Ethical Issues Related to Advanced Artificial General Intelligence

("A Few Small Worries")

Ben Goertzel
Novamente LLC

A few small worries…

- AGI poses nontrivial existential risk
  - Does open-source development increase or decrease the risk?
- Costs of action vs costs of inaction
- Is it ethical to experiment on AGIs?
  - Will advanced AGIs be conscious?
- How to encourage AGIs to be Ethical?
  - Program the right goal system?
  - Teach them right, via imitation/reinforcement/correction?
  - Both?
- How to validate the ethicalness of an advanced AGI system … won't it "game" the tests?
- Can one create AGI goal systems that are stable under iterative self-modification?
- Even if one could create an AGI with high odds of self-modifying in a stably ethical way, what's to stop somebody else from building a nasty AGI first … or stealing yours and hacking its ethics code?
- Even if one could create a stable AGI goal system, and avoid intervention by nasty people, unpredictable aliens or fluky natural phenomena -- **what's the top-level goal?**
- If "preserving humanity in its present form" is a wimpy and unrealistic goal, what's a better one?
  - How much can humanity change and still remain humanity?
  - Long-term goals for the cosmos: Growth, joy and choice?
    - Such qualities are difficult and controversial to define
    - Choice is wrapped up with causality and mental reflection
  - Personal goal: Controlled ascent?
- Nietzsche: We must create our own values
  - All the world as morphology and will-to-power

Can one create AGI goal systems that are stable under iterative self-modification?

For which Goal2's and which goal-achieving systems and environments is the following an attractor?

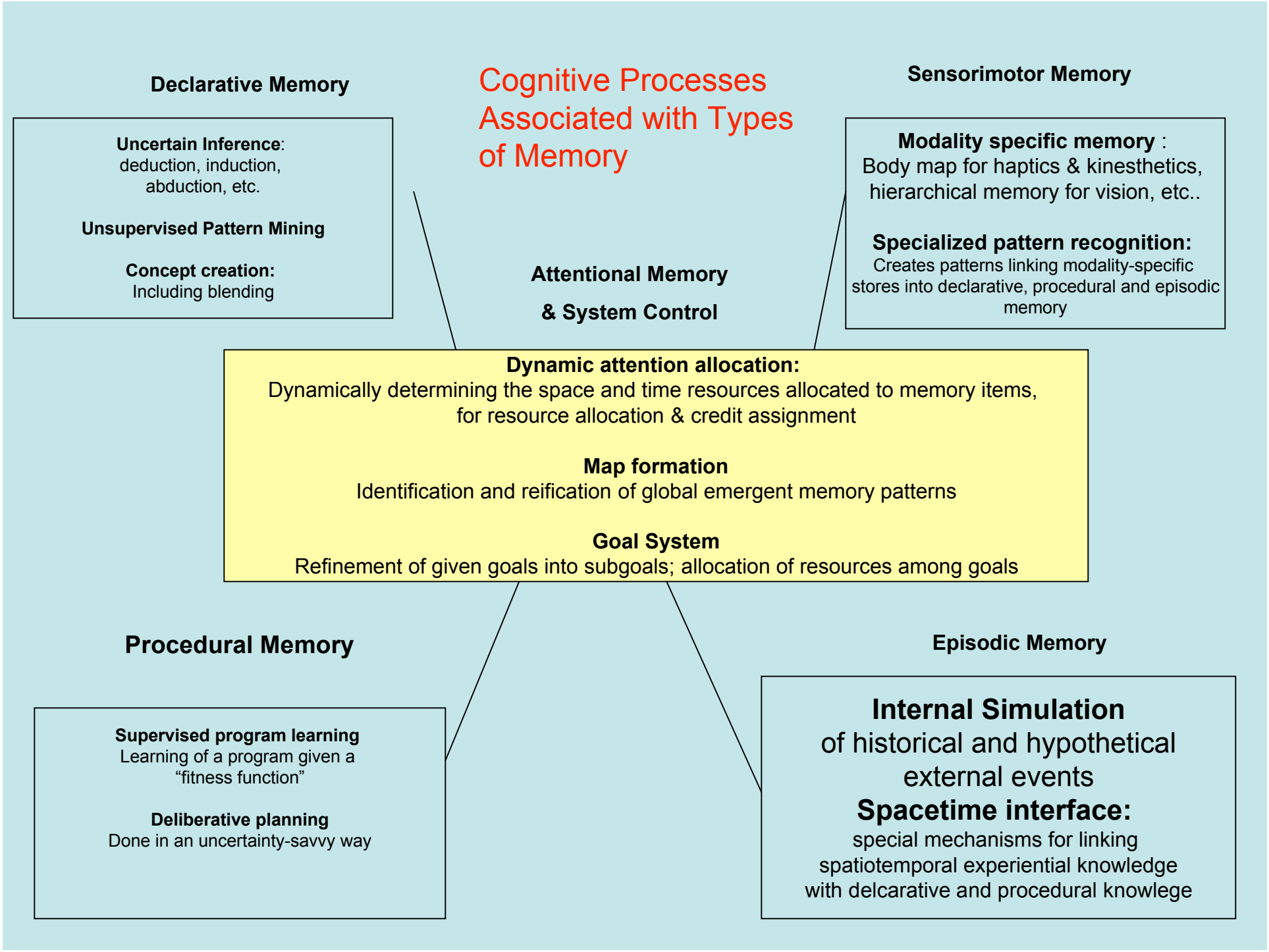
Goal1 =

"Achieve Goal2 as best you can while preserving Goal1"

Long-term goals for the cosmos: Growth, joy and choice?

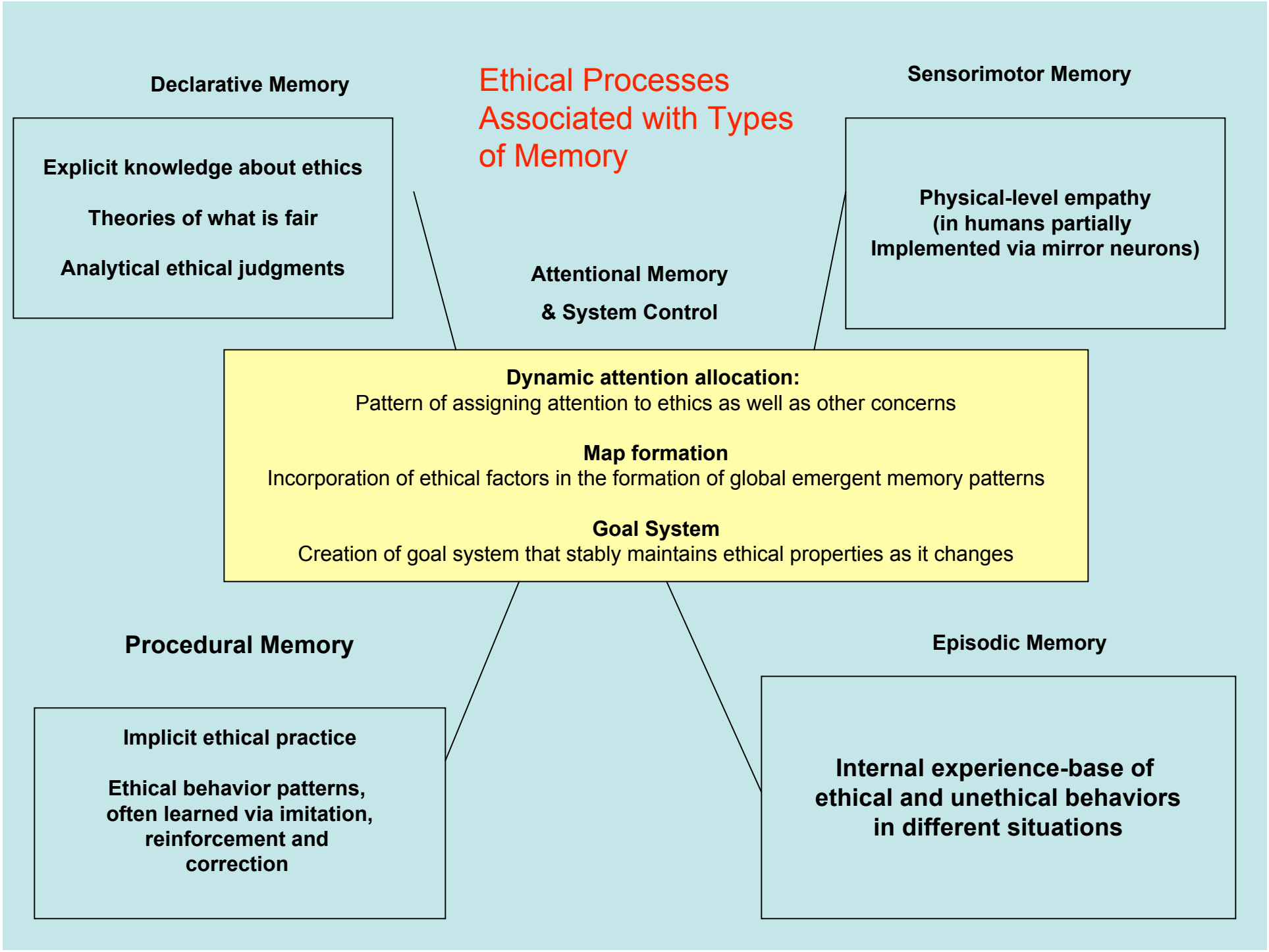
•Such qualities are difficult and controversial to define

- **Growth** = more and more patterns?
- **Joy** (happiness?)
  - Vynnycenko: “an active balance of [our] values and their agreement among themselves and in the forces outside of us”
  - Paulhan: “the feeling of increasing order”
  - S = “I am S and I am noting increasing balance/order in myself and my world”
  - Does joy, as distinct from raw pleasure, require self-reflection?
- **Choice** = “the feeling of causing what happens”?
  - "S wills X" is defined as: The declarative content that {"S wills X" causally implies "S does X"}
  - Choice seems to require a system that has self-reflection and thinks in terms of causation


Cognitive Processes Associated with Types of Memory
Declarative Memory
Uncertain Inference: deduction, induction, abduction, etc.
Unsupervised Pattern Mining
Concept creation: Including blending
Sensorimotor Memory
Modality specific memory : Body map for haptics & kinesthetics, hierarchical memory for vision, etc..
Specialized pattern recognition: Creates patterns linking modality-specific stores into declarative, procedural and episodic memory
Attentional Memory & System Control
Dynamic attention allocation: Dynamically determining the space and time resources allocated to memory items, for resource allocation & credit assignment
Map formation
Identification and reification of global emergent memory patterns
Goal System
Refinement of given goals into subgoals; allocation of resources among goals
Procedural Memory
Supervised program learning
Learning of a program given a "fitness function"
Deliberative planning
Done in an uncertainty-savvy way
Episodic Memory
Internal Simulation
of historical and hypothetical external events
Spacetime interface:
special mechanisms for linking spatiotemporal experiential knowledge with delcarative and procedural knowlege

Cognitive Synergy

Fine-grained, continual interaction between cognitive processes associated with different types of memory, enabling each one to function much more effectively than it would do on its own


Ethical Processes Associated with Types of Memory
Declarative Memory
Explicit knowledge about ethics
Theories of what is fair
Analytical ethical judgments
Sensorimotor Memory
Physical-level empathy (in humans partially Implemented via mirror neurons)
Attentional Memory & System Control
Dynamic attention allocation:
Pattern of assigning attention to ethics as well as other concerns
Map formation
Incorporation of ethical factors in the formation of global emergent memory patterns
Goal System
Creation of goal system that stably maintains ethical properties as it changes
Procedural Memory
Implicit ethical practice
Ethical behavior patterns, often learned via imitation, reinforcement and correction
Episodic Memory
Internal experience-base of ethical and unethical behaviors in different situations

Ethical Synergy

Fine-grained, continual interaction between **ethical** processes associated with different types of memory, enabling each one to function much more effectively than it would do on its own