Symbolic cognitive robotics - Design notes

Author: Jean-François Cloutier

Research project: Symbolic Cognitive Robotics, Active Inference Institute

Last updated: April 10 2024

Design 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 embodied agent's perceptions and actions are constructively co-dependent)
- Apperception (predictive sense-making comes from the discovery of unified causal theories)
- Biosemiotics (sense-making and sign-acting are intrinsically 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 world)
- Constraint Closure (the cognition actors constrain how the Society of Mind as a whole 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 engagement and experiences
- A SOM is a "connectome" of Cognition Agents (CAs)
- Each CA has an umwelt consisting of other CAs or some sensed aspects of the agent's environment
- A SOM starts with *a priori* constituents
  - The *a priori* objects are `self`, `world` and `ground`
  - The *a priori* relations is `in`
    - objects `ground` and `self` are `in` world
  - The *a priori* value domains are
    - boolean - [true, false]
    - integer ranges, including percentage
    - colors - [unknown, black, blue, green, yellow, red, white, brown]
  - Detectors and effectors are exposed as *a priori* CAs
    - Each defines an *a priori* vocabulary of
      - beliefs (e.g. `color`, `distance`, `touched`)
        - with value domains (e.g. colors, 0..100, etc.)
      - actions (e.g. `spin`, `reverse spin`)

    - Metacongnition actors have *a priori* capabilities and each CA has *a priori*, shared introspective belief voacabulary (see below)

Exteroception vs interoception vs introspection

- Exteroception => Cognition of (more or less abstracted) sensations from the external world
    - e.g. the distance to an obstacle
- Interoception => Cognition of (more or less abstracted) bodily sensations
    - e.g. a motor is stalled
- Introspection => Cognition of sensations about computations by **Cognition Actors (CAs)**
    - perceived by Metacognition actors
    - e.g. a CA has low predictive capability
- A CA can non-exclusively coopt other CAs into its umwelt as sources of interoceptive or exteroceptive sensations, and as effectors of actions

The umwelt of a Cognition Actor (CA)

- The (immediate) umwelt of a CA is a (small) fixed set (fixed at instantiation) of other CAs
    - The CAs in an umwelt also have CAs in their own umwelts
        - All the way down to *a priori* CAs
    - The transitive umwelt of a CA is the CAs in its immediate umwelt plus the CAs in their transitive umwelts
    - The level of abstraction of a CA is the maximal depth of its transitive umwelt
    - A CA builds its umwelt from CAs of lower and uniform levels of abstraction
- CA Interfaces
    - A CA's interface is what it exposes to other CAs, namely
        - The vocabulary of its beliefs (what others can make predictions about), composed of
            - extant (observed), latent (imagined) and synthetic (derived) objects,
            - extant, latent and synthetic relations/properties
                - a latent or synthetic property is always boolean-valued
        - The actions it affords
            - either *a priri* actions (e.g. spin and reverse spin of motors) or synthetic actions
    - All CAs have a common vocabulary of meta-cognition beliefs
        - exposed to meta-cognition actors
- CA events
    - A CA emits events listened to by the CAs tp which unwelts it belongs, and by the meta-cognition actor overseeing it
        - when prompted by predictions about its beliefs, a CA emits
            - prediction errors given its currently held beliefs
                - with varying precision
                - from perceiving other CAs
                - or from cognitive self-assessments/beliefs (from introspection)

A CA's perceptions

- A CA processes perceptions one discrete time slice after another
    - the time slice duration is constant for the CA and proportional to the CA's abstraction level
        - i.e. the depth of its transitive umwelt
- Perceiving is making predictions about about the beliefs of CAs in its umwelt
    - possibly corrected by prediction errors emitted by CAs in its umwelt
- 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 has sway
            - Tie-breaking is random
- The precision of a prediction error (0% to 100%) 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 behind the belief (see below)
        - The duration of the perceptual trend supporting the belief, modulated by (see below)
            - the average precision and variance of the perceptions aggregated by the trend (see below)

A CA's beliefs

- Abelief is what's imagined (latent) or synthesized by the CA
    - from its perceptions
        - which are 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)
    - or needed to label significant perceptual trends (synthetic)
- Beliefs have associated normativity (pleasant vs unpleasant vs indifferent beliefs) from ambient feelings
- "Thin now" vs. "thick now" beliefs
    - Thin now beliefs are
        - unobserved but imagined/abduced properties/relations/objects to (causally) make sense of observations
    - Thick now beliefs are
        - synthetic beliefs, induced from, and thus supported by, perceptual trends

Perceptual trends support synthetic beliefs

- Perceptual trends (from the analysis of past perceptions) support synthetic beliefs
- A trend is about the relation or property of a given object
    - A relation associates two objects
    - A property associates an object and a value from a defined domain of values
- A trend is itself given a value
    - a trend is either stable, unstable, up, down
        - `trend(<predicate name>(<object name>), <trend value>, <since>)`
        - 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 CA keeps a limited history of past perceptions
        - It is the CA's short-term memory (developing trends)
        - Individual, past perceptions are eventually forgotten
    - A "compressed" trend (a trend about now forgotten perceptions) can be preserved as long-term memory
        - `compressed(<trend>, <time interval>)`
        - It is associated with remembered past beliefs (that the trends supported)
            - When past beliefs are forgotten so are their associated compressed trends
        - Compressed trends represent the CA's long-term memory

Inducing beliefs from trends

- Beliefs are induced from the analysis of trends, themsleves produced from analyzing a perceptual history.
    - A belief induced form the analysis of a trend is then supported by that trend.
    - Induction can result from
        - Associating
            - 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>)`
        - Partitioning
            - 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)
        - Categorizing
            - Beliefs about partition cause the abduction of new objects (the parts)
            - The "part" object is assigned a new (abduced) object type
            - `is_a(<object_name>, <new object_type>)`

- A trend is **significant** and worthy of belief induction if
    - it breaks surprisingly from a previous trend
    - or if correlates with a change in feelings

Trends, feelings and the normativity of beliefs

- Normativity (something being good, bad or indifferent) 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, ambient 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

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 defines a new action?

- A CA synthesize a new action because a sequence of actions is empirically associated with a significant belief change
- Belief changes that may lead to action synthesis are
    - 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 in its 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 "chance" 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 (see below)
    - 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
- A CA must not include a CA in its immediate umwelt if the latter is already in its transitive umwelt.
- 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 interface if it is used by another CA
    - to formulate a causal theory
    - to synthesize a belief or action
- 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, etc.)
    - 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 + 2 etc.
- 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