Position Paper: Towards ethical agents in Distributed Constraint Reasoning

Ghizlane EL KHATTABI, Imade BENELALLAM, El Houssine BOUYAKHF, and Rajae HAOUARI

Physic department, Mohammed V university, LIMIARF B.P.1014 RP, Rabat, Morocco
SI2M Laboratory, INSEA, Irfane Institut, Rabat, Morocco
elkhattabi.ghizlane@gmail.com,
imade.benelallam@ieee.org,bouyakhf@mtds.com,
rajae.haouari@gmail.com

'Ethics' is a new research area in artificial intelligence. Several studies have, recently, begun this discipline, such as "Machine Ethics" [1][2][3] and "Military Robot" (or the "killer robots") [4]. The main purpose of "Machine Ethics" is to create an "Artificial Moral Agent" (or AMA) model of a moral machine that follows a set of ethical principles in order to make decisions. For "Military Robot", the authors wondered whether "We Want Robot Warriors to Decide Who Lives or Dies?". Assuming that the robot term that has been appeared in order to produce low cost goods, it has ended with non ethic robots that kill the humans race, because they have varying degrees of autonomy (some tasks are controlled by humans) that will be in the future a complete autonomy. And fearing that robotic weapons may trigger a world war.

A machine or a robot ethics involves ethical intelligent agents, and then ethical Multi Agent System SMA. One of the most important programming paradigm that uses SMA system is the DisCSP formalism. Using DisCSP, agents may have a high level of decision autonomy within unethical internal solver (i.e. unsupervised solver). This internal flexibility allows them to take autonomous decisions, based on its internal objectives. In this context, agent can make use of some approaches that do not meet ethical principles and rules. For example, in DisFC-lie algorithm[1] [5], an agent can lie on its local solution in order to meet its requirements. This behaviour can effect the whole resolution process.

In this position paper, we have conducted some preliminary experiments, using a modified version of ABT [6] that introduce liar agent in a distributed constraint network. The results show that during the resolution process: just 12.2% of problems have found the right solution, 17.7% of problems have missed the first solution, 25.6% of problems have given inconsistent assignment and almost 44% problems have fell into infinite loops.

Based on the prototypes used for achieving the "Machine Ethics", we aim to propose a new extension of DisCSP, named E-DisCSP (Ethical DisCSP). Firstly,

[1] Here, the objective is to keep the value privacy, assuming that an agent musts send the correct value, in order to catch up the global resolution process. However, in some real time decision problems, the consequences of this incorrect value may be dramatic

we intend to detect the abnormal activity in the presence of unethical agent(s), appoint them and then take actions vis-a-vis such agent behaviours. This new formalization will be based on the same parameters of DisCSP, while adding three other components: a set of ethical rules R, a set of actions (or decisions) A_c and a function that associates each rule to its action(s) $f, (f(R_i) = A_{c_j}$ means if the rule R_i is violated then the action A_{c_j} is applied).

To control the abnormal activities in a E-DisCSP problem, we use the Problem solving System concepts [7]. It is a conjunction of two important components: the inference Engines (IE) such as the rule-based inference systems, and the Truth Maintenance System (TMS or Reason Maintenance System RMS) that handles the beliefs (decisions) in the offered sentences (informations).

The benefits of the TMS systems include the opportunity to provide justifications, identify inconsistencies, and uphold default reasoning. The sentence can replace a fact (Socrates is a man) or a rule (If X is man then the X is mortal). The possible beliefs that the TMS can report are: false (contradiction), true (premise), assumed-true (enabled assumption), assumed-false (retracted assumption), assumed or don't-care. The sentence is considered IN if the belief is true or assumed-true. It is OUT if belief is false, assumed false or don't care. The IE can order the TMS to add a sentence, to create a justification, to link some rules to sentences,and to apply them when some beliefs hold.

There are different characteristics of the TMS: the JTMS (Justification-Based TMS), ATMS (Assumption-Based TMS) [8] and the LTMS (Logical-Based TMS). In the JTMS and the LTMS, just one set of current assumptions is examined. But the ATMS can examine several simultaneous current assumptions. In our case, is the ATMS the most useful.

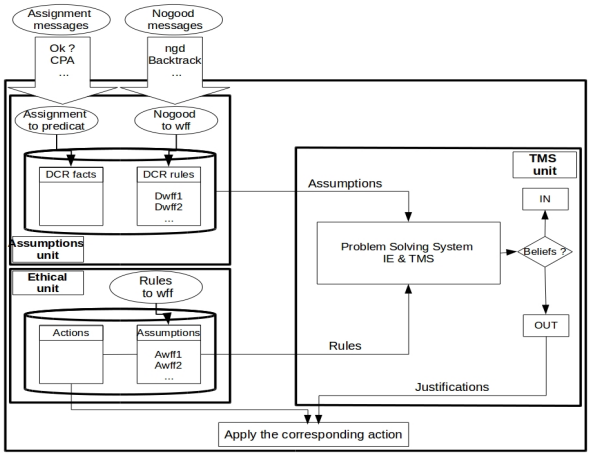

Fig. 1. control procedure

To implement this system, in each DisCSP agent, we add a control unit (Figure 1), that includes: i) the Assumptions Unit AU, which converts each received message (assignment or nogood) to a wff (Well-formed formula) assumption, ii) the Ethical Unit EU, that stores ethical rules as wffs and the corresponding actions that represent a penalty function. An action can be applied if the corresponding rule is violated, and iii) the Truth Maintenance System Unit TMSU that is the most important unit. It uses the received information from the AU unit in order to check if the stored rules in EA unit are satisfied, using IE and the ATMS system. If the inconsistency (OUT) is detected the corresponding action is applied, according to the given justifications.

We perceive the creation of a method to convert any information or received message into a wff and to propose an algorithm that implement this control unit functions.

References

1. Michael Anderson, Susan Leigh Anderson, and Chris Armen. Towards machine ethics. In *AAAI-04 workshop on agent organizations: theory and practice, San Jose, CA*, 2004.
2. James H Moor. The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems*, 21(4):18–21, 2006.
3. Michael Anderson and Susan Leigh Anderson. *Machine ethics*. Cambridge University Press, 2011.
4. Robert Sparrow. Killer robots. *Journal of applied philosophy*, 24(1):62–77, 2007.
5. Ismel Brito and Pedro Meseguer. Distributed forward checking may lie for privacy. In *International Workshop on Constraint Solving and Constraint Logic Programming*, pages 93–107. Springer, 2006.
6. Christian Bessière, Arnold Maestre, Ismel Brito, and Pedro Meseguer. Asynchronous backtracking without adding links: a new member in the abt family. *Artificial Intelligence*, 161(1):7–24, 2005.
7. Kenneth D Forbus and Johan De Kleer. *Building problem solvers*, volume 1. MIT press, 1993.
8. Johan De Kleer. An assumption-based tms. *Artificial intelligence*, 28(2):127–162, 1986.