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

Josh Joseph, Dhaval Adjodah, Joichi Ito

Massachusetts Institute of Technology

jmjoseph@mit.edu

mit media lab

MIT Quest for Intelligence

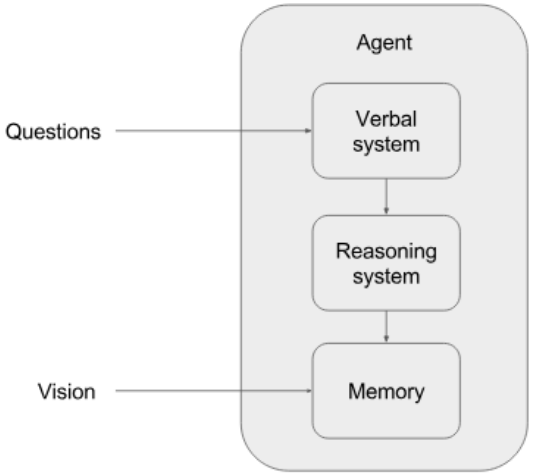
Why attempt to reify philosophy with code

- 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

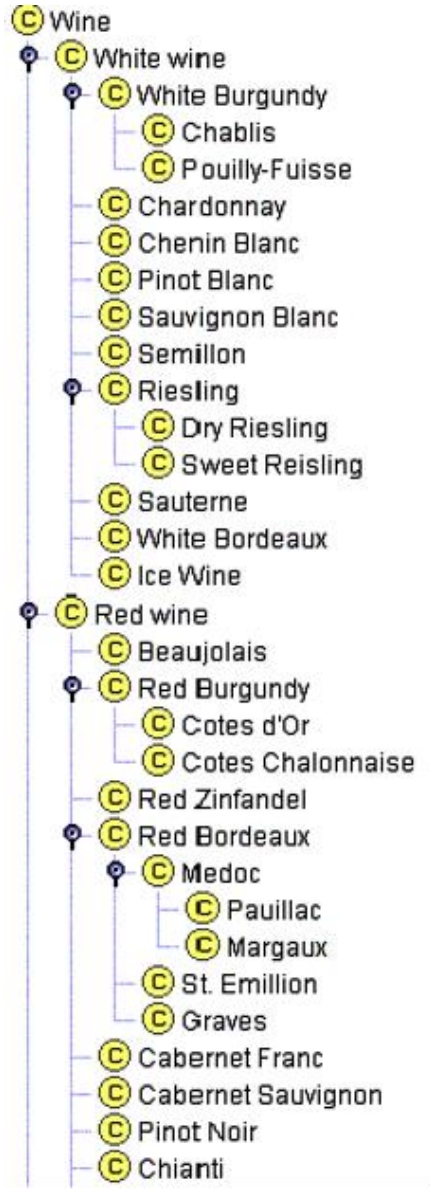
Ontologies in Computer Science

- 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

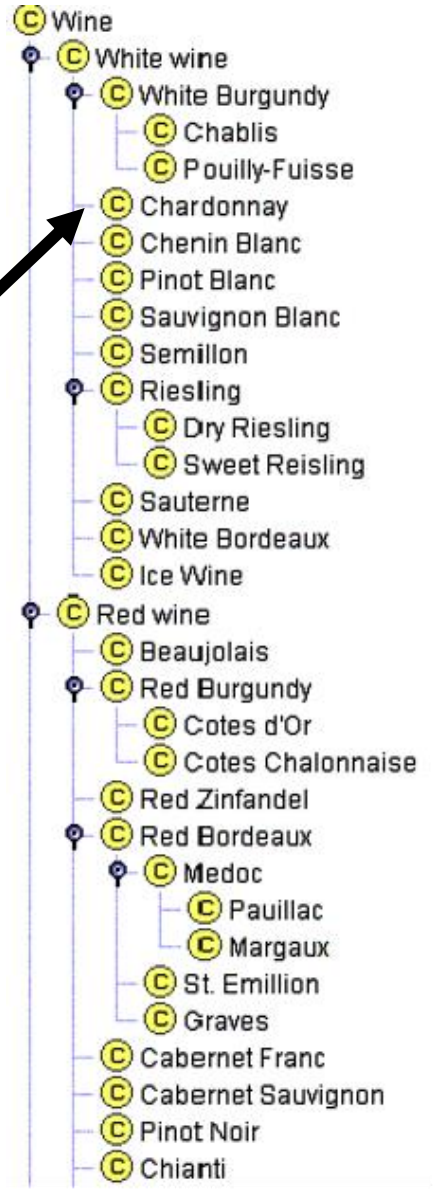
- Class-instance distinction



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



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

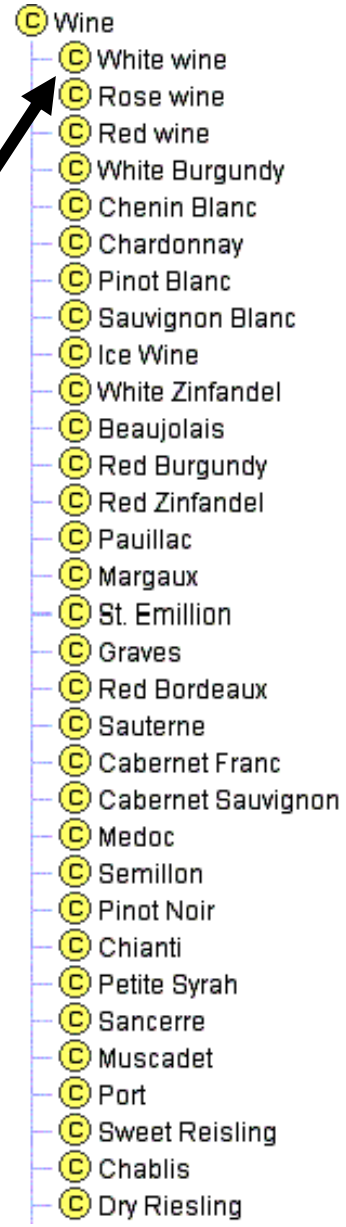
- Class-instance distinction

- 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

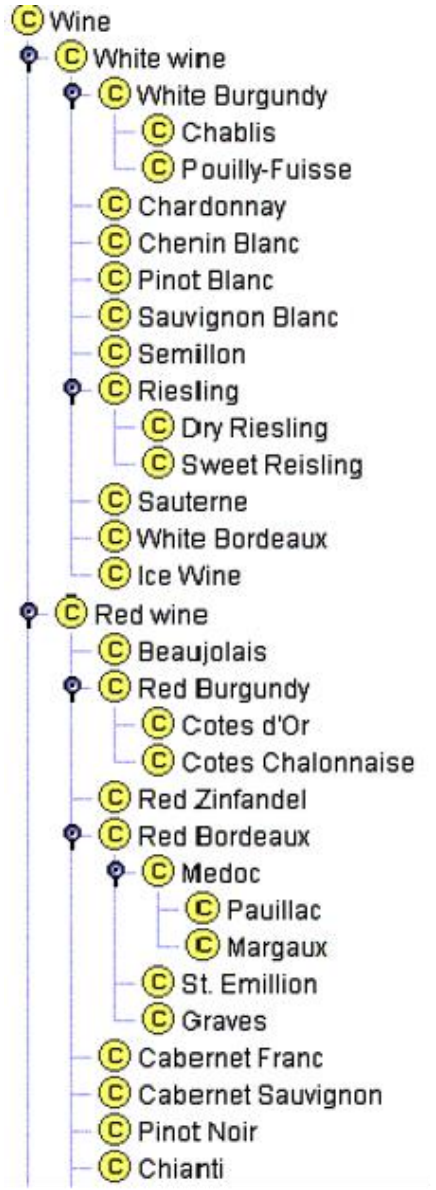
- Class-instance distinction



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

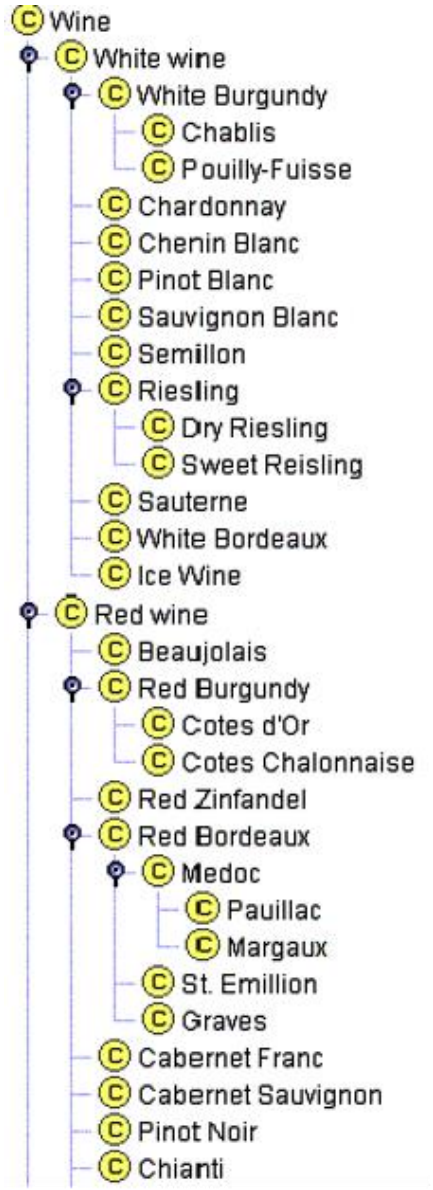
- Class-instance distinction
- Type-token distinction



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

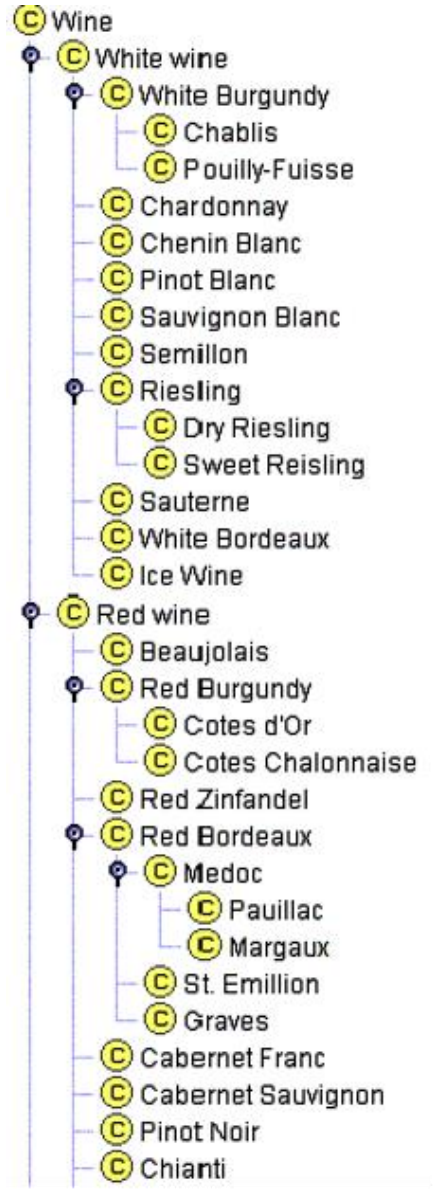
- Class-instance distinction
- Type-token distinction
  - "They drive the same car“
    - They drive the same car type
      - (a Toyota)
    - They drive the same car token
      - (the 2003 Toyota Corolla with VIN: 2QFBORHE4KP911561)



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



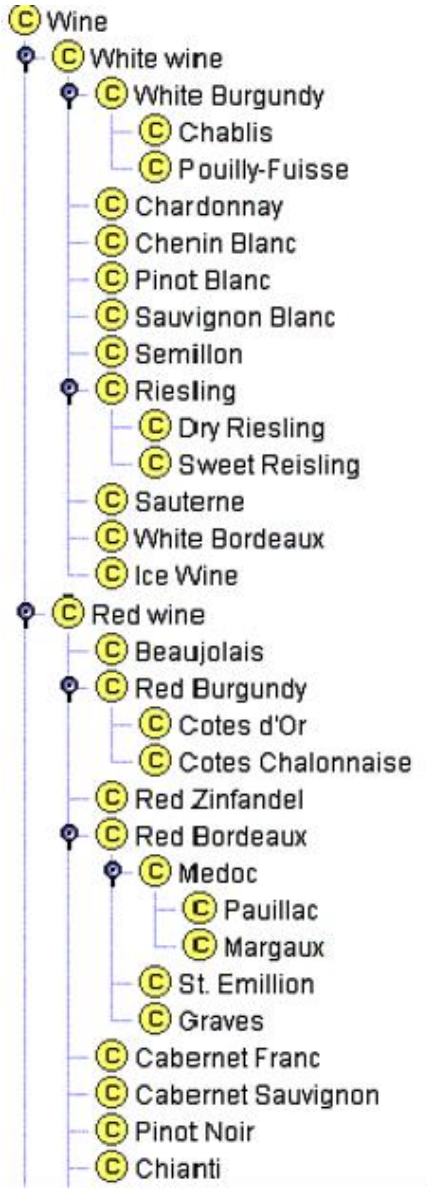
(C) A set of wine bottles

(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



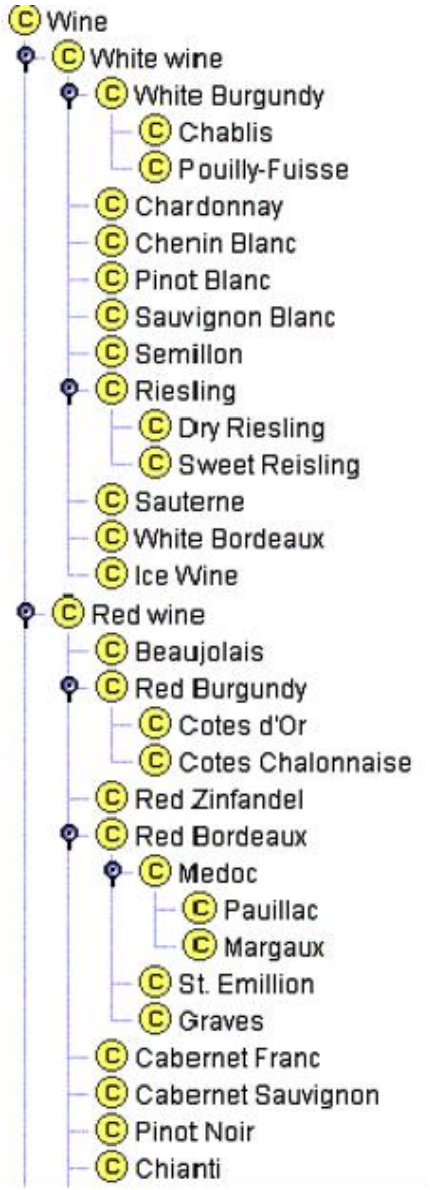
(C) A set of wine bottles

(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



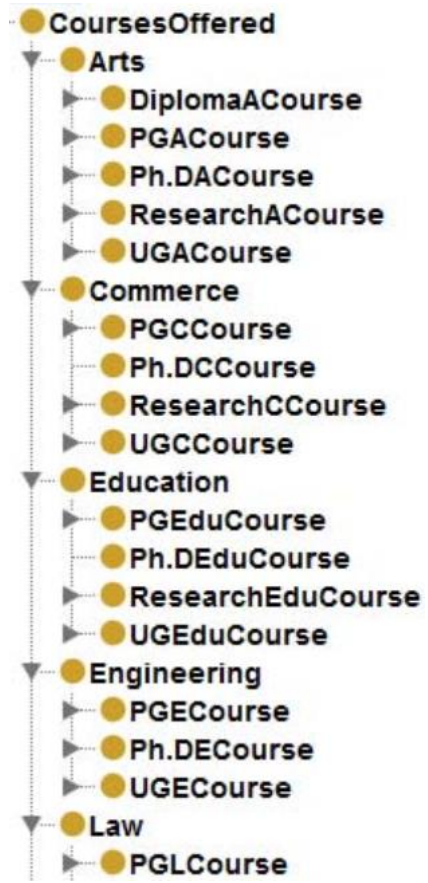
(C) A set of wine bottles

(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

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

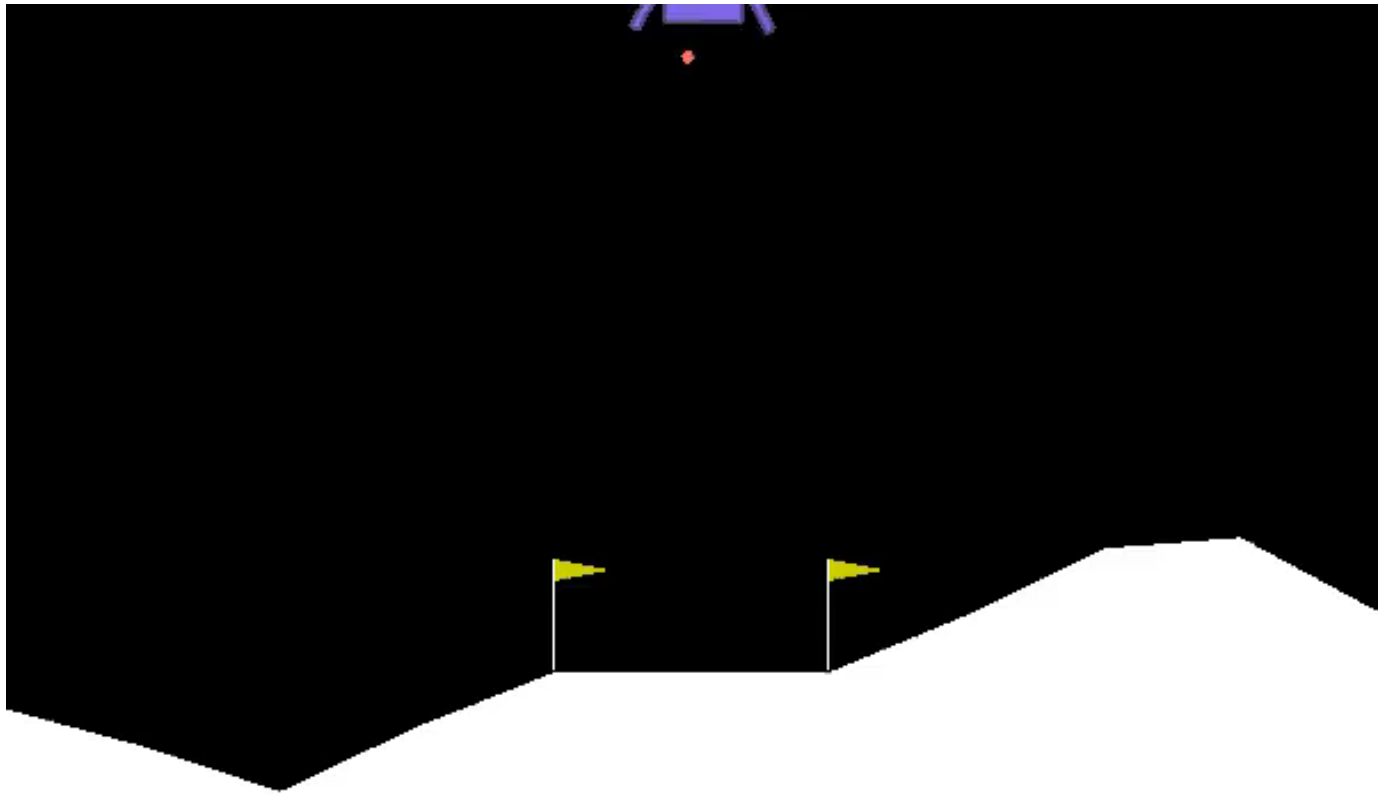
- Design decisions

Design, V0

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

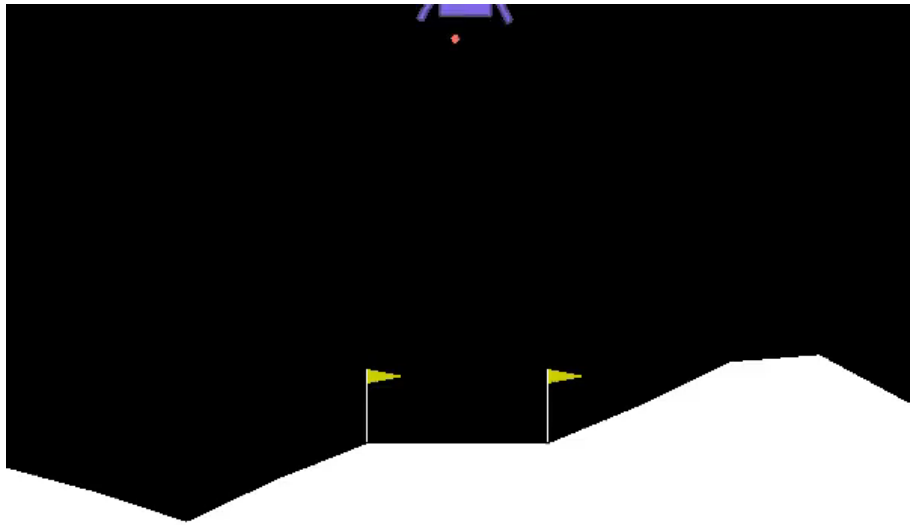
Design, V0

- OpenAI's LunarLander-v2



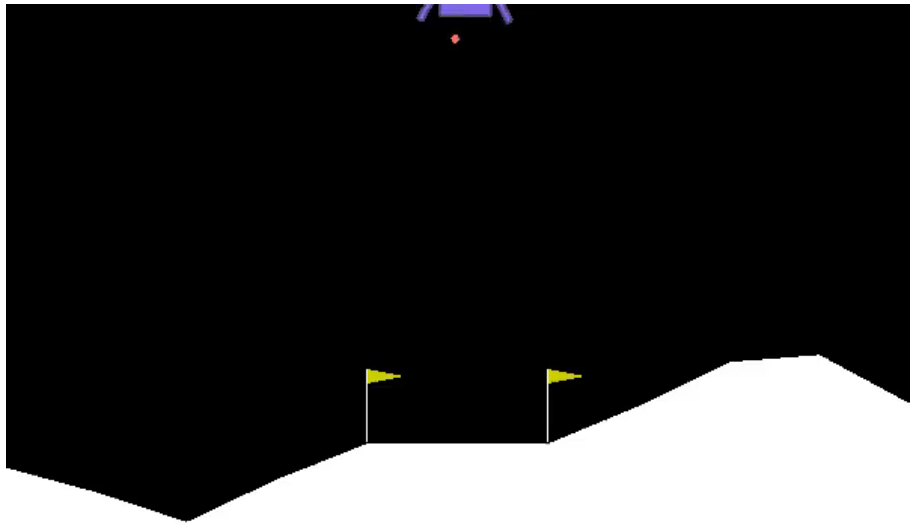
Design, V0

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



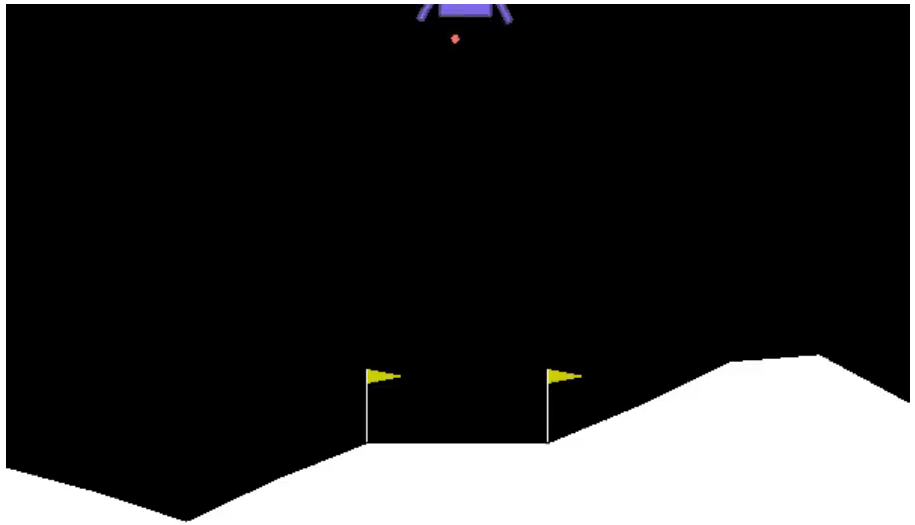
Design, V0

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

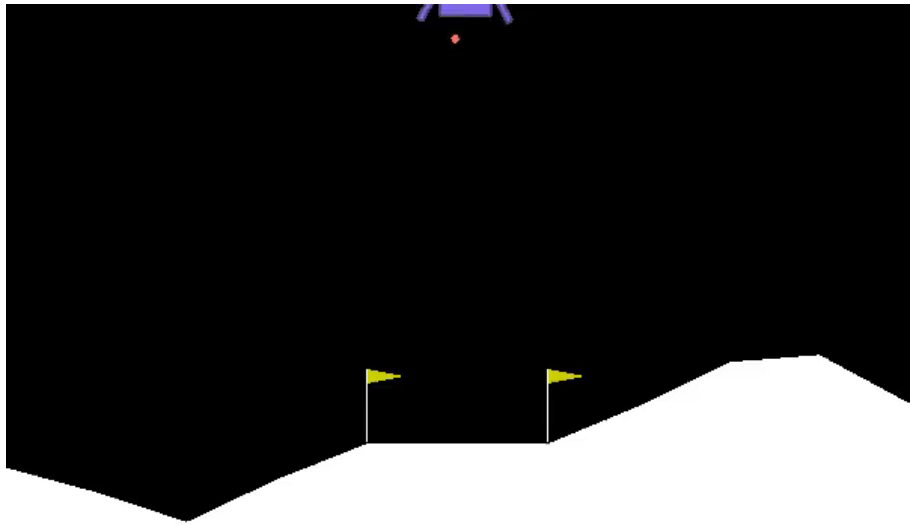


Design, V0

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

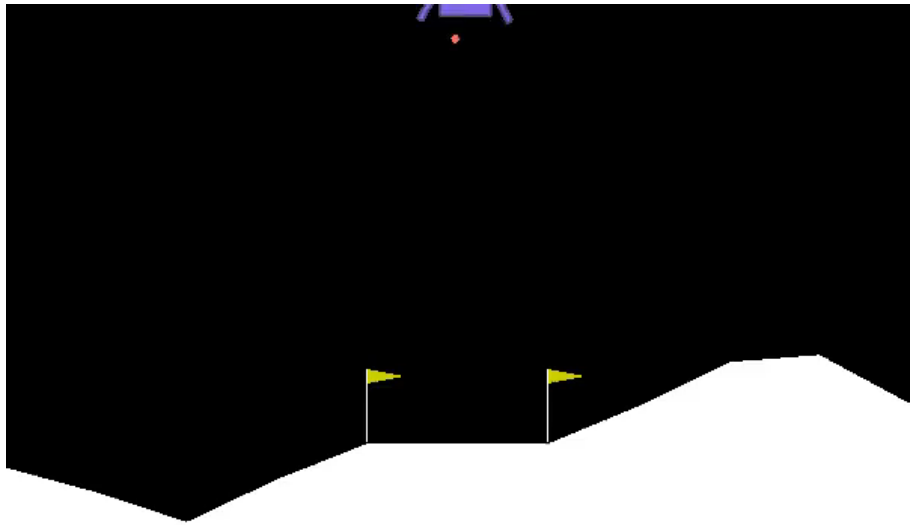


Design, V0



- 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

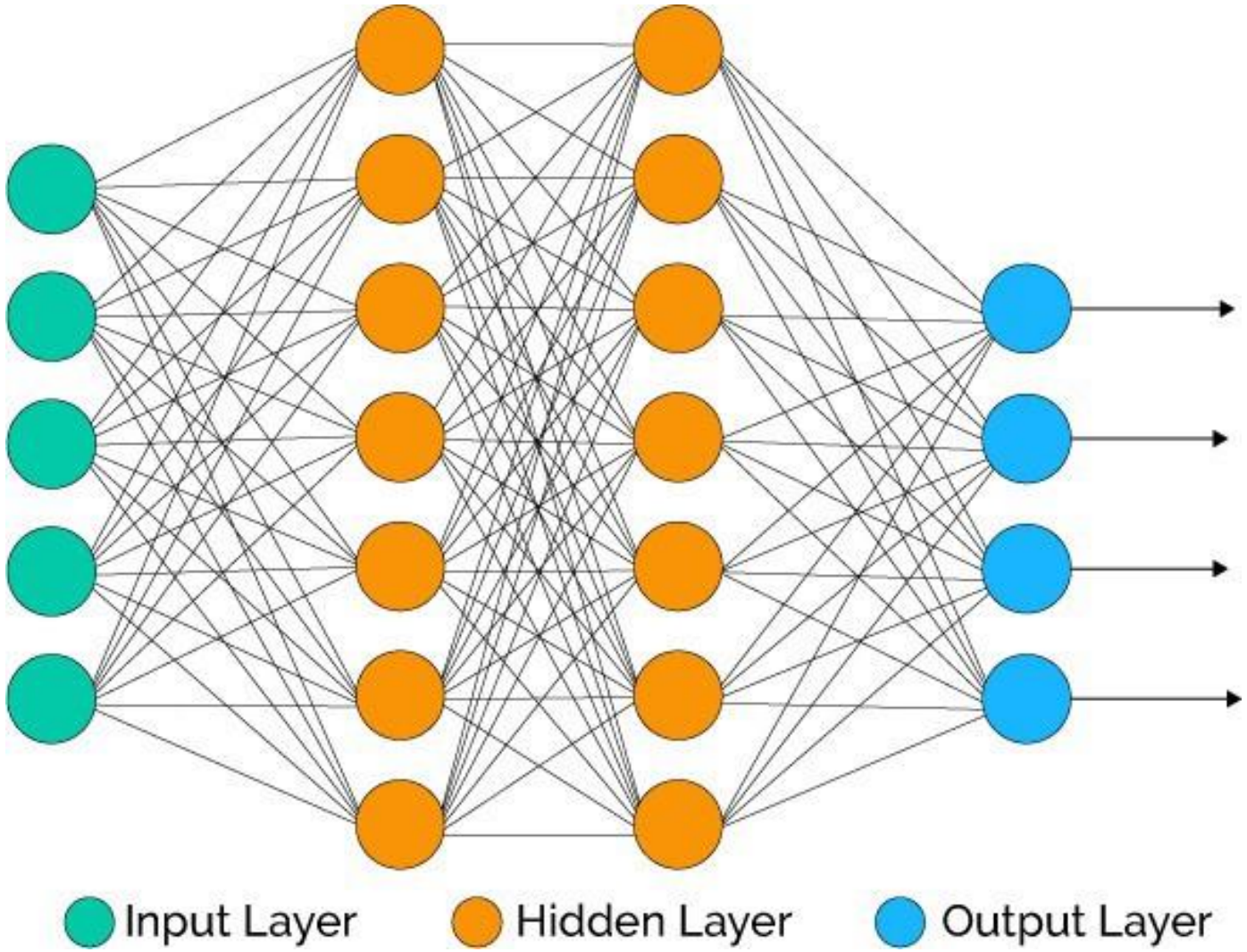


Image from:
https://medium.com/datadriveninvestor/when-not-to-use-neural-networks-89fb50622429

Neural networks

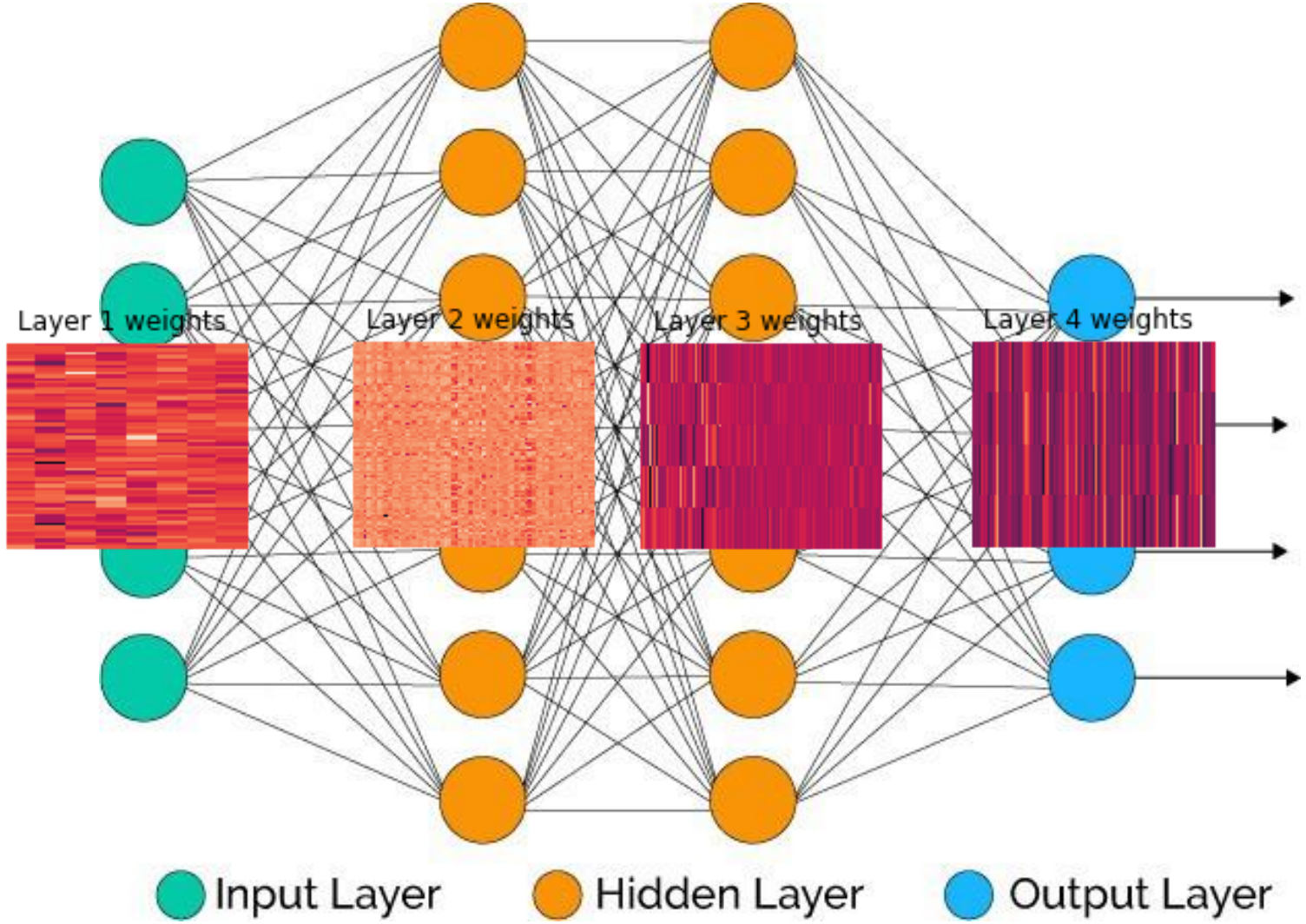


Image from:
https://medium.com/datadriveninvestor/when-not-to-use-neural-networks-89fb50622429

Neural networks

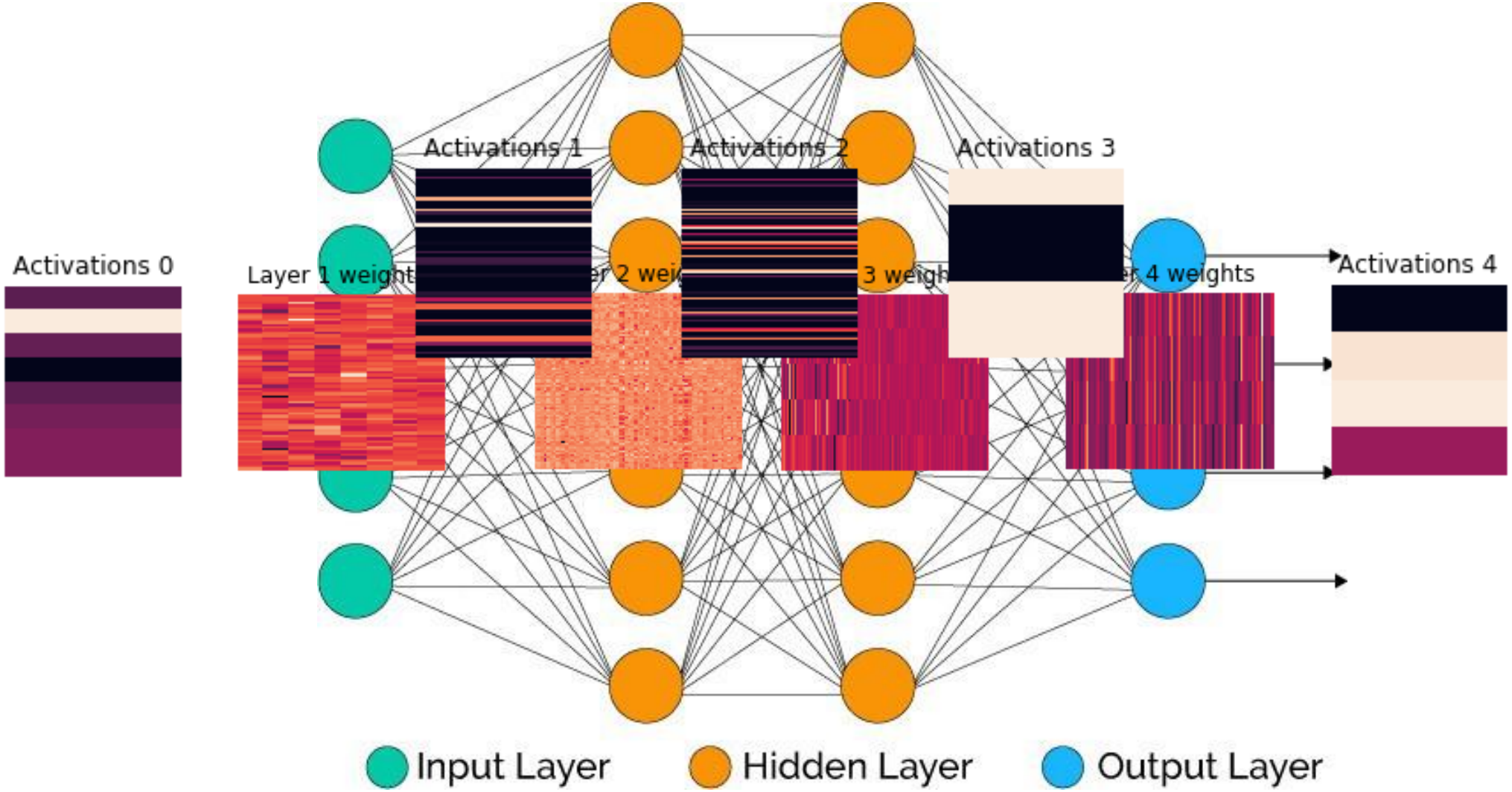
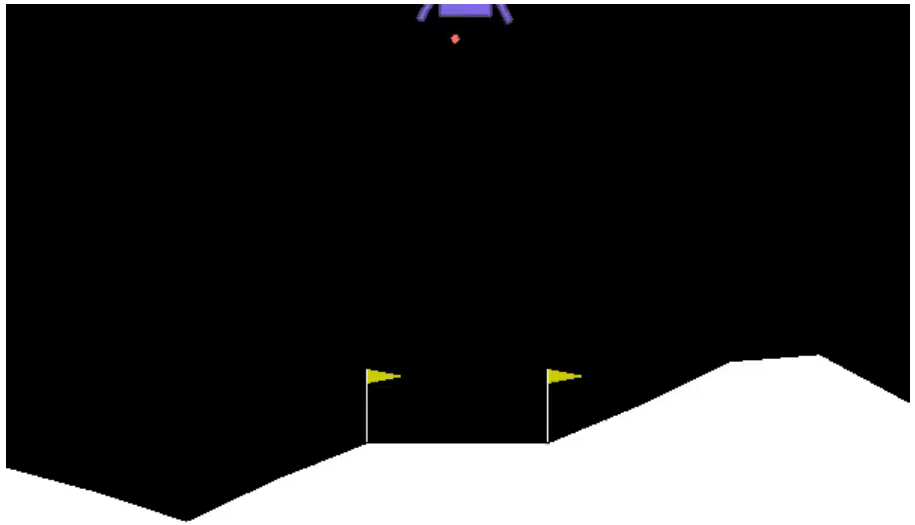


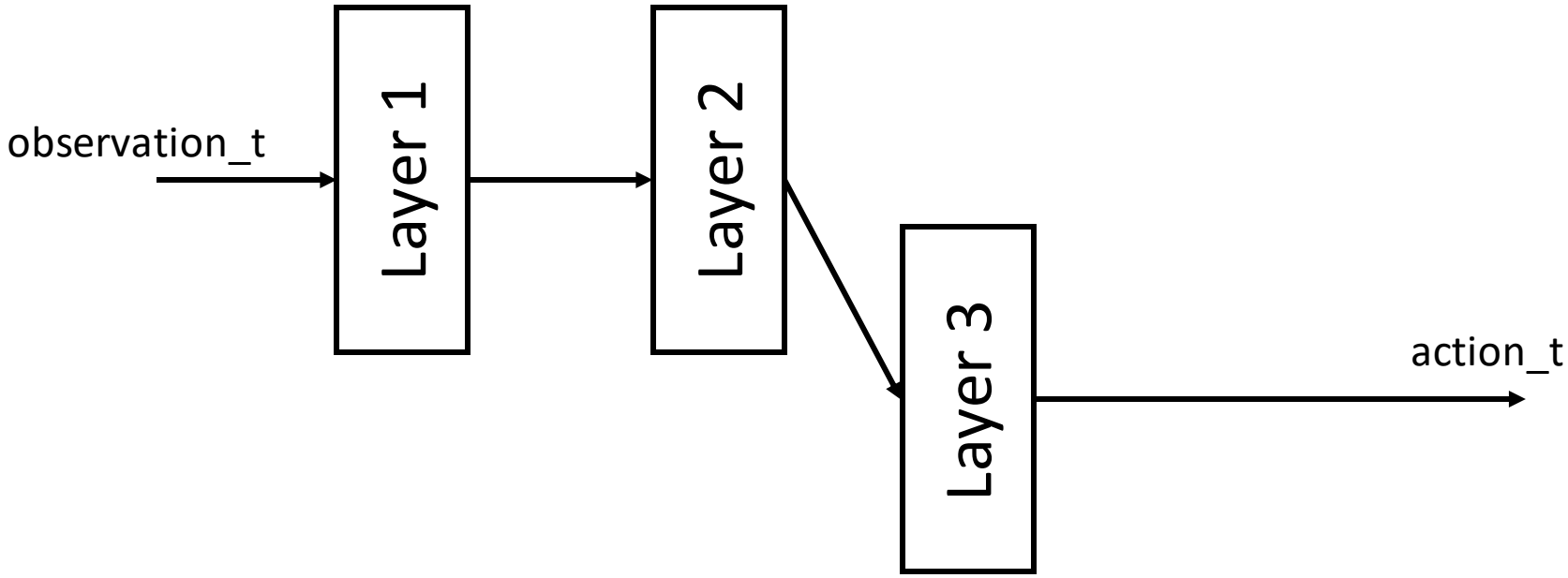
Image from:
https://medium.com/datadriveninvestor/when-not-to-use-neural-networks-89fb50622429

Design, V0

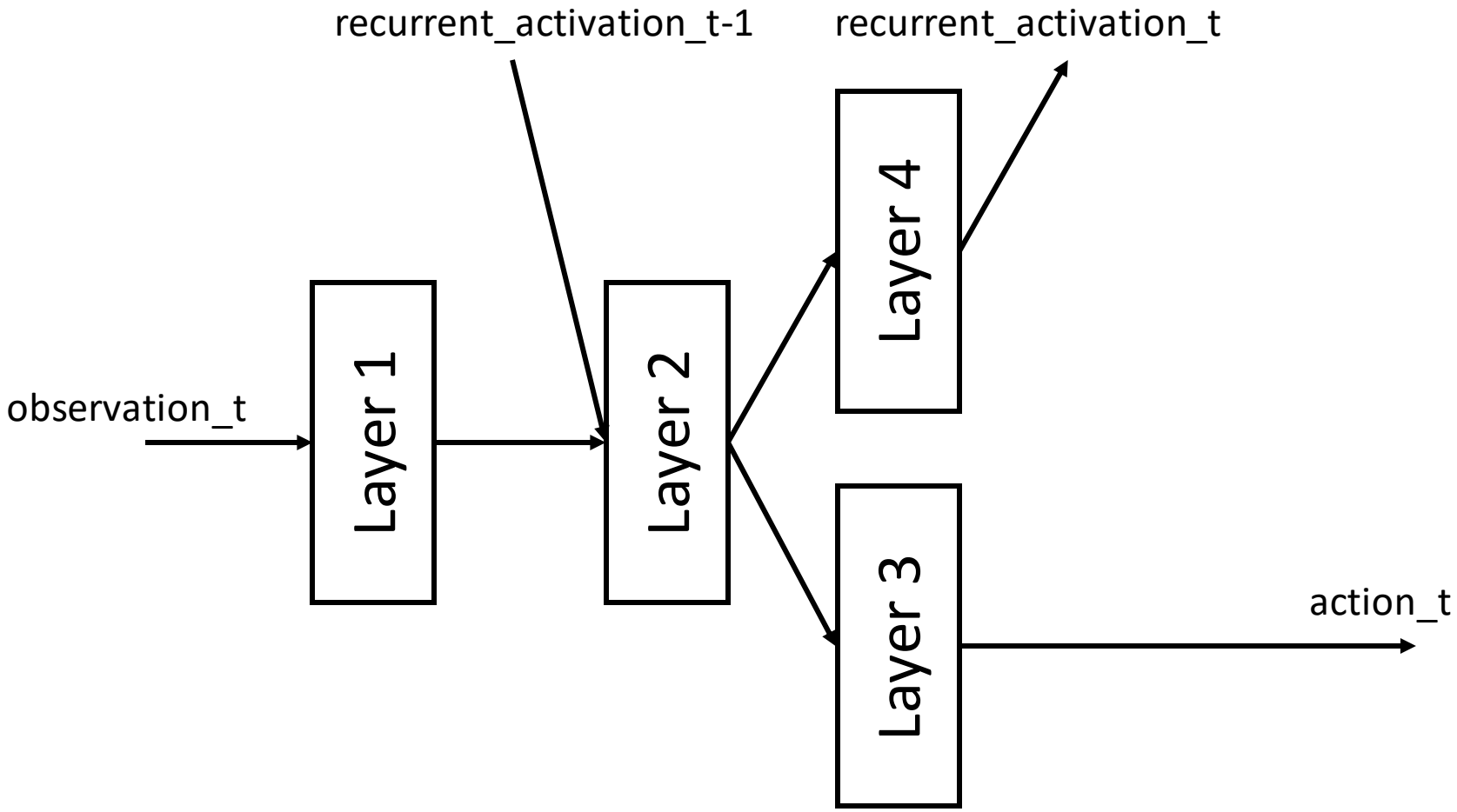


- Design decisions
  - Environment and the agent's "physical" form
  - Internal state of the agent
    - Beliefs about itself relative to semantically important regions
      - Left of the flags, right of the flags, high above the ground, close to the ground, falling too fast
  - Brain state of the agent

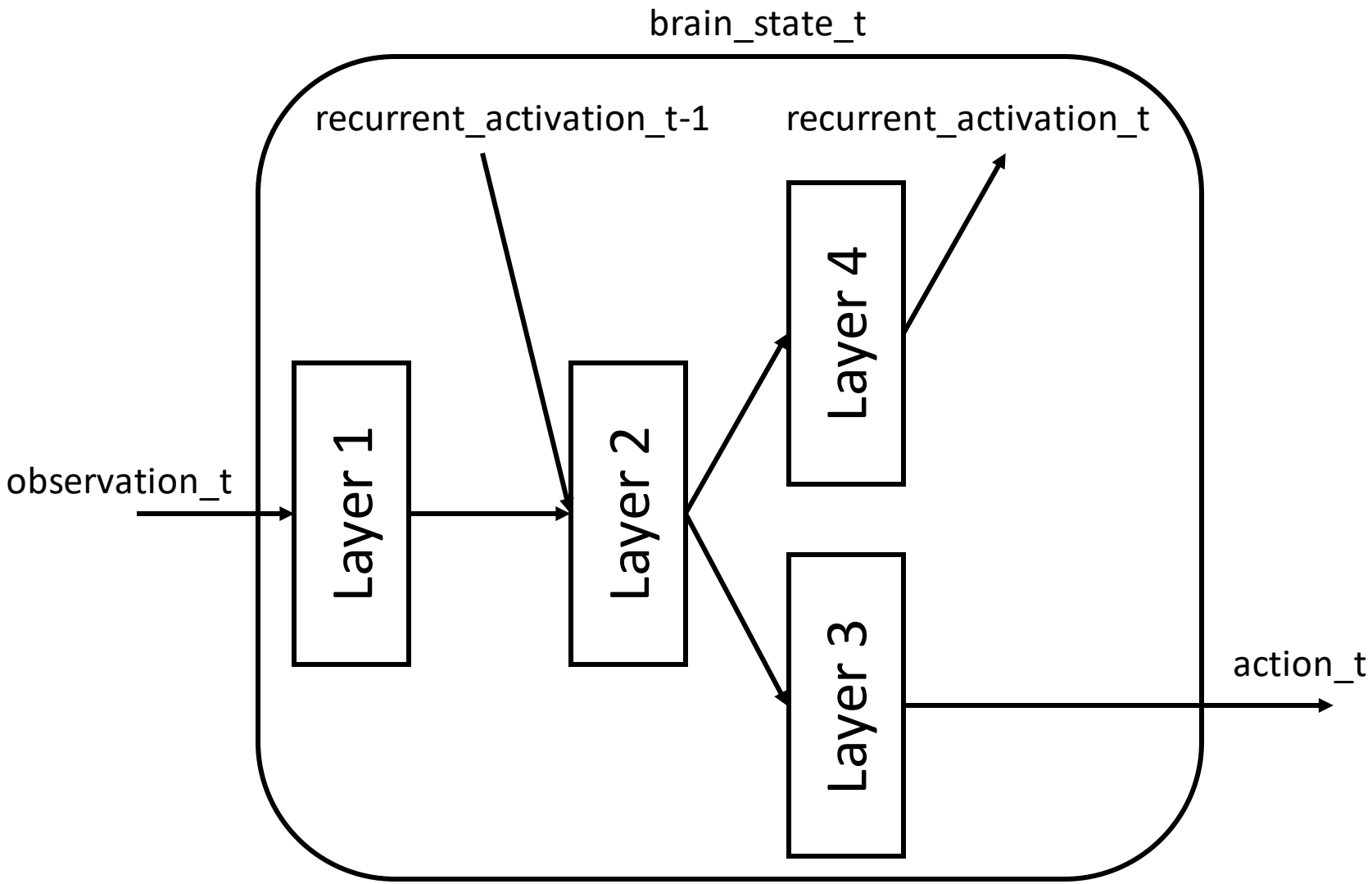
Design, V0



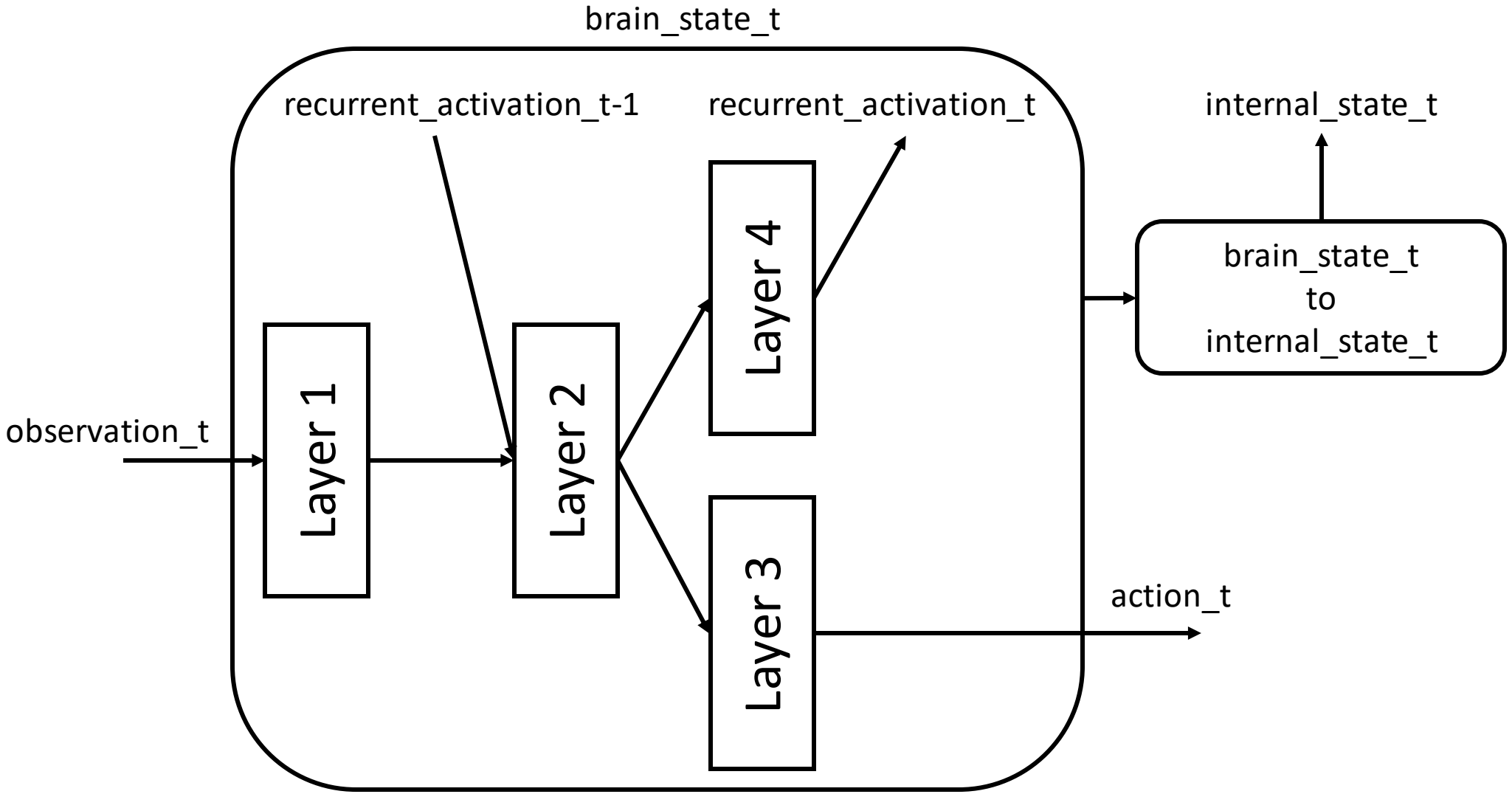
Design, V0


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

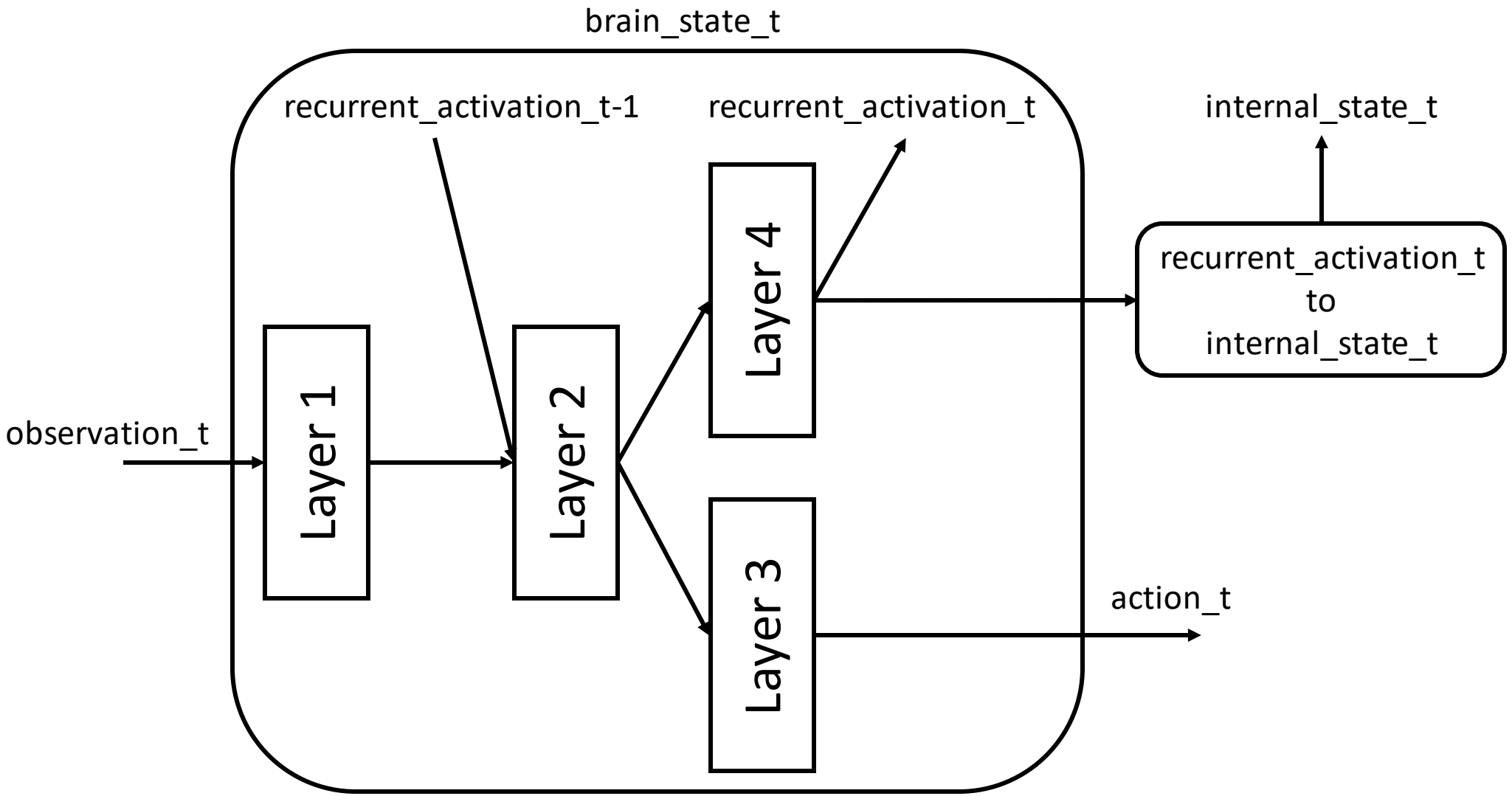
Design, V0


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

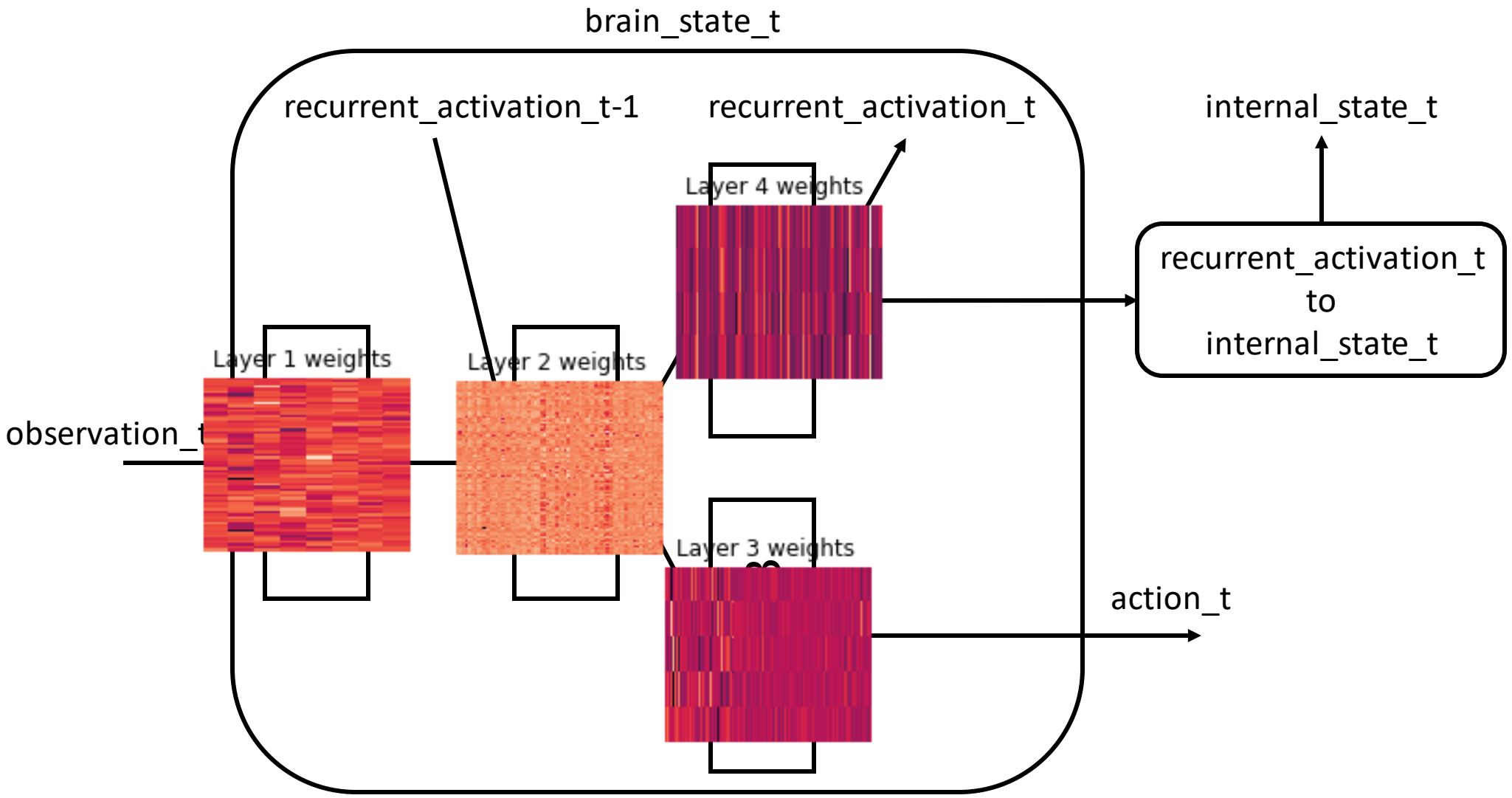
Design, V0


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

Design, V0

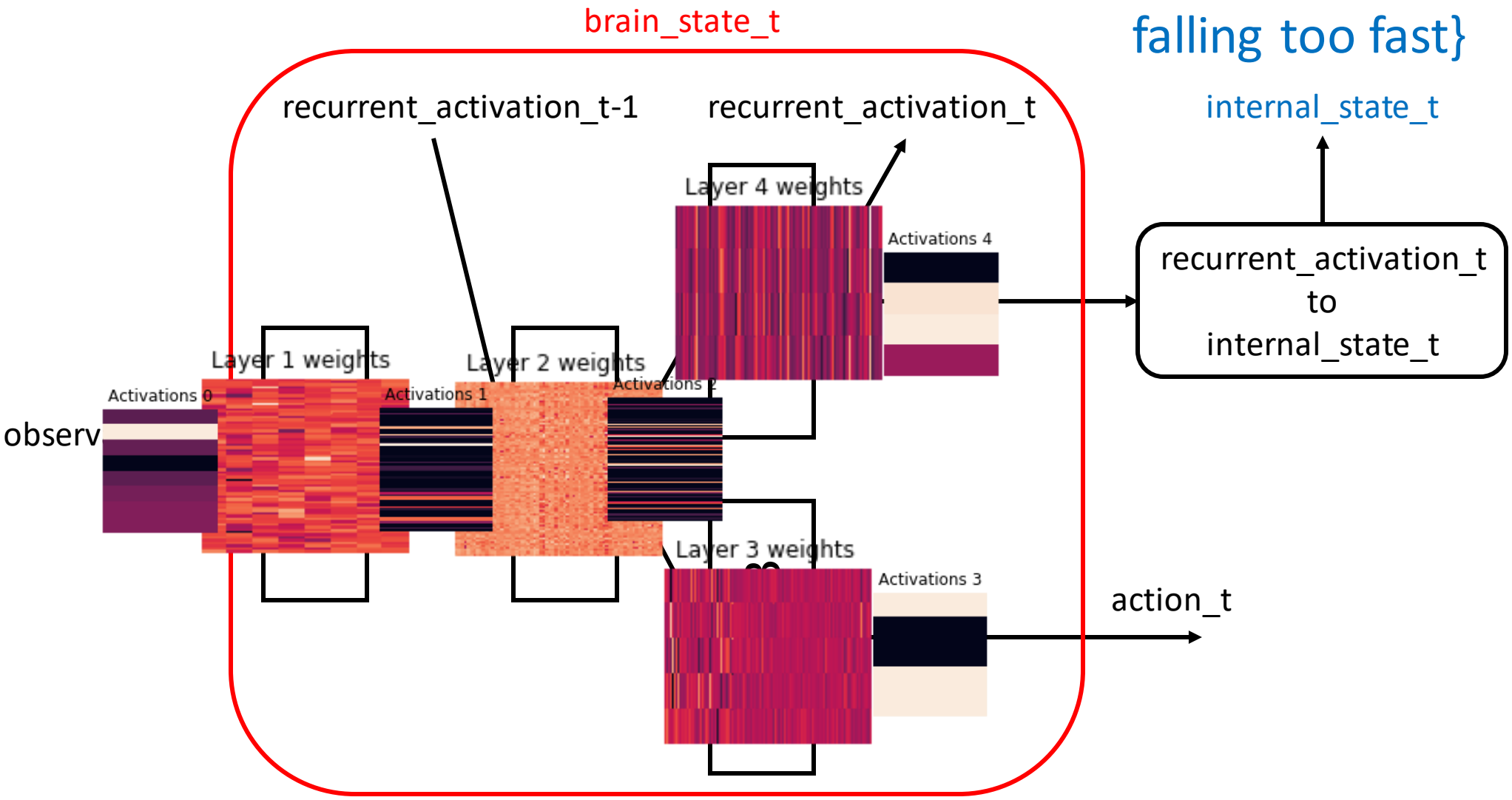

brain_state_t
recurrent_activation_t-1
recurrent_activation_t
internal_state_t
Layer 4
recurrent_activation_t
to
internal_state_t
observation_t
Layer 1
Layer 2
Layer 3
action_t

Design, V0

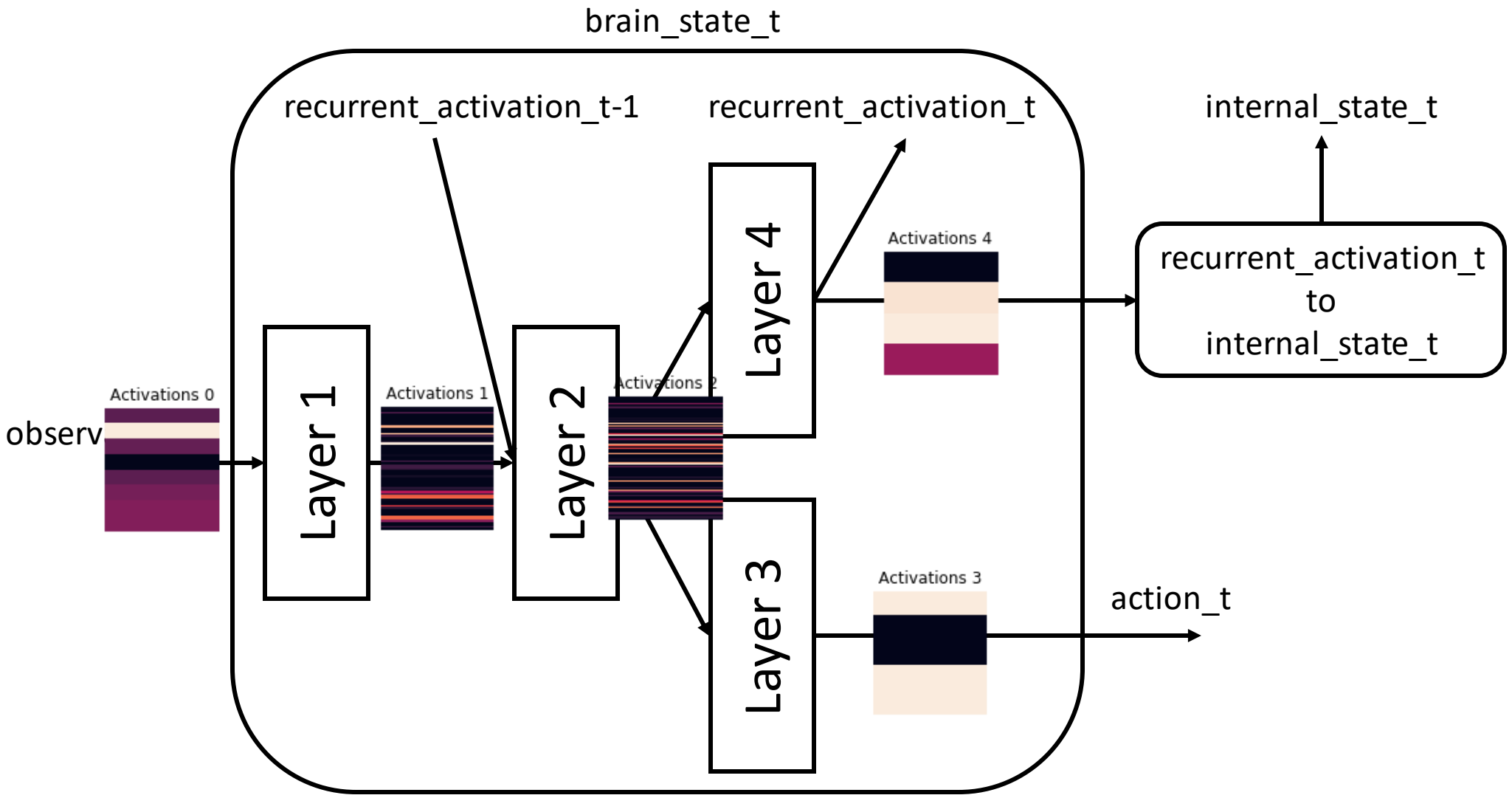

brain_state_t
recurrent_activation_t-1
recurrent_activation_t
internal_state_t
Layer 4 weights
recurrent_activation_t
to
internal_state_t
Layer 1 weights
Layer 2 weights
observation_t
Layer 3 weights
action_t

Design, V0

{High above the ground, right of the center falling too fast}

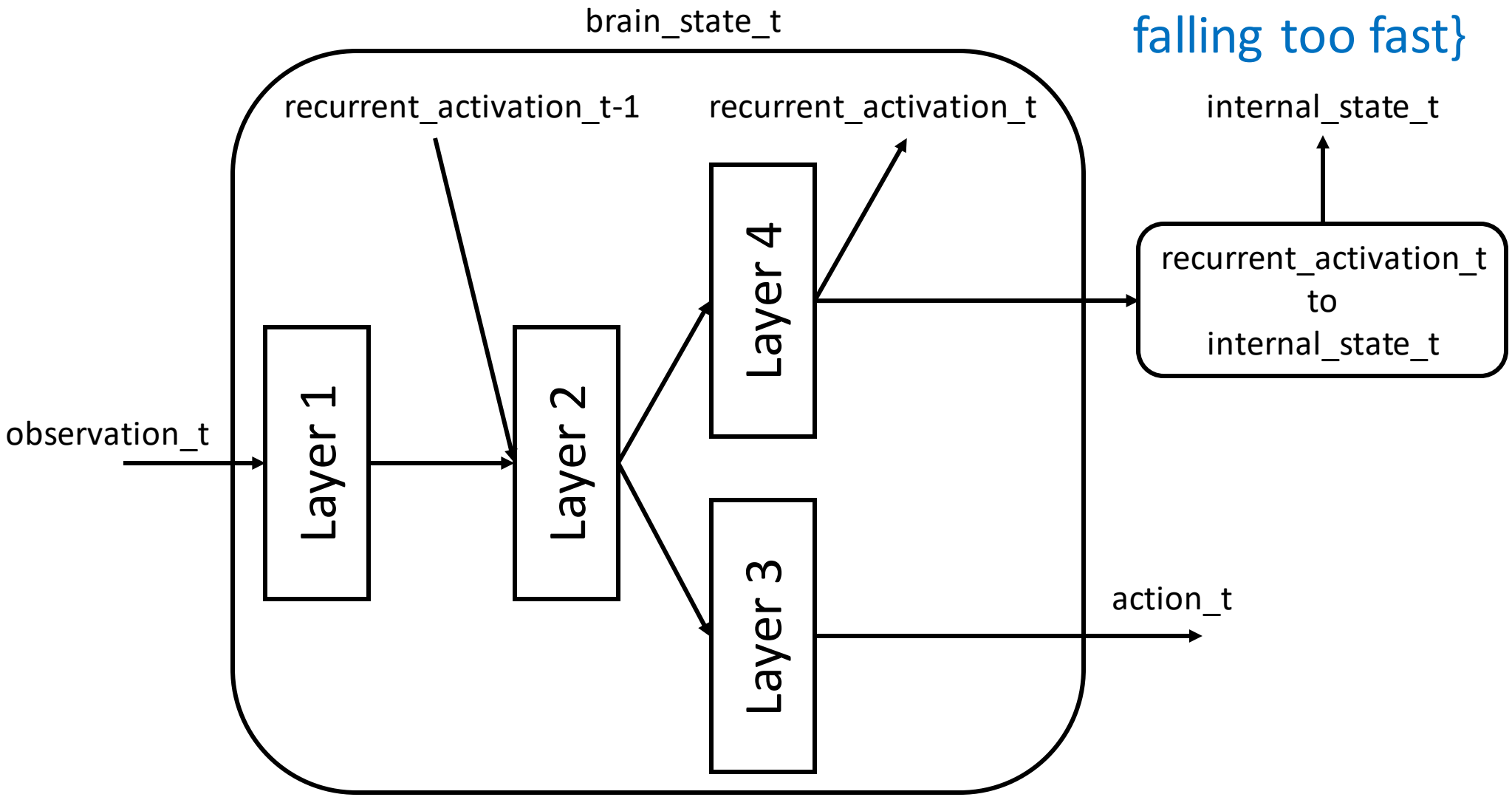


Design, V0

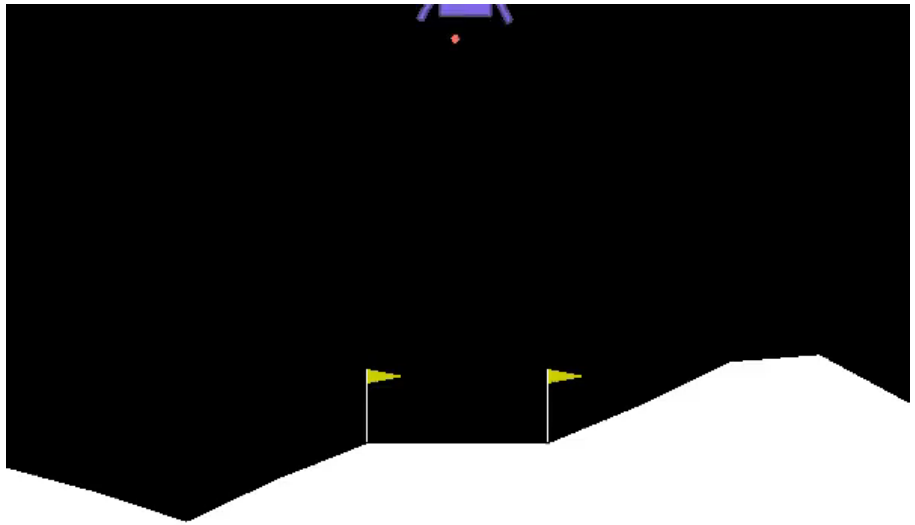

brain_state_t
recurrent_activation_t-1
recurrent_activation_t
internal_state_t
Activations 4
recurrent_activation_t
to
internal_state_t
Layer 4
Activations 0
observ
Layer 1
Activations 1
Layer 2
Activations 2
Layer 3
Activations 3
action_t

Design, V0

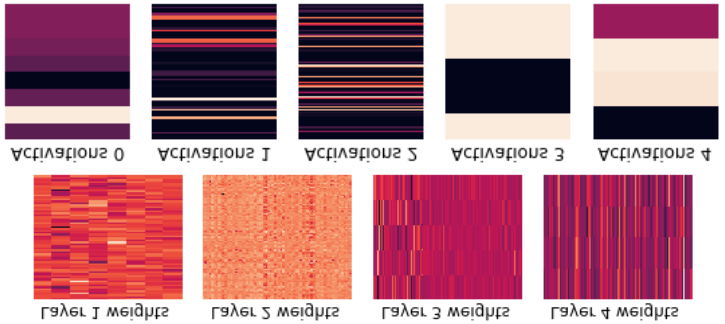
{High above the ground, right of the center falling too fast}



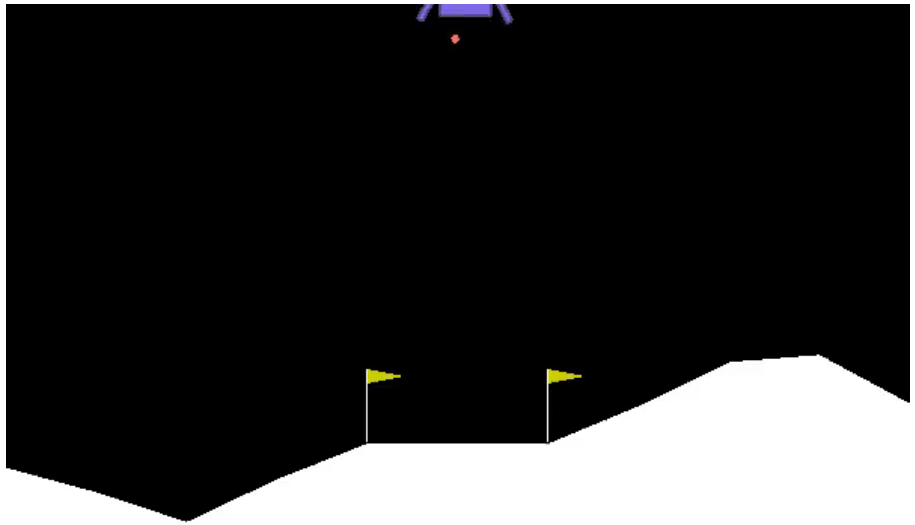
Design, V0



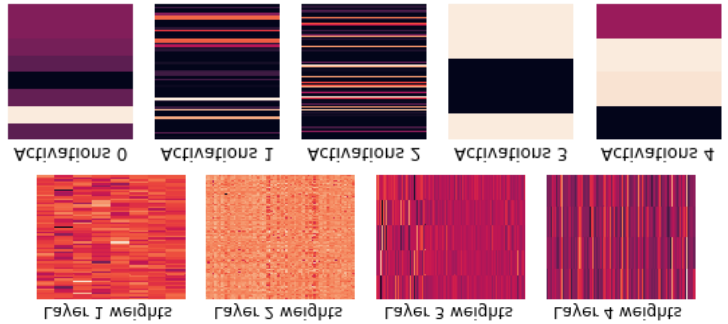
- Design decisions
  - Environment and the agent's "physical" form
  - Internal state of the agent
    - Beliefs about itself relative to semantically important regions
      - Left of the flags, right of the flags, high above the ground, close to the ground, falling too fast
  - Brain state of the agent



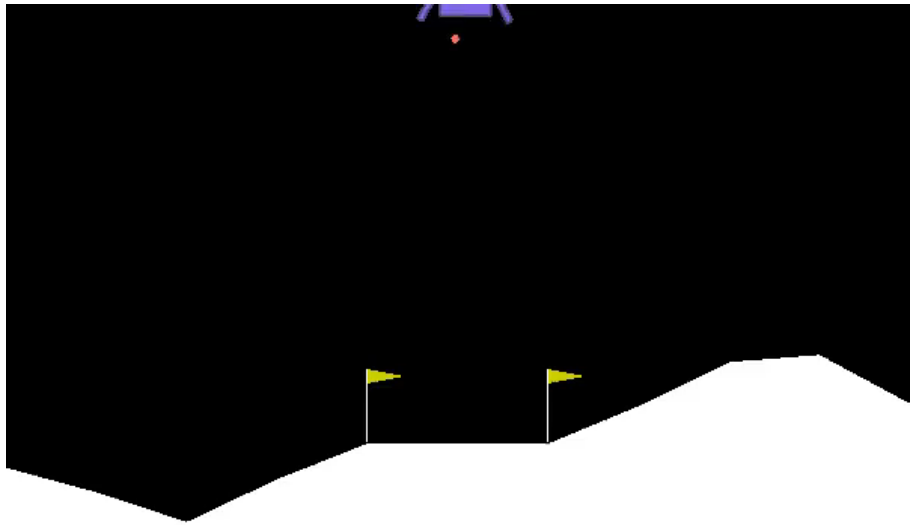
Design, V0



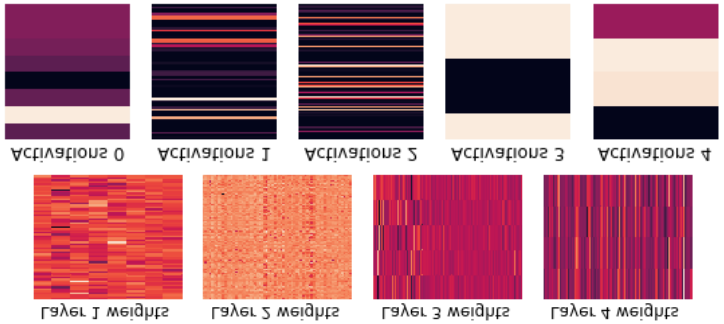
- Design decisions
  - Environment and the agent's "physical" form
  - Internal state of the agent
    - Beliefs about itself relative to semantically important regions
      - Left of the flags, right of the flags, high above the ground, close to the ground, falling too fast
  - Brain state of the agent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



Design, V0

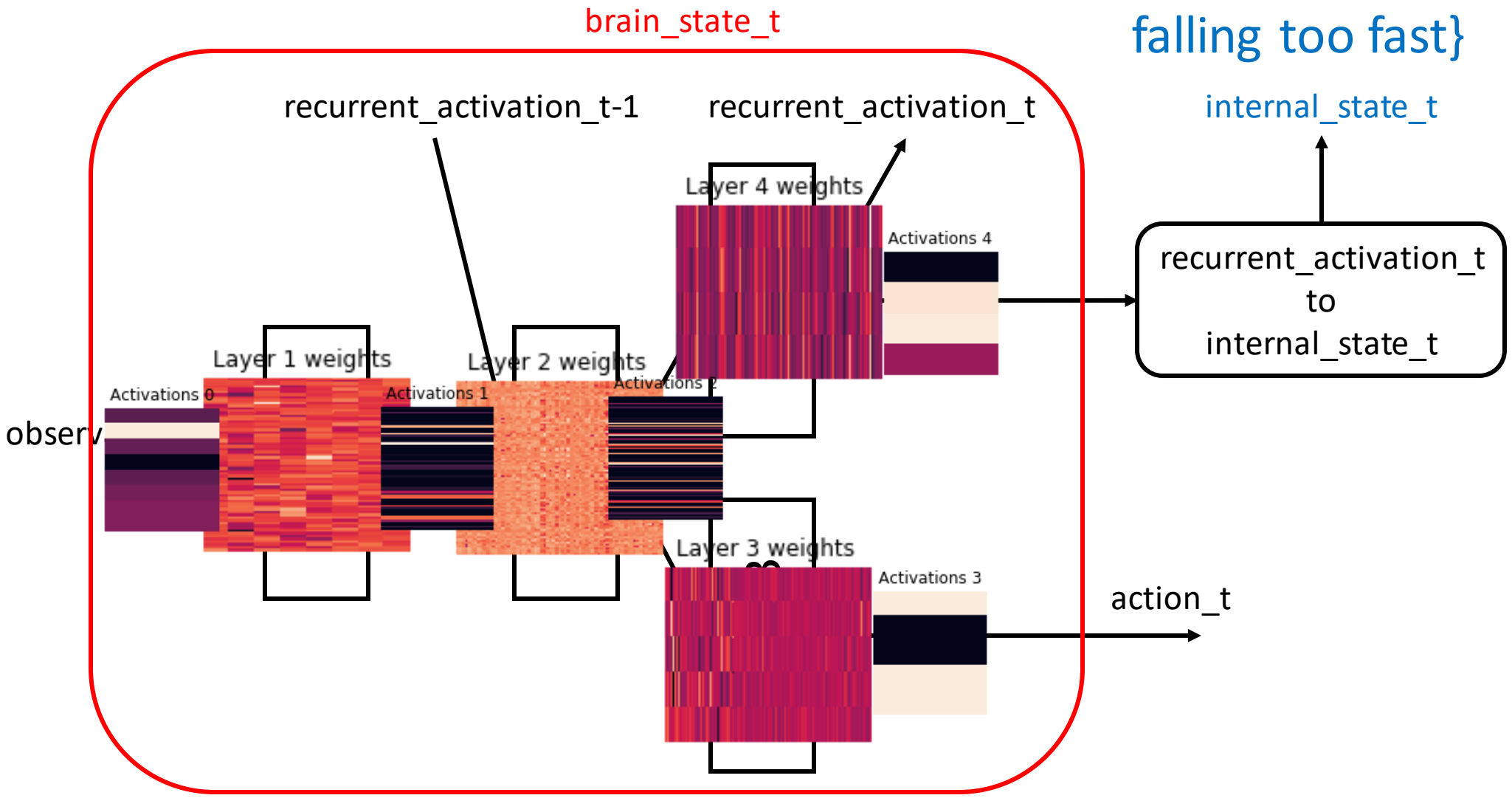


- Design decisions
  - Environment and the agent's "physical" form
  - Internal state of the agent
    - Beliefs about itself relative to semantically important regions
      - Left of the flags, right of the flags, high above the ground, close to the ground, falling too fast
  - Brain state of the agent
  - 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 at time t (set of layer weights, activations, and connectivity)
    - A region the agent believes it's in
    - Internal state at time t (set of regions the agent believes it's in)



Design, V0

{High above the ground, right of the center 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

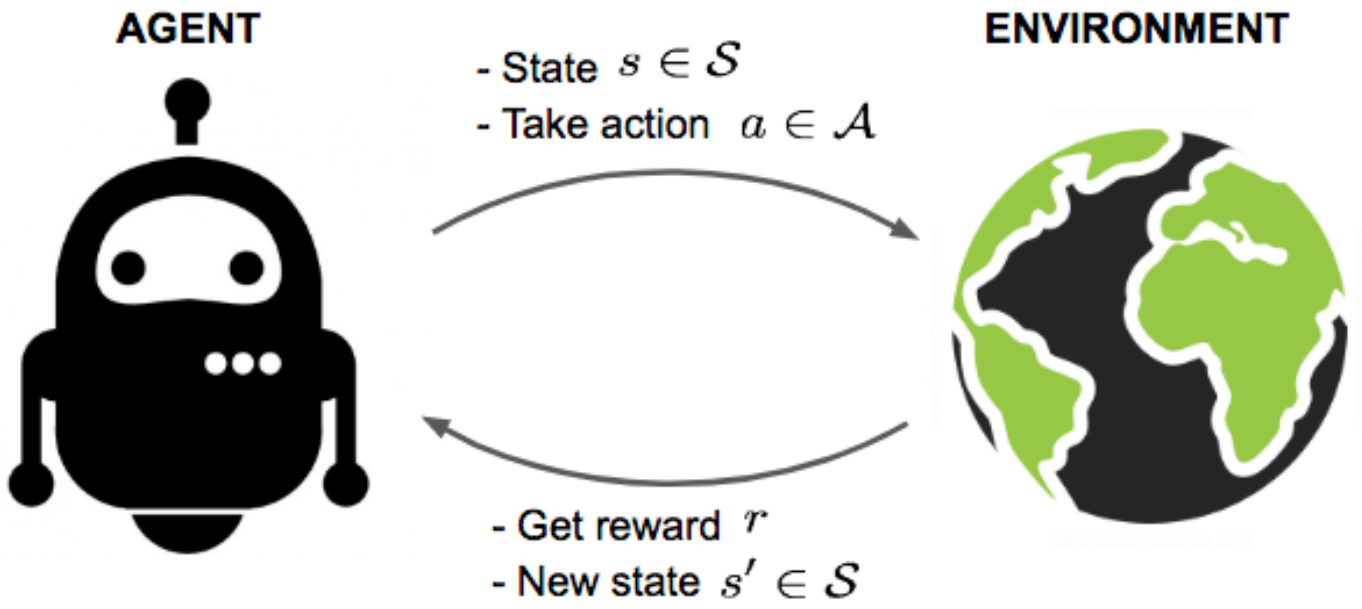
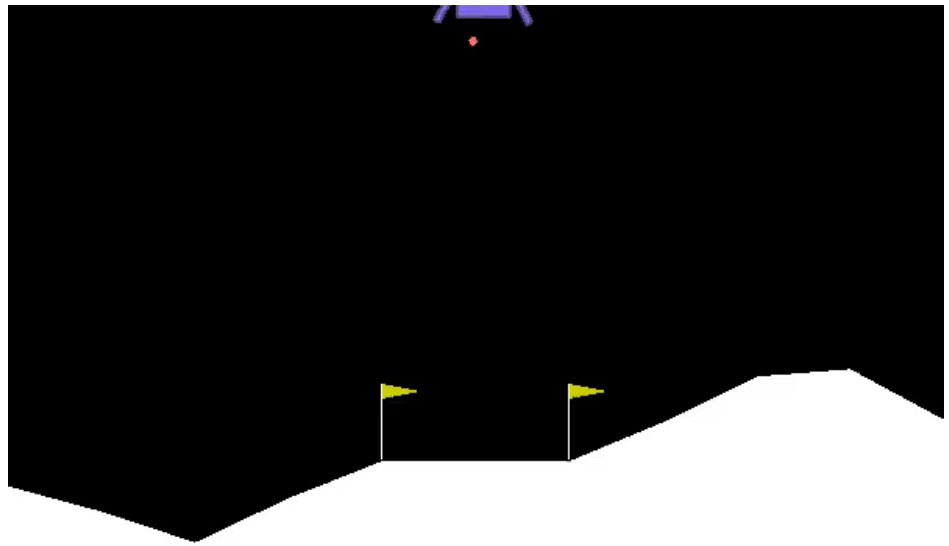


Image from:
https://lilianweng.github.io/lil-log/2018/02/19/a-long-peek-into-reinforcement-learning.html

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

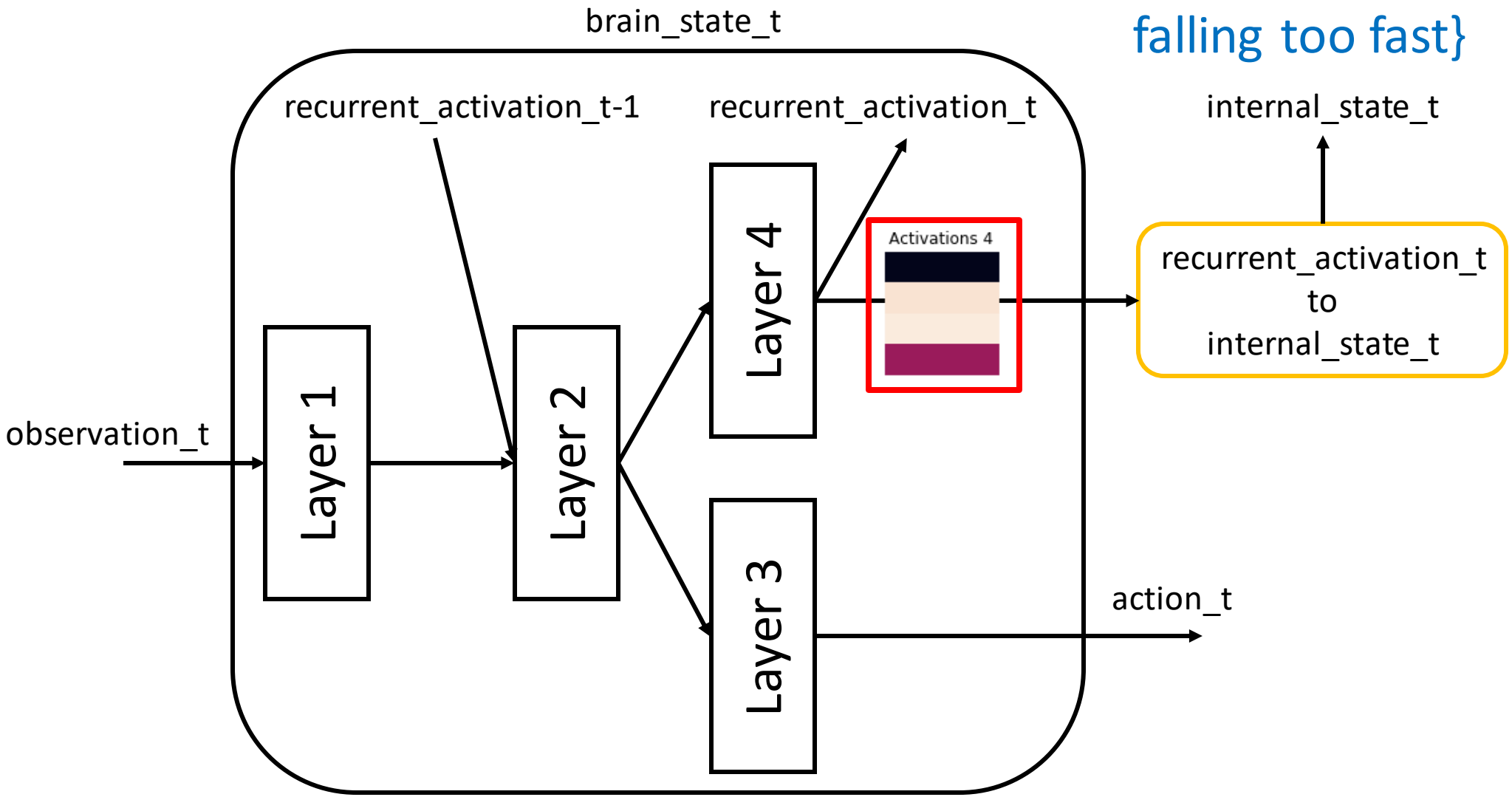


Design, V0

- 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

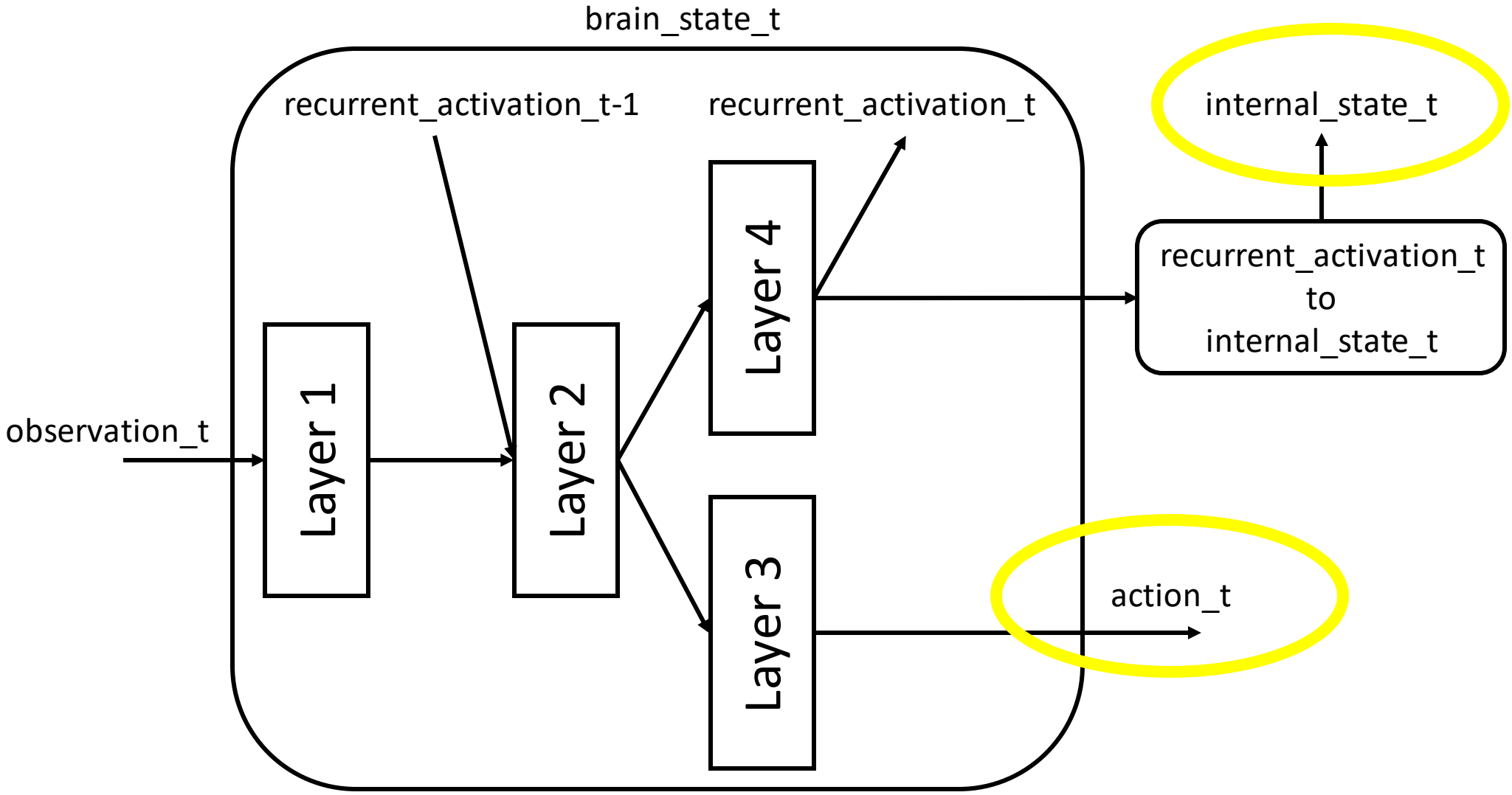
{High above the ground, right of the center falling too fast}



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!

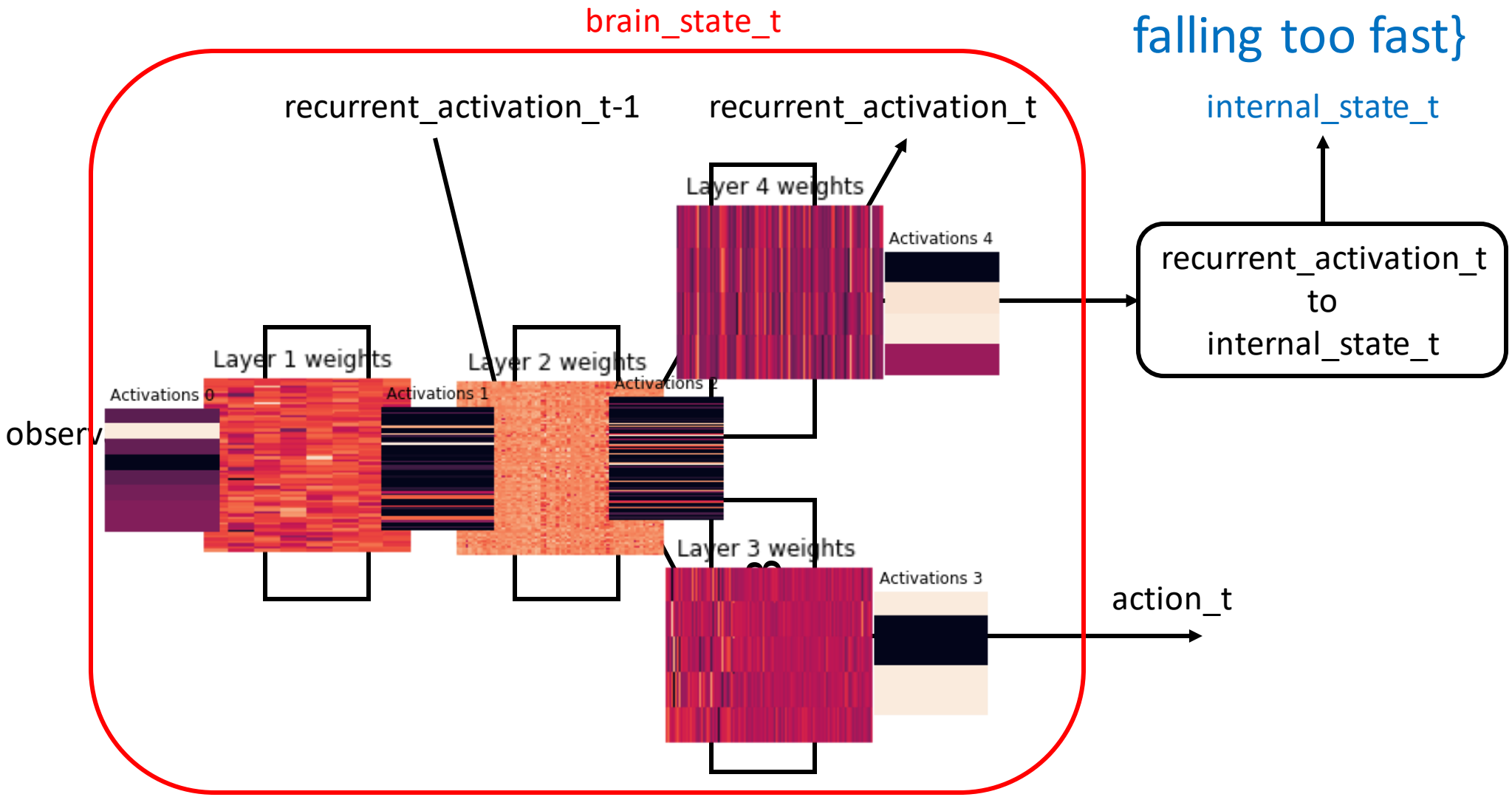
Did we satisfy our requirements?

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

Did we satisfy our requirements?

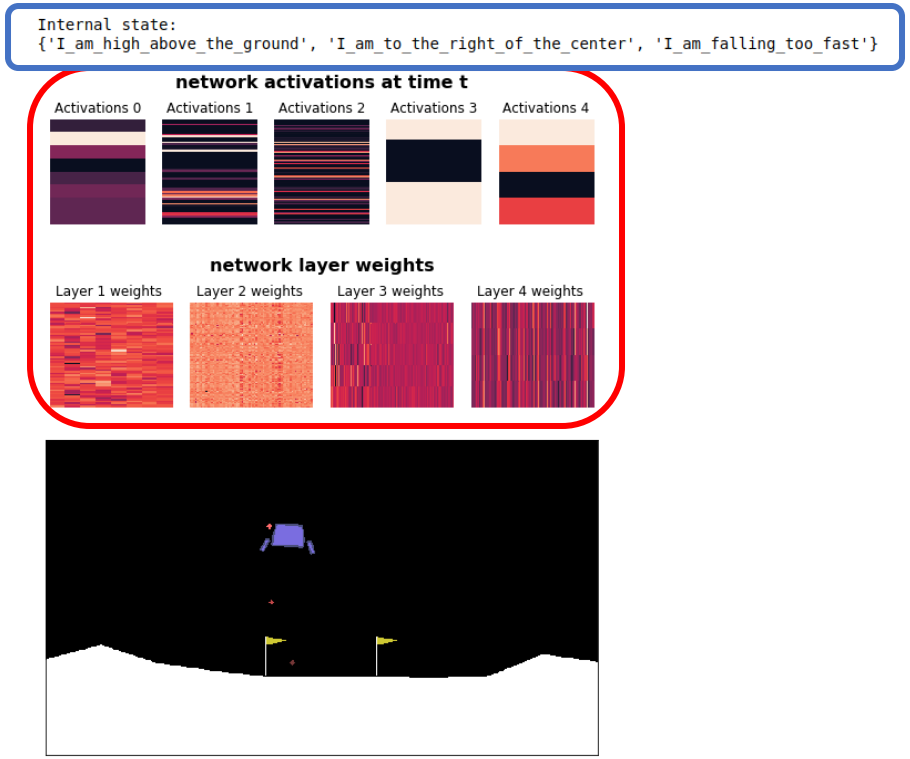
- v0
  - Internal states are casually reducible to brain states
  - Internal states are ontologically irreducible to brain states

{High above the ground, right of the center falling too fast}



Did we satisfy our requirements?

- v0
  - Internal states are casually reducible to brain states
  - Internal states are ontologically irreducible to brain states



Did we satisfy our requirements?

- v0
  - Internal states are casually reducible to brain states
  - Internal states are ontologically irreducible to brain states

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

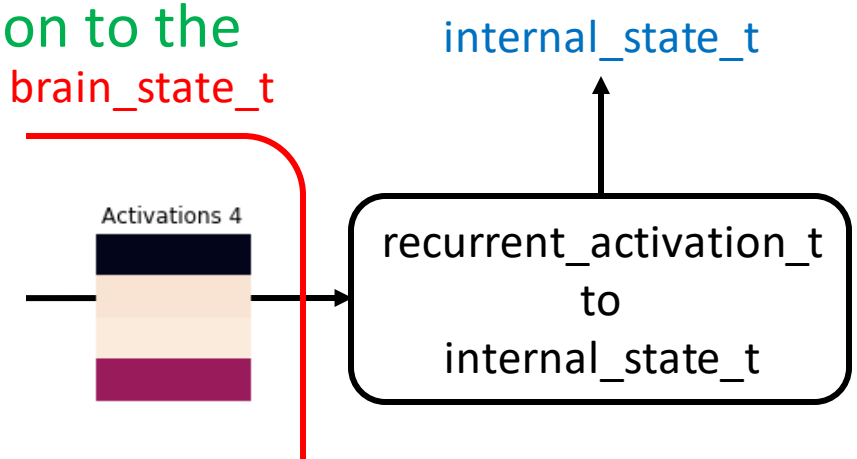
Did we satisfy our requirements?

- v0
  - Internal states are casually reducible to brain states
  - Internal states are ontologically irreducible to brain states

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

{High above the ground, right of the center falling too fast}



Did we satisfy our requirements?

- v0
  - Internal states are casually reducible to brain states
  - Internal states are ontologically irreducible to brain states

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

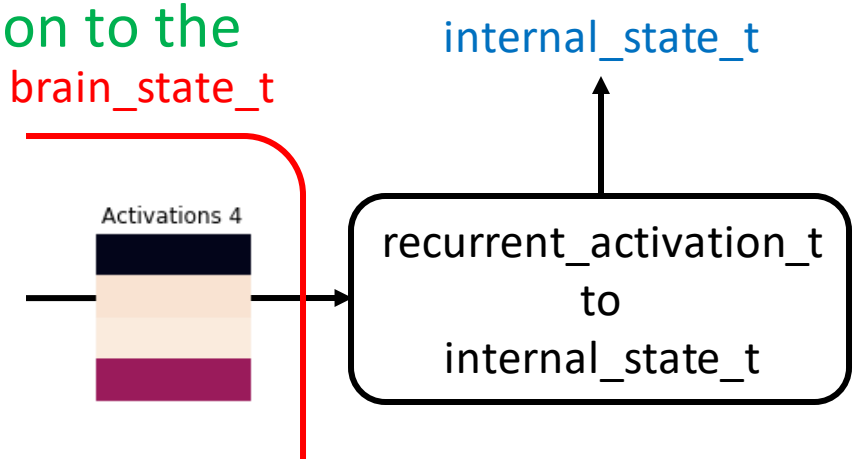
{High above the ground, right of the center falling too fast}

```
def recurrent_activations_to_internal_state(recurrent_activations):

    internal_state = set()

    for activation, region in zip(recurrent_activations, regions):
        if activation > 0.5:
            internal_state.add(region.__name__)

    return internal_state
```



Did we satisfy our requirements?

- v0
  - ✓ Internal states are casually reducible to brain states
    - Internal states are ontologically irreducible to brain states

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

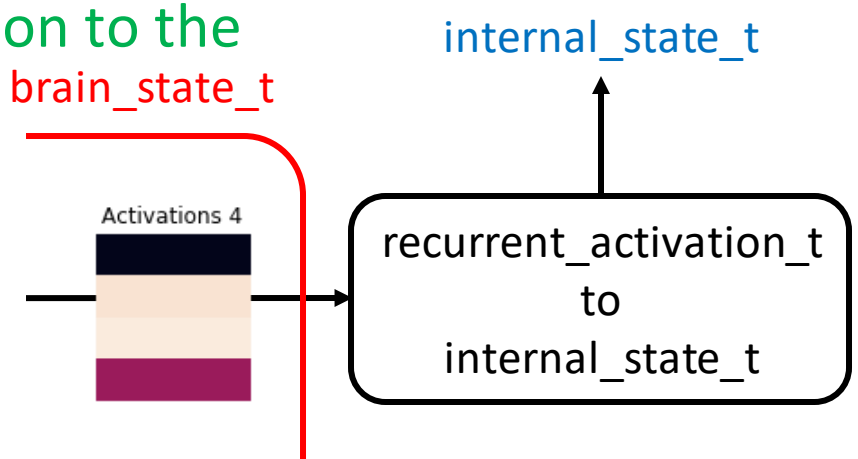
{High above the ground, right of the center falling too fast}

```python
def recurrent_activations_to_internal_state(recurrent_activations):

    internal_state = set()

    for activation, region in zip(recurrent_activations, regions):
        if activation > 0.5:
            internal_state.add(region.__name__)

    return internal_state
```



Did we satisfy our requirements?

- v0
  - ✓ Internal states are casually reducible to brain states
    - Internal states are ontologically irreducible to brain states

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

Did we satisfy our requirements?

- v0
  - ✓ Internal states are casually reducible to brain states
    - Internal states are ontologically irreducible to brain states

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

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 at time t (set of layer weights, activations, and connectivity)
- A region the agent believes it's in
- Internal state at time t (set of regions the agent believes it's in)

Did we satisfy our requirements?

- v0
  - ✓ Internal states are casually reducible to brain states
    - Internal states are ontologically irredu tates

Instances of class A are ontologically reduci class B
if and only if instances of A's are nothing bu

- 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

(C) Case of wine

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 at time t (set of layer weights, activations, and connectivity)
- A region the agent believes it's in
- Internal state at time t (set of regions the agent believes it's in)

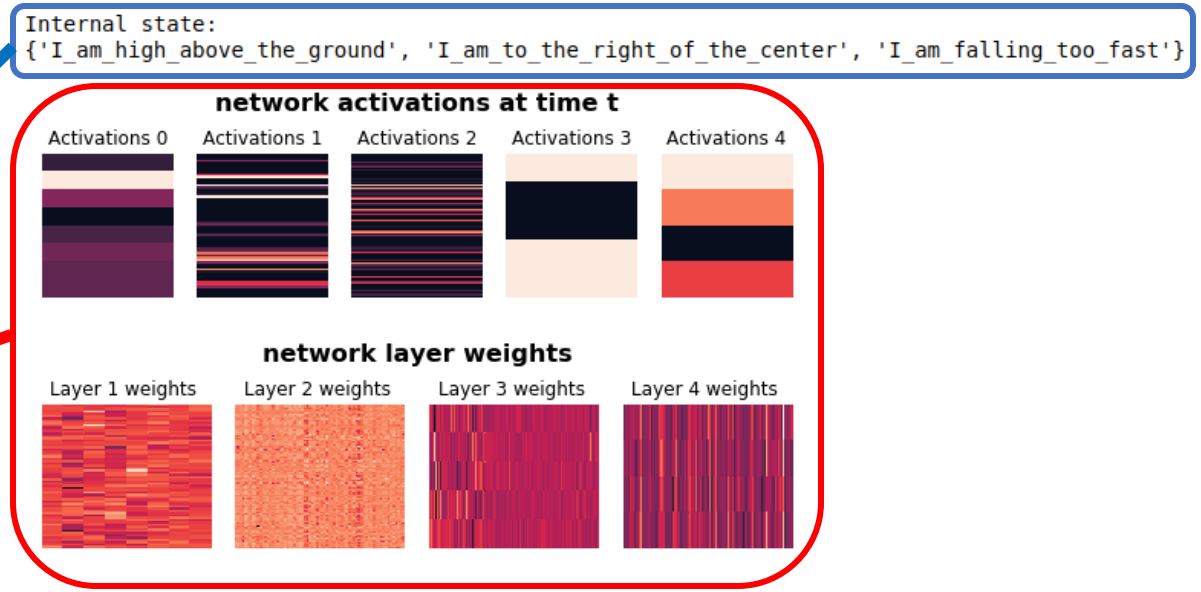
Did we satisfy our requirements?

- v0
  - ✓ Internal states are casually reducible to brain states
    - Internal states are ontologically irreducible to brain states

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

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 at time t (set of layer weights, activations, and connectivity)
- A region the agent believes it's in
- Internal state at time t (set of regions the agent believes it's in)



Did we satisfy our requirements?

- v0
  - ✓ Internal states are casually reducible to brain states
    - Internal states are ontologically irreducible to brain states

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

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 at time t (set of layer weights, activations, and connectivity)
- A region the agent believes it's in
- Internal state at time t (set of regions the agent believes it's in)

```
Internal state:
{'I_am_high_above_the_ground', 'I_am_to_the_right_of_the_center', 'I_am_falling_too_fast'}
```

```
def I_am_high_above_the_ground(observation):
    return observation[1] > 0.5

def I_am_low_to_the_ground(observation):
    return observation[1] <= 0.5

def I_am_to_the_left_of_the_center(observation):
    return observation[0] > 0.

def I_am_to_the_right_of_the_center(observation):
    return observation[0] <= 0.

def I_am_falling_too_fast(observation):
    return observation[3] < -0.2
```

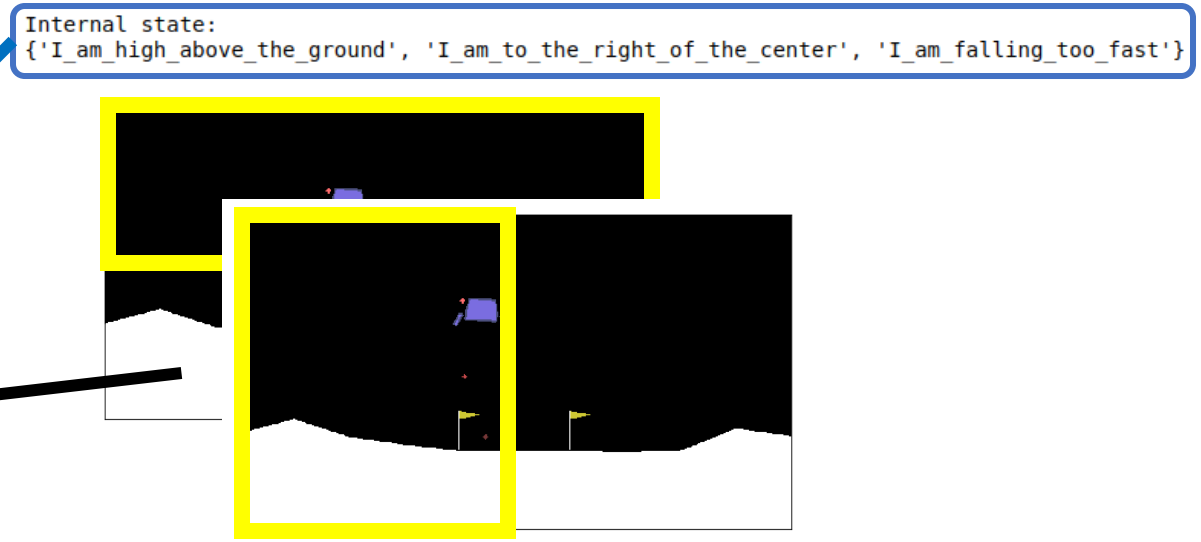
Did we satisfy our requirements?

- v0
  - ✓ Internal states are casually reducible to brain states
    - Internal states are ontologically irreducible to brain states

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

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 at time t (set of layer weights, activations, and connectivity)
- A region the agent believes it's in
- Internal state at time t (set of regions the agent believes it's in)



Did we satisfy our requirements?

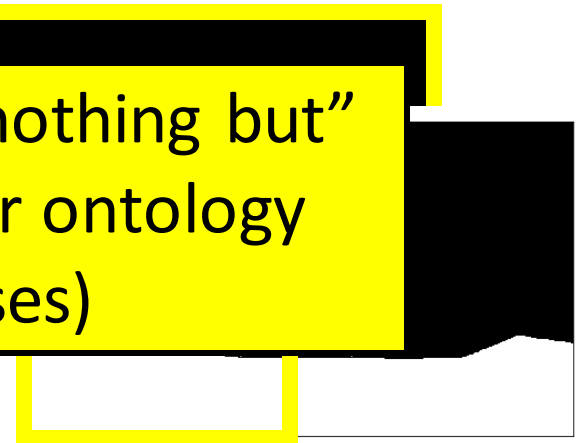
- v0
  - ✓ Internal states are casually reducible to brain states
    - Internal states are ontologically irreducible to brain states

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

Our ontology

- Layer weights of the neural network
- Connectivity of the neural network
- Activations of the neural ne
- The agent's observation at
- The agent's action at time t
- The position and velocity o
- Brain state at time t (set of
- A region the agent believes it's in
- Internal state at time t (set of regions the agent believes it's in)

Internal state:
{'I_am_high_above_the_ground', 'I_am_to_the_right_of_the_center', 'I_am_falling_too_fast'}



Mental state instances are not "nothing but" brain state instances under our ontology (they are different classes)

Did we satisfy our requirements?

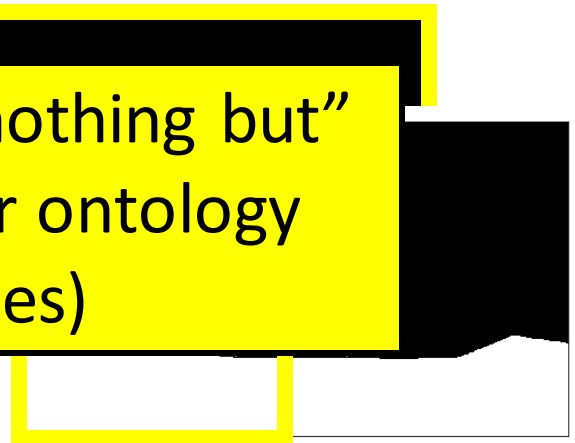
- v0
  - ✓ Internal states are casually reducible to brain states
  - ✓ Internal states are ontologically irreducible to brain states

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

Our ontology

- Layer weights of the neural network
- Connectivity of the neural network
- Activations of the neural ne
- The agent's observation at
- The agent's action at time t
- The position and velocity o
- Brain state at time t (set of
- A region the agent believes it's in
- Internal state at time t (set of regions the agent believes it's in)

Internal state:
{'I_am_high_above_the_ground', 'I_am_to_the_right_of_the_center', 'I_am_falling_too_fast'}



Mental state instances are not "nothing but" brain state instances under our ontology (they are different classes)

Is that the “real” ontology though?

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

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

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 at time t (set of layer weights, activations, and connectivity)
- A region the agent believes it's in
- Internal state at time t (set of regions the agent believes it's in)

Is that the “real” ontology though?

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

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

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 at time t
- A region the agent believes it's in
- Internal state at time t (set of regions the agent believes it's in)

- Bits
- Python objects
- Electrons
- Quarks
- …

Is that the “real” ontology though?

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

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

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 at time t (all of the bits contained in my computer)
- A region the agent believes it's in
- Internal state at time t (set of regions the agent believes it's in)

- Bits
- Python objects
- Electrons
- Quarks
- …

Is that the “real” ontology though?

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

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

Our ontology

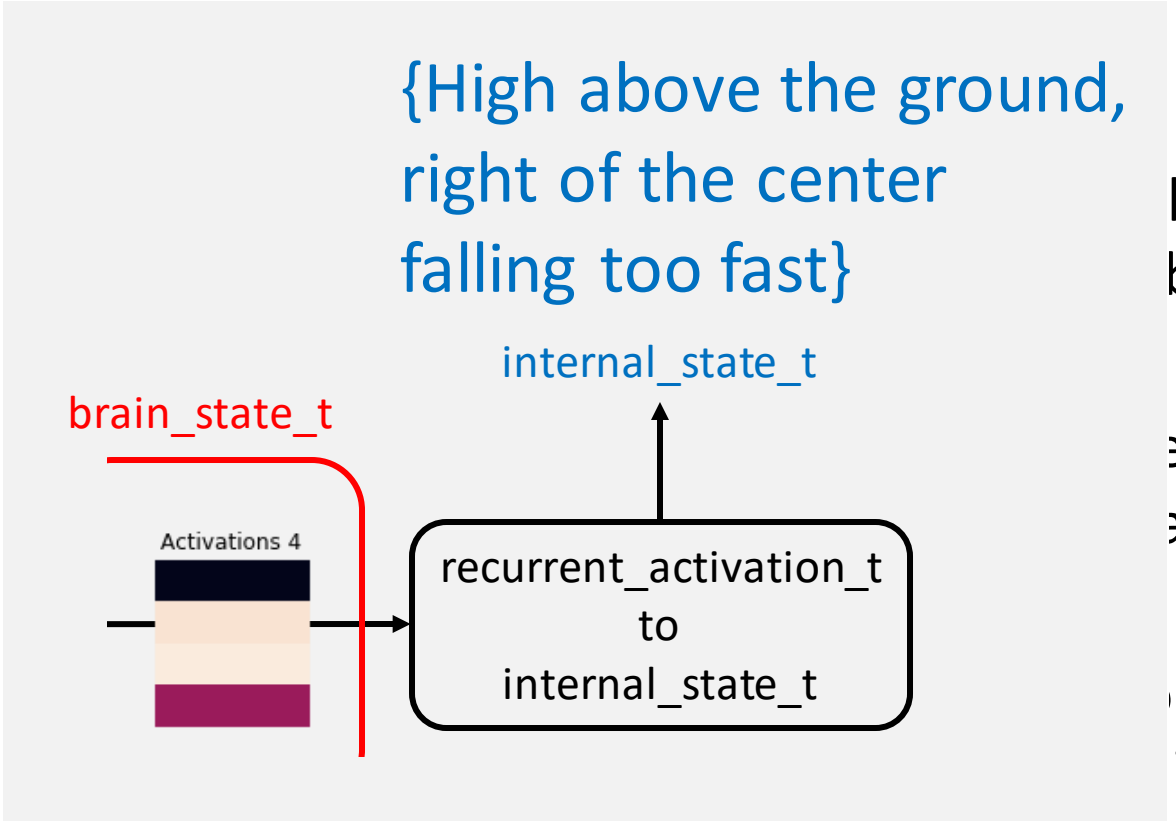
- Layer weights of the neural network
- Connectivity of the neural network
- Activations of the neural network at time t
- The agent's observation at time t
- The agent's action at time t
- The position and velocity of the agent at time t
- Brain state at time t (all of the bits contained in my computer)
- A region the agent believes it's in
- Internal state at time t (set of regions the agent believes it's in)

- Bits
- Python objects
- Electrons
- Quarks
- …

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



brain states

bl

```
def recurrent_activations_to_internal_state(recurrent_activations):

    internal_state = set()

    for activation, region in zip(recurrent_activations, regions):
        if activation > 0.5:
            internal_state.add(region.__name__)

    return internal_state
```

ec

all

brain states

ible 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?
  - Great! Open an issue and/or submit a pull request in GitHub
- Thoughts on other theories of mind/consciousness that may be particularly well suited for this type of approach?