Computational Modelling of Metacognition in Emotion Regulation

Catriona M.Kennedy[0000−0002−1699−9328]

School of Computer Science, University of Birmingham, Birmingham, B15 2TT
catm.kennedy@gmail.com

Abstract. Computational metacognition is an AI paradigm that enables an agent to inspect and modify its own reasoning. This paper describes work-in-progress research on metacognition for emotion regulation in simulated agents. We are using example scenarios where negative emotions interfere with the agent's goal. The agent should override these emotions using reappraisal as a regulation strategy.
Our initial work has identified the following challenges: (a) for general and robust recognition of the "involuntary" nature of emotions, the agent needs to develop a model of its own volition; (b) to implement reappraisal, metacognition needs to initiate new reasoning strategies so that the emotion-causing situation can be evaluated more positively; (c) to enhance robustness and adaptability, the architecture needs distributed and parallel processes within the deliberative and metacognitive layers.

Keywords: cognitive architecture · emotion models · metacognition.

1 Introduction

This paper presents work-in-progress research into computational metacognition as a paradigm for emotion regulation modelling. Metacognition is the capability of an agent to inspect and modify its own reasoning and decisions [1].

For emotion regulation, our focus is on re-direction of reasoning processes so that unwanted emotion-driven thoughts (and subsequent behaviour) are avoided. This kind of regulation is called reappraisal [2]. Our purpose in modelling is to enable rapid prototyping of agents to improve conceptual precision and to inform the design of assistant agents for e-therapy or behaviour change.

The paper is structured as follows. Sections 2 and 3 summarise related work and background respectively. Section 4 defines some key (prototypical) aspects of emotion as it occurs within a cognitive architecture, using H-CogAff [3] as a starting point. Our working definition emphasises the involuntary nature of emotions [4]. We also use concepts from appraisal theory [5] and action readiness [6]. Section 5 defines some example scenarios involving emotion regulation, and then considers the information processing requirements for an agent to cope in these scenarios. In Section 6, we present our current work in building a metacognitive agent to satisfy the requirements generated by the scenarios. Finally, Section 7 summarises future challenges.

2 Related work

Our work is most related to computational cognitive architectures which model human cognition and emotion at a conceptual level. These include H-CogAff [3] and the critic-selector architecture of Minsky [7]. In Minky's model, networks of "critics" and "meta-critics" evaluate a situation using different methods (or "ways to think"), while "selectors" determine what actions (or alternative thinking) is required.

Other related work includes MAMID [8], which models the effects of emotion on decision-making and EMA [9], which is an appraisal model that can activate coping responses, some of which can modify cognitive processes to regulate emotions. The main difference between these two models and ours is that we are emphasising the role of emotion as an involuntary disturbance to the agent's current deliberation, which it must recognise and counteract.

Theories of behaviour change and human motivation are also relevant. In particular, the concept of "affective force" of a motivation or emotion is important in the CEOS model of hard-to-maintain behaviour change [10] and in the PRIME model of motivation [11].

3 Background: simplified H-CogAff

We use a simplified version of H-CogAff [3] as a starting point. This is shown in Figure 1 as a three-layer structure: reactive, deliberative and metacognitive. Two different views (a and b) of this architecture are used. The first (a) is shown in Figure 1 and shows a dual process architecture which separates "executive control" from the reactive layer. Executive control includes both deliberation and metacognition because they are both involved in top-down control and hypothetical reasoning (evaluating non-actual possibilities).

The second view (b) is shown in Figure 2 and separates "meta-level" from "object-level". This view distinguishes monitoring and control of *the world* (object-level) from monitoring and control of the agent *mental states* (meta-level). For example, metacognition may detect that the agent has a gap in its knowledge (due to unexpected outcomes) and initiate a plan to learn the required knowledge. Our architecture for emotion regulation is a specialist form of metacognition in H-CogAff.

All three layers in H-CogAff run in parallel. In the simplest model, the internal processing within the deliberative and metacognitive layers is largely sequential, while the reactive layer is internally highly parallel. More complex and biologically plausible models are possible, which include parallel and distributed processing within each of the two upper layers.

4 Emotions and affective processes: working definitions

We will build on the definition of emotion in [3]. Namely, *emotions* are affective states that *interrupt or modulate the current processing*. For example, attention

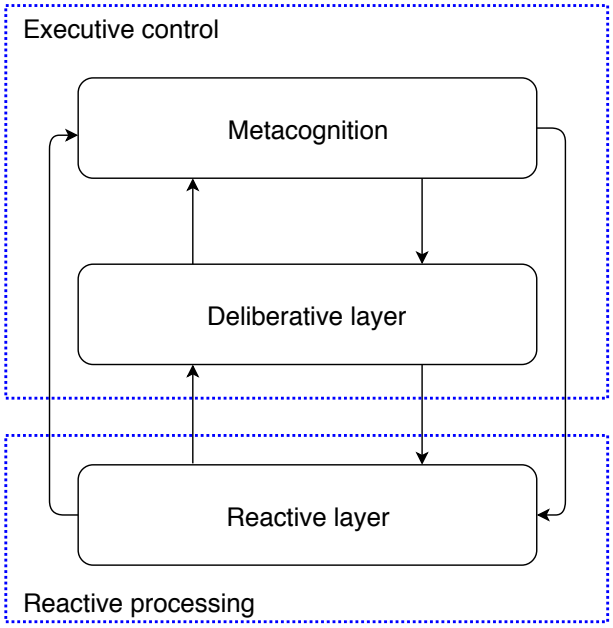

Fig. 1. A simplified version of H-CogAff, showing view (a): the distinction between executive control components and reactive layer.

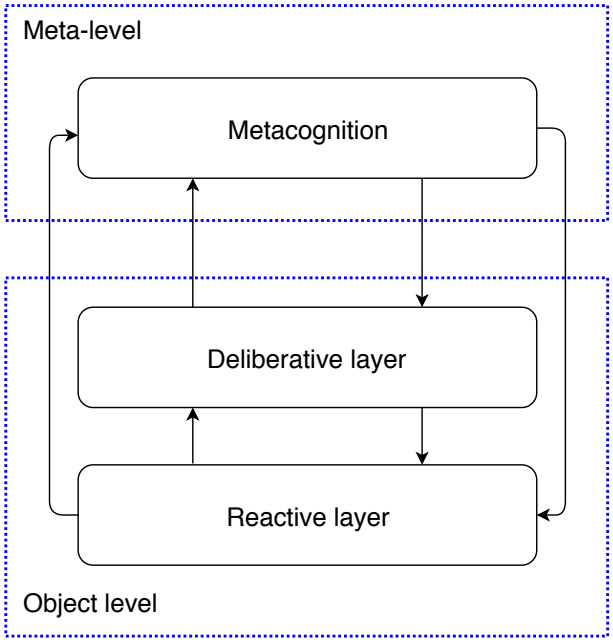

Fig. 2. View (b) showing the distinction between meta-level and object-level

may be directed involuntarily towards an emotion-activating situation. Since the term "current processing" implies a sequential process, we assume that the current deliberation is interrupted (as it has a sequential nature). The disturbance may also affect the metacognition or the reactive layer.

Affective states are defined by [3] as follows: a negative state tends to cause avoidance of the state (e.g. pain), while a positive state tends to cause persistence of the state (e.g. pleasure). For our purposes, it is important to identify the *process* by which an affective state causes change.

4.1 Affective process

Our definition of *affective process* is based approximately on the stages in Gross's process model [2], where each stage represents a point where emotion can be regulated. An affective process can originate from the executive control layer or from the reactive layer. It involves the following steps:

1. direction of attention to perceive a situation;
2. evaluation of the situation (which may lead to further attention focus);
3. preparing for action (if required).

Evaluation is a key concept in appraisal theory [5, 13]. The "situation" evaluated can be any kind of state, process or action, and can be actual or non-actual (e.g. hypothetical or remembered). The evaluation can be extremely simple (good or bad) or it can involve more complex forms of cognitive appraisal (e.g. beliefs about relevance of events to goals). In the case of metacognition, the situation being evaluated can be the agent's own reasoning or affective processes.

4.2 Action preparation

Another important component is preparation for action (emphasised by [6]). We define this as a process that generates *potential actions* that are ready to be executed. In a deliberative system, this process can be divided into steps which address the following questions:

1. what states would be better than the current one (what is desirable)?
2. what desires should be pursued as goals? This involves reasoning about what is achievable, or compatible with other goals.
3. what goals or actions should be chosen right now?

Each step produces potential actions at increasing levels of priority, with (3) being the most immediate. In a reactive system, a predefined action may be triggered immediately (e.g. move away from a speeding vehicle) or it may be generated as a highest priority action (e.g. find food).

5 Emotion regulation: requirements

To address emotion regulation, we consider some scenarios where regulation is required and then determine what capabilities are needed.

Scenario 1: Overriding alarm signals: Jane wants to tidy up the childrens bedroom quickly because she has an appointment later in the morning. Outside, she hears loud and angry voices and becomes anxious. She goes outside to investigate. However, it appears only to be a heated argument with no threat of violence, so she resumes her work while attempting to ignore the raised voices.

Scenario 2: Smoking cessation: Tom has decided to stop smoking and has so far succeeded. However, after a row with a colleague at the office, he feels angry and disappointed. These negative emotions interfere with his resolve not to smoke and he feels that a cigarette can help him calm down. To counteract these emotions, he thinks about the long-term advantages of persisting as a non-smoker and the need to put into perspective the relatively minor issue of the office confrontation.

The type of regulation required in both can be classed as "reappraisal", which is one of the strategies in [2].

6 Building a metacognitive agent

Metacognition is highlighted in Figure 2 and can be divided into *monitoring* and *control*. Monitoring evaluates the performance of the object-level. Control makes necessary adjustments to object-level processes. Effectively, meta-level monitoring is a form of *appraisal* of thought processes.

In the tidying up scenario, the meta-level needs to detect that the object-level is being disrupted by anxiety about violence. In the non-smoker scenario, it needs to detect the disruption to the object-level caused by negative emotions resulting from the office experience.

Currently, we are making the simplifying assumption that the disturbance only affects the deliberation, not the metacognition. In more complex models, the disturbance can also affect some metacognitive processes. This needs a distributed architecture on the metacognitive level. An early version of distribution is given in [14]. The challenge is to apply these principles to human-level cognitive models.

6.1 Monitoring: importance of a reasoning trace

For the meta-level to detect problems, the object-level needs to leave a *trace* of its reasoning for analysis by the meta-level. Such a trace can include important steps in the object-level processing (like an audit trail). Some traces can also allow an agent to *explain* its reasoning introspectively [12]. Different kinds of trace may be generated. Examples are below, along with examples of introspective narratives:

T1: changes in the agent's attention focus over time. E.g. what scenes, people or actions were being considered? An example narrative using T1 and labels X, Y might be: "I was thinking about X; suddenly I remembered Y".

T2: within each context in T1, what were the agent's beliefs and evaluations (appraisals) and how did they change? These would contain statements, not just labels. Example narrative: "I saw that (the room was in a mess)", and "I hoped that (there would be time to tidy it)". Statements are in brackets.

T3: deliberative decisions: The following information might be recorded for each decision: (a) Current goal and the options that were being considered; (b) How were the options evaluated (what criteria/reasons)? (c) Which option was chosen, and why? Example narrative: "I planned to put the toys in the box first because they were on the floor and a trip hazard".

Agent motivational state: Varying levels of detail may be recorded in the above traces, depending on the requirements of the metacognitive process at the time. For example, in the tidying up scenario, the agent will be planning sequences of actions (such as first putting all toys in a box) and then executing them. Two important decisions are taken repeatedly: (1) what to focus attention on next? (e.g. toys on floor); (2) what is the next action? (The action may be the next step in a pre-defined plan, or there may be a decision made on the basis of priority or urgency). The deliberative component records these events in the trace. A redirection of attention is recorded in T1; the next action is recorded in T3 (possibly along with reasons, which might involve beliefs in T2).

We define the agent's *current motivational state* as a collection of variables that might be recorded in a trace. This includes current attention focus and current potential actions.

6.2 Simulating disturbances

An agent simulation for the room tidying scenario is under development. Emotions can be activated by creating random events in this simulation which activate an affective process in the agent. The event can be in the external world (e.g. loud voices) or a memory (as in the office confrontation scenario).

In the room tidying scenario, the event can initiate an affective process, which directs attention to outside the house and changes the "next action" involuntarily to "investigate". This changes the agent's motivational state in a way that was not planned.

6.3 Recognising disturbances

Currently we are building a meta-level decision process which determines if a potential action (or goal) was generated by a process of deliberation or whether

it was generated reactively. The meta-level determines whether the current motivational state (attention focus, potential actions etc.) is consistent with the intended one resulting from the deliberative process recorded in the trace. If there is no record of deliberative reasoning leading up to the current state, then an involuntary disturbance has been detected.

However, this definition of "involuntary" is very simplistic, since it is based on a predefined trace pattern. It is important to think more generally about how an agent distinguishes between its own volition and outside influences. This requires a sense of self. In particular, since emotions can interfere with deliberation in complex ways (e.g. by introducing biases), it may be useful to learn the difference between the effects of a "negative" emotion and a "positive" one. In this way, the agent can develop models of how it is affected by emotions, along with its own strengths and vulnerabilities. This is a long-term challenge. Early work on learning a sense of identity for very simplified agents is in [15]. This needs to be applied to emotion models.

6.4 Meta-level control and "affective force"

Computationally, it is easy to delete the agent's current (disruptive) motivational state and restore the original state that was generated by the deliberative process, but this would not be a model of human emotion regulation. To stop a negative emotion from causing unwanted thoughts or behaviour, the meta-level needs to generate an "affective force" which is greater than the opposing one [10, 11].

Current work is investigating the modelling of "affective force" as the generation of *reasons* why a disruptive emotion should be overridden. Minsky's critic-selector architecture [7] can provide a library of alternative "ways to think" about a situation, which can be useful for reappraisal. Each "way to think" can be selected and evaluated (by a meta-critic) according to how much counteracting "force" it can generate. This is similar to decision-making using debate and argumentation.

7 Challenges and future work

This work has identified a number of challenges. First, the current model only recognises involuntary disturbances that do not match a predefined deliberative trace. A future challenge is to learn general patterns of volitional processes and how they can be affected by emotion.

Secondly, current meta-level control overrides emotional interference by restoring the agent's attention and priorities to their previous state. To implement reappraisal, the meta-level control needs to select new reasoning strategies so that re-evaluation is possible. Furthermore, to actually make a difference to the agent's motivation, it needs to generate sufficient "affective force" using reasons and arguments.

Finally, to address both these challenges, distributed and parallel architectures are required, both for metacognition and deliberation.

References

1. Cox, M. T. and Raja, A.: Metareasoning: an Introduction. In: Cox, M. T. and Raja, A. (eds.) *Metareasoning: Thinking about Thinking*, pp 3–14, MIT Press (2011).
2. Gross, J. and Thompson, R.: Emotion Regulation: Conceptual Foundations. In: Gross, J. (ed.) *Handbook of Emotion Regulation*, New York: Guilford Publications (2007)
3. Sloman, A., Chrisley, R., Scheutz, M.: The Architectural Basis of Affective States and Processes. In: Fellous, J.-M., Arbib, M.A. (eds.) *Who Needs Emotions?* New York: Oxford University Press (2005)
4. Simon, H. A.: Motivational and Emotional Controls of Cognition. Psychological Review, **74**(1), 29–39 (1967)
5. Scherer, K. R.: Appraisal theory. In T. Dalgleish, T. and Power, M. (eds.), *Handbook of Cognition and Emotion*, pp. 637–663. Chichester, UK: John Wiley and Sons. (1999)
6. Frijda, N. H.: Emotion, cognitive structure, and action tendency. *Cognition and Emotion*, **1**, 115–143. (1987)
7. Minsky, M.: The Emotion Machine: Commonsense thinking, artificial intelligence, and the future of the human mind. New York: Simon and Schuster. (2007)
8. Hudlicka, E.: Reasons for emotions: Modeling emotions in integrated cognitive systems. In W. Gray (Ed.), *Integrated models of cognitive systems*, 137. New York: Oxford University Press. (2007)
9. Marsella, S. C., Gratch, J.: EMA: A process model of appraisal dynamics. *Cognitive Systems Research*, **10**, 70–90. (2009)
10. Borland, R.: CEOS Theory: A Comprehensive Approach to Understanding Hard to Maintain Behaviour Change. *Applied Psychology: Health and WellBeing*, **9**: 3–35. (2017) https://doi.org/10.1111/aphw.12083
11. West, R.: The PRIME Theory of motivation as a possible foundation for addiction treatment. In: Henningfield, J., Santora, P. and West, R. (eds.) *Drug Addiction Treatment in the 21st Century: Science and Policy Issues*. Baltimore: Johnś Hopkins University Press (2007)
12. Cox, M. T.: Metareasoning, Monitoring, and Self-Explanation. In: Proceedings of the First International Workshop on Metareasoning in Agent-based Systems at AAMAS-07, pp. 46–60, (2007).
13. Ortony, A., Clore, G., and Collins, A.: The Cognitive Structure of Emotions. Cambridge, MA: Cambridge University Press.(1988)
14. Kennedy, C. M.: Distributed Reflective Architectures for Anomaly Detection and Autonomous Recovery, Ph.D. dissertation, University Of Birmingham, Birmingham, UK (2003)
15. Kennedy, C. M.: Distributed Meta-Management for Self-Protection and Self-Explanation. In: Cox, M. T. and Raja, A. (eds.) *Metareasoning: Thinking about Thinking*, pp 3–14, Cambridge, MA: MIT Press (2011).