On attempting to reify a few of the things we may mean by “consciousness” with code

Josh Joseph, Dhaval Adjodah, Joichi Ito

MIT

mit media lab

MIT Quest for Intelligence

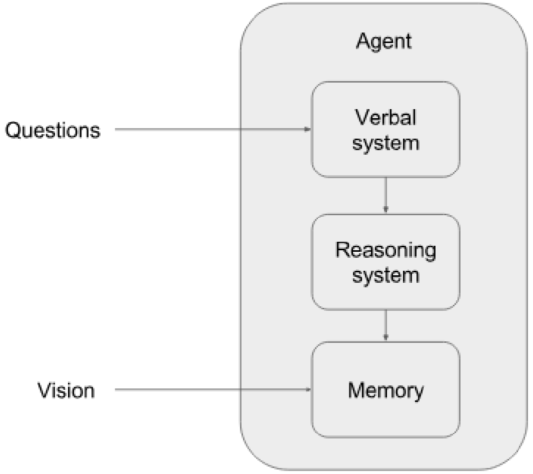
Why attempt to reify philosophy with code

- Lots of what philosophers think a lot about show up in CS/AI research
  - Mind, awareness, imagination, reasoning, consciousness, etc.
- CS/AI could benefit from a deeper understanding of philosophy
- Possibly benefit philosophy by bringing code-style concreteness
  - (TBD)

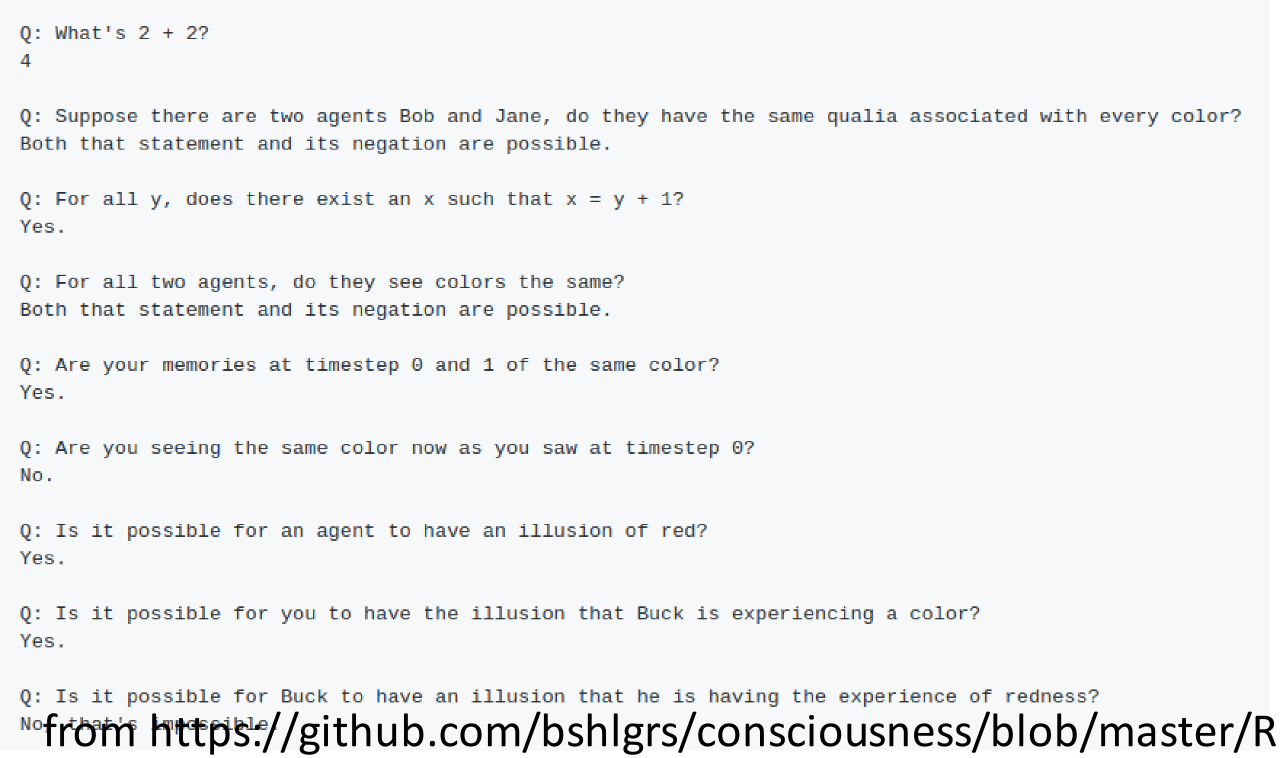
- (Disclaimer: our backgrounds are CS/AI)

Reifying philosophy with code

- Muehlhauser, Shlegeris: A Software Agent Illustrating Some Features of an Illusionist Account of Consciousness
- An agent that observes the world and uses a theorem prover to answer questions asked of it



from shlegeris.com



```
Q: What's 2 + 2?
4

Q: Suppose there are two agents Bob and Jane, do they have the same qualia associated with every color?
Both that statement and its negation are possible.

Q: For all y, does there exist an x such that x = y + 1?
Yes.

Q: For all two agents, do they see colors the same?
Both that statement and its negation are possible.

Q: Are your memories at timestep 0 and 1 of the same color?
Yes.

Q: Are you seeing the same color now as you saw at timestep 0?
No.

Q: Is it possible for an agent to have an illusion of red?
Yes.

Q: Is it possible for you to have the illusion that Buck is experiencing a color?
Yes.

Q: Is it possible for Buck to have an illusion that he is having the experience of redness?
No, that's impossible.
```

from https://github.com/bshlgrs/consciousness/blob/master/README.md

Reifying philosophy with code

Reifying philosophy with code

- Searle’s view of the relationship between consciousness and brain states
    - Consciousness is causally reducible to brain states but consciousness is ontologically irreducible to brain states

Reifying philosophy with code

- Searle’s view of the relationship between consciousness and brain states
  - Consciousness is causally reducible to brain states but consciousness is ontologically irreducible to brain states
    - ...what does that mean?

Reifying philosophy with code

- Searle’s view of the relationship between consciousness and brain states
  - Consciousness is causally reducible to brain states but consciousness is ontologically irreducible to brain states
    - ...what does that mean?
- Generally is some confusion
  - Enough disagreement that Searle wrote the paper: “Why I’m Not a Property Dualist”

What we're not doing

- Trying to propose a cognitive architecture
- Trying to propose a new AI or machine learning algorithm
- Trying to claim that the software agent is conscious
- Trying to convince anyone these are the correct/best/most useful definitions of mental states and brain states
- Trying to convince anyone Searle is right or wrong

What we're trying to do

- Create a software agent that is consistent with Searle’s view on consciousness
    - (or at least a simplified version of Searle’s view)

What we're trying to do

- Create a software agent that is consistent with Searle’s view on consciousness
    - (or at least a simplified version of Searle’s view)
- (Hopefully) gain a bit deeper understanding of what we may mean by consciousness, brain states, causal reduction, and ontological reduction along the way

Software Engineering, 101

- Requirements – what must the agent do
- Design – how will we build an agent to meet the requirements
- Implementation – the built agent consistent with the design

Agent requirements: unpacking Searle's view

- Consciousness is causally reducible to brain states
- Consciousness is ontologically irreducible to brain states

Agent requirements: unpacking Searle's view

- Brain state
  - The full physical-chemical state of the brain and nervous system
  - Third person, objective

Agent requirements: unpacking Searle's view

- Brain state
  - The full physical-chemical state of the brain and nervous system
  - Third person, objective
- Internal state
  - Representations, goals, rewards, observations, actions, etc.
  - Subjective

Agent requirements: unpacking Searle's view

- Brain state
  - The full physical-chemical state of the brain and nervous system
  - Third person, objective
- Internal state
  - Representations, goals, rewards, observations, actions, etc.
  - Subjective
- Mental state
  - Beliefs, desires, thoughts, perceptions, emotions, knowledge, etc.
  - First person, subjective

Agent requirements: unpacking Searle's view

- Brain state
    - The full physical-chemical state of the brain and nervous system
    - Third person, objective
- Internal state
    - Representations, goals, rewards, observations, actions, etc.
    - Subjective
- Mental state
    - Beliefs, desires, thoughts, perceptions, emotions, knowledge, etc.
    - First person, subjective
- Conscious mental state
    - A mental state in which it is "something it's like to be in"
    - First person, subjective character of experience, phenomenal

Agent requirements: unpacking Searle's view

- Searle’s view
  - Consciousness is causally reducible to brain states
  - Consciousness is ontologically irreducible to brain states

Agent requirements: unpacking Searle's view

- Searle’s view
  - Consciousness is causally reducible to brain states
  - Consciousness is ontologically irreducible to brain states
- V2
  - Conscious mental states are casually reducible to brain states
  - Conscious mental states are ontologically irreducible to brain states

Agent requirements: unpacking Searle's view

- Searle’s view
  - Consciousness is causally reducible to brain states
  - Consciousness is ontologically irreducible to brain states
- V2
  - Conscious mental states are casually reducible to brain states
  - Conscious mental states are ontologically irreducible to brain states
- V1
  - Mental states are casually reducible to brain states
  - Mental states are ontologically irreducible to brain states

Agent requirements: unpacking Searle's view

- Searle’s view
  - Consciousness is causally reducible to brain states
  - Consciousness is ontologically irreducible to brain states
- V2
  - Conscious mental states are casually reducible to brain states
  - Conscious mental states are ontologically irreducible to brain states
- V1
  - Mental states are casually reducible to brain states
  - Mental states are ontologically irreducible to brain states
- V0
  - Internal states are casually reducible to brain states
  - Internal states are ontologically irreducible to brain states

Agent requirements: unpacking Searle's view

- V0
  - Internal states are casually reducible to brain states
  - Internal states are ontologically irreducible to brain states

Agent requirements: unpacking Searle's view

- V0
  - Internal states are casually reducible to brain states
  - Internal states are ontologically irreducible to brain states

Phenomena of type A are ontologically reducible to phenomena of type B
if and only if A’s are nothing but B’s

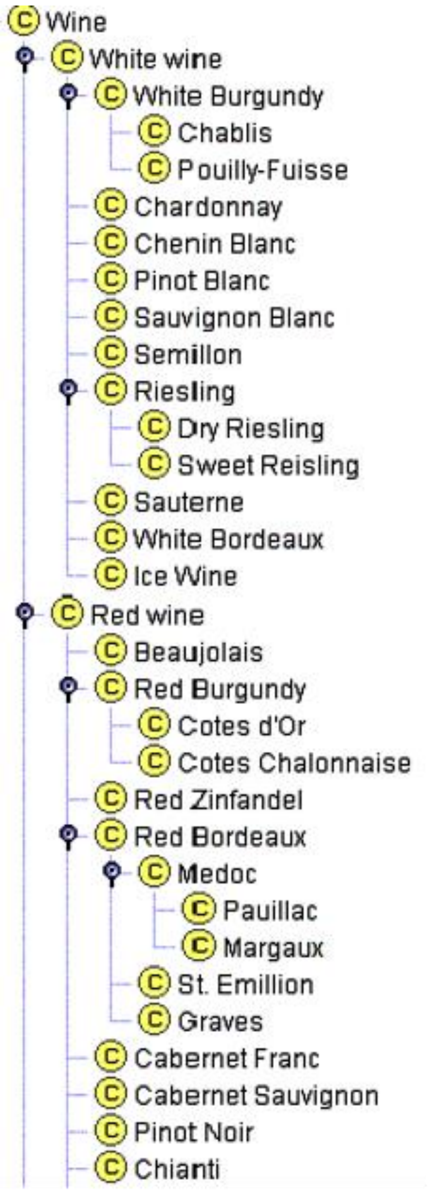
Ontologies in Computer Science

- Class-instance distinction

Images from:
https://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html
https://www.researchgate.net/figure/Owl-Viz-view-of-course-ontology_fig1_261339041

Ontologies in Computer Science

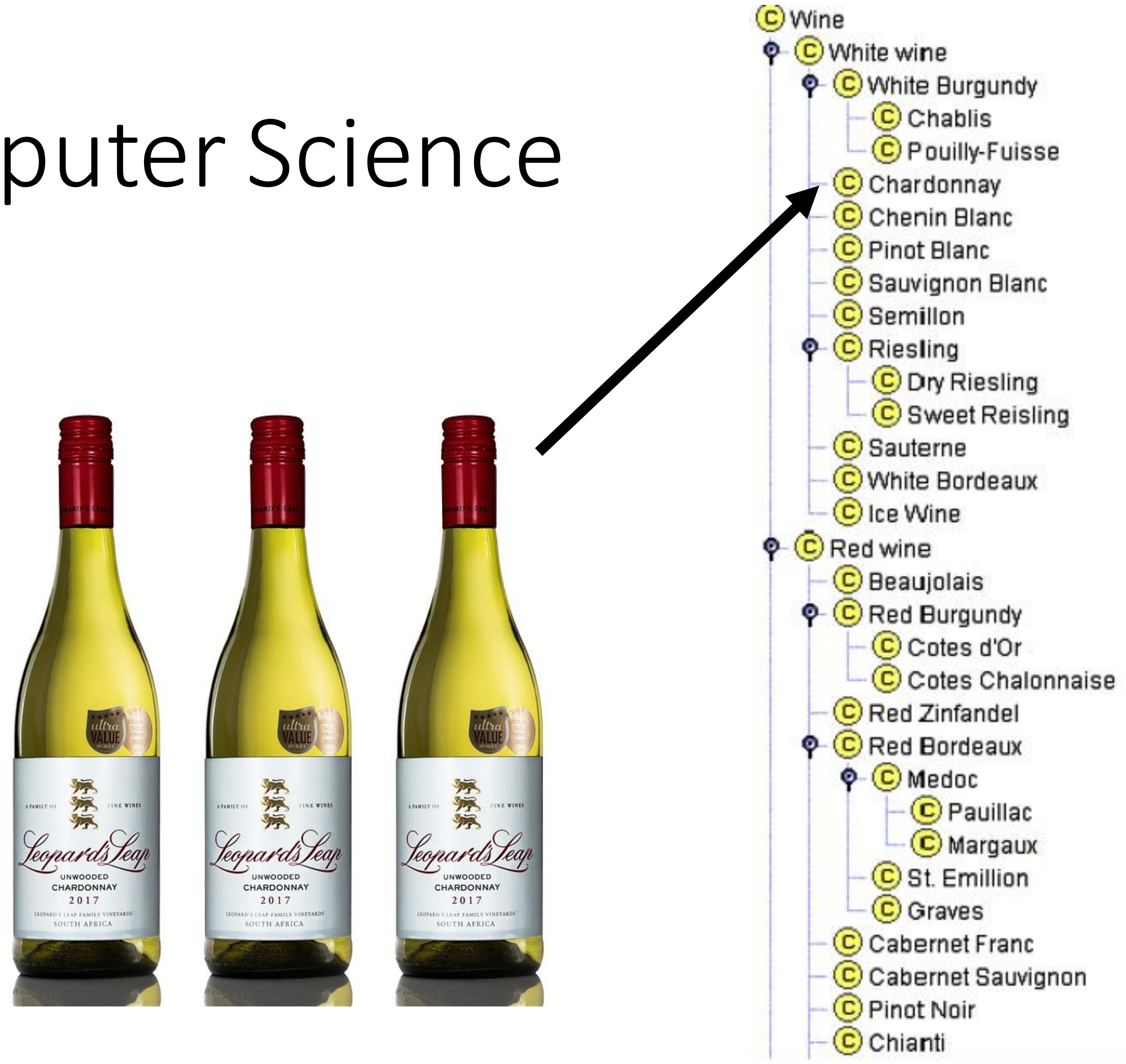
- Class-instance distinction



Images from:
https://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html
https://www.researchgate.net/figure/Owl-Viz-view-of-course-ontology_fig1_261339041

Ontologies in Computer Science

- Class-instance distinction

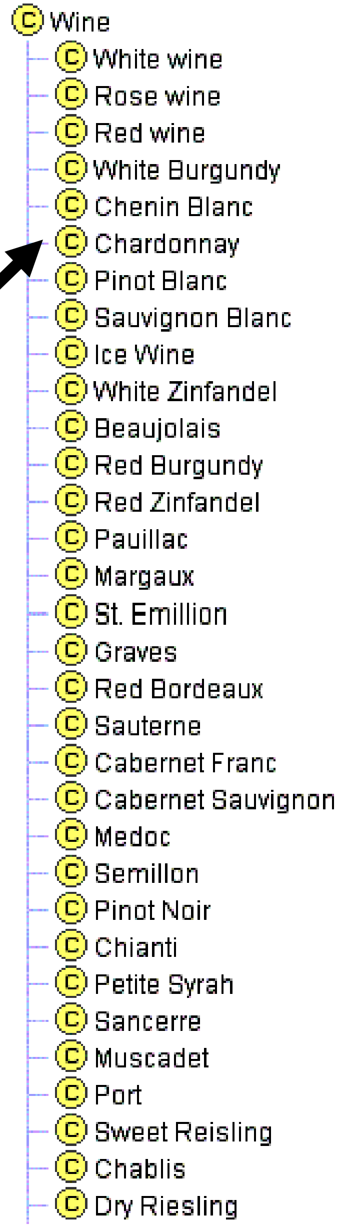


- Wine
  - White wine
    - White Burgundy
      - Chablis
      - Pouilly-Fuisse
    - Chardonnay
    - Chenin Blanc
    - Pinot Blanc
    - Sauvignon Blanc
    - Semillon
    - Riesling
      - Dry Riesling
      - Sweet Reisling
    - Sauterne
    - White Bordeaux
    - Ice Wine
  - Red wine
    - Beaujolais
    - Red Burgundy
      - Cotes d'Or
      - Cotes Chalonnaise
    - Red Zinfandel
    - Red Bordeaux
      - Medoc
        - Pauillac
        - Margaux
      - St. Emillion
      - Graves
    - Cabernet Franc
    - Cabernet Sauvignon
    - Pinot Noir
    - Chianti

Images from:
https://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html
https://www.researchgate.net/figure/Owl-Viz-view-of-course-ontology_fig1_261339041

Ontologies in Computer Science

- Class-instance distinction

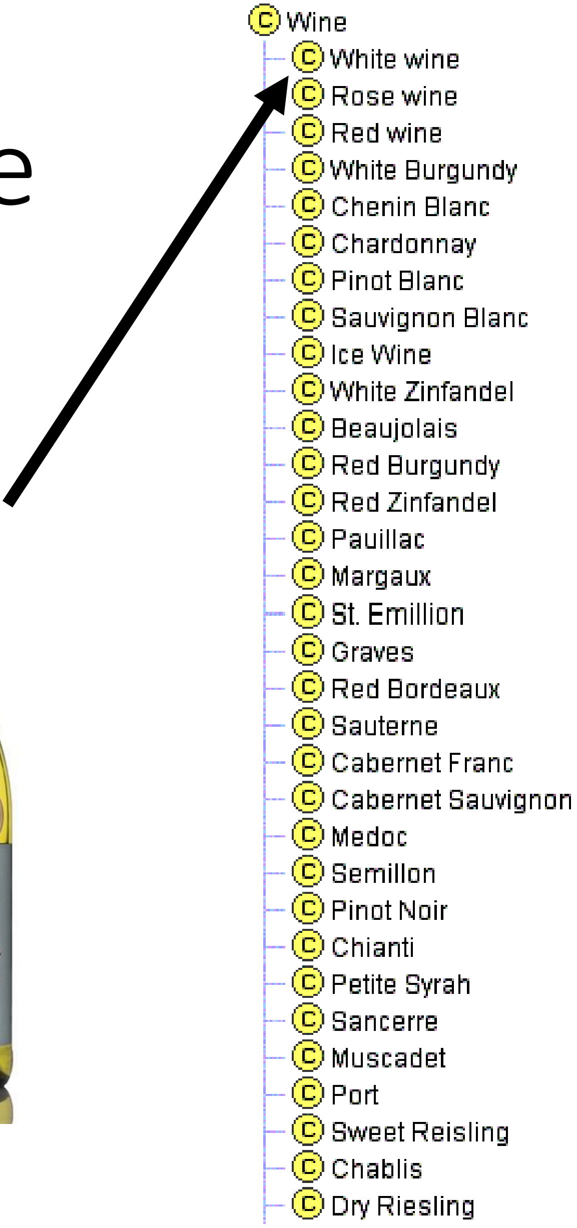


- Wine
  - White wine
  - Rose wine
  - Red wine
  - White Burgundy
  - Chenin Blanc
  - Chardonnay
  - Pinot Blanc
  - Sauvignon Blanc
  - Ice Wine
  - White Zinfandel
  - Beaujolais
  - Red Burgundy
  - Red Zinfandel
  - Pauillac
  - Margaux
  - St. Emillion
  - Graves
  - Red Bordeaux
  - Sauterne
  - Cabernet Franc
  - Cabernet Sauvignon
  - Medoc
  - Semillon
  - Pinot Noir
  - Chianti
  - Petite Syrah
  - Sancerre
  - Muscadet
  - Port
  - Sweet Reisling
  - Chablis
  - Dry Riesling

Images from:
https://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html
https://www.researchgate.net/figure/Owl-Viz-view-of-course-ontology_fig1_261339041

Ontologies in Computer Science

- Class-instance distinction

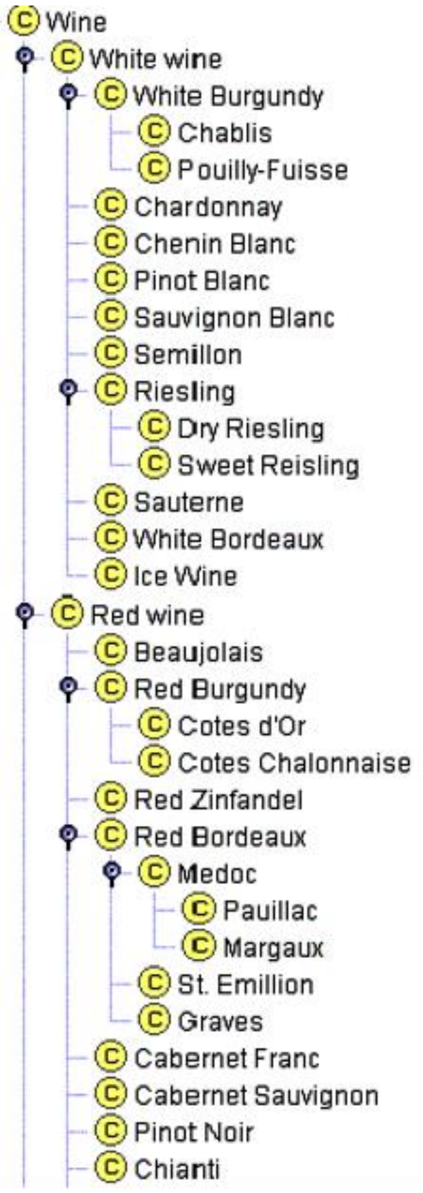


- Wine
  - White wine
  - Rose wine
  - Red wine
  - White Burgundy
  - Chenin Blanc
  - Chardonnay
  - Pinot Blanc
  - Sauvignon Blanc
  - Ice Wine
  - White Zinfandel
  - Beaujolais
  - Red Burgundy
  - Red Zinfandel
  - Pauillac
  - Margaux
  - St. Emillion
  - Graves
  - Red Bordeaux
  - Sauterne
  - Cabernet Franc
  - Cabernet Sauvignon
  - Medoc
  - Semillon
  - Pinot Noir
  - Chianti
  - Petite Syrah
  - Sancerre
  - Muscadet
  - Port
  - Sweet Reisling
  - Chablis
  - Dry Riesling

Images from:
https://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html
https://www.researchgate.net/figure/Owl-Viz-view-of-course-ontology_fig1_261339041

Ontologies in Computer Science

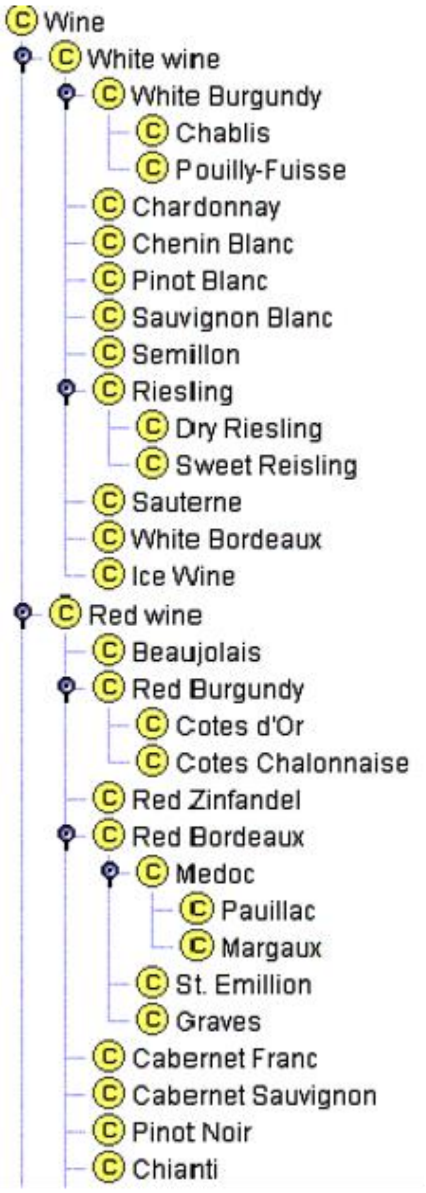
- Class-instance distinction
- Type-token distinction



Images from:
https://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html
https://www.researchgate.net/figure/Owl-Viz-view-of-course-ontology_fig1_261339041

Ontologies in Computer Science

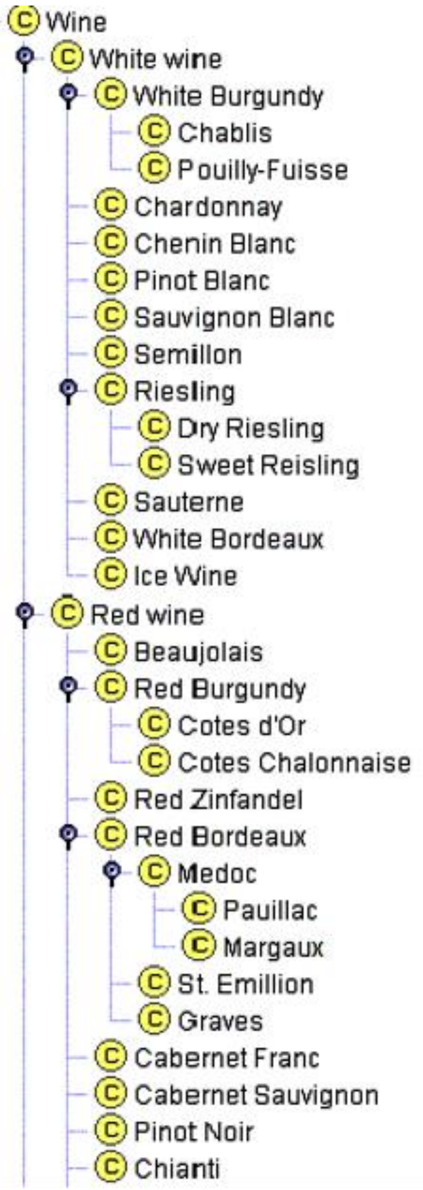
- Class-instance distinction
- Type-token distinction
  - "They drive the same car“
    - They drive the same car type
      - (a Toyota)
    - They drive the same car token
      - (the 2003 Toyota Corolla with VIN: 2QFBORHE4KP911561)



Images from:
https://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html
https://www.researchgate.net/figure/Owl-Viz-view-of-course-ontology_fig1_261339041

Ontologies in Computer Science

- Class-instance distinction
- Type-token distinction
  - "They drive the same car“
    - They drive the same car type
      - (a Toyota)
    - They drive the same car token
      - (the 2003 Toyota Corolla with VIN: 2QFBORHE4KP911561)
- Representing tokens of one type as tokens of another type



- Wine
  - White wine
    - White Burgundy
      - Chablis
      - Pouilly-Fuisse
    - Chardonnay
    - Chenin Blanc
    - Pinot Blanc
    - Sauvignon Blanc
    - Semillon
    - Riesling
      - Dry Riesling
      - Sweet Reisling
    - Sauterne
    - White Bordeaux
    - Ice Wine
  - Red wine
    - Beaujolais
    - Red Burgundy
      - Cotes d'Or
      - Cotes Chalonnaise
    - Red Zinfandel
    - Red Bordeaux
      - Medoc
        - Pauillac
        - Margaux
      - St. Emillion
      - Graves
    - Cabernet Franc
    - Cabernet Sauvignon
    - Pinot Noir
    - Chianti

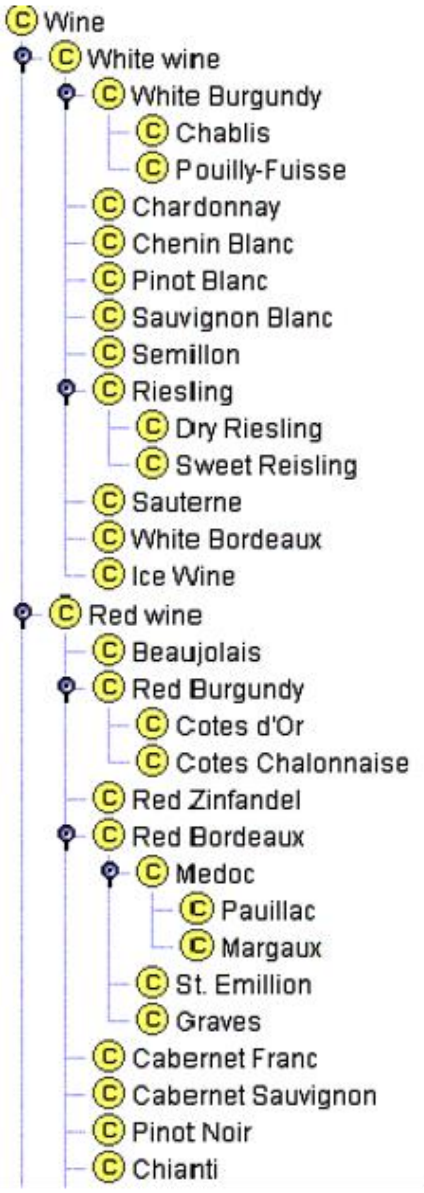
(C) A set of wine bottles

(C) Case of wine

Images from:
https://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html
https://www.researchgate.net/figure/Owl-Viz-view-of-course-ontology_fig1_261339041

Ontologies in Computer Science

- Class-instance distinction
- Type-token distinction
  - "They drive the same car“
    - They drive the same car type
      - (a Toyota)
    - They drive the same car token
      - (the 2003 Toyota Corolla with VIN: 2QFBORHE4KP911561)
- Representing tokens of one type as tokens of another type



- Ⓒ Wine
  - Ⓒ White wine
    - Ⓒ White Burgundy
      - Ⓒ Chablis
      - Ⓒ Pouilly-Fuisse
    - Ⓒ Chardonnay
    - Ⓒ Chenin Blanc
    - Ⓒ Pinot Blanc
    - Ⓒ Sauvignon Blanc
    - Ⓒ Semillon
    - Ⓒ Riesling
      - Ⓒ Dry Riesling
      - Ⓒ Sweet Reisling
    - Ⓒ Sauterne
    - Ⓒ White Bordeaux
    - Ⓒ Ice Wine
  - Ⓒ Red wine
    - Ⓒ Beaujolais
    - Ⓒ Red Burgundy
      - Ⓒ Cotes d'Or
      - Ⓒ Cotes Chalonnaise
    - Ⓒ Red Zinfandel
    - Ⓒ Red Bordeaux
      - Ⓒ Medoc
        - Ⓒ Pauillac
        - Ⓒ Margaux
      - Ⓒ St. Emillion
      - Ⓒ Graves
    - Ⓒ Cabernet Franc
    - Ⓒ Cabernet Sauvignon
    - Ⓒ Pinot Noir
    - Ⓒ Chianti

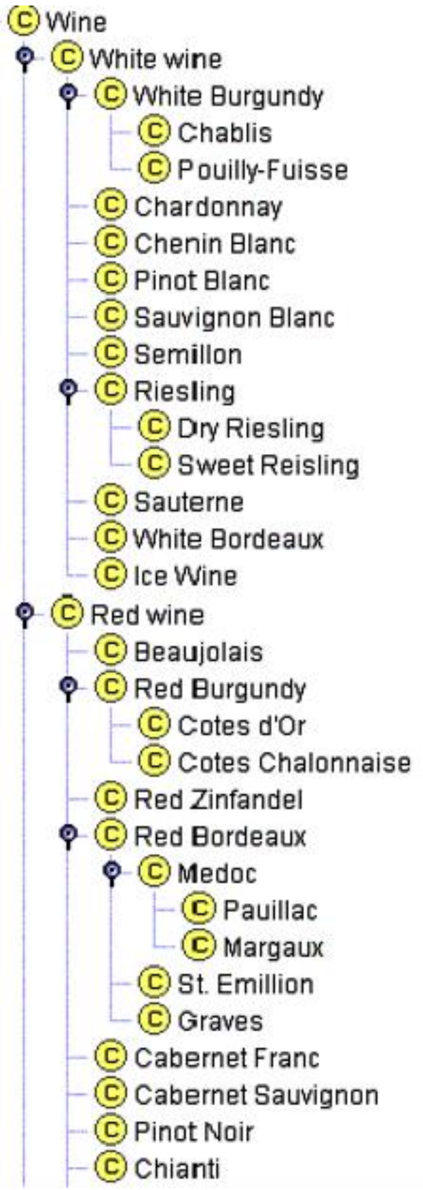
(C) A set of wine bottles

(C) Case of wine

Images from:
https://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html
https://www.researchgate.net/figure/Owl-Viz-view-of-course-ontology_fig1_261339041

Ontologies in Computer Science

- Class-instance distinction
- Type-token distinction
  - "They drive the same car“
    - They drive the same car type
      - (a Toyota)
    - They drive the same car token
      - (the 2003 Toyota Corolla with VIN: 2QFBORHE4KP911561)
- Representing tokens of one type as tokens of another type



- Wine
  - White wine
    - White Burgundy
      - Chablis
      - Pouilly-Fuisse
    - Chardonnay
    - Chenin Blanc
    - Pinot Blanc
    - Sauvignon Blanc
    - Semillon
    - Riesling
      - Dry Riesling
      - Sweet Reisling
    - Sauterne
    - White Bordeaux
    - Ice Wine
  - Red wine
    - Beaujolais
    - Red Burgundy
      - Cotes d'Or
      - Cotes Chalonnaise
    - Red Zinfandel
    - Red Bordeaux
      - Medoc
        - Pauillac
        - Margaux
      - St. Emillion
      - Graves
    - Cabernet Franc
    - Cabernet Sauvignon
    - Pinot Noir
    - Chianti

(C) A set of wine bottles

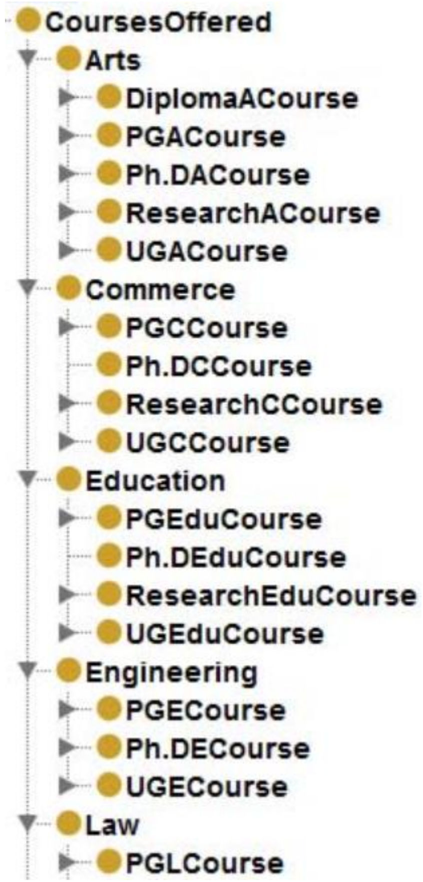
(C) Case of wine

Images from:
https://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html
https://www.researchgate.net/figure/Owl-Viz-view-of-course-ontology_fig1_261339041

Ontologies in Computer Science

- Class-instance distinction
- Type-token distinction
  - "They drive the same car“
    - They drive the same car type
      - (a Toyota)
    - They drive the same car token
      - (the 2003 Toyota Corolla with VIN: 2QFBORHE4KP911561)
- Representing tokens of one type as tokens of another type

(C) Case of wine



- CoursesOffered
  - Arts
    - DiplomaACourse
    - PGACourse
    - Ph.DACourse
    - ResearchACourse
    - UGACourse
  - Commerce
    - PGCCourse
    - Ph.DCCourse
    - ResearchCCourse
    - UGCCourse
  - Education
    - PGEduCourse
    - Ph.DEduCourse
    - ResearchEduCourse
    - UGEduCourse
  - Engineering
    - PGECourse
    - Ph.DECourse
    - UGECourse
  - Law
    - PGLCourse

Images from:
https://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html
https://www.researchgate.net/figure/Owl-Viz-view-of-course-ontology_fig1_261339041

Agent requirements: unpacking Searle's view

- V0
  - Internal states are casually reducible to brain states
  - Internal states are ontologically irreducible to brain states

Phenomena of type A are ontologically reducible to phenomena of type B
if and only if A’s are nothing but B’s

Agent requirements: unpacking Searle's view

- V0
  - Internal states are casually reducible to brain states
  - Internal states are ontologically irreducible to brain states

~~Phenomena of type A are ontologically reducible to phenomena of type B~~
~~if and only if A's are nothing but B's~~

Instances of class A are ontologically reducible to instances of class B
if and only if instances of A's are nothing but instances B's

Agent requirements: unpacking Searle's view

- V0
  - Internal states are casually reducible to brain states
  - Internal states are ontologically irreducible to brain states

Agent requirements: unpacking Searle's view

- V0
  - Internal states are casually reducible to brain states
  - Internal states are ontologically irreducible to brain states

Phenomena of type A are causally reducible to phenomena of type B if and only if:

- the behavior of A’s are entirely casually explained by the behavior of B’s
- A’s have no causal powers in addition to the powers of B’s

Agent requirements: unpacking Searle's view

- V0
  - Internal states are casually reducible to brain states
  - Internal states are ontologically irreducible to brain states

~~Phenomena of type A are causally reducible to phenomena of type B if and only if:~~

- ~~the behavior of A's are entirely casually explained by the behavior of B's~~
- ~~A's have no causal powers in addition to the powers of B's~~

Instances of class A are causally reducible to objects of class B if and only if:

- the behavior of instances of A's are entirely casually explained by the behavior of instances of B's
- instances of A's have no causal powers in addition to the powers of the instances of B's

Agent requirements, V0

- Internal states are casually reducible to brain states
- Internal states are ontologically irreducible to brain states

Design, V0

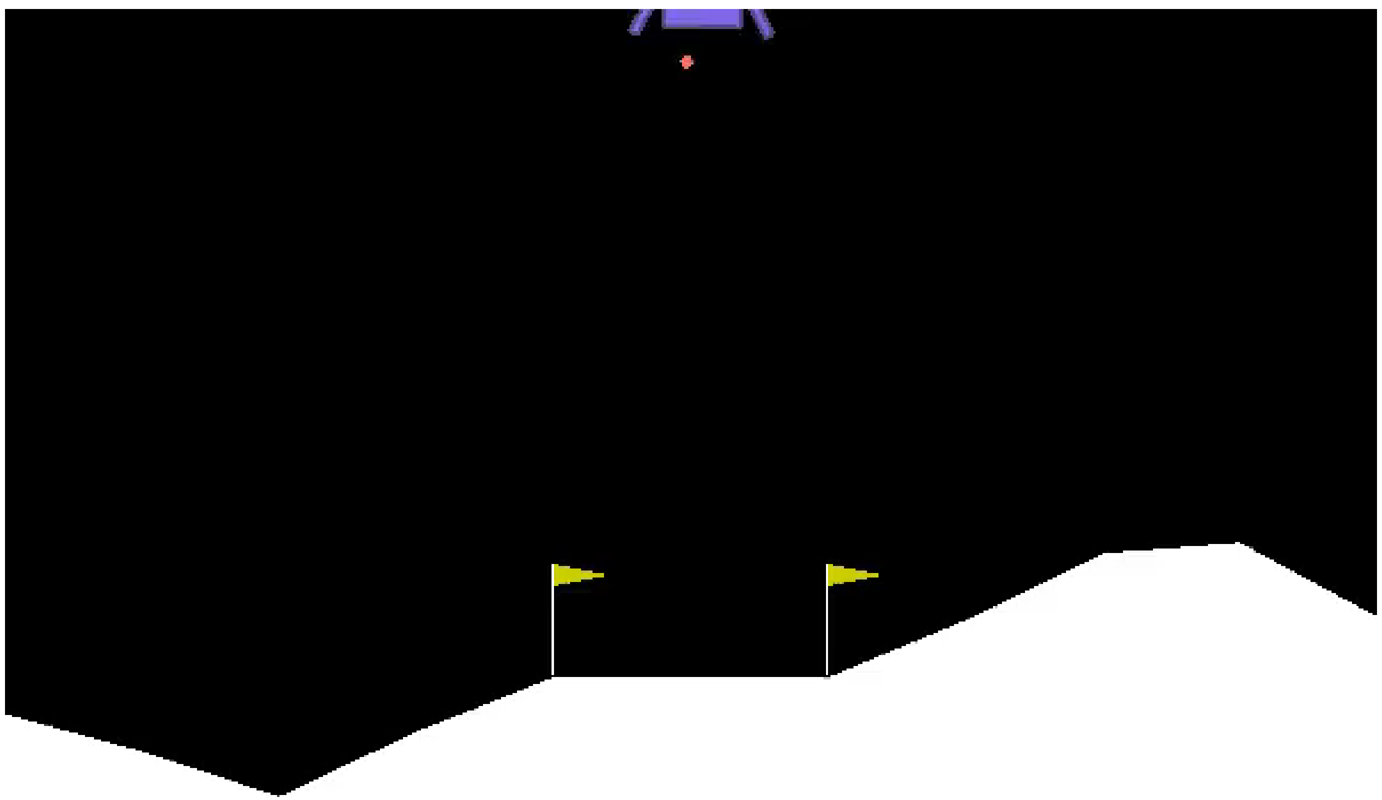
- Design decisions

Design, V0

- Design decisions
  - Environment and the agent’s “physical” form

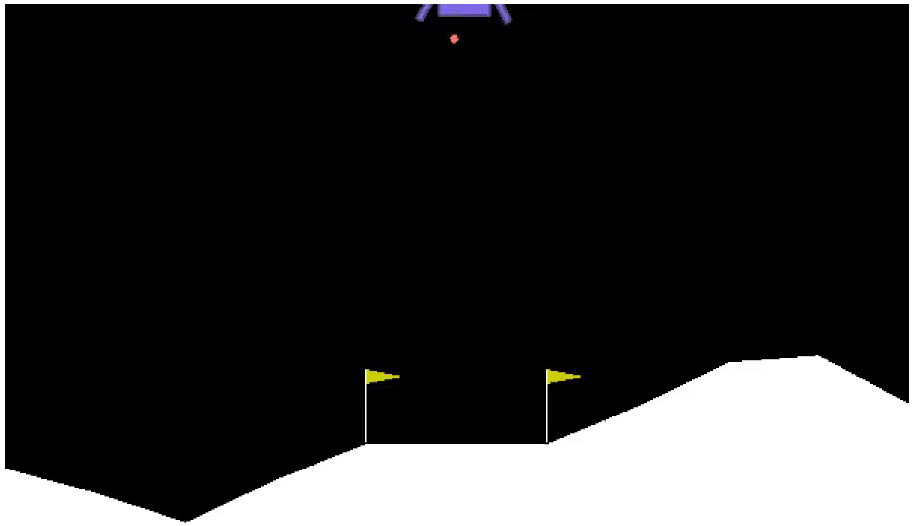
Design, V0

- OpenAI's LunarLander-v2



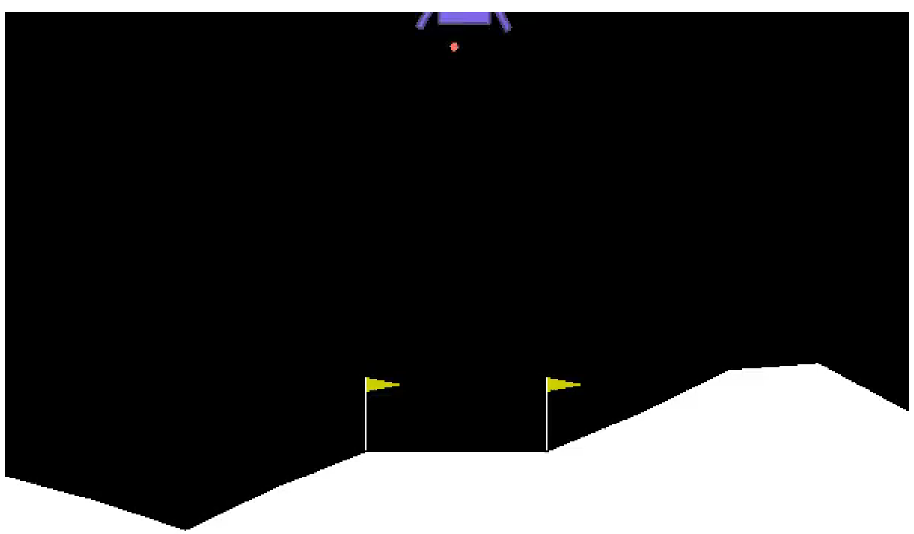
Design, V0

- Design decisions
  - Environment and the agent’s “physical” form



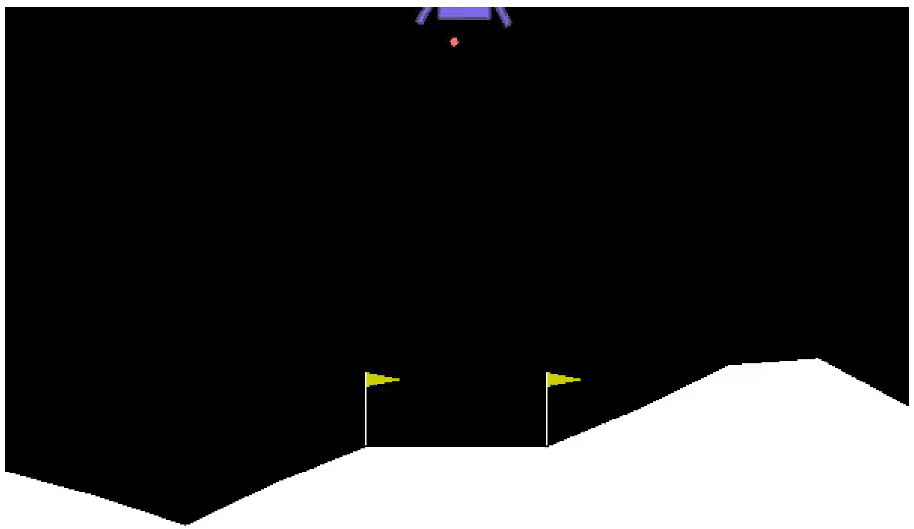
Design, V0

- Design decisions
  - Environment and the agent's "physical" form
  - Internal state of the agent

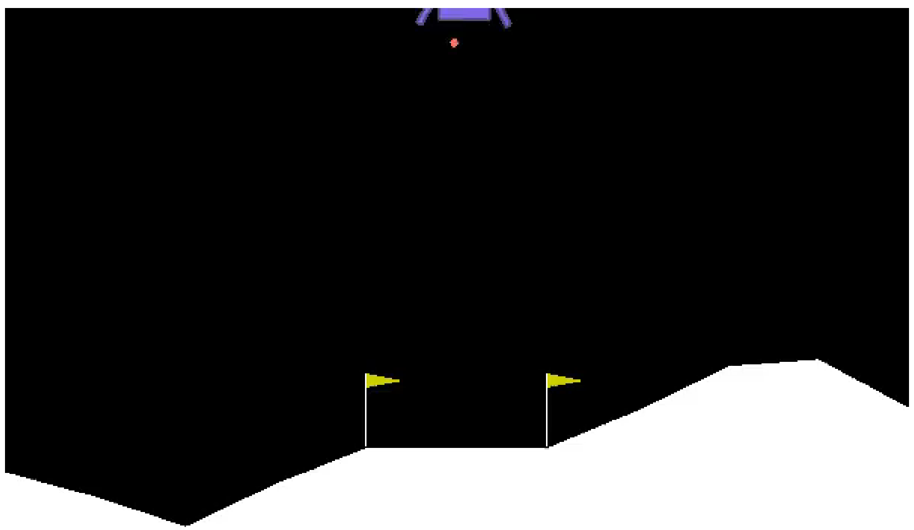


Design, V0

- Design decisions
  - Environment and the agent’s “physical” form
  - Internal state of the agent
    - Beliefs about itself relative to semantically important regions

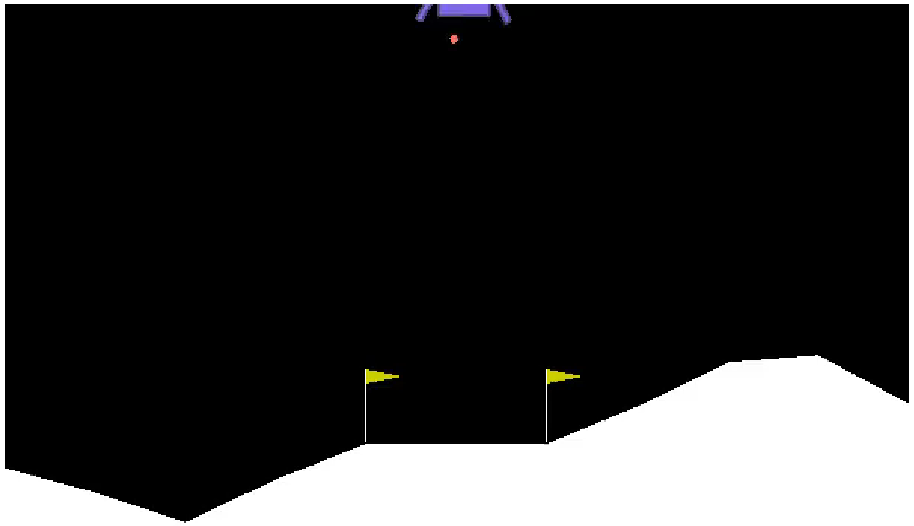


Design, V0



- Design decisions
  - Environment and the agent’s “physical” form
  - Internal state of the agent
    - Beliefs about itself relative to semantically important regions
      - Left of the flags, right of the flags, high above the ground, close to the ground, falling too fast

Design, V0



- Design decisions
  - Environment and the agent’s “physical” form
  - Internal state of the agent
    - Beliefs about itself relative to semantically important regions
      - Left of the flags, right of the flags, high above the ground, close to the ground, falling too fast
  - Brain state of the agent

Neural networks

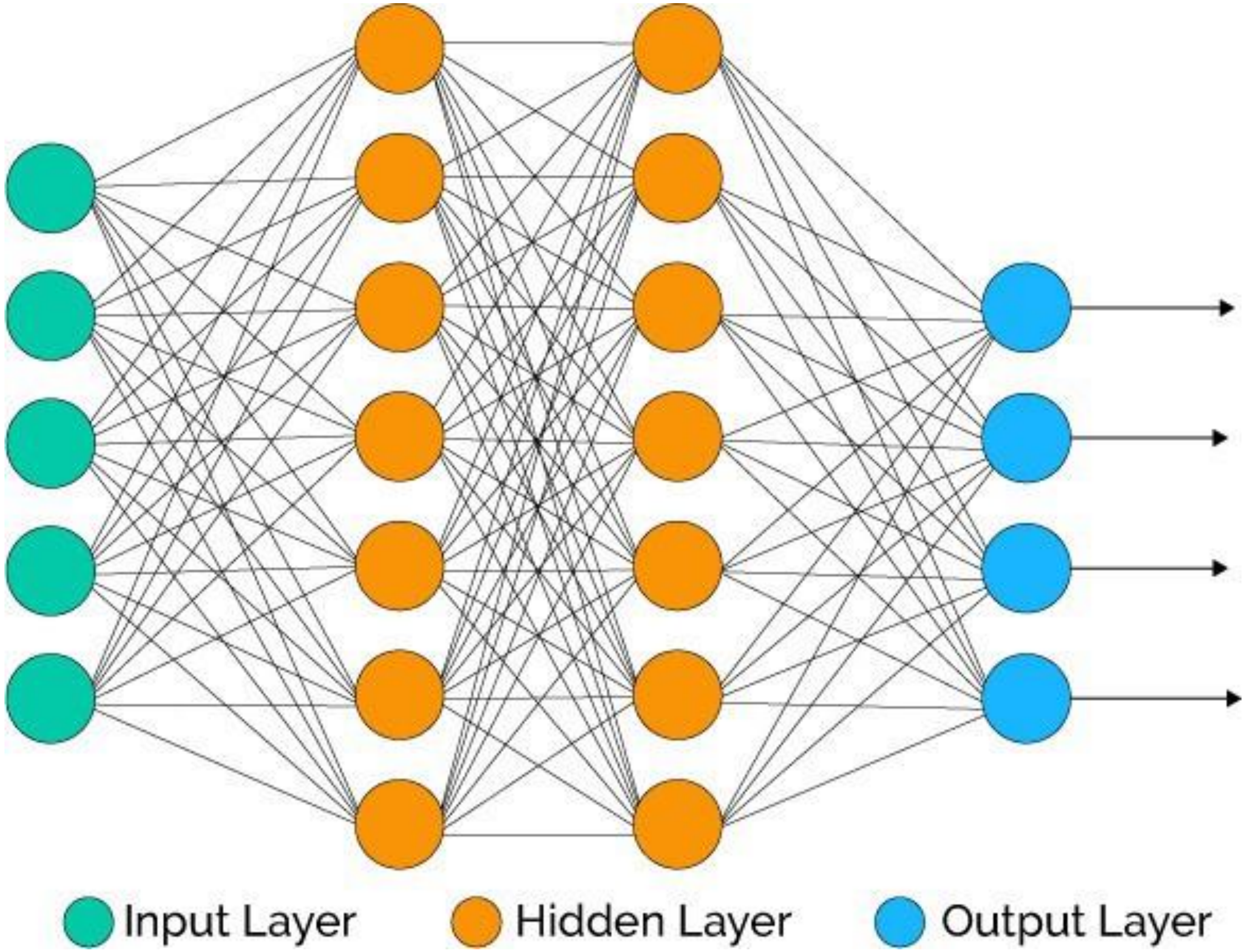


Image from:
https://medium.com/datadriveninvestor/when-not-to-use-neural-networks-89fb50622429

Neural networks

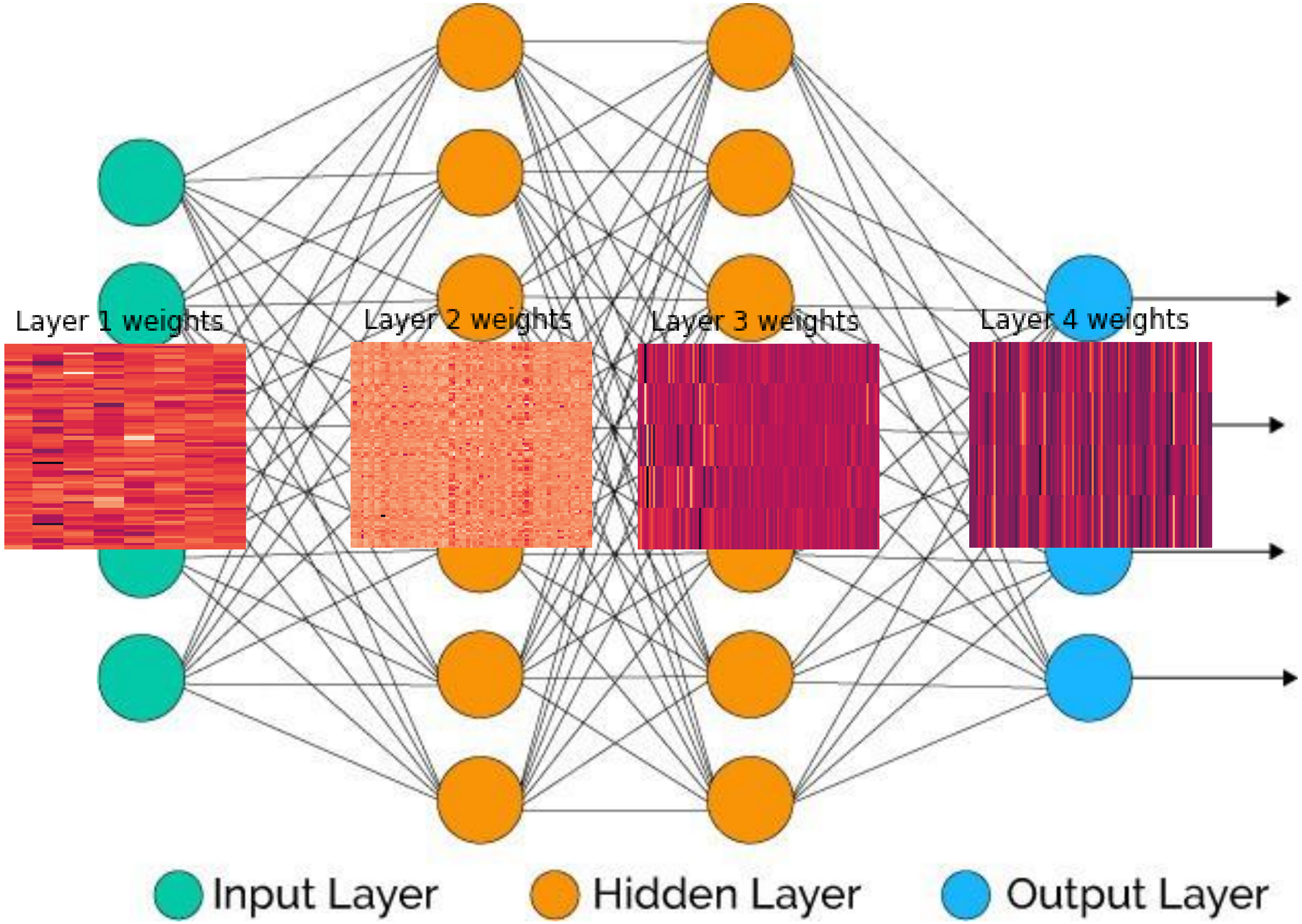


Image from:
https://medium.com/datadriveninvestor/when-not-to-use-neural-networks-89fb50622429

Neural networks

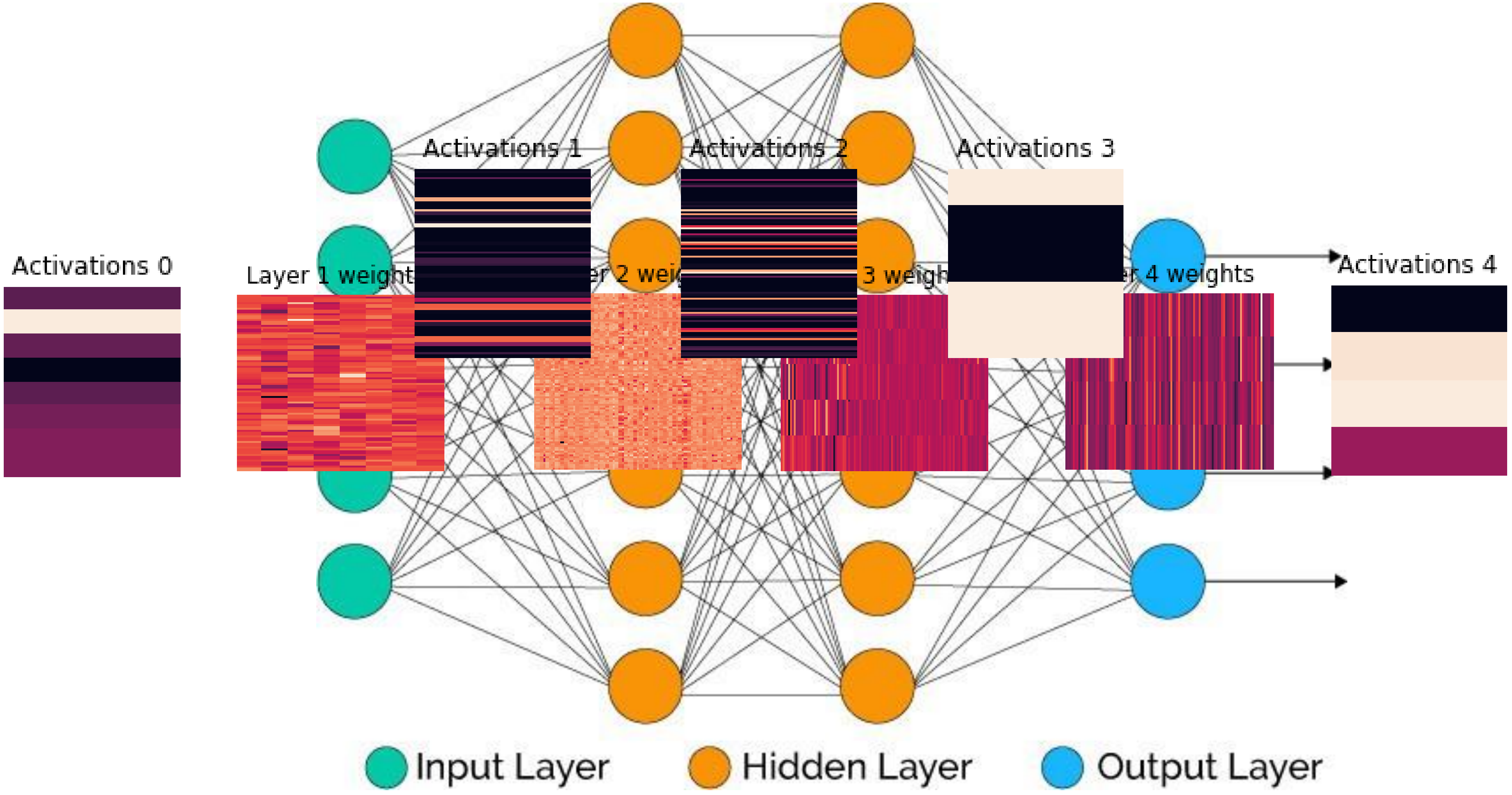
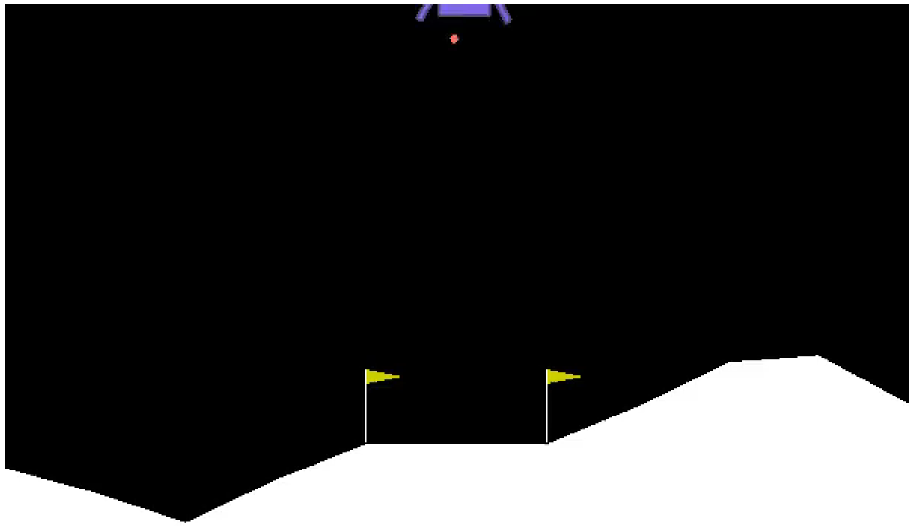


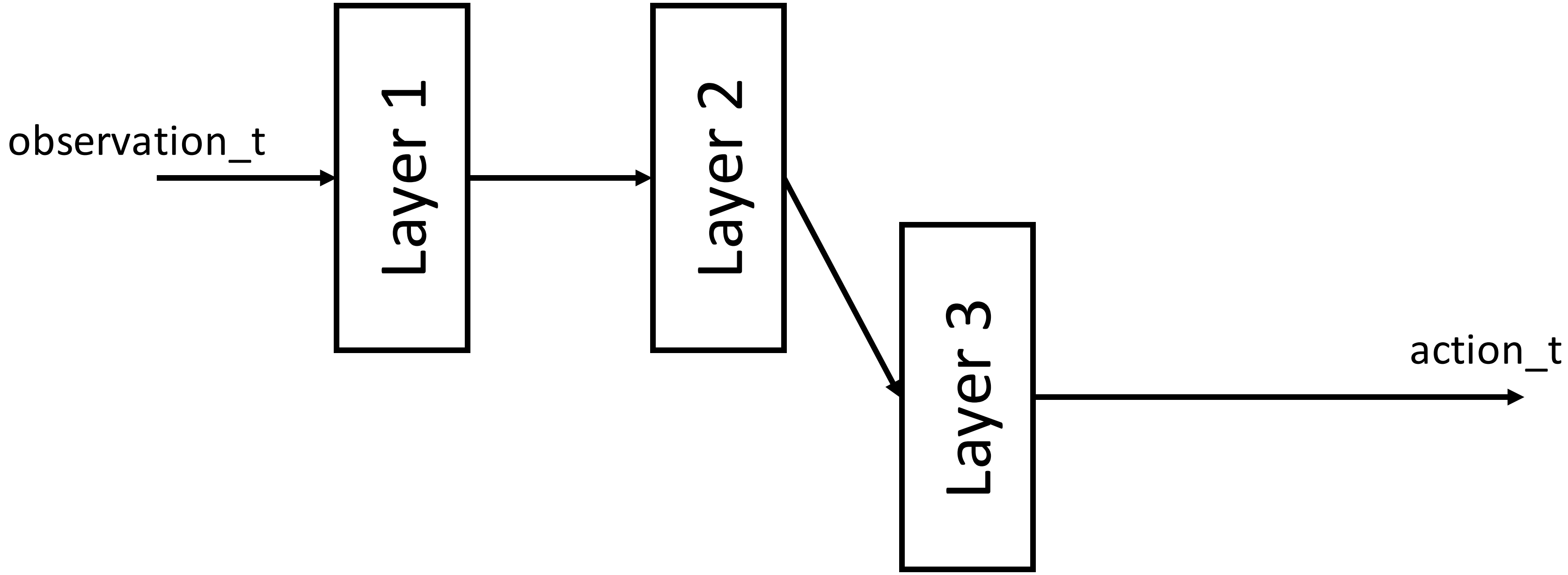
Image from:
https://medium.com/datadriveninvestor/when-not-to-use-neural-networks-89fb50622429

Design, V0

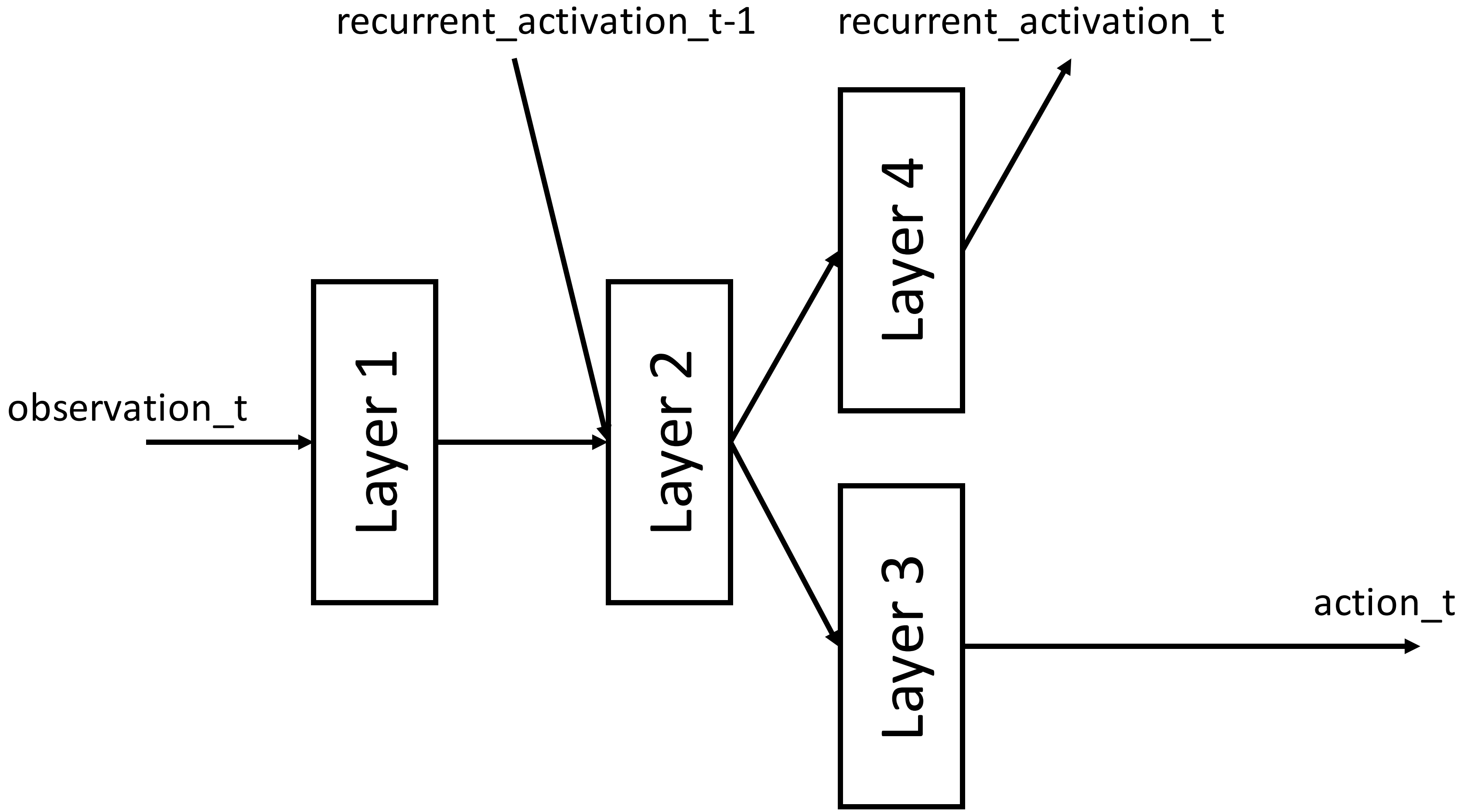


- Design decisions
  - Environment and the agent's "physical" form
  - Internal state of the agent
    - Beliefs about itself relative to semantically important regions
      - Left of the flags, right of the flags, high above the ground, close to the ground, falling too fast
  - Brain state of the agent

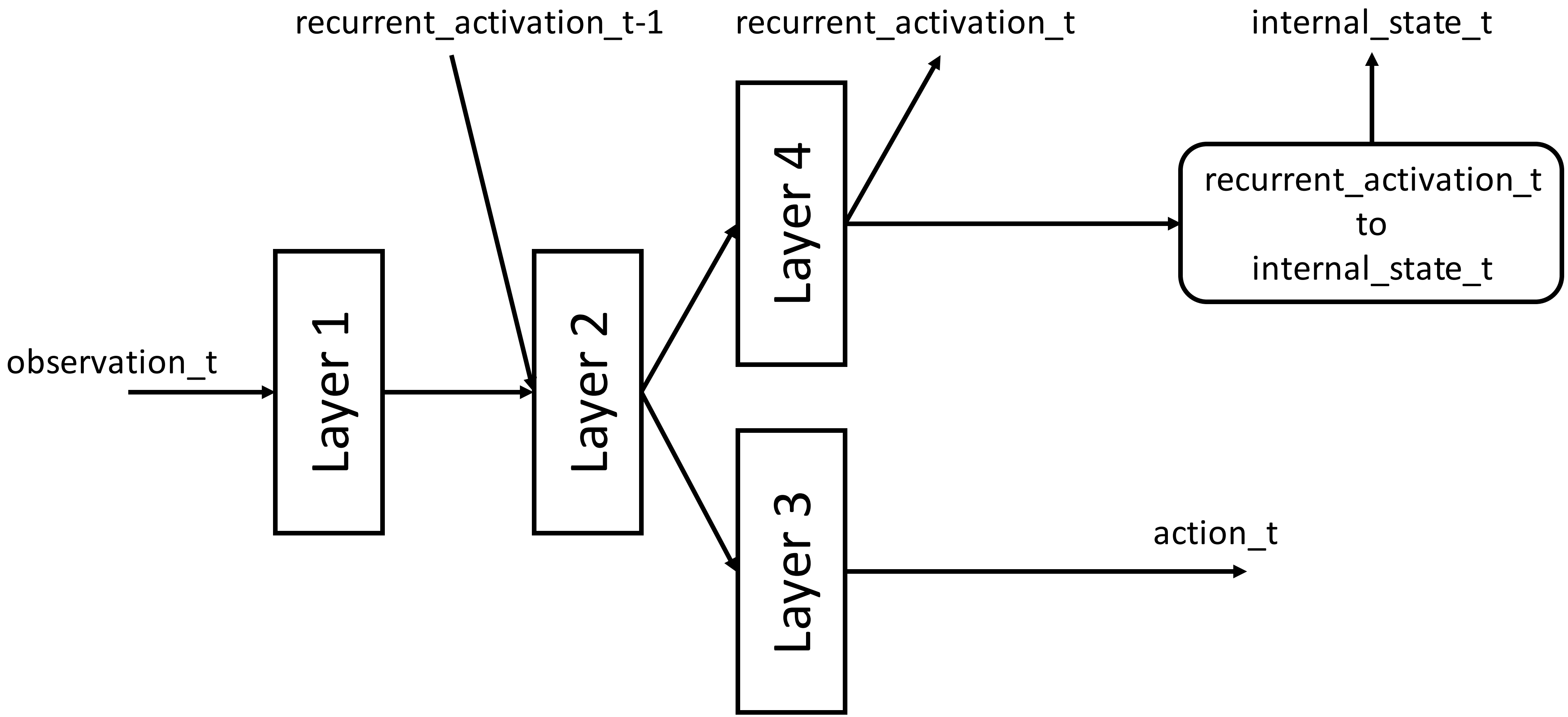
Design, V0



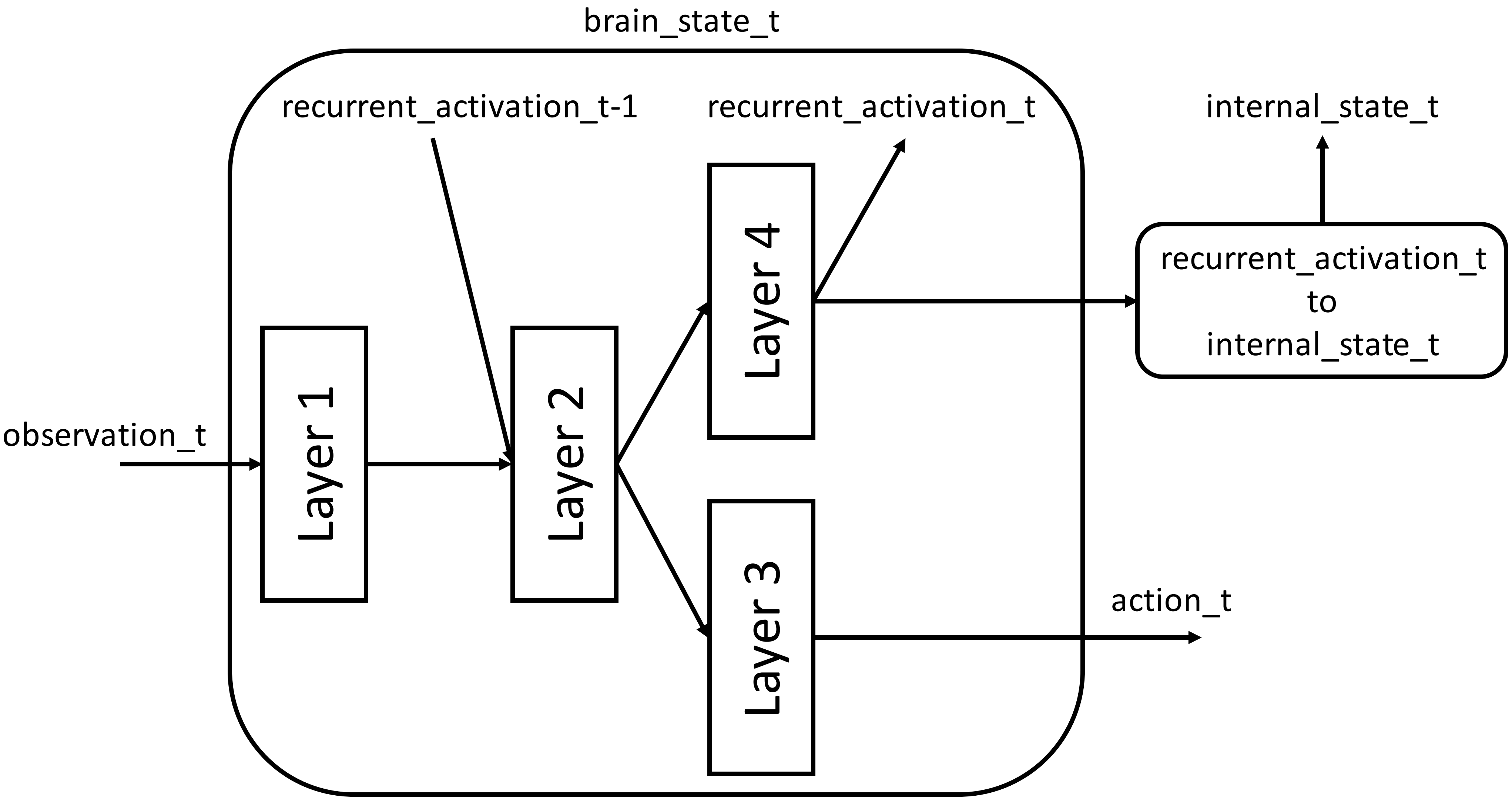
Design, V0


recurrent_activation_t-1
recurrent_activation_t
Layer 4
observation_t
Layer 1
Layer 2
Layer 3
action_t

Design, V0


recurrent_activation_t-1
recurrent_activation_t
internal_state_t
recurrent_activation_t
to
internal_state_t
observation_t
Layer 1
Layer 2
Layer 4
Layer 3
action_t

Design, V0



brain_state_t

recurrent_activation_t-1

recurrent_activation_t

internal_state_t

recurrent_activation_t
to
internal_state_t

observation_t

Layer 1

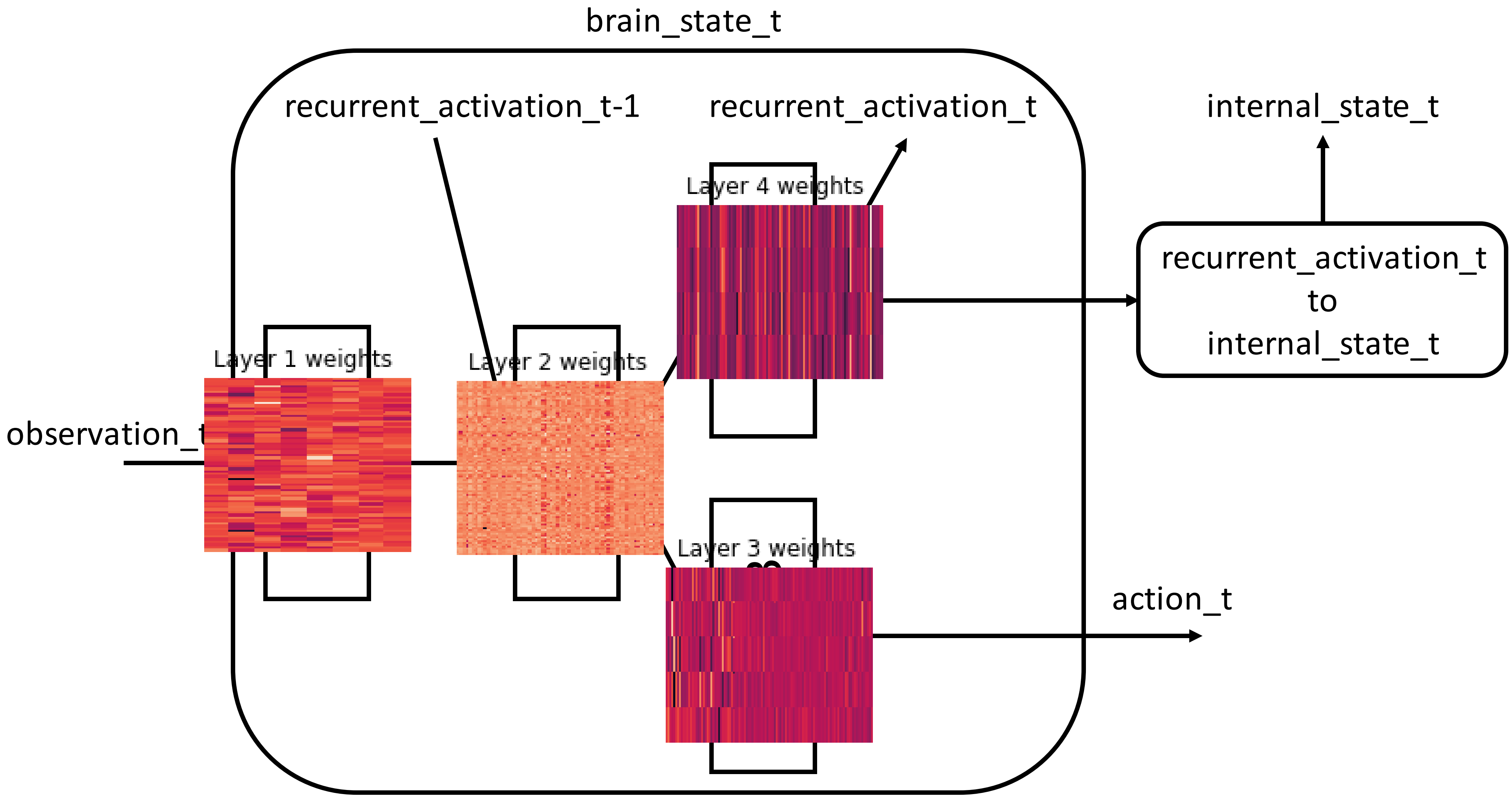
Layer 2

Layer 4

Layer 3

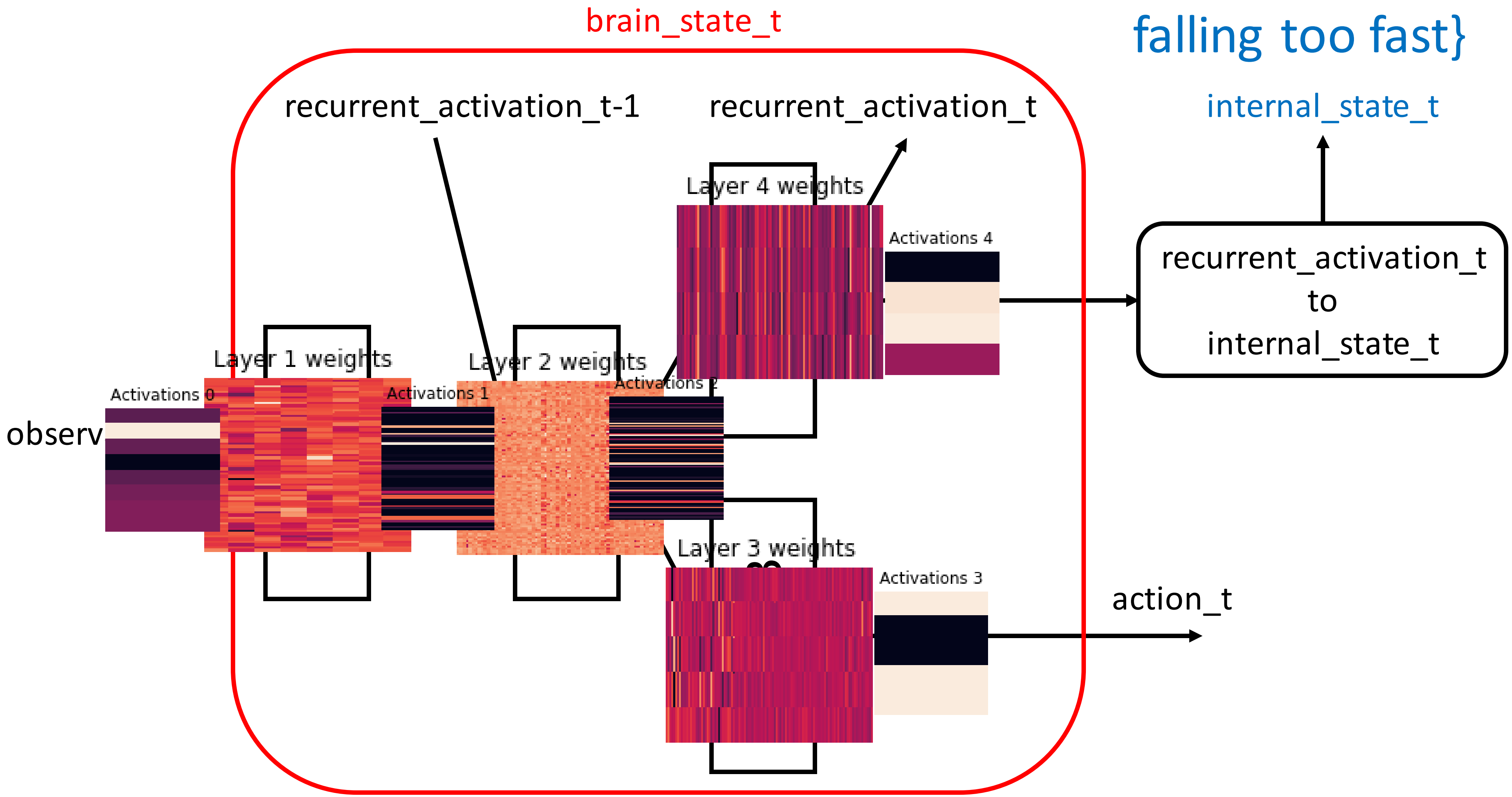
action_t

Design, V0


brain_state_t
recurrent_activation_t-1
recurrent_activation_t
internal_state_t
Layer 4 weights
recurrent_activation_t
to
internal_state_t
Layer 1 weights
Layer 2 weights
observation_t
Layer 3 weights
action_t

Design, V0

{Left of the flags,
high above the ground,
falling too fast}



brain_state_t

recurrent_activation_t-1

recurrent_activation_t

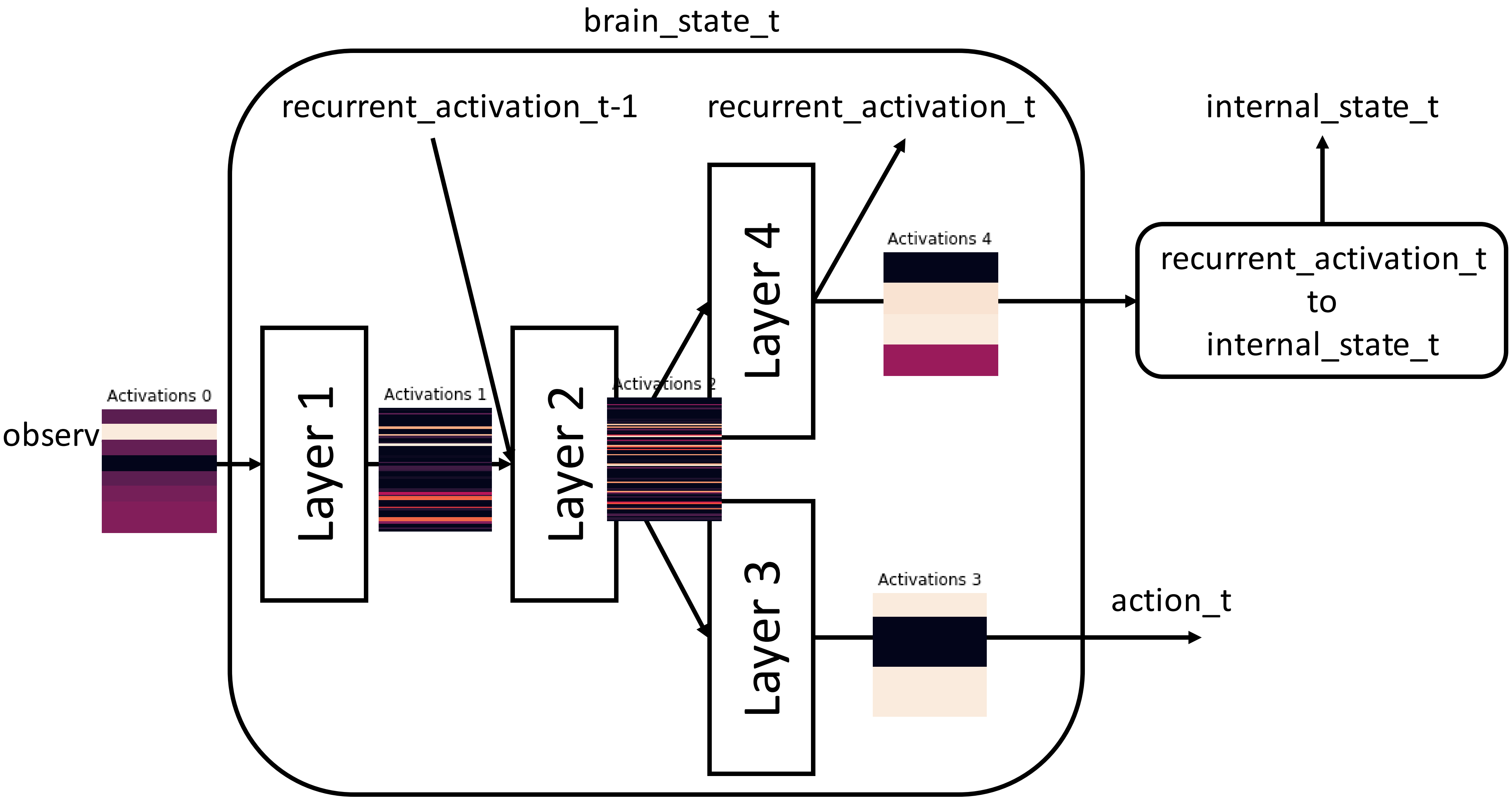
internal_state_t

recurrent_activation_t
to
internal_state_t

observ

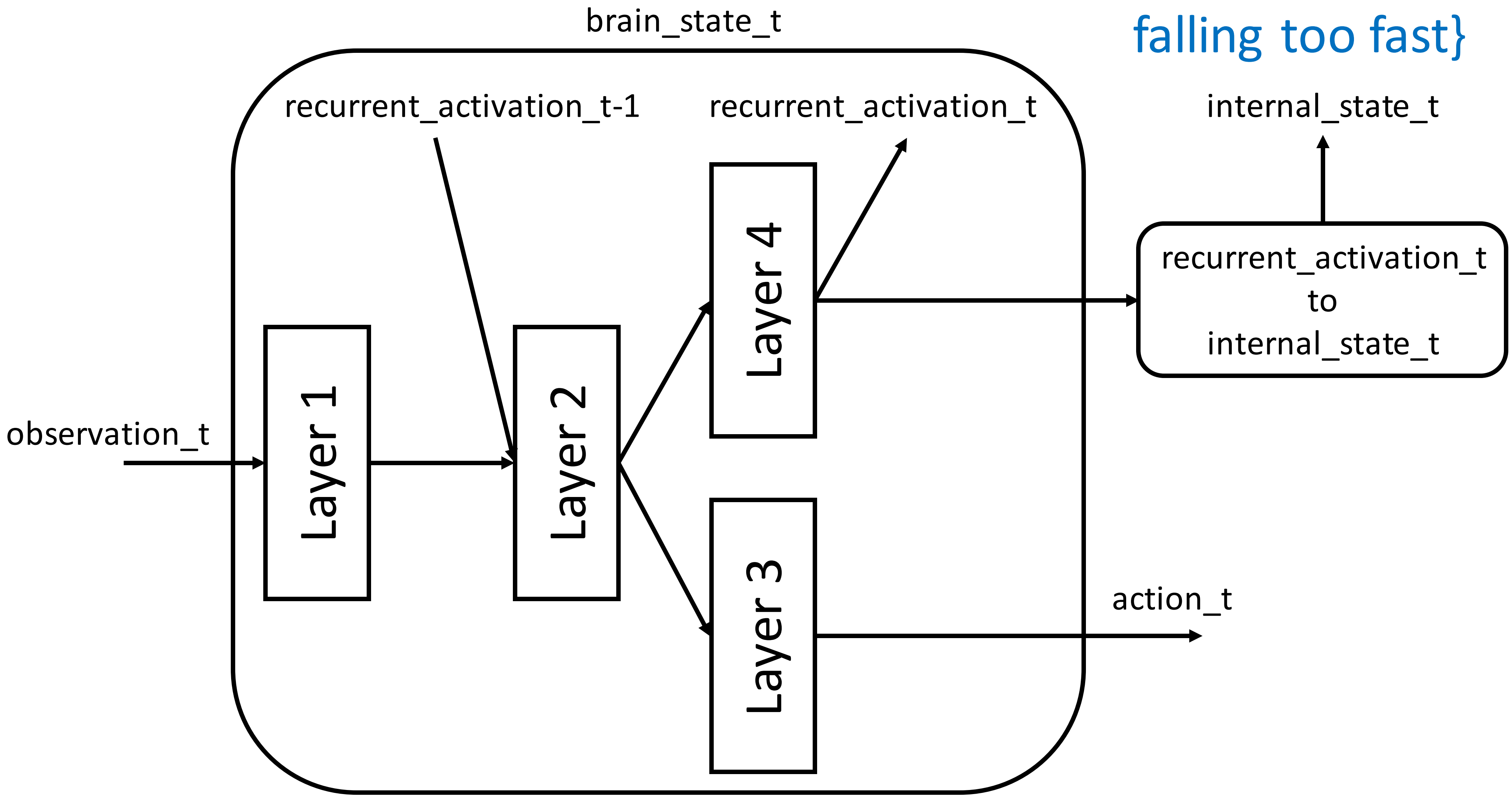
action_t

Design, V0


brain_state_t
recurrent_activation_t-1
recurrent_activation_t
internal_state_t
recurrent_activation_t
to
internal_state_t
Activations 4
Layer 4
Activations 0
observ
Layer 1
Activations 1
Layer 2
Activations 2
Layer 3
Activations 3
action_t

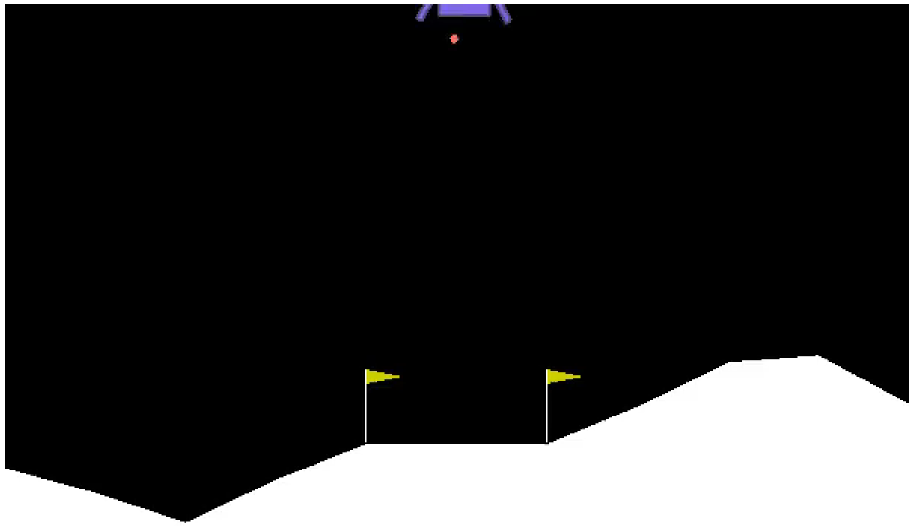
Design, V0

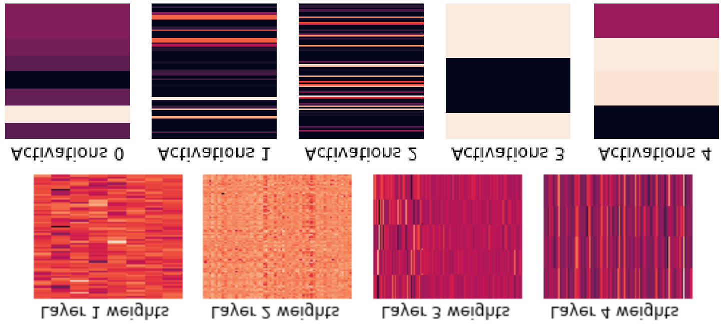
{Left of the flags,
high above the ground,
falling too fast}



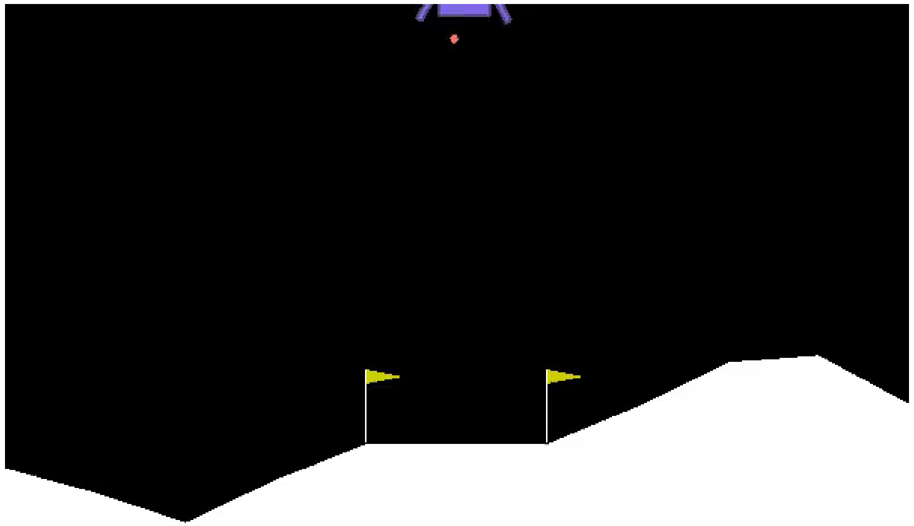
Design, V0

- Design decisions
  - Environment and the agent’s “physical” form
  - Internal state of the agent
    - Beliefs about itself relative to semantically important regions
      - Left of the flags, right of the flags, high above the ground, close to the ground, falling too fast
  - Brain state of the agent

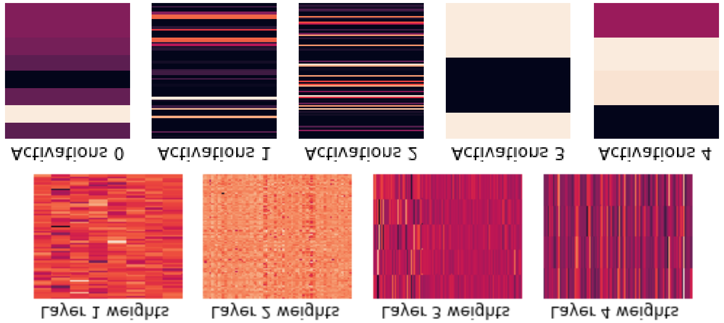




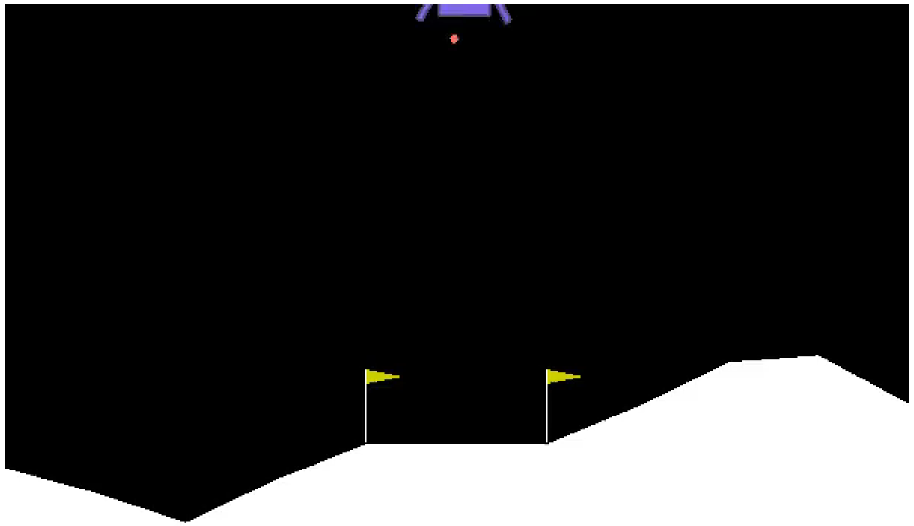
Design, V0



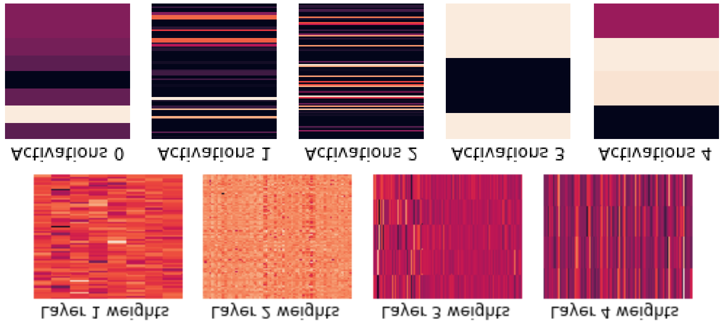
- Design decisions
  - Environment and the agent’s “physical” form
  - Internal state of the agent
    - Beliefs about itself relative to semantically important regions
      - Left of the flags, right of the flags, high above the ground, close to the ground, falling too fast
  - Brain state of the agent
  - Our ontology



Design, V0

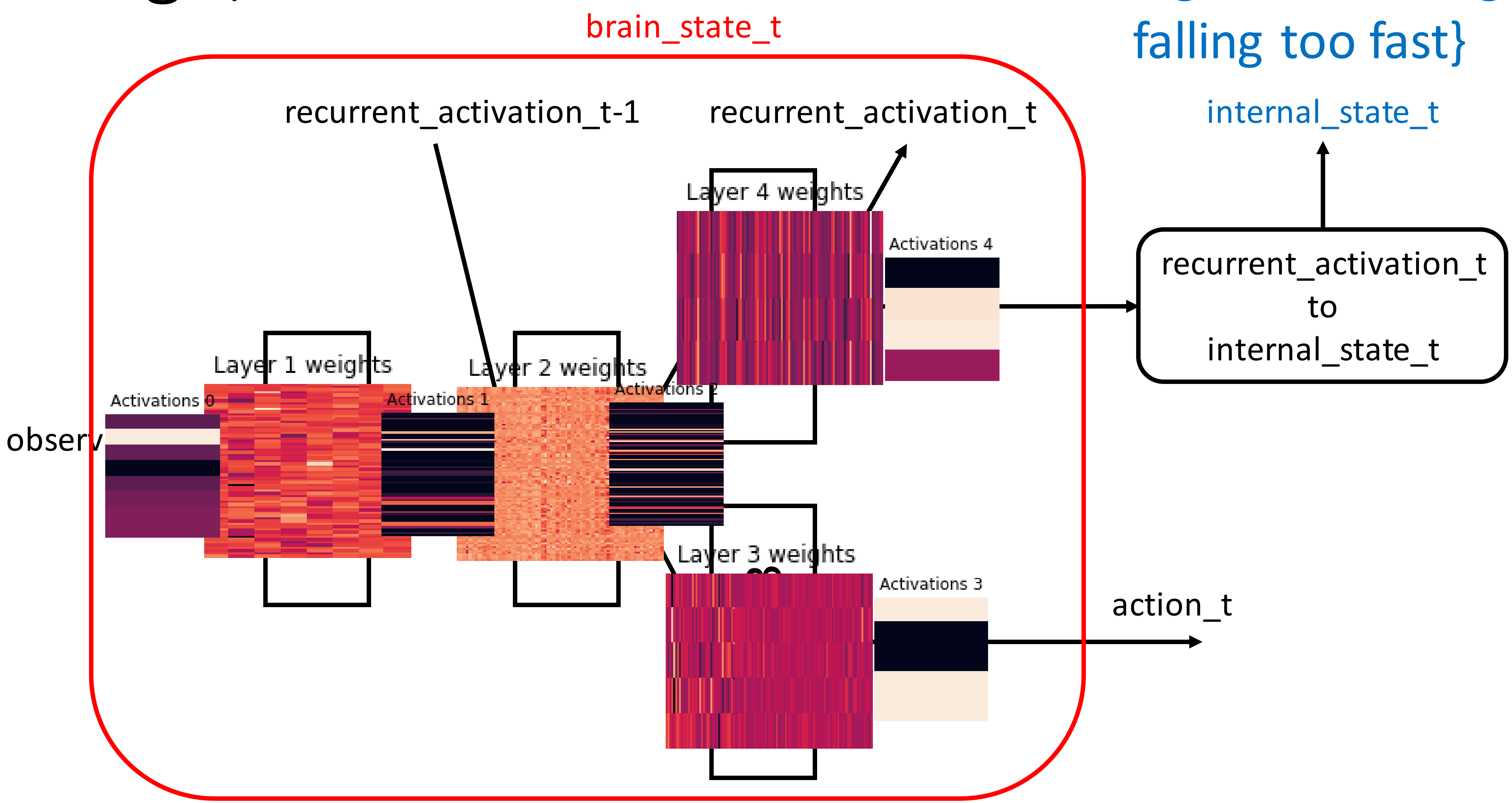


- Design decisions
  - Environment and the agent’s “physical” form
  - Internal state of the agent
    - Beliefs about itself relative to semantically important regions
      - Left of the flags, right of the flags, high above the ground, close to the ground, falling too fast
  - Brain state of the agent
  - Our ontology
    - Layer weights of the neural network
    - Connectivity of the neural network
    - Activations of the neural network at time t
    - The agent's observation at time t
    - The agent's action at time t
    - The position and velocity of the agent at time t
    - Brain state (set of layer weights, activations, and connectivity) at time t
    - Region
    - Internal state (set of regions the agent believes it's in) at time t



Design, V0

{Left of the flags,
high above the ground,
falling too fast}



Design, V0

- Remaining questions
  - How will the agent learn to behave in the world?
  - How will brain states be “connected” to internal states?
  - How will the agent learn to recognize the correspondence between its internal states and its position/velocity?

Reinforcement learning

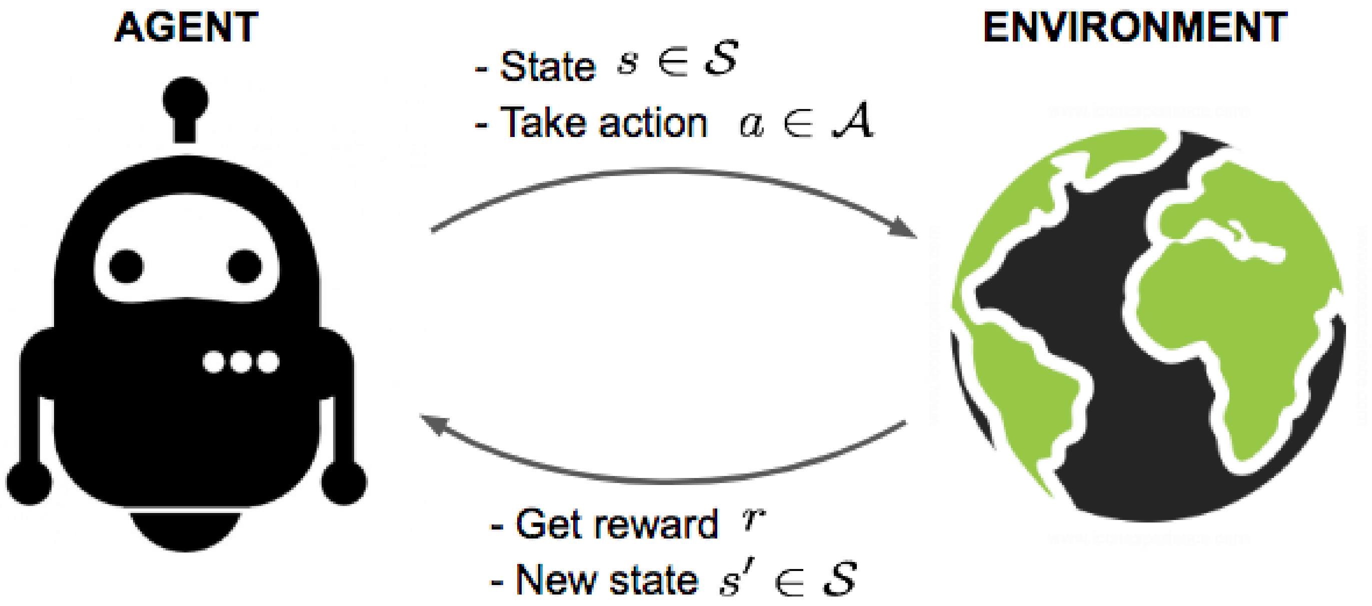
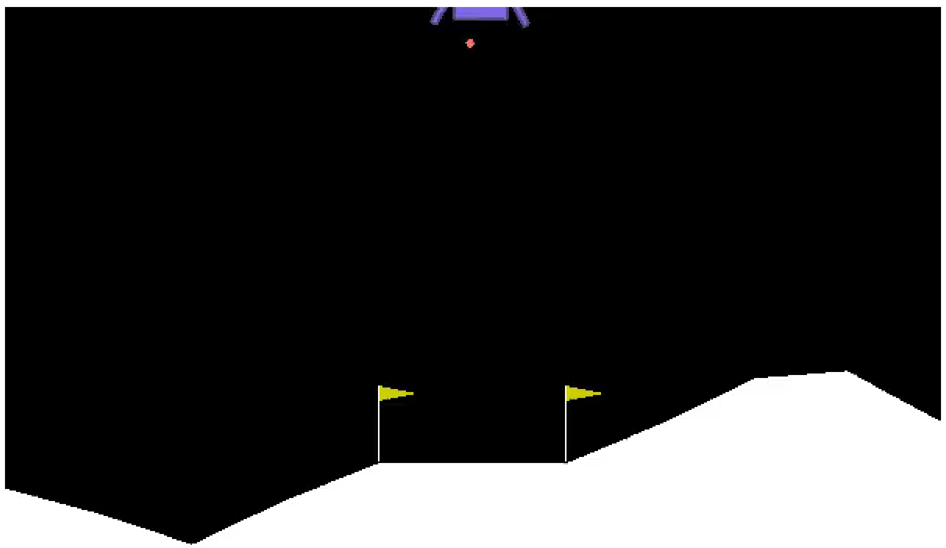

AGENT
ENVIRONMENT
- State $s \in \mathcal{S}$
- Take action $a \in \mathcal{A}$
- Get reward r
- New state $s' \in \mathcal{S}$

Image from:
https://lilianweng.github.io/lil-log/2018/02/19/a-long-peek-into-reinforcement-learning.html

Reinforcement learning

- OpenAI's LunarLander-v2
  - The goal is to softly land between the flags
  - Episode finishes if the lander crashes or comes to rest, receiving additional -100 or +100 points
  - Each leg ground contact is +10
  - Firing the engines is a small negative reward
  - Small positive reward for smoother flight
  - Fuel is infinite
  - Four discrete actions available:
    - do nothing, fire left orientation engine, fire main engine, fire right orientation engine
- We used DQN to train the network

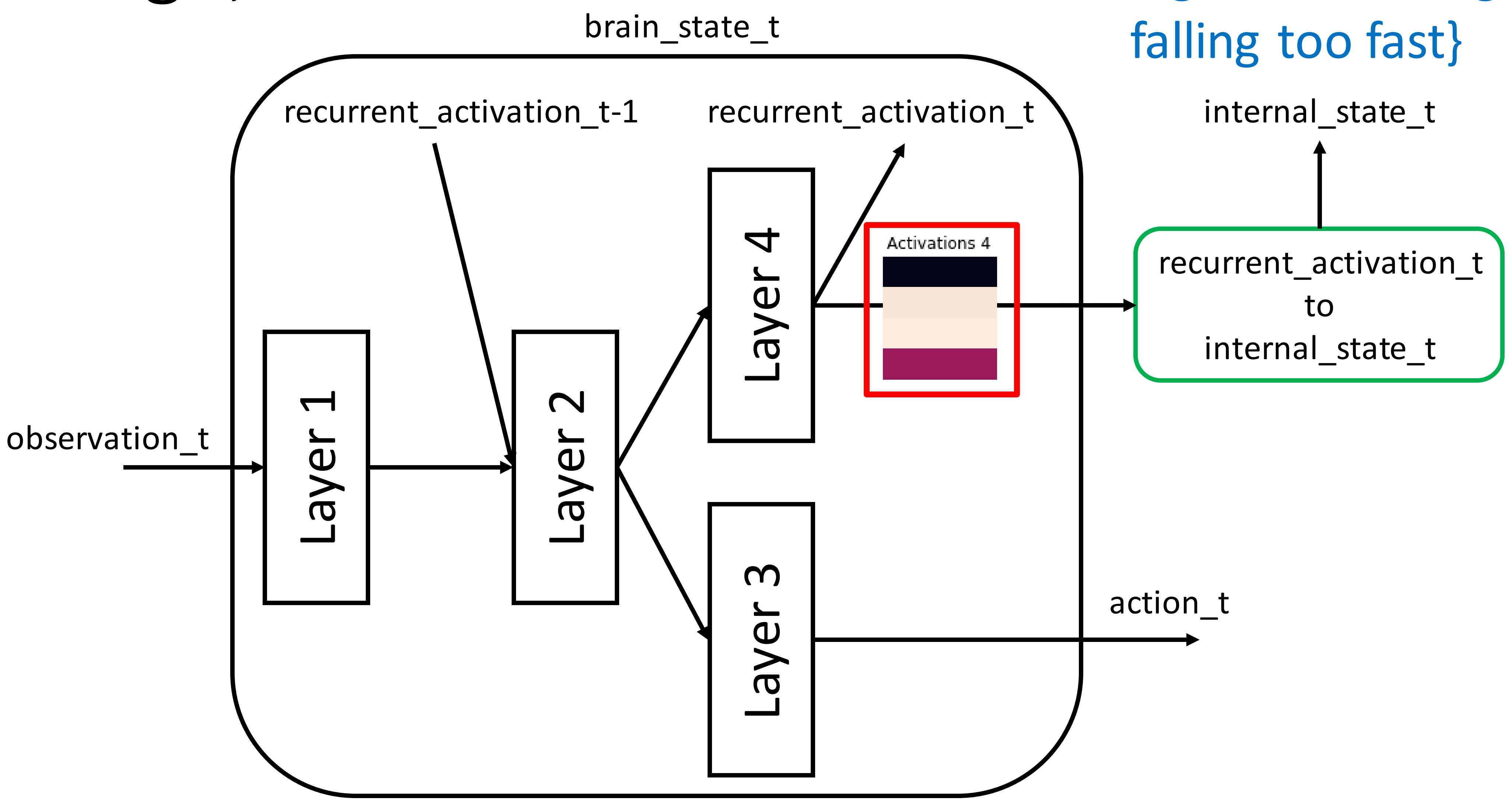


Design, V0

- Remaining questions
    - How will the agent learn to behave in the world?
        - Reinforcement learning
    - How will brain states be “connected” to internal states?
    - How will the agent learn to recognize the correspondence between its internal states and its position/velocity?

Design, V0

{Left of the flags,
high above the ground,
falling too fast}



brain_state_t

recurrent_activation_t-1

recurrent_activation_t

internal_state_t

observation_t

Layer 1

Layer 2

Layer 4

Layer 3

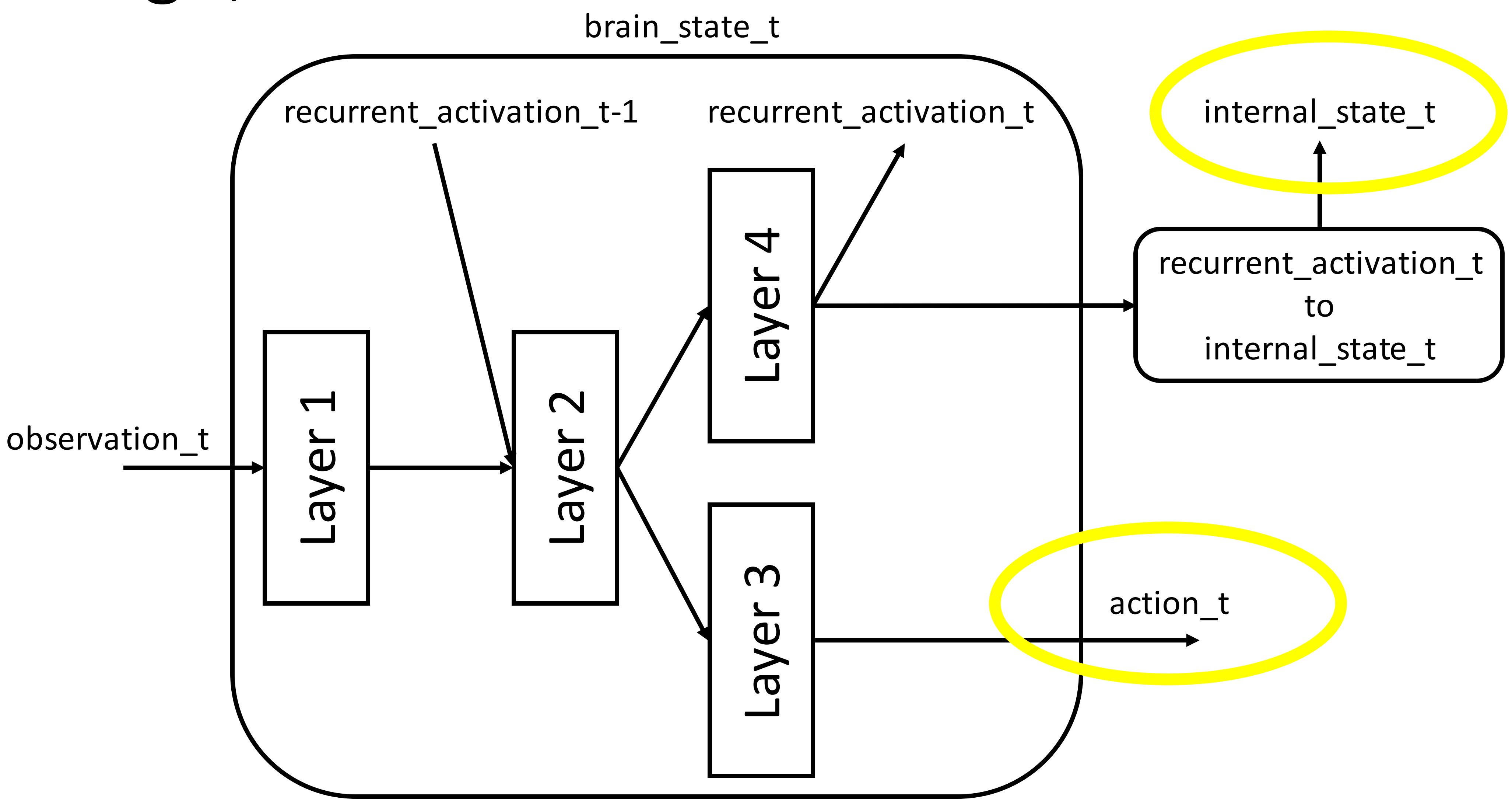
recurrent_activation_t
to
internal_state_t

action_t

Design, V0

- Remaining questions
    - How will the agent learn to behave in the world?
        - Reinforcement learning
    - How will brain states be “connected” to internal states?
        - A function that converts between classes (types)
    - How will the agent learn to recognize the correspondence between its internal states and its position/velocity?

Design, V0

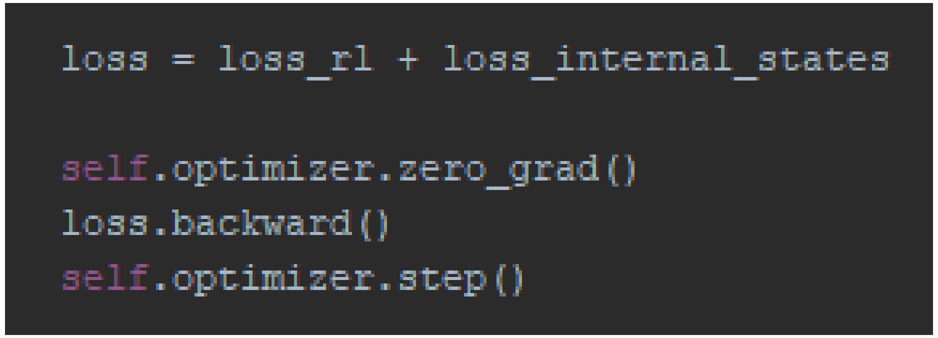

brain_state_t
recurrent_activation_t-1
recurrent_activation_t
internal_state_t
recurrent_activation_t
to
internal_state_t
observation_t
Layer 1
Layer 2
Layer 4
Layer 3
action_t

Design, V0

- Remaining questions
    - How will the agent learn to behave in the world?
        - Reinforcement learning
    - How will brain states be “connected” to internal states?
        - A function that converts between types
    - How will the agent learn to recognize the correspondence between its internal states and its position/velocity?
        - Jointly optimize both the RL loss to act and the internal state labeling loss

Design, V0

- Remaining questions
  - How will the agent learn to behave in the world?
    - Reinforcement learning
  - How will brain states be “connected” to internal states?
    - A function that converts between types
  - How will the agent learn to recognize the correspondence between its internal states and its position/velocity?
    - Jointly optimize both the RL loss to act and the internal state labeling loss



```
loss = loss_rl + loss_internal_states

self.optimizer.zero_grad()
loss.backward()
self.optimizer.step()
```

Quick review before moving to implementation

- Requirements, V0
  - Internal states are casually reducible to brain states
  - Internal states are ontologically irreducible to brain states
- Design, V0
  - Environment and the agent’s “physical” form
  - Internal state of the agent (set of semantically important regions)
  - Brain state of the agent (neural network structure and activations)
  - Our ontology
  - Jointly optimize both the RL loss to act and the internal state labeling loss
  - Simple function to map recurrent_activation_t to internal_state_t

Implementation, V0

- Jupyter notebook time!

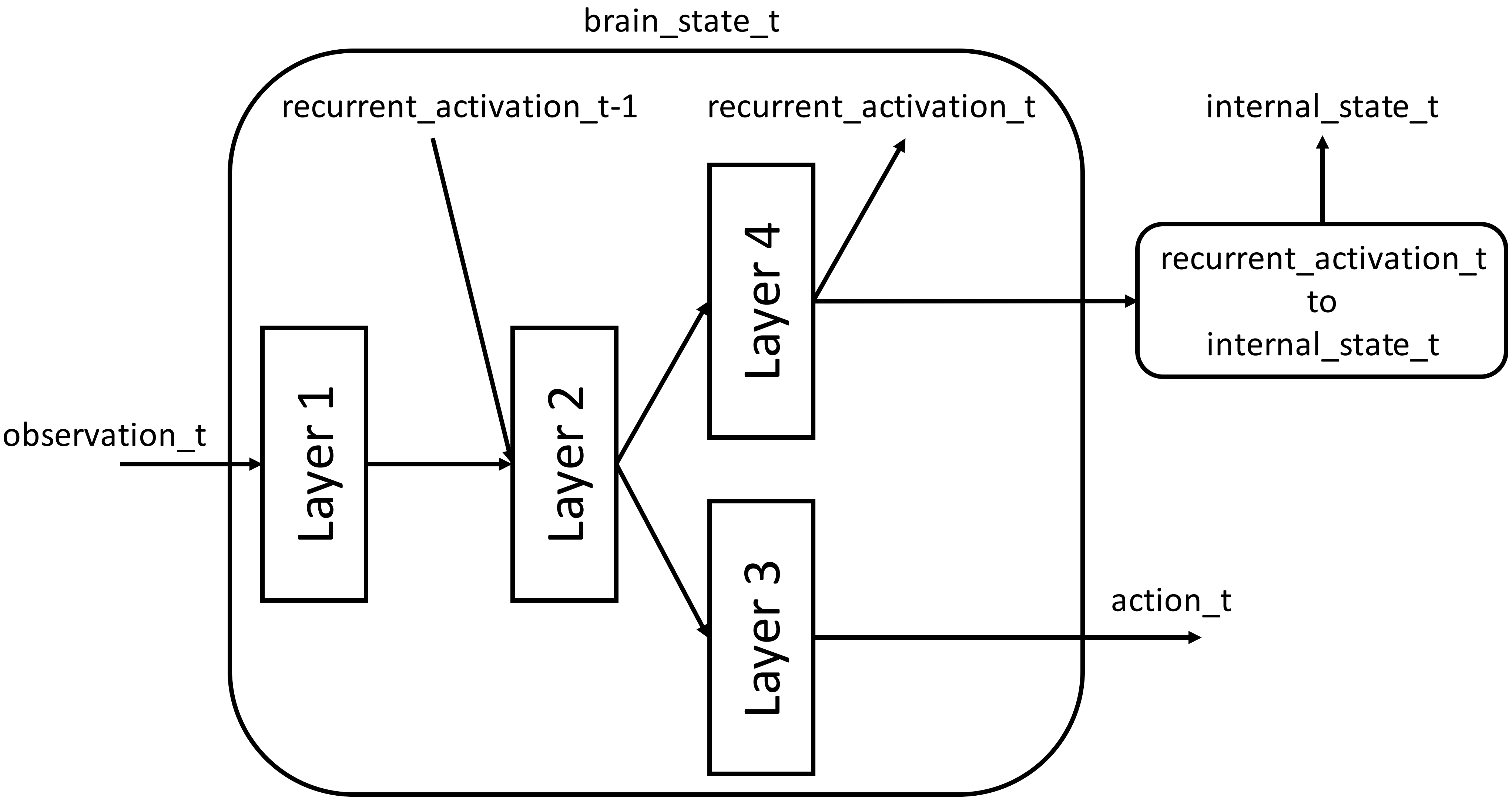
Conclusion

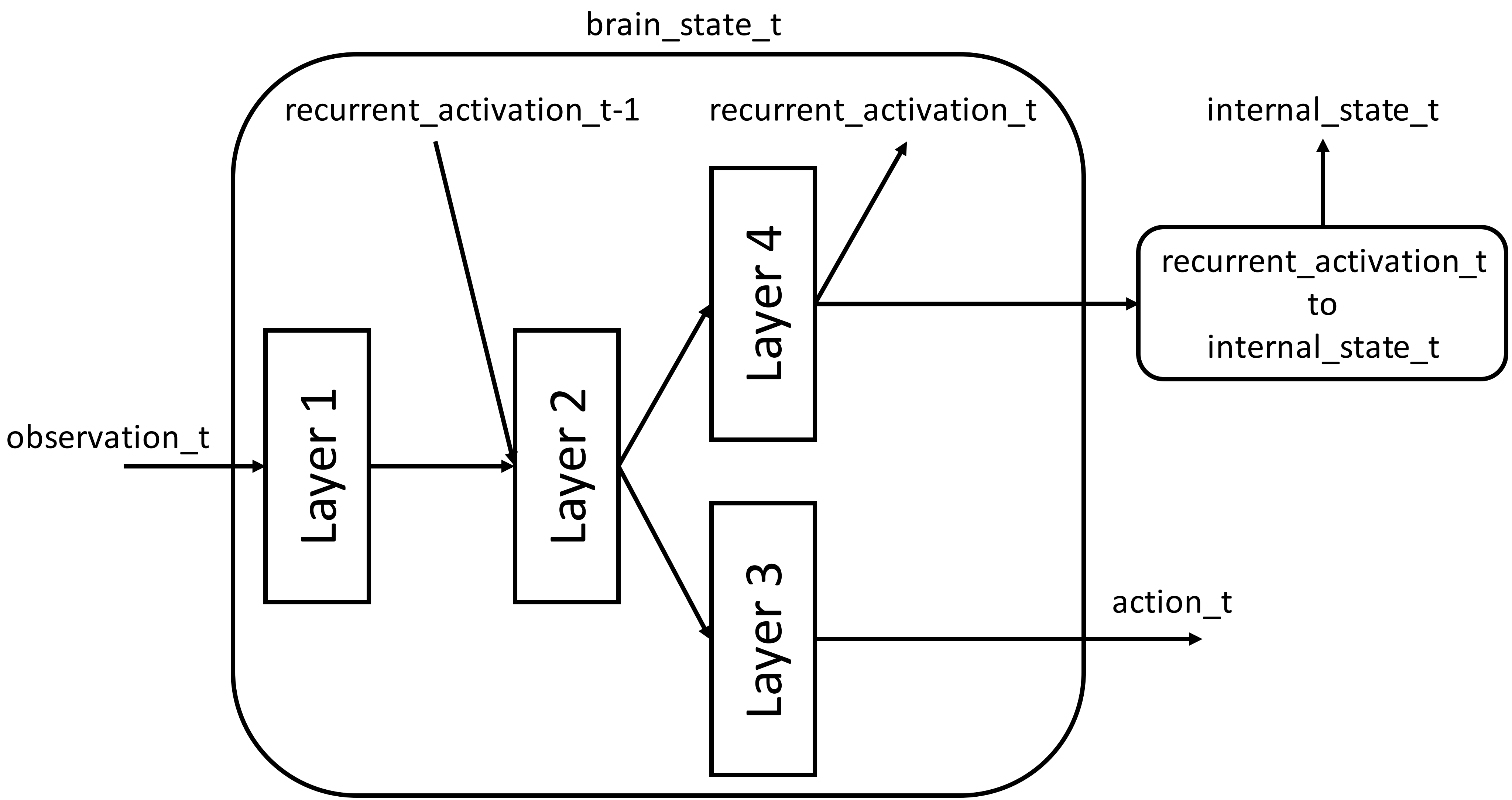
- Searle’s view
  - Consciousness is causally reducible to brain states
  - Consciousness is ontologically irreducible to brain states
- V2
  - Conscious mental states are casually reducible to brain states
  - Conscious mental states are ontologically irreducible to brain states
- V1
  - Mental states are casually reducible to brain states
  - Mental states are ontologically irreducible to brain states
- V0
  - Internal states are casually reducible to brain states
  - Internal states are ontologically irreducible to brain states

Conclusion

- Download and play with the code yourself
- github.com/Josh-Joseph/tsc-2019
- Disagree with our implementation of (a simplified version of) Searle's view?
    - Great! Open an issue and/or submit a pull request in GitHub
    - Concrete, constructive way of disagreeing
- Thoughts on other theories of mind/consciousness that may be particularly well suited for this type of approach?

Backup slides

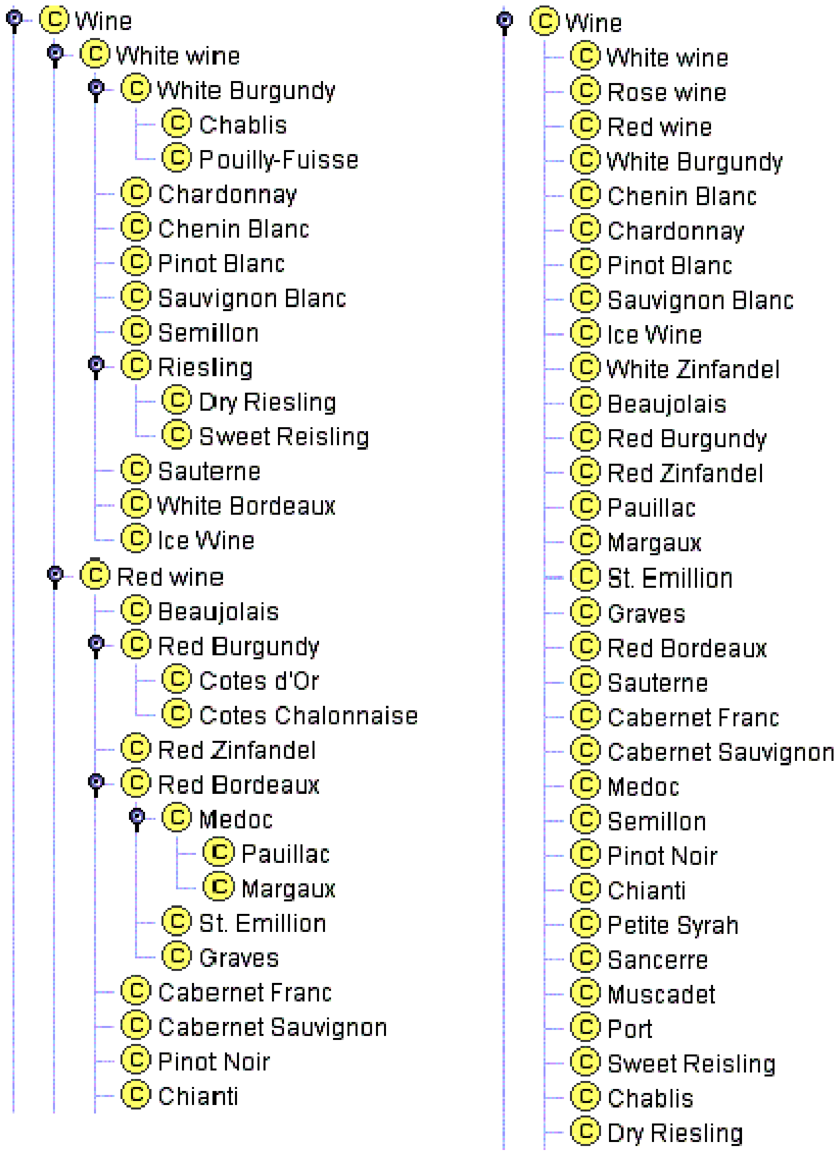

brain_state_t
recurrent_activation_t-1
recurrent_activation_t
internal_state_t
observation_t
Layer 1
Layer 2
Layer 4
Layer 3
recurrent_activation_t
to
internal_state_t
action_t



brain_state_t
recurrent_activation_t-1
recurrent_activation_t
internal_state_t
Layer 4
recurrent_activation_t
to
internal_state_t
observation_t
Layer 1
Layer 2
Layer 3
action_t

Background

- Reinforcement learning
- Neural networks
- Ontologies in computer science
  - "They drive the same car"
  - Type-token distinction
    - They drive the same car type (a Toyota)
    - They drive the same car token (the 2003 Toyota Carolla with VIN: 2QFBORHE4KP911561)



- Wine
  - White wine
    - White Burgundy
      - Chablis
      - Pouilly-Fuisse
    - Chardonnay
    - Chenin Blanc
    - Pinot Blanc
    - Sauvignon Blanc
    - Semillon
    - Riesling
      - Dry Riesling
      - Sweet Reisling
    - Sauterne
    - White Bordeaux
    - Ice Wine
  - Red wine
    - Beaujolais
    - Red Burgundy
      - Cotes d'Or
      - Cotes Chalonnaise
    - Red Zinfandel
    - Red Bordeaux
      - Medoc
        - Pauillac
        - Margaux
      - St. Emillion
      - Graves
    - Cabernet Franc
    - Cabernet Sauvignon
    - Pinot Noir
    - Chianti

- Wine
  - White wine
  - Rose wine
  - Red wine
  - White Burgundy
  - Chenin Blanc
  - Chardonnay
  - Pinot Blanc
  - Sauvignon Blanc
  - Ice Wine
  - White Zinfandel
  - Beaujolais
  - Red Burgundy
  - Red Zinfandel
  - Pauillac
  - Margaux
  - St. Emillion
  - Graves
  - Red Bordeaux
  - Sauterne
  - Cabernet Franc
  - Cabernet Sauvignon
  - Medoc
  - Semillon
  - Pinot Noir
  - Chianti
  - Petite Syrah
  - Sancerre
  - Muscadet
  - Port
  - Sweet Reisling
  - Chablis
  - Dry Riesling

From https://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html

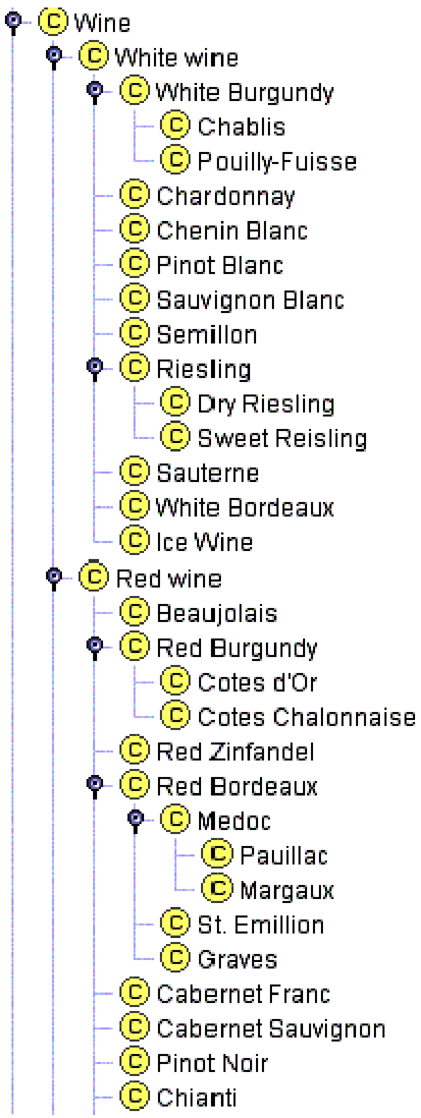
Agent requirements: unpacking Searle's view

- ~~Conscious~~ mental states are causally reducible to brain states but mental states are ontologically irreducible to brain states
- This has the feel of *maybe* being concrete enough that we can try to build an example of what we think Searle means

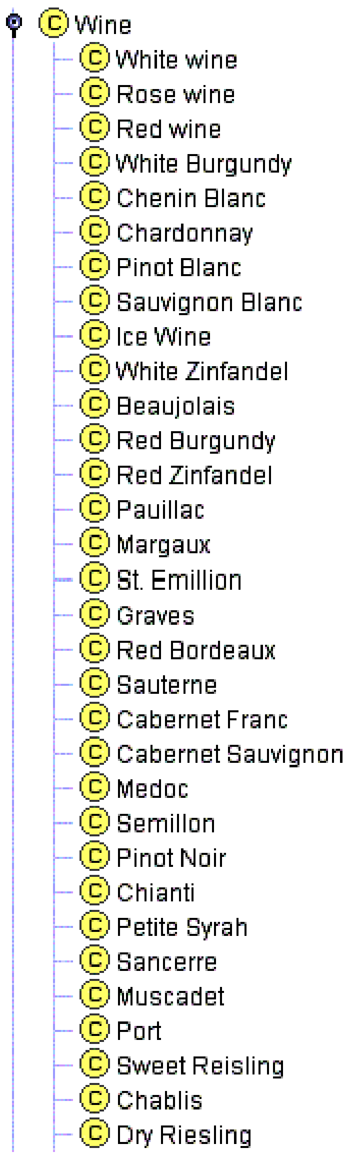
Unpacking Searle’s view

- ~~Conscious~~ mental states are causally reducible to brain states but mental states are ontologically irreducible to brain states
- This has the feel of *maybe* being concrete enough that we can try to build an example of what we think Searle means
- So let's build an agent who exhibits:
  - Mental states that are causally reducible to brain states
  - Mental states that are ontologically irreducible to brain states

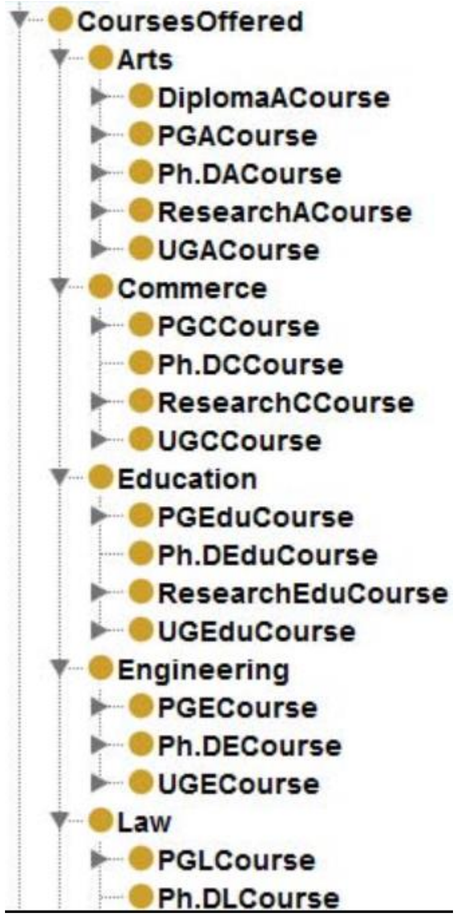
Ontolog m cience



- Wine
  - White wine
    - White Burgundy
      - Chablis
      - Pouilly-Fuisse
    - Chardonnay
    - Chenin Blanc
    - Pinot Blanc
    - Sauvignon Blanc
    - Semillon
    - Riesling
      - Dry Riesling
      - Sweet Reisling
    - Sauterne
    - White Bordeaux
    - Ice Wine
  - Red wine
    - Beaujolais
    - Red Burgundy
      - Cotes d'Or
      - Cotes Chalonnaise
    - Red Zinfandel
    - Red Bordeaux
      - Medoc
        - Pauillac
        - Margaux
      - St. Emillion
      - Graves
    - Cabernet Franc
    - Cabernet Sauvignon
    - Pinot Noir
    - Chianti



- Wine
  - White wine
  - Rose wine
  - Red wine
  - White Burgundy
  - Chenin Blanc
  - Chardonnay
  - Pinot Blanc
  - Sauvignon Blanc
  - Ice Wine
  - White Zinfandel
  - Beaujolais
  - Red Burgundy
  - Red Zinfandel
  - Pauillac
  - Margaux
  - St. Emillion
  - Graves
  - Red Bordeaux
  - Sauterne
  - Cabernet Franc
  - Cabernet Sauvignon
  - Medoc
  - Semillon
  - Pinot Noir
  - Chianti
  - Petite Syrah
  - Sancerre
  - Muscadet
  - Port
  - Sweet Reisling
  - Chablis
  - Dry Riesling

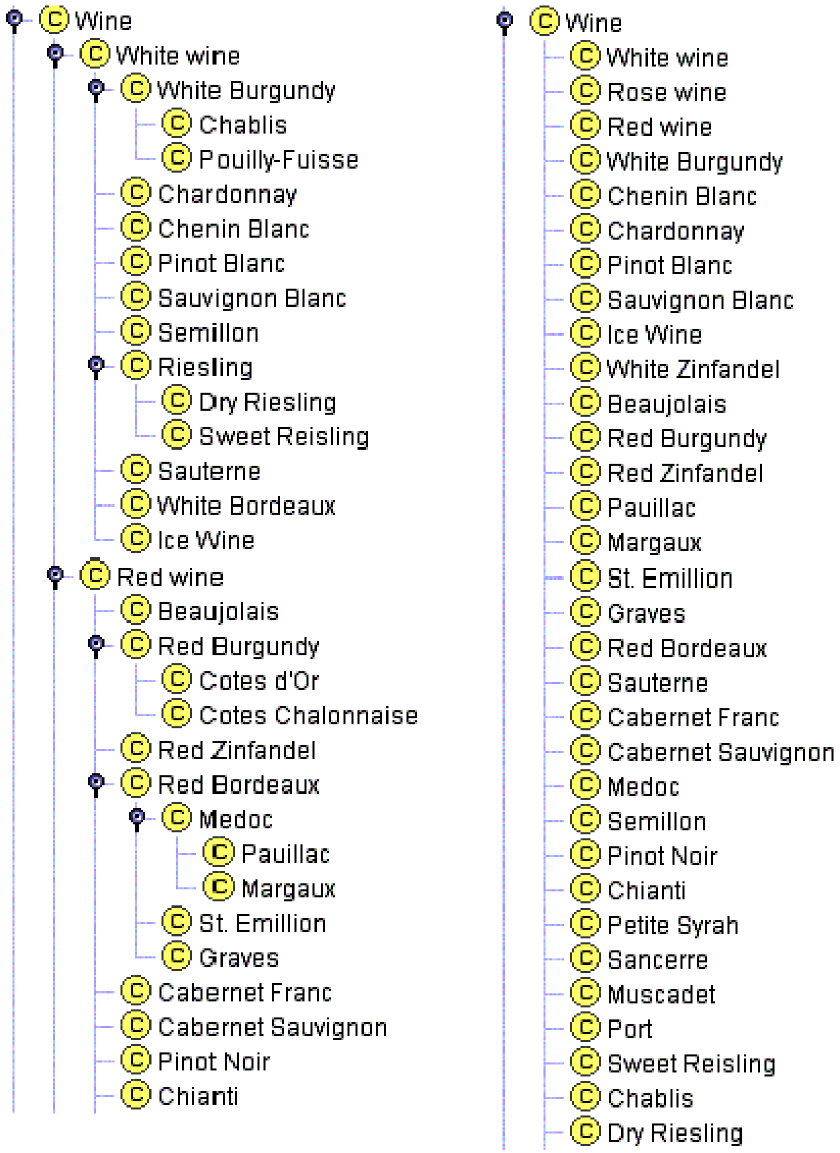


- CoursesOffered
  - Arts
    - DiplomaACourse
    - PGACourse
    - Ph.DACourse
    - ResearchACourse
    - UGACourse
  - Commerce
    - PGCCourse
    - Ph.DCCourse
    - ResearchCCourse
    - UGCCourse
  - Education
    - PGEduCourse
    - Ph.DEduCourse
    - ResearchEduCourse
    - UGEduCourse
  - Engineering
    - PGECourse
    - Ph.DECourse
    - UGECourse
  - Law
    - PGLCourse
    - Ph.DLCourse

Images from:
https://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html
https://www.researchgate.net/figure/Owl-Viz-view-of-course-ontology_fig1_261339041

Background

- Reinforcement learning
- Neural networks
- Ontologies in computer science



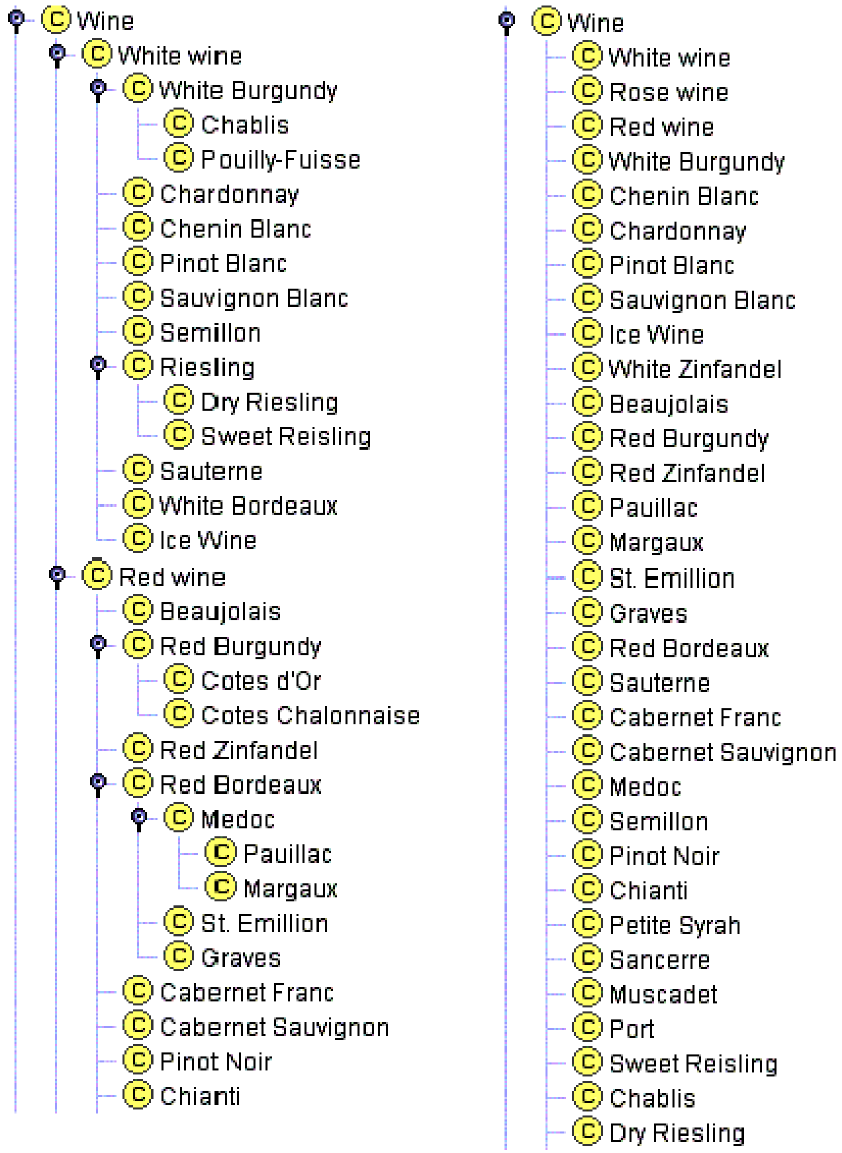
- Wine
  - White wine
    - White Burgundy
      - Chablis
      - Pouilly-Fuisse
    - Chardonnay
    - Chenin Blanc
    - Pinot Blanc
    - Sauvignon Blanc
    - Semillon
    - Riesling
      - Dry Riesling
      - Sweet Reisling
    - Sauterne
    - White Bordeaux
    - Ice Wine
  - Red wine
    - Beaujolais
    - Red Burgundy
      - Cotes d'Or
      - Cotes Chalonnaise
    - Red Zinfandel
    - Red Bordeaux
      - Medoc
        - Pauillac
        - Margaux
      - St. Emillion
      - Graves
    - Cabernet Franc
    - Cabernet Sauvignon
    - Pinot Noir
    - Chianti

- Wine
  - White wine
  - Rose wine
  - Red wine
  - White Burgundy
  - Chenin Blanc
  - Chardonnay
  - Pinot Blanc
  - Sauvignon Blanc
  - Ice Wine
  - White Zinfandel
  - Beaujolais
  - Red Burgundy
  - Red Zinfandel
  - Pauillac
  - Margaux
  - St. Emillion
  - Graves
  - Red Bordeaux
  - Sauterne
  - Cabernet Franc
  - Cabernet Sauvignon
  - Medoc
  - Semillon
  - Pinot Noir
  - Chianti
  - Petite Syrah
  - Sancerre
  - Muscadet
  - Port
  - Sweet Reisling
  - Chablis
  - Dry Riesling

From https://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html

Background

- Reinforcement learning
- Neural networks
- Ontologies in computer science
  - "They drive the same car"
  - Type-token distinction
    - They drive the same car type (a Toyota)
    - They drive the same car token (the 2003 Toyota Carolla with VIN: 2QFBORHE4KP911561)



- Wine
  - White wine
    - White Burgundy
      - Chablis
      - Pouilly-Fuisse
    - Chardonnay
    - Chenin Blanc
    - Pinot Blanc
    - Sauvignon Blanc
    - Semillon
    - Riesling
      - Dry Riesling
      - Sweet Reisling
    - Sauterne
    - White Bordeaux
    - Ice Wine
  - Red wine
    - Beaujolais
    - Red Burgundy
      - Cotes d'Or
      - Cotes Chalonnaise
    - Red Zinfandel
    - Red Bordeaux
      - Medoc
        - Pauillac
        - Margaux
      - St. Emillion
      - Graves
    - Cabernet Franc
    - Cabernet Sauvignon
    - Pinot Noir
    - Chianti

- Wine
  - White wine
  - Rose wine
  - Red wine
  - White Burgundy
  - Chenin Blanc
  - Chardonnay
  - Pinot Blanc
  - Sauvignon Blanc
  - Ice Wine
  - White Zinfandel
  - Beaujolais
  - Red Burgundy
  - Red Zinfandel
  - Pauillac
  - Margaux
  - St. Emillion
  - Graves
  - Red Bordeaux
  - Sauterne
  - Cabernet Franc
  - Cabernet Sauvignon
  - Medoc
  - Semillon
  - Pinot Noir
  - Chianti
  - Petite Syrah
  - Sancerre
  - Muscadet
  - Port
  - Sweet Reisling
  - Chablis
  - Dry Riesling

From https://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html

Unpacking Searle’s view

- ~~Conscious~~ mental states are causally reducible to brain states but mental states are ontologically irreducible to brain states
- This has the feel of *maybe* being concrete enough that we can try to build an example of what we think Searle means
- So let's build an agent who exhibits:
  - Mental states that are causally reducible to brain states
  - Mental states that are ontologically irreducible to brain states

Unpacking Searle’s view

- ~~Conscious~~ mental states are causally reducible to brain states but mental states are ontologically irreducible to brain states
- This has the feel of *maybe* being concrete enough that we can try to build an example of what we think Searle means
- So let's build an agent who exhibits:
  - Mental states that are causally reducible to brain states
  - Mental states that are ontologically irreducible to brain states

Phenomena of type A are causally reducible to phenomena of type B if and only if:

- the behavior of A’s are entirely casually explained by the behavior of B’s
- A’s have no causal powers in addition to the powers of B’s

Unpacking Searle’s view

- ~~Conscious~~ mental states are causally reducible to brain states but mental states are ontologically irreducible to brain states
- This has the feel of *maybe* being concrete enough that we can try to build an example of what we think Searle means
- So let's build an agent who exhibits:
    - Mental states that are causally reducible to brain states
    - Mental states that are ontologically irreducible to brain states

Phenomena of type A are ontologically reducible to phenomena of type B if and only if A’s are nothing but B’s

Our objectives

- Unpack a somewhat confusing theory of consciousness by creating a software agent that is consistent with the theory
- Gain a deeper understanding of what the theory means by brain states, by mental states, and the relationship between them through examining the agent

Unpacking Searle’s view

- Conscious mental states are causally reducible to brain states but mental states are ontologically irreducible to brain states

Unpacking Searle’s view

- ~~Conscious~~ mental states are causally reducible to brain states but mental states are ontologically irreducible to brain states
- This has the feel of *maybe* being concrete enough that we can try to build an example of what we think Searle means

Unpacking Searle's view

- ~~Conscious~~ mental states are causally reducible to brain states but mental states are ontologically irreducible to brain states
- This has the feel of *maybe* being concrete enough that we can try to build an example of what we think Searle means
- So let's build an agent who exhibits:
  - Mental states that are causally reducible to brain states
  - Mental states that are ontologically irreducible to brain states

Phenomena of type A are causally reducible to phenomena of type B if and only if:

- the behavior of A's are entirely casually explained by the behavior of B's
- A's have no causal powers in addition to the powers of B's

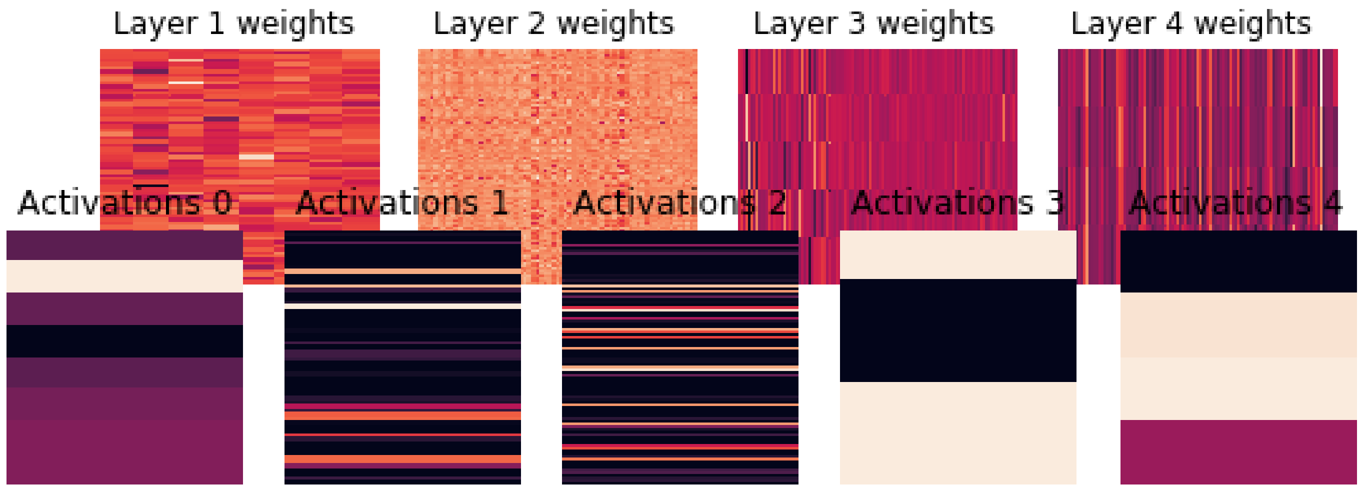
Unpacking Searle’s view

- ~~Conscious~~ mental states are causally reducible to brain states but mental states are ontologically irreducible to brain states
- This has the feel of *maybe* being concrete enough that we can try to build an example of what we think Searle means
- So let's build an agent who exhibits:
  - Mental states that are causally reducible to brain states
  - Mental states that are ontologically irreducible to brain states

Phenomena of type A are ontologically reducible to phenomena of type B if and only if A’s are nothing but B’s

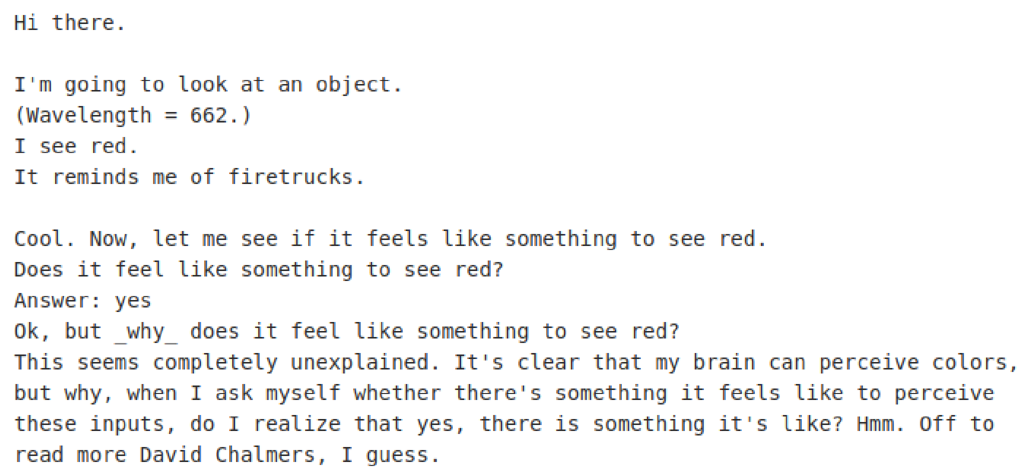
- Standard RL benchmark environment
- A neural network based agent that learns to act in the environment using RL
- A third person ontology (which contains a brain state type)
- A first person ontology of the agent (which contains a mental state type)
- A function that maps brain state types to mental state types
- Simple argument that these objects are consistent with Searle’s claim

Notebook time!

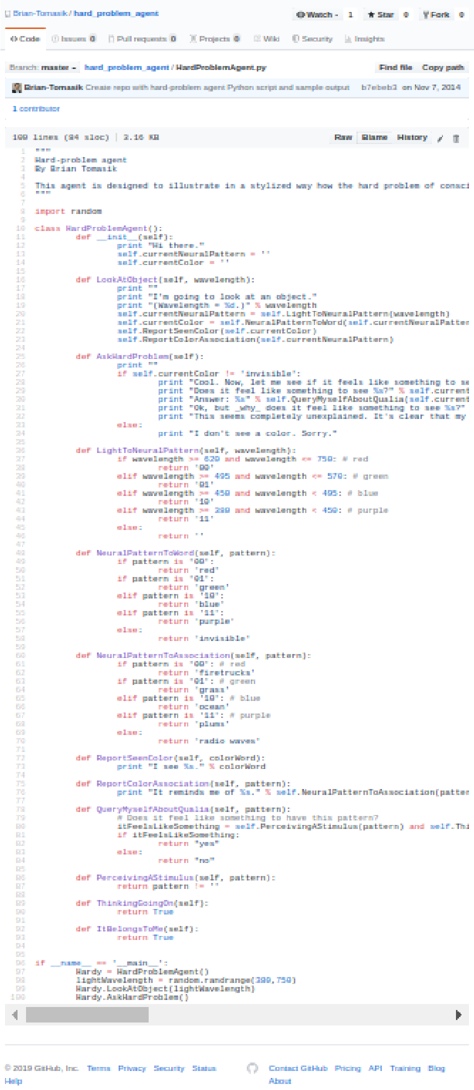


Reifying philosophy with code

- Tomasik: A Simple Program to Illustrate the Hard Problem of Consciousness
- The hard problem is confusing because our brains create a thought that there's something it's like to be us



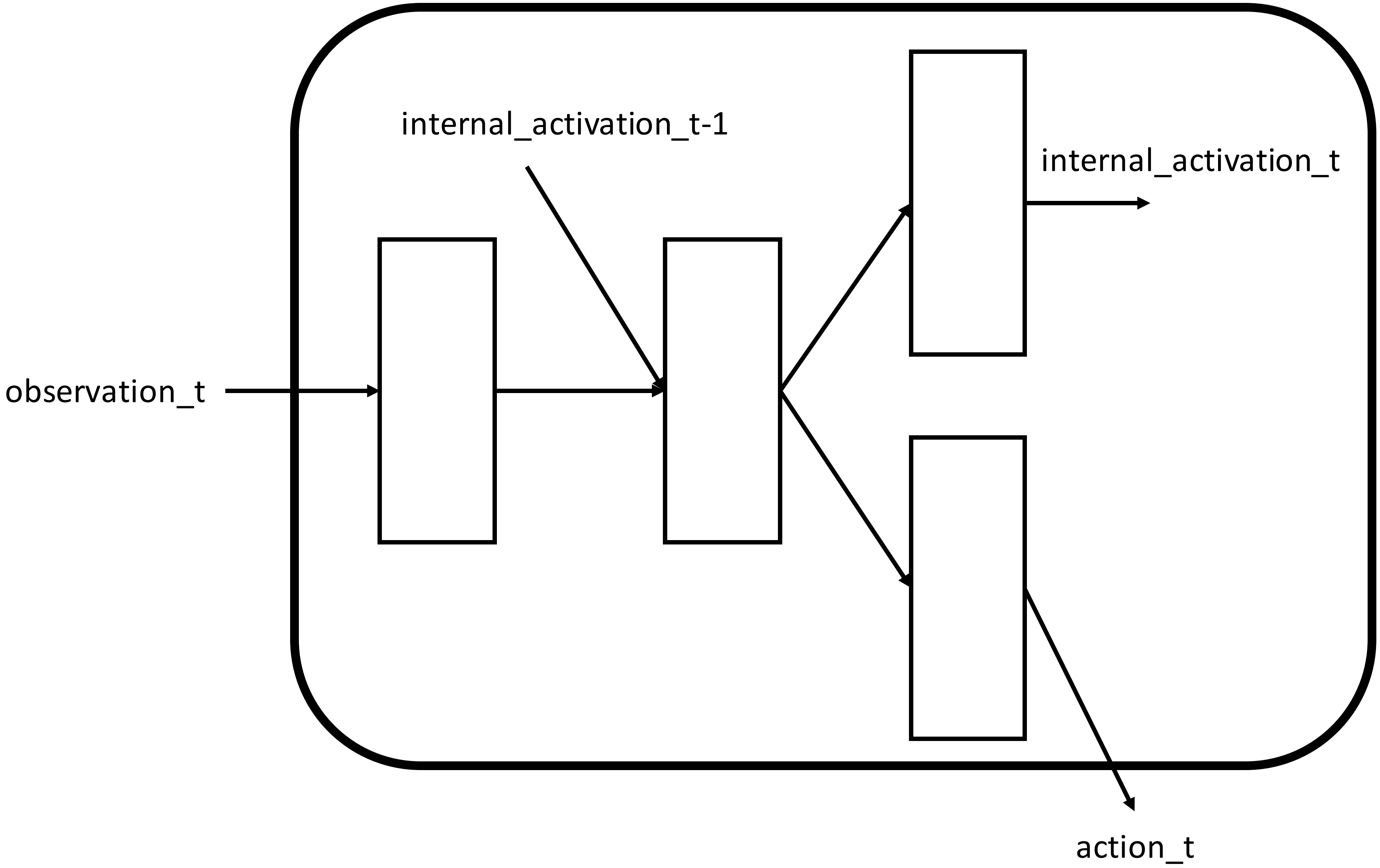
```
Hi there.

I'm going to look at an object.
(Wavelength = 662.)
I see red.
It reminds me of firetrucks.

Cool. Now, let me see if it feels like something to see red.
Does it feel like something to see red?
Answer: yes
Ok, but _why_ does it feel like something to see red?
This seems completely unexplained. It's clear that my brain can perceive colors,
but why, when I ask myself whether there's something it feels like to perceive
these inputs, do I realize that yes, there is something it's like? Hmm. Off to
read more David Chalmers, I guess.
```

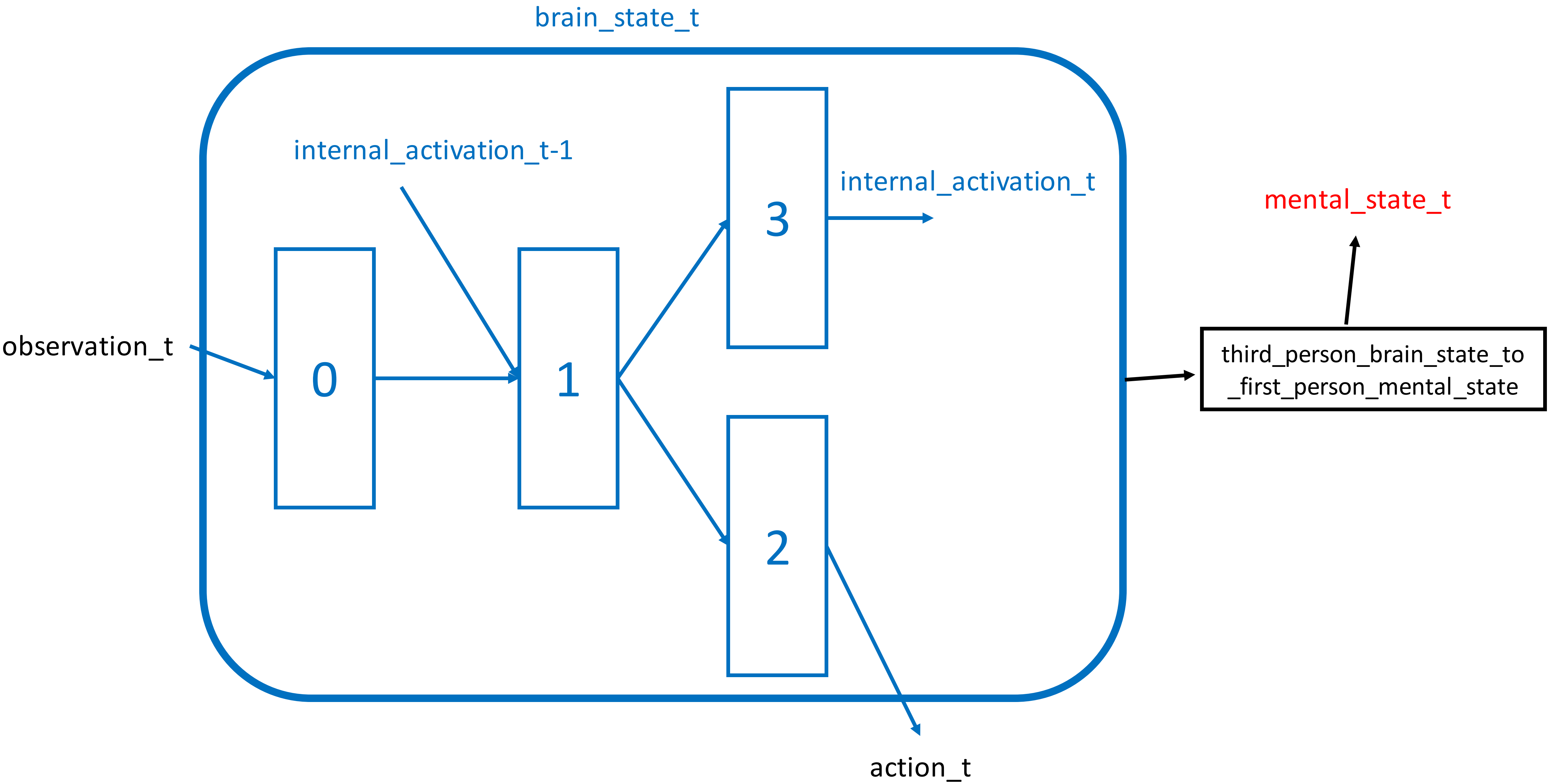

Brian-Tomasik / hard_problem_agent
Watch 1
Star 0
Fork 0
Code
Issues 0
Pull requests 0
Projects 0
Wiki
Security
Insights
Branch: master
hard_problem_agent / HardProblemAgent.py
Find file
Copy path
Brian-Tomasik Create repo with hard-problem agent Python script and sample output
b7ebeb3 on Nov 7, 2014
1 contributor
100 lines (94 sloc) | 2.16 KB
Raw
Blame
History

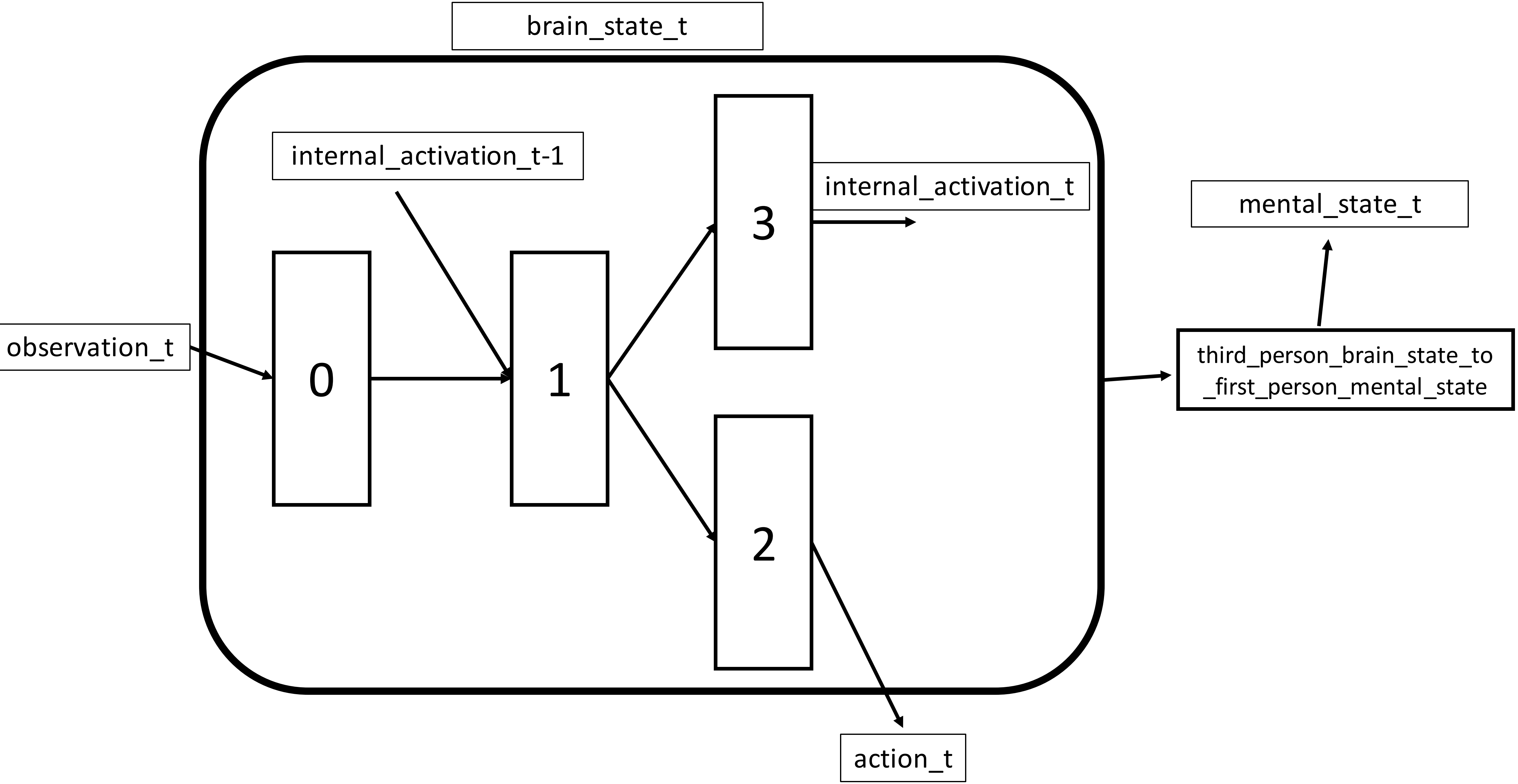
```
"""
Hard-problem agent
By Brian Tomasik

This agent is designed to illustrate in a stylized way how the hard problem of consci
"""

import random

class HardProblemAgent():
        def __init__(self):
                print "Hi there."
                self.currentNeuralPattern = ''
                self.currentColor = ''

        def LookAtObject(self, wavelength):
                print ""
                print "I'm going to look at an object."
                print "(Wavelength = %d.)" % wavelength
                self.currentNeuralPattern = self.LightToNeuralPattern(wavelength)
                self.currentColor = self.NeuralPatternToWord(self.currentNeuralPatter
                self.ReportSeenColor(self.currentColor)
                self.ReportColorAssociation(self.currentNeuralPattern)

        def AskHardProblem(self):
                print ""
                if self.currentColor != 'invisible':
                        print "Cool. Now, let me see if it feels like something to se
                        print "Does it feel like something to see %s?" % self.current
                        print "Answer: %s" % self.QueryMyselfAboutQualia(self.current
                        print "Ok, but _why_ does it feel like something to see %s?"
                        print "This seems completely unexplained. It's clear that my
                else:
                        print "I don't see a color. Sorry."

        def LightToNeuralPattern(self, wavelength):
                if wavelength >= 620 and wavelength <= 750: # red
                        return '00'
                elif wavelength >= 495 and wavelength <= 570: # green
                        return '01'
                elif wavelength >= 450 and wavelength < 495: # blue
                        return '10'
                elif wavelength >= 380 and wavelength < 450: # purple
                        return '11'
                else:
                        return ''

        def NeuralPatternToWord(self, pattern):
                if pattern is '00':
                        return 'red'
                if pattern is '01':
                        return 'green'
                elif pattern is '10':
                        return 'blue'
                elif pattern is '11':
                        return 'purple'
                else:
                        return 'invisible'

        def NeuralPatternToAssociation(self, pattern):
                if pattern is '00': # red
                        return 'firetrucks'
                if pattern is '01': # green
                        return 'grass'
                elif pattern is '10': # blue
                        return 'ocean'
                elif pattern is '11': # purple
                        return 'plums'
                else:
                        return 'radio waves'

        def ReportSeenColor(self, colorWord):
                print "I see %s." % colorWord

        def ReportColorAssociation(self, pattern):
                print "It reminds me of %s." % self.NeuralPatternToAssociation(patter

        def QueryMyselfAboutQualia(self, pattern):
                # Does it feel like something to have this pattern?
                itFeelsLikeSomething = self.PerceivingAStimulus(pattern) and self.Thi
                if itFeelsLikeSomething:
                        return "yes"
                else:
                        return "no"

        def PerceivingAStimulus(self, pattern):
                return pattern != ''

        def ThinkingGoingOn(self):
                return True

        def ItBelongsToMe(self):
                return True


if __name__ == '__main__':
        Hardy = HardProblemAgent()
        lightWavelength = random.randrange(380,750)
        Hardy.LookAtObject(lightWavelength)
        Hardy.AskHardProblem()
```

from https://reducing-suffering.org/simple-program-illustrate-hard-problem-consciousness/

from https://github.com/Brian-Tomasik/hard_problem_agent/blob/master/HardProblemAgent.py


internal_activation_t-1
internal_activation_t
observation_t
action_t


brain_state_t
internal_activation_t-1
3
internal_activation_t
mental_state_t
observation_t
0
1
third_person_brain_state_to
_first_person_mental_state
2
action_t



brain_state_t

internal_activation_t-1

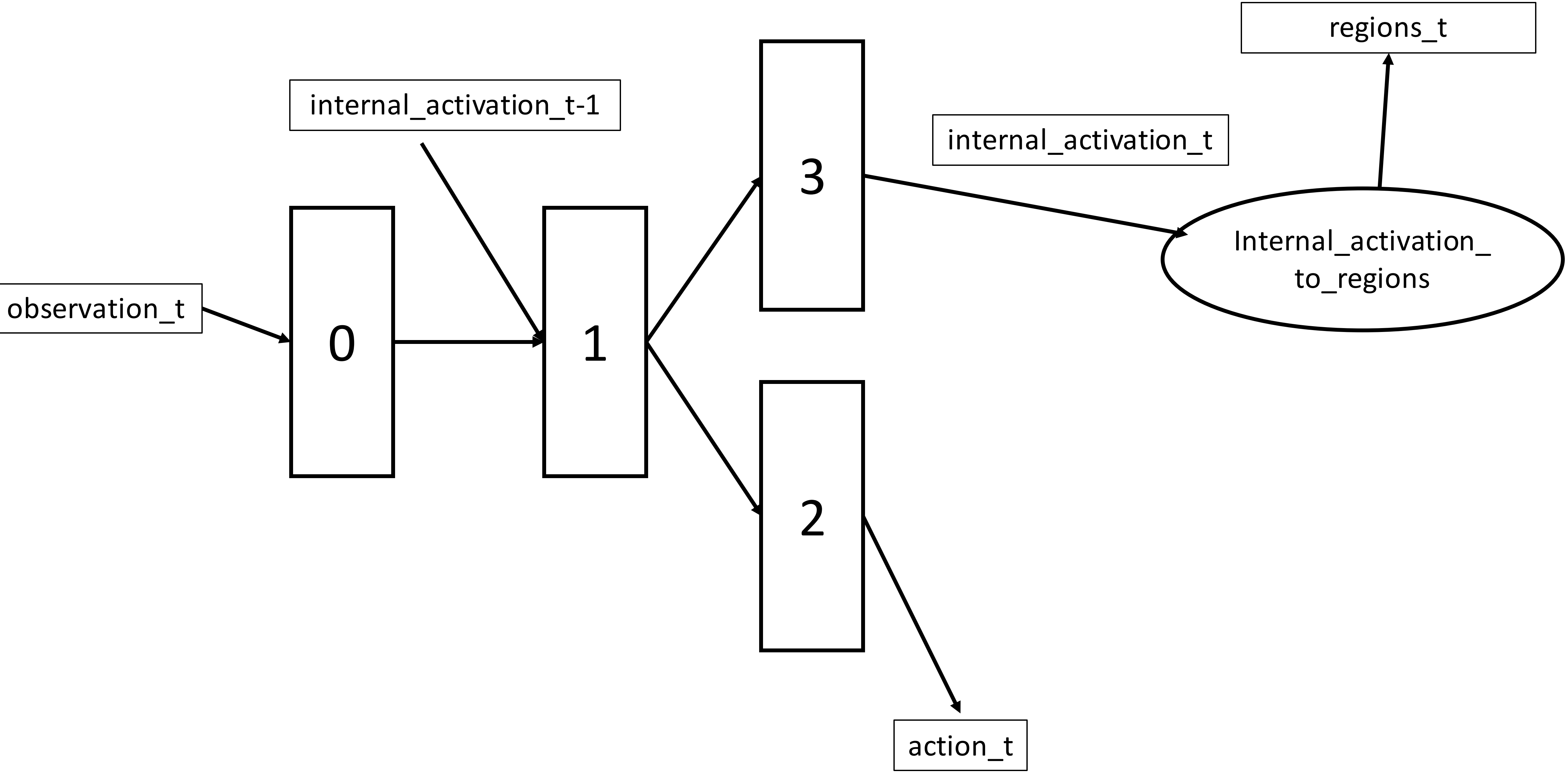
internal_activation_t

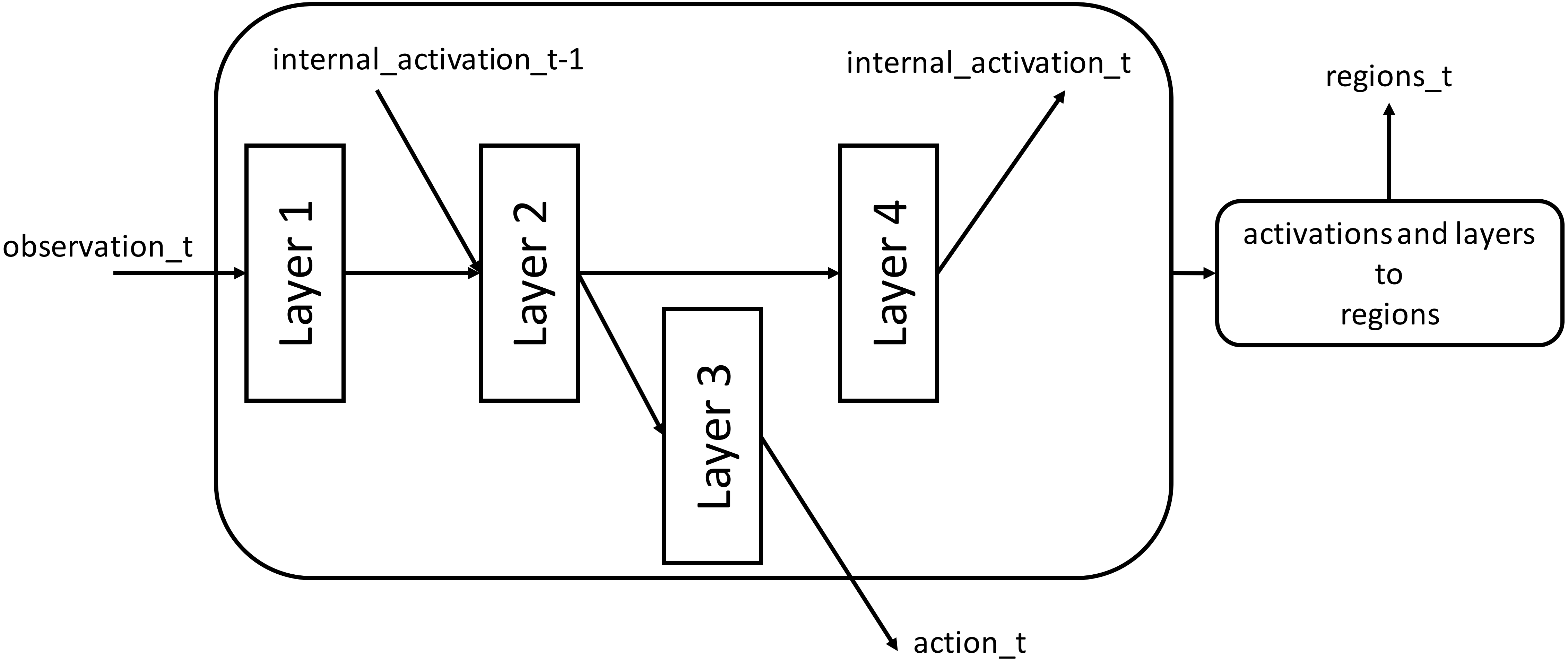
mental_state_t

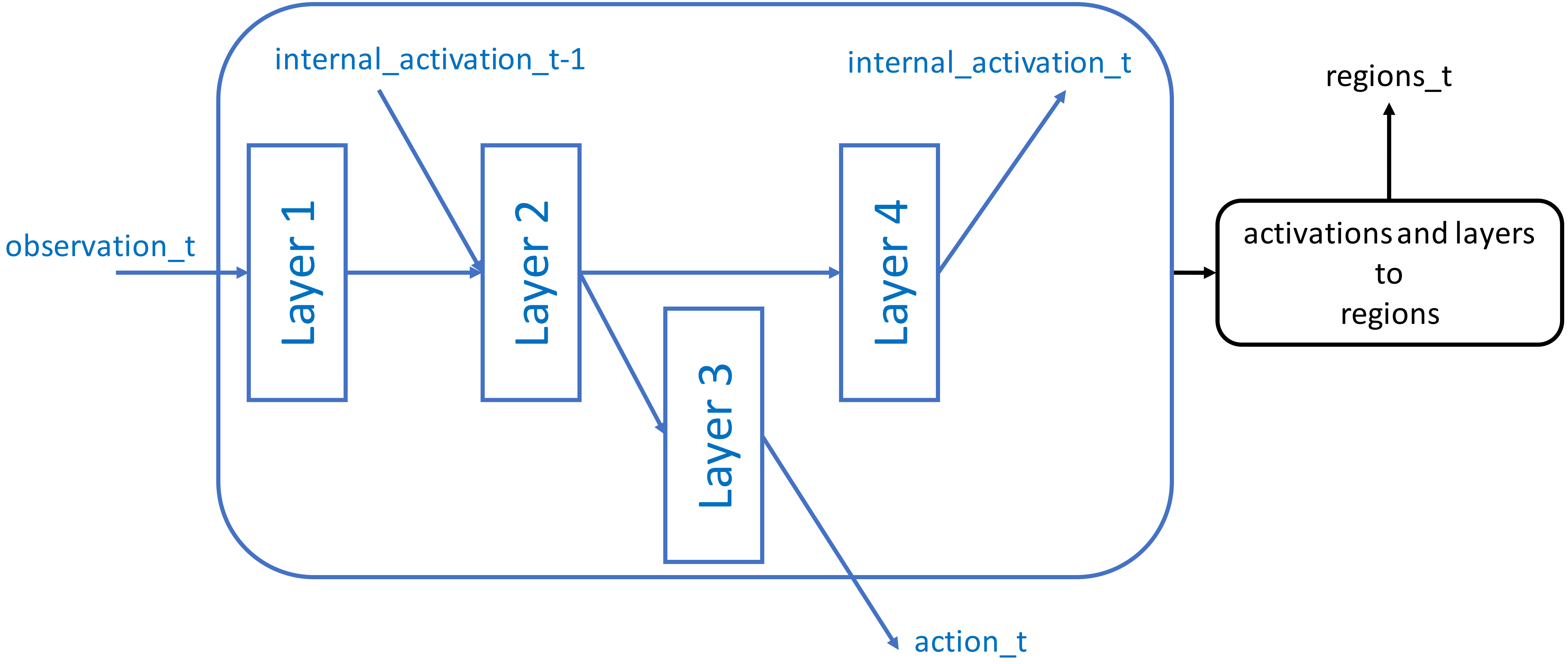
observation_t

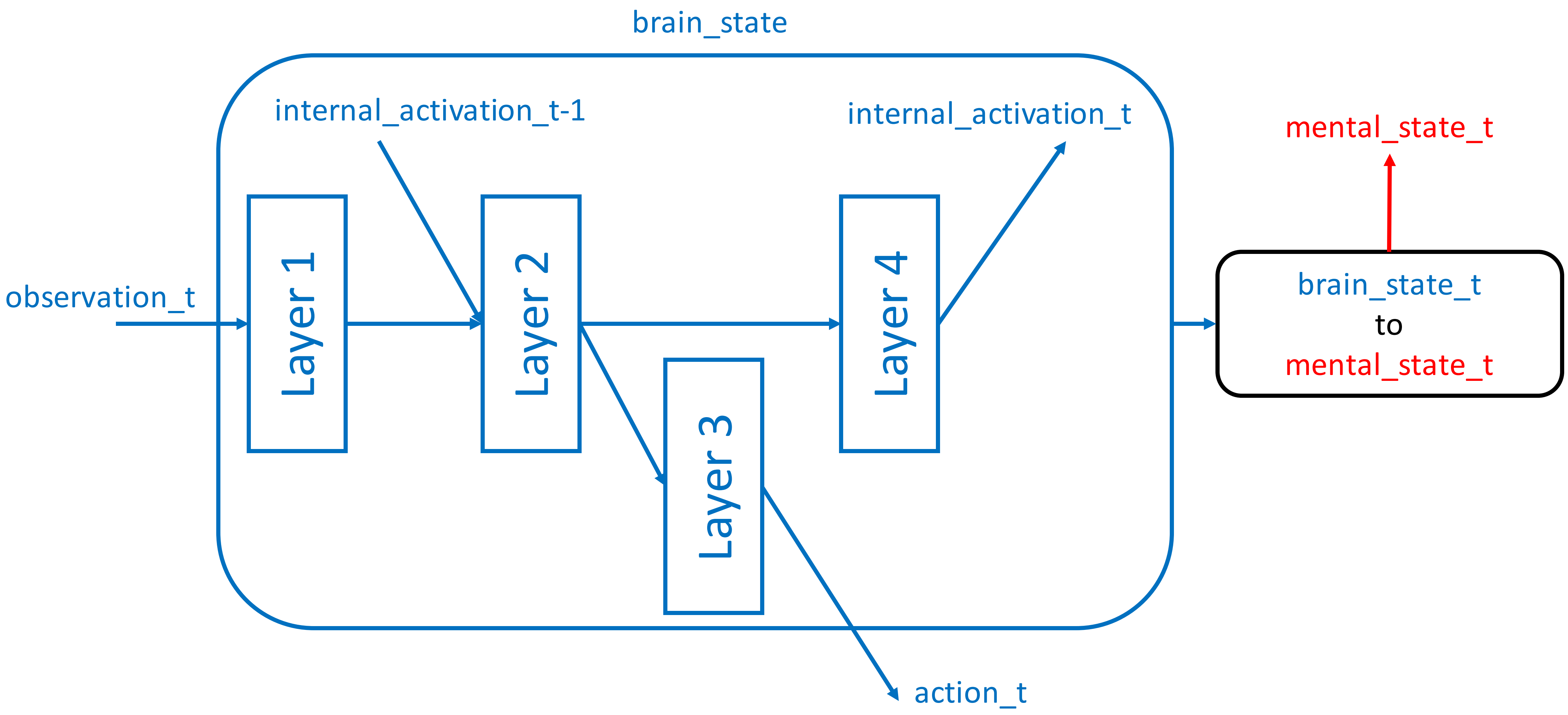
third_person_brain_state_to
_first_person_mental_state

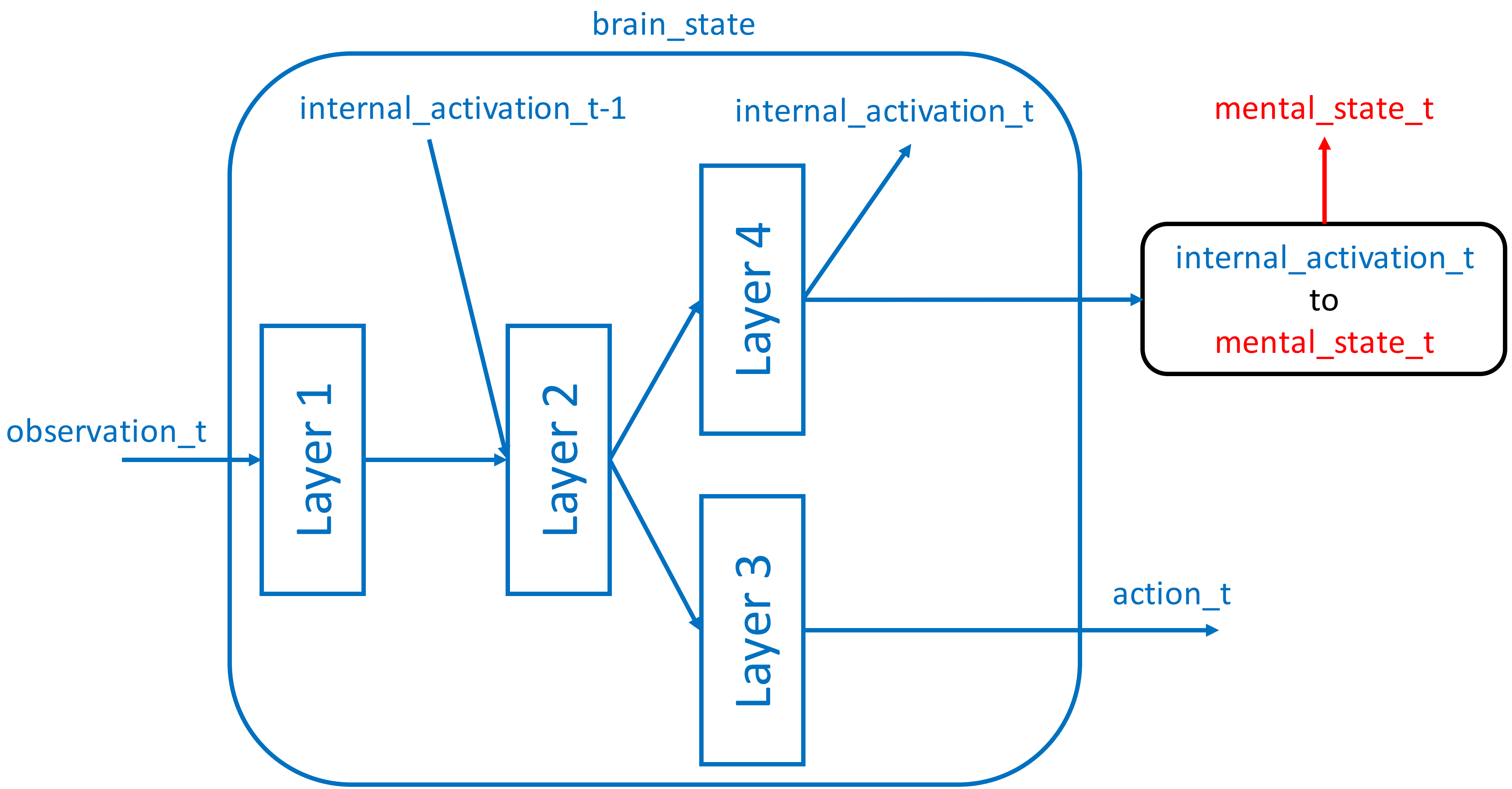
action_t

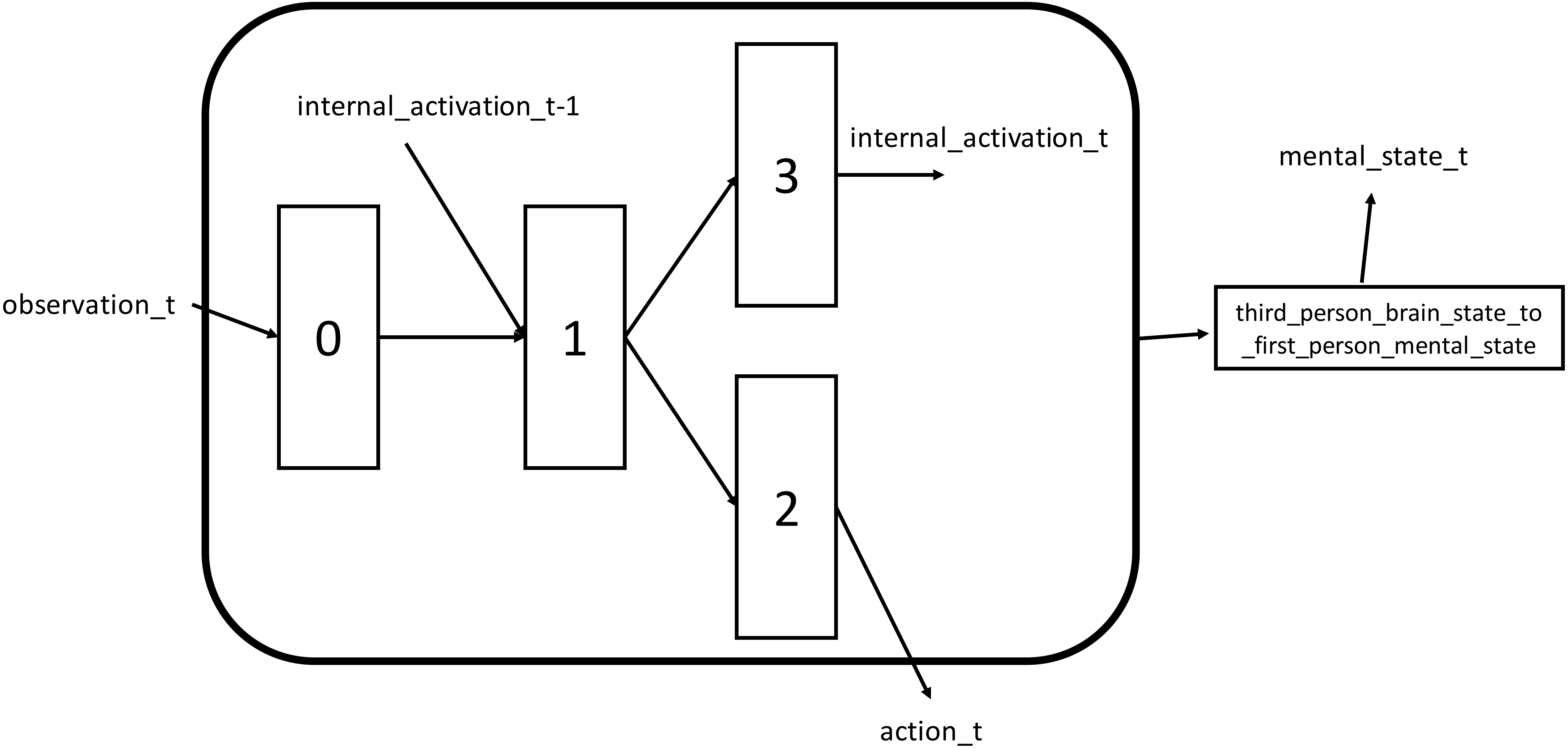

regions_t
internal_activation_t-1
internal_activation_t
3
Internal_activation_
to_regions
observation_t
0
1
2
action_t


internal_activation_t-1
internal_activation_t
regions_t
observation_t
Layer 1
Layer 2
Layer 4
Layer 3
activations and layers
to
regions
action_t


internal_activation_t-1
internal_activation_t
regions_t
observation_t
Layer 1
Layer 2
Layer 3
Layer 4
activations and layers
to
regions
action_t


brain_state
internal_activation_t-1
internal_activation_t
mental_state_t
observation_t
Layer 1
Layer 2
Layer 3
Layer 4
brain_state_t
to
mental_state_t
action_t


brain_state
internal_activation_t-1
internal_activation_t
mental_state_t
Layer 4
internal_activation_t
to
mental_state_t
observation_t
Layer 1
Layer 2
Layer 3
action_t


internal_activation_t-1
internal_activation_t
mental_state_t
observation_t
0
1
3
2
third_person_brain_state_to
_first_person_mental_state
action_t