse" [12] or as that expressed in the postulate of a dynamic logic of contradiction: "for each phenomenon, or element, or logical event, as also for the judgment that it thinks, for the proposition that expresses it or the sign that symbolizes it ... have always to be associated with it, structural and functional, an anti-phenomenon, or anti-element, or logical anti-event, and therefore a contradictory judgment, proposition, sign ..." [6].

In the sense of these affirmations, beside the fact that we can note that some machine learning techniques imply to provide positive examples as well as negative examples during the learning process of concepts [8], decision-making problems for triggering or inhibiting certain actions, can be approached. For an intelligent agent, we can raise the problem of a simultaneous consideration of some assessment processes for direct actions as well as for contrary or different actions with regard to the same verb (if the action is that of 'going forward', it is obvious that between 'not going forward' and 'going back' there is a difference whose consequence may be important to assess). A suggestive enough example could be: an autonomous agent has to move to a certain position. The locomotory subsystem receives and executes the commands of movement in the direction given by the coordinates of the position but, after covering a certain distance, an obstacle detected by the proximity sensors appears. This is a contradictory situation: on the one hand the agent has the task of reaching the given position (which implies the movement in that direction) and, on the other hand, the agent must avoid collision with the obstacle which it cannot overpass (which can imply stopping or moving in another direction while trying to avoid the obstacle). These two tasks could be performed by distinct processes. The activity of learning and planning could be improved by considering contradictory phenomena/actions, including the creation of hypothetical scenarios. We could implement processes/modules able to provide answers different from one another or even contrary to one another (in the sense of action or non-action, of action in contrary

directions or in the sense of establishing different control parameters) the selection being done (following assessments) according to the criteria determined by the context and/or by a value system through a process of arbitration (the calculus of the gravity centre for a fuzzy system can be said to achieve such an arbitration by averaging the values obtained by applying fuzzy rules).

REFERENCES

1. ARISTOTLE, **Organon, Topics** (II, 7, 113a), IRI, 1998.
2. ARISTOTLE, **Nicomachean Ethics** (VI, 12, 1144a), IRI, 1998.
3. ENCIU, B., **Artificial Intelligence from Perception to Reasoning (A Value_System-based Architecture)**, STUDIES IN INFORMATICS AND CONTROL, Vol. 7, No. 3, 1998, pp. 221-225.
4. FLONDOR, P., **A Presentation at the ICI-based Seminar on Fuzzy Sets**, Bucharest,1999.
5. HUGHES, G.E. and CRESSWELL, M.J., **An Introduction to Modal Logic**., METHUEN AND CO LTD, 1968.
6. LUPASCO, S., **Dynamic Logic of the Contradictory**, Political Publishing House, 1982 (in Romanian).
7. MC CAIN, N. and TURNER, H., **Causal Theories of Action and Change**, Proceedings of AAAI-97, 1997.
8. MITCHELL, T.M., **Machine Learning**, WCB/MCGRAW-HILL, 1997.
9. POPA, C., **Action Theory and Formal Logic**, Scientific and Encyclopedic, 1984 (in Romanian).
10. PRIGOGINE, I. and STENGERS, I., **La nouvelle alliance - Metamorphose de la science**, EDITIONS GALLIMARD, 1979.
11. TURNER, H., **A Logic of Universal Causation**, University of Texas at Austin, USA, 1998.
12. WITTGENSTEIN, L., **Tractatus Logico-philosophicus** (5.5151), HUMANITAS, Bucharest, 1991.
13. WOOLDRIDGE, M. and JENNINGS, N.R., **Intelligent Agents: Theory and Practice**, submitted to KNOWLEDGE ENGINEERING REVIEW, October 1994, revised version January 1995.
14. WRIGHT VON, G. H., **Norm and Action**, ROUTLEDGE AND KEGAN PAUL, 1963.
15. WRIGHT VON, G.H., **Explanation and Understanding**, CORNELL UNIVERSITY PRESS, 1971.
16. ***. **Alien Hand Syndrome**. http: //www.indiana.edu/~pietsch/alienhand-psy.html.