

# Metacognition and the Process of Switching Between Reasoning Frames

James L Eilbert

AP Technologies  
6 Forrest Central Dr, Titusville, NJ 08560  
jleilbert@verizon.net

## Abstract

Evidence is offered to support two claims about human metacognition. First, metacognition is a ubiquitous part of cognition and has deep evolutionary roots. It is claimed that two driving forces for the evolution of metacognition are: awareness that the senses are error prone, and the competing objectives involved in carrying out even a simple task. Reasoning and metareasoning are needed to select among behaviors that on one hand have outcomes that depend on the situation estimate, and on the other hand have different impacts on each of the competing objectives.

Second, metacognition has two important components – metareasoning within a single reasoning frame, and switching between reasoning frames. Fauconnier’s linguistic studies suggest that Mental Spaces act as frames in which humans can do reasoning. Fauconnier has studied the construction and initialization of Mental Spaces extensively, but has been less specific about the process of switching between Mental Spaces. A case is presented for emotions playing a critical role in this process by triggering and directing the switches between Mental Spaces. Affective cognitive models that can already perform emotional computations can readily be adapted to direct switching between reasoning frames.

## The Need for Metacognition in Simple Tasks

Simple navigation has at least three objectives competing for the control of the motor system. There is the low-level objective of not colliding with anything; there is a recognition objective to move to locations where it is possible to recognize landmarks and other relevant objects; and there is a high-level objective of moving toward the ultimate goal. Any behavior selected and executed by the navigator will have an effect on all three objectives. In the case when the navigation is being done by a human, there will be explicit or at least implicit metareasoning about the consequences of each decision. If reaching the final navigation goal is particularly critical, then the human may select whatever behavior gets them to the goal most quickly, implicitly deciding to ignore any other consequences of the behavior selected except a damaging collision. On the other hand, if the main navigation goal is to return home eventually, and the recognition goal is to explore an unfamiliar area, then reasoning is likely to

select behaviors that get the human into positions where unexplored areas can be seen.

Competing objectives can act as a driver for human meta-reasoning within a single reasoning frame. Introspective monitoring is important in exploration, since the person must decide whether the selected behavior is getting the desired information and should continue, or whether a new behavior should be selected. Metareasoning would also be needed to determine when resources should be committed to the competing navigation goal of returning home by a certain time. It would have to decide how often to run the time-to-get-home calculation. This would require introspective monitoring to determine when the person was near the limit, and to check if conditions that might substantially change the time to return home significantly had occurred. Thus, the execution of even simple tasks with competing objectives can benefit from both meta-level control and introspective monitoring, which are described in detail in the Metareasoning Manifesto (Cox & Raja 2007) and shown in Figure 1.

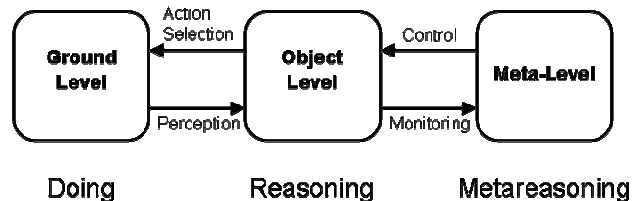


Figure 1: Metacognitive Reasoning in a Fixed Reasoning Frame

Uncertainty about what is being sensed and resulting uncertainty in estimating the current situation causes humans to frequently switch between reasoning frames. Deciding when to switch reasoning frames and which frame to move into requires a different set of metacognitive capabilities than the ones needed to deal with competing objectives. As described below, there is evidence that this switching is closely tied to emotional response and the appraisal of motive satisfaction. To include metareasoning about the switching between reasoning frames, Figure 1 would need to have many interacting combinations of Ground Level and Object Level processing, rather than just one. Metareasoning

would have to decide on which combination to focus attention.

## Evolutionary Roots of Metacognition

Continuing with our navigation example, consider the case where the navigator is a cat. There are still competing objectives and behavior selections being made, but is there real metareasoning going on? A cat won't chase a squirrel once the squirrel reaches the trees, although it can climb pretty well. This behavior shows that the cat considers a squirrel on the ground a different sort of thing than a squirrel in a tree. Is there metareasoning going on that tells the cat that it is now in an environment that gives a big advantage to the squirrel, or is giving up just a conditioned response? When a cat is stalking a mouse, the cat is fairly oblivious to other objects in its environment. Is there metareasoning going on that helps the cat evaluate whether the risk in turning off its normal monitoring for larger predators outweighs the risk of losing the mouse if it splits its attention? Sophisticated metareasoning in cats is unlikely, however a cat has many behaviors available to it and multiple objectives driving its behavior selection so it has at least as much need for rudimentary metareasoning as it has for reasoning. Our hypothesis is that in higher animals, reasoning and metareasoning will have the same level of sophistication and utilize similar reasoning mechanisms.

In addition to the type of anecdotal evidence given above, there are several studies of learning and memory indicating that mammals and birds may have some rudimentary reasoning and metareasoning capabilities. Memory is a prerequisite for reasoning, since reasoning requires a trustworthy internal model of the world and the ability to anticipate the changes in the model that will result when a behavior is executed. Cats do have pretty sophisticated internal models of the world, particularly of spatial memory, which includes landmarks and objects of interest or value. Storing a spatial memory of the environment that is independent of an animal's focus of attention is a biologically old capability seen in reptiles and birds, as well as mammals (Gallistel 1990). Clayton, et.al. (1999) have studied episodic-like memory in scrub jays and shown that, in addition to the 'what' and 'where' information stored in spatial memory, jays seem able to remember 'when' information as well. Dere, et.al. (2006) have collected behavioral evidence showing that animal species show behavioral manifestations of various aspects of episodic memory, e.g. 'metacognition', 'conscious recollection' of past events, and 'temporal order memory.'

There is also evidence from recent work on the learning curve for individual animals during conditioning experiments suggesting that mammals may be able to anticipate the effects of behaviors. These experiments show that once the mammal "gets it," they complete learning in just a few trials (Gallistel, Fairhurst, and

Balsam, 2004). This speed-up of conditioning compared to lower animals could be based on a match between the mammal's expectations and its experience, a trick used by Barto and Rosenstein (2004) to speed reinforcement learning.

If the state of the world is uncertain and there are many alternative models, then the reasoner must have the time and computational resources to consider the effects of behaviors on a reasonable fraction of the alternatives. People are not inherently very good at this (Heuer 1999), and need to be trained in order to be good analysts. Uncertainty about the world, uncertainty about the effect of behaviors, the ability to interrupt and resume behaviors, and the large number of behaviors available at any time, makes reasoning about future situations arbitrarily hard. Thus, research into more effective reasoning systems can probably continue indefinitely. As discussed above, uncertainty plus multiple objectives also creates a need for metacognition. A metareasoning system needs to be as sophisticated as the reasoner it is associated with in order to analyze what the reasoner is doing. However, in a fixed reasoning frame, there is no clear requirement for a metareasoner to have more sophisticated capabilities than its associated reasoner.

We next consider the nature of reasoning frames in human metacognition.

## Mental Spaces as Reasoning Frames

According to Fauconnier and Sweetser (1996),

Mental spaces are very partial assemblies constructed as we think and talk, for purposes of local understanding and action. They contain elements and are structured by frames and cognitive models. Mental Spaces are connected to long-term schematic knowledge, such as the frame for walking along a path, and to long-term specific knowledge, such as a memory of the time you climbed Mount Rainier in 2001.

We can return to previously constructed Mental Spaces to do additional reasoning. Mental Spaces are models detached from the world, in the sense that a person can change a Mental Space in counterfactual ways and reason about the dynamics of this alternate world (Fauconnier, Turner 2002). These Mental Spaces are just what is needed to do introspective monitoring in an uncertain world, and to do metareasoning about the impact of behavioral choices in different possible worlds. Fauconnier has claimed that the ability to construct Mental Spaces and do detailed reasoning in them is a purely human capability. We will return to the question of how the cat is able to anticipate the future without Mental Spaces below.

## Storage of Complex Relational Information

Animals as simple as snails can establish links between events through classical conditioning, and between objects with operant conditioning. Beyond these basic association mechanisms that have been extensively modeled in neural networks (McClelland, Rumelhardt 1986), humans have evolved four mechanisms for storing more complex relational information:

1. Internal maps of 2D and 3D structure or spatial memory (Allen 2004).
2. Functional groupings that capture “what” must come together to do a task, i.e. things that can be used for the same task. For example, saws, axes, and logs are all associated with getting firewood (Luria 1974). People can also recognize negative examples; e.g. a microwave does not belong with logs and a saw.
3. Episodic memory (Tulving 2001) captures not just what, where and when; but the causal sequences involved in the execution of a task, i.e. how things are used to carry out a task. This mechanism seems specific to humans and complex languages usage, as does the next mechanism
4. Category hierarchies or ontologies. Saws, axes, screwdrivers and can openers are all tools, although they are unlikely to be used on the same task. However, logs are not normally considered tools. Note that ontologies generally include information on functional groupings as well as category hierarchies.

Together these mechanisms form a tightly coupled and vast relational web. It is impossible for people to pay attention to the whole vast relational web they have constructed over their life at once. A person’s sense of “current context” is tied to the active portion of the relational web. The active portion of the relational web is in turn determined by the sensations or expectations that have drawn attention and act as starting points for activating specific geospatial, episodic, and functional grouping memories. In this view, everything made active when attention is focused on a starting point is context information for the current situation. Since attention can be focused on anything, everything is current context at some point and a starting point at others. Weather, terrain, and other operating conditions are often drawn into the current context through their many connections within the relational web.

Cats like humans have extensive relational knowledge about the world. A cat’s web of relationships certainly includes spatial memory, and probably includes functional groupings of objects. It is unlikely to include causal narrative structure, and certainly does not include hierarchical categories. The benefit that the cat derives from storing relational information is the ability to fill in information that was not sensed directly while estimating the current situation. Jui and Shah (2007) have shown that

is possible to use the detection of known objects in a video sequence that belong to the same functional group as an activity to infer the presence of the activity even though the system has no perceptual model of the activity. An analogous filling-in process was used by Eilbert, et.al. (2002) who were able to infer future situations based on the occurrence of part of a narrative sequence or plan. In practice, functional groupings and narrative structure can be encoded in a set of link tables in a relational database or in an ontology.

These results suggest that just recognizing parts of a functional grouping or an episodic-like memory would allow the cat to anticipate an associated event (without reasoning about exactly how it would occur), and respond emotionally to the event. This could trigger a “metacognitive” response in the cat’s attentional settings and its response patterns.

Category hierarchies allow people to manipulate their web of relational information in dramatic ways that go well beyond what other animals can do. They can “blend” functional groups or narrative structures in order to create new structures that they have never seen before (Fauconnier and Turner 2002). For example, a person with a functional grouping corresponding to the concept of a soldier and one corresponding to a horseman can imagine a cavalryman by drawing attributes from both concepts. They can also imagine impossible concept blends, such as the Grim Reaper. Once constructed, these blended concepts remain in the web of relationships.

People can also create a new “blended narrative or episode” from several different approaches to accomplishing an overall task. For example, people can easily imagine a race between two people who have never really raced each other. In the Olympics skiing, the position of the current racer relative to the leader is often shown. The key to blending several different episodic memories into a blended concept is compressing them over time. For example, by compressing both a direct handoff of an item between two people, and the drop-off of the item by one person followed by the pick up of the item at a later time by a second person, it can be understood that both are ways to accomplish the exchange of an item.

## Switching Mental Spaces

People can switch among many overlapping Mental Spaces of various sizes and scales (Eilbert and Hicinbothom 2006). Analysts are aware that people often get stuck in the wrong Mental Space. For example, in explaining an analytic error Shlaim (1976) stated that:

Since the facts do not speak for themselves but need to be interpreted, it is inevitable that the individual human propensities of an intelligence officer will enter into the process of evaluation.

On occasions when people are dropped into situations where they have no appropriate Mental Space, they encounter problems estimating the situation and making appropriate decisions. When this occurs, people will quickly search for a more appropriate Mental Space.

Some Mental Spaces are much simpler than others. When surprised or terrified, people tend to drop into simpler reactive states, where everything except what they are scared of drops into the background, and the only behaviors accessible are simple variants of fight or flight reactions. After time and distance are placed between the observer and the frightening object, people will switch back to a richer Mental Space. However, their analysis of the situation, and their choice of behaviors, is likely to be biased by the frightening experience well after they are able to do more normal situation estimation.

Games with fixed rules and solving mathematical problems are about the only places where people can reason in a fixed Mental Space or frame for very long. Therefore, a complete model of metareasoning should examine the jumps between Mental Spaces, as well as the reasoning within a single Mental Space. The next section describes appraisal models of emotion and how emotions can act as triggers for transitions between Mental Spaces.

### Appraisal Models of Emotion

In psychological appraisal models of emotion, the emotional responses are the result of appraisals of what happens to an individual's goals and why. Probably, the psychological appraisal model most widely used as a basis for computational models of emotion was developed by Ortony, Clore and Collins (1988); one example of a computational model based on it is found in Elliott's (1992) affective reasoner. Roseman's (2001) model was used in Velasquez' (1997) Cathexis model. The motive for constructing these computational models was either to construct an autonomous agent with more human-like responses or to model a particular set of psychological data. Here we are suggesting a new role for emotion models in autonomous agents, i.e. emotion can act as a metacognitive mechanism directing the switching between reasoning frames.

The personality-enabled Architecture for Cognition (PAC) (Read, et.al. 2006), an affective cognitive model, is described in some detail in order to show how computations already incorporated in that model can be used to direct switching between reasoning frames. PAC's emotional model is based on Roseman's appraisal model of emotion. The motivation for this selection was the distinction between approach and avoidance motivational systems made in Roseman's model, but not in other appraisal models. Roseman's model has two fundamental dimensions, motivational state which distinguishes

appetitive and aversive motives (i.e., what we want versus what we want to avoid), and situational state which distinguishes what happens to those motives, (i.e., do we succeed or fail). The two dimensions lead to four possible combinations of motive relevant events: getting what you want (appetitive, success), not getting what you want (appetitive, failure), not getting what you don't want (aversive, success) and getting what you don't want (aversive, failure).

Roseman (2001) argues that several secondary dimensions also impact emotion:

- (1) unexpectedness of the event: whether or not one's expectations are violated;
- (2) probability of the motive relevant event: is the outcome uncertain or is it definite?
- (3) agency: what or who caused the motive-relevant event?
- (4) control potential: Can one do something about the motive relevant outcome?, and
- (5) problem type: some negative emotions are due to inherent characteristics of an object, e.g. disgust at something rotten, whereas others are the result of a motive inconsistent outcome.

Roseman (2001) identifies how 17 distinct emotions can result from different combinations of these factors. In PAC, the emotional state of an agent is continuously calculated based on numerical approximations of the levels of each factor. New emotion levels are typically triggered by a change in the appraisal dimensions, and their value is subsequently allowed to decay over time. Several examples of emotion activation functions are defined below (Read, et.al. 2006):

$$\text{SURPRISE} = (1 - \text{Expectancy})$$

$$\text{HOPE} = \text{Expectancy} \cdot (1 - \text{Certainty}) \cdot \text{MotiveConsistency} \text{ (iff Agency} = \emptyset \text{)}$$

$$\text{FEAR} = \text{Expectancy} \cdot (1 - \text{Certainty}) \cdot (1 - \text{MotiveConsistency}) \cdot (1 - \text{Controllability}) \text{ (iff Agency} = \emptyset \text{)}$$

$$\text{SADNESS} = \text{Expectancy} \cdot \text{Certainty} \cdot (1 - \text{MotiveConsistency}) \cdot \text{Appetitiveness} \cdot (1 - \text{Controllability}) \text{ (iff Agency} = \emptyset \text{)}$$

$$\text{ANGER} = \text{Expectancy} \cdot (1 - \text{MotiveConsistency}) \cdot \text{Appetitiveness} \cdot \text{Controllability} \text{ (iff Agency} = \text{other})$$

The meaning of the last two equations can be described as follows. Sadness occurs when one has an appetitive motive, and there is a motive inconsistent outcome, and the outcome is certain, and there is low control potential for changing the outcome. Anger results when an appetitive motive is blocked (motive inconsistent) by someone else

and there is high control potential for changing the outcome.

Each of the five secondary dimensions provides a basis for Mental Space or reasoning frame switching. The unexpectedness of the event and the surprise emotion is the easiest switching trigger to explain. Surprise occurs when there is a massive discrepancy between expectations and sensed information in the current Mental Space. The current Mental Space is dropped in favor of a new Mental Space constructed around some of the most salient features causing the surprise.

When probability of the motive relevant event is low there is a reason to switch among similar Mental Spaces to find better or worse alternatives. Hope and fear are both emotions that result from things that might occur in the future rather than from an actual experience. They can cause the current Mental Space to be expanded to include other actors whose actions might reduce the certainty of a feared situation looming in the future.

An outside agent's actions provide a basis for estimating its goals and traits. If an agent's actions cause a change in the basic characterization of that agent, a switch to a new Mental Space with the revised model of the agent is necessary. Anger directed at an agent could provide a basis for making this type of Mental Space switching. Other emotions can cause various contractions, expansions, and changes in beliefs about character traits that effectively change the Mental Space where reasoning operates.

## Conclusion

We presented a case for multiple objectives acting as a driver for metareasoning within a single frame. We also presented a case for untrustworthy sensing and uncertain real-world environments acting as drivers for metareasoning leading to switching between reasoning frames. We also argued that the sophistication of the metareasoning capability should be matched to the sophistication of the reasoning it monitors.

Fauconnier's work on Mental Spaces and blending demonstrates that people have developed flexible models of the world in which they can do sophisticated reasoning. There is a separate line of evidence that humans have developed a variety of mechanisms for remembering complex relational information about the world. We suggest that Mental Spaces are built on top of the vast web of relational information stored in the mind using mechanisms that select and manipulate the stored information. A cat also has a large store of relational information in its brain, although there is no evidence that it can store detailed narratives or category hierarchies the way people can. However, the cat seems able to anticipate

events and do rudimentary metacognition with the web of relational information that it does have.

Finally, we pointed out that switching between Mental Spaces is a standard part of human cognition. There is evidence showing that emotions are correlated with Mental Space switching, and even suggesting they may trigger it. If Mental Space switching is a standard part of human cognition, then it is something that metacognition should be able to reason about. A review of Roseman's appraisal model of emotion showed that many of the dimensions considered in appraising emotion are also motivations for switching reasoning frames. The PAC affective cognitive model, which is based on Roseman's theory, already has most of the computational apparatus needed to do metareasoning about when to do 'Mental Space' switching and even how to construct the new Mental Space. Thus, it could form the basis of a more complete model of metacognition and metareasoning.

## References

- Allen, G.L. 2004. *Human Spatial Memory: Remembering Where*. Mahwah, NJ :Lawrence Erlbaum Associates.
- Barto, A.G., and Rosenstein, M.T. 2004. Supervised Actor-Critic Reinforcement Learning. In Si, J., Barto, A.G., Powell, W.B., and Wunsch, D., editors, *Handbook of Learning and Approximate Dynamic Programming*, Chapter 14, pages 359 - 380. Wiley-IEEE Press, Piscataway, NJ.
- Clayton, N.S., and Dickinson, A. 1999. Scrub jays remember the relative time of caching as well as the location and content of their caches. *J. Comp. Psychol.* 113, 403–416.
- Cox, M. T., and Raja, A. 2007. Metareasoning: A Manifesto, Technical Report, BBN TM-2028, BBN Technologies. [www.mcox.org/Metareasoning/Manifesto](http://www.mcox.org/Metareasoning/Manifesto)
- Dere, E, E Kart-Teke, J P Huston, and M A De Souza Silva. 2006. The case for episodic memory in animals. *Neuroscience and Biobehavioral Reviews* 30: 1206–1224.
- Eilbert, J.L., Carmody, D.M, Fu, D., Santarelli, T., Wischusen, D., Donmoyer, J. 2002. Reasoning About Adversarial Intent in Asymmetric Situations. Proceedings of Intent Inference for Users, Teams, and Adversaries. AAI 2002 Fall Symposium Series.
- Eilbert, J.L., Hicinbothom, J. 2006. A Cognitive Framework for Modeling Mental Space Construction and Switching During Situation Assessment. FLAIRS06.

- Elliott, C. 1992. The affective reasoner: A process model of emotions in a multi-agent system (Ph.D. Dissertation No. 32). Northwestern, IL: Northwestern University Institute for the Learning Sciences.
- Fauconnier, G. & Sweetser, E. 1996. *Spaces, Worlds, and Grammar*. Chicago: University of Chicago Press.
- Fauconnier, G. & Turner, M. 2002. *The Way We Think*. New York: Basic Books.
- Gallistel, C.R. 1990. *The Organization of Learning*. The MIT Press, Cambridge, MA.
- Gallistel, C. R., Fairhurst, S., Balsam, P. 2004. The learning curve: Implications of a quantitative analysis. PNAS. 101(36):13124–13131.
- Heuer, R. 1999. The Psychology of Intelligence Analysis. Center for the Study of Intelligence, CIA. <http://www.cia.gov/csi/books/19104/index.html>.
- Lazarus, R. 1991. *Emotion and Adaptation*. NY: Oxford University Press.
- Lehnert, W.G., Loiselle, C.L. 1989. An introduction to plot units. In Waltz, D.L. (ed.) *Semantic Structures: Advances in Natural Language Processing*. Lawrence Erlbaum Associates, Inc. Hillsdale, NJ.
- Luria, A.R. 1976. *Cognitive Development: Its Cultural and Social Foundations*. Cambridge, MA. Harvard University Press.
- Ortony, A., Clore, G., & Collins, A. 1988. *The Cognitive Structure of Emotions*. Cambridge University Press.
- Pew, R. & Mavor, A. (eds.) 1998. *Modeling Human and Organizational Behavior: Application to Military Simulations*. Washington, DC: National Academy Press.
- Read, S., Miller, L., Rosoff, A., Eilbert, J.L., Jordanov, V., LeMentec, J.C., Zachary, W. 2006. Integrating Emotional Dynamics into the PAC Cognitive Architecture. Proceedings of BRIMS06.
- Roseman, J. J. 2001. A model of appraisal in the emotion system: Integrating theory, research, and applications. In K. R. Scherer, A. Schorr, & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, methods, research* (pp. 68-91). Oxford: Oxford University Press.
- Liu, J., and Shah, M. 2007. Scene Modeling Using Co-clustering. ICCV 2007.
- Shlaim, A. 1999. *The Iron Wall: Israel and the Arab World*. W.W. Norton.
- Tulving, E. 1985. *Memory and Consciousness*. Oxford. Clarendon Press.
- Velasquez, J., 1997. Modeling emotions and other motivations in synthetic agents. In: Proceedings of the 1997 National Conference on Artificial Intelligence (AAAI97). Providence, RI, pp. 10-15.