Symbolic cognitive robotics - Design notes

Author: Jean-François Cloutier

Research project: Symbolic Cognitive Robotics, Active Inference Institute

Last updated: January 19 2024

Pre-implementation notes on the next iteration of a cognitive architecture for Lego robots.

Formative concepts

- Active Inference (an agent actively minimizes surprise to survive)
- Enactivism (an agent's perceptions and actions are constructively co-dependent)
- Apperception (sense-making as discovery of unified causal theories)
- Mortal Computing (meaning is grounded in the agent's drive to survive)
- Society of Mind (an agent is animated by a collective of cognition actors interacting with each other and the environment)
- Constraint Closure (the cognition actors constrain how the Society of Mind can change, and vice-versa)
- Kantian Whole (the parts -cognition actors- exist for and by means of the whole -the Society of Mind-)

Society of Mind (SOM)

- An agent grows and evolves a Society of Mind from its experiences
- A SOM is a "connectome" of Cognition Agents (CAs)
- Each CA has an umwelt consisting of other CAs

Introspection vs extrospection

- Extrospection => Cognition of (more or less abstracted) sensations from the external world
    - The objects named "world" and "ground" are primordial
    - object ground is `in` world
- Introspection => Cognition of sensations from the computations by **Cognition Actors (CAs)**
    - The object named "self" (self is `in` world) is primordial
        - self is in world
- Detectors and effectors are exposed as extrospective primitive CAs
- Every other CA can be coopted into one or more umwelts as a source of introspective or extrospective sensations, and as an effector of actions

The umwelt of a Cognition Actor (CA)

- A set of other CAs are seen by others through what they expose (their API)
    - What a CA exposes to all other CAs
        - the vocabulary of their beliefs (what others can make predictions about)

        - extant, latent and synthetic objects (typed with extant or latent object types),
        - extant, latent and synthetic relations/properties
            - a latent or synthetic property is always boolean-valued
        - all CAs have a common vocabulary of meta-cognition beliefs
    - what they emit when prompted by predictions about their beliefs:
        - prediction errors from their beliefs
            - from perceiving other CAs
            - from their cognitive self-assessments/beliefs (from introspection)
            - with varying precision

A CA's perceptions

- A CA processes perceptions one discrete time slice after another
    - its duration is constant and proportional to the CA's abstraction level
        - i.e. the depth of its umwelt
- Perceiving is making predictions about about the beliefs of CAs in its umwelt
    - and getting prediction errors or not
- Perceptions are
    - Uncontradicted predictions
    - Prediction errors can be emitted in response to predictions, with attached precision
        - If multiple CAs respond to a prediction with prediction errors
            - The prediction error with the highest precision is picked
            - Tie-breaking is random
- The precision of a prediction error (a float between 0 and 1) is a function of:
    - The confidence of the emitting CA in the contradicting belief, which is a function of:
        - The accuracy of the supporting causal model
        - The duration of the supporting trend modulated by
            - the average precision and variance of the perceptions aggregated by the trend

A CA's beliefs

- What's imagined, analyzed, partitioned and categorized by the CA
    - from its perceptions
        - unrefuted predictions + prediction errors about the beliefs of CAs in its immediate umwelt
- Beliefs are available to other CA's as *synthetic or latent* and thus *novel* perceptions
- Beliefs are abduced predicates
    - needed to formulate a causal theory (latent)
    - needed to label significant perceptual trends (synthetic)
- Beliefs have associated normativity (pleasant vs unpleasant vs indifferent beliefs)
- "Thin now" vs. "thick now" beliefs

  - Thin now beliefs
    - Unobserved but imagined/abduced properties/relations/objects to (causally) make sense of observations - the thin now -
  - Thick now beliefs
  - Synthetic. induced from, and thus supported by, perceptual trends - the thick now -

Perceptual trends support synthetic beliefs

- Trends support the synthesis of beliefs in the thick now
- A trend is given a value
  - one of stable, unstable, up or down
- Specific vs generic trend
  - Specific trend-
    - `trend(<predicate name>(<object name>, <object name> | <domain value>), <trend value>, <since>)` - a trend on an instance of a property/relation (stable, unstable)
  - Generic trend - `trend(<predicate name>(<object name>), <trend value>, <since>)` - a trends on a type of property/relation for an object (stable, unstable, up, down)
    - for relations, up/down describes a count of related objects,
    - for properties, up/down describes the rise/fall in values (property value domains are ordered from lesser to greater)
- Memorizing trends
  - A compressed trend and associated normativity can be preserved as long-term memory
    - `compressed(<trend>, <time interval>)`
    - and associated with past beliefs (that the trends supported)
  - Uncompressed trends represent short-term memory (developing trends)

Inducing beliefs from trends

- By association
  - Synthetic properties/relations are supported by attention-worthy (strongly felt or surprising) trends
    - `<synthetic property name>(<object_name>, true | false)`
    - `<synthetic relation name>(<object name>, <object name>)`
- By partition
  - Parts-whole beliefs are induced by detecting boundaries in an observed object.
    - `in(<new object name>, <object name>)`
  - How are boundaries detected?
    - An object has differentiable, stable sub-trends that coincide in time
    - This might indicate that different parts of the object were being observed at different times
      - e.g. "patch of food" in the "ground" in the "world" ("self" is always in the "world")
  - A part is not of the same object type as the whole (assuming no fractal objects)
- By categorization
  - Beliefs about partition cause the abduction of new objects (the parts)
  - The "part" object ia assigned a new (abduced) object type
  - `is_a(<object_name>, <new object_type>)`

Trends, feelings and the normativity of beliefs

- Normativity (from association with current feelings) is **always** about trends
    - It exists in the "thick now"
- Normative valuation comes from associating trends with feelings (see below)
- A trend takes its (normative) value from the intensity of concurrent feelings
    - `trend_value(<trend>, good | bad | neutral)`
- A belief supported by a trend takes the normative value of that trend
    - A belief associated with a bad feeling is unpleasant, else it's pleasant (good) or indifferent (neutral)
    - Since trends have lengths, the normative values of trends have duration
        - e.g. a long-lasting unpleasant belief are worse than a short-lasting one
    - A trend is **significant** (and worthy of belief synthesis) if
        - it breaks surprisingly from a previous trend
        - or if correlates with a change in feelings

CA actions

- Changes in properties/relations observed by a CA are either caused by latent processes or by actions.

    - **In a static environment, they are caused entirely by actions!**
    - No perception without action and no action without perception

- To make sense of/apperceive the consequences of actions, they must be observed together with the property/relation changes they (may) cause

- A CA exposes, by name, the actions it can execute

- A CA must always be capable of acting

    - i.e. it has at least one effector CA in its transitive umwelt

- The action repertoire of a CA consists of

    - the actions it synthesized
    - plus the distinct actions exposed by CAs in its umwelt

- The CA of an effector exposes atomic actions

    - For example, a wheel CA exposes the atomic actions "spin" and "reverse spin"

- A CA syntesizes actions from the actions exposed by CAs in its umwelt, names them and exposes them in turn

- A synthetic action is a named list of actions

    - e.g. action_2 = [action(ca_2, action_1), action(ca_2, action_1), action(ca_3, action_2)]
        - an action can be repeated
    - a synthetic action is, via closure, a sequence of atomic actions

Why does a CA synthesize a new action?

- Because a sequence of actions is empirically associated with a significant belief change
- Belief changes
    - Abduced object, property or relation from a causal model (thin now belief)
    - Correlation with a belief-supporting trend starting/ending/enduring (long now belief)
        - The sequence of actions that runs before/through the trend is extracted
    - Babbling
        - A CA synthesizes an action to see what would happen if executed
            - As a variation on an action already n the repertoire
                - Amplify sub-sequences via action duplication
                - Tone down sub-sequences by reducing duplication
                - Splice and recombine a synthetic action

Action intents

- An **intent** names an action that a CA wants executed.
    - A CA can intend any action in its repertoire
- What motivates action intents by a CA (from less to most motivated)
    - Babbling
        - to maybe cause a "random" belief
    - Evidencing
        - to impact confidence in a belief (thus the precsion of reported prediction errors)
    - Eliminating
        - to terminate an unpleasant belief
- A CA intends at most one action per time slice
    - It intends the most motivated action in its repertoire
        - favoring, but not always, actions of the most successful policies (seebleow)
    - If multiple actions are considered
        - A motivation tie is randomly broken

Action execution

- Execution of an intended action is inhibited if another CA concurrently intends an action that
    - covers it (is a super-sequence)
    - or is identical and has higher normative motivation
- All actions taken are observable by all CAs
    - The atomic actions from the closure of synthetic actions are observed
        - During time slice T of the CA
        - If a sub-sequence of the observed atomic actions recreates a synthetic action in the repertoire of the CA
            - then the longest synthetic action is what is observed, plus the second longest etc.

CA action policies

- A policy is an action associated by a CA with a belief, a goal (verification, elimination) and a success rating from its executions

Feelings

- Feelings are agent-wide signals about detected existential risks
- Feeling types
    - Hunger
        - Depleted energy/resource stores
    - Pain
        - Damage - loss of structural integrity
    - Fear
        - Lack of foresight - Inability to predict
- `feeling(<feeling type>, good | bad | neutral)`
- Motivational ranking
    - Hunger > Pain > Fear
    - The agent dies when energy/resources are depleted
    - The agent is immobilized when pain is too high
- Feelings are centrally computed from
    - detector sensations
        - touch - pain increases
        - color - resources increase if color == food type
    - effector sensations
        - work done - energy decreases
    - CA cognitive sensations
        - mental effort - energy decreases
        - prediction success rate - fear increases/decreases
        - relevance (rate of received predictions, intended composited actions)
    - The passing of time
        - healing - pain decreases
        - base metabolism - resources/energy decreases
- Any change in hunger/pain/fear intensity is signaled to all CAs
- For each CA, for each time slice, there's an average intensity of each feeling type

Constraints

- Umwelts (when closed) must be acyclic directed graphs but not necessarily trees
- Abstraction must be monotonic
    - A CA must not include a CA in its immediate umwelt if the latter is already in its transitiveumwelt.
- A CA must be either cognitive or meta-cognitive, never both
- Only one synthetic action in a conflicting set can be executing at any given point in time
    - A synthetic action conflicts with another if their closed sequences have any simple action type * incommon.
    - Practically speaking, only one synthetic action is allowed to execute at any time
- A * CA must not remove an element from its API if it is used by another CA
    - to formulate a causal theory
    - to synthesize a belief or action
- If a CA must archive a belief (without normativity) or action and its (compressed - abstracted)support when the support is gone but the belief or action is still used by other CAs
- A new belief must not be created if its support is subsumed by the compressed support of an archivedbelief

    - The archive belief is ressucitated and given the current support
- An archived belief/action must be deleted if the belief/action is no longer used by another CA.

Initial state of the Society of Mind (SOM)

- Initial CAs
    - One primitive CA per effector (wheel_1, wheel_2)
    - One primitive CA per detector (color_sensor, touch_sensor, obstacle_sensor, beam_sensor)
    - One meta-cognition CA with as umwelt all the primitive CAs *Initial steady state variables (sources of feelings)
    - Integrity 100%
    - Energy 100%
    - Foresight 100%

Meta-cognition CAs (MCAs)

- Every CA of level N belongs to the umwelt of one MCA associated to that level
    - The level of a CA is the number of edges from the CA to primitive CAs
    - For a Society of Mind (SOM) with N levels, there must be one MCA per level 1..N, plus one MCA for level N+1 with an empty umwelt
    - Once a CA is added at level N + 1, an MCA is immediately created for the empty level N + 2etc.
- An MCA observes only cognitive sensations from its umwelt

Meta-cognitive actions

- An MCA is exploring the "connectome space" of CAs at one level of abstraction, looking for a beneficial organization
- An MCA at level N can
    - Create a CA at level N
        - And add level N-1 CAs to its umwelt at creation
    - Remove a CA

Cognitive sensations

- The cognitive sensations are
    - effort as
        - apperception engine usage
        - memory load
    - foresight as
        - prediction success rate
    - stability as
        - rate of change in beliefs, actions, causal theories
    - relevance as
        - rate of received predictions,
        - percent of actions composited by other CAs