

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



Why attempt to reify philosophy with code

- Lots of the words philosophers use describing aspects of consciousness tends shows up in CS/AI research
 - Mind, awareness, imagination, reasoning, consciousness, etc.

Why attempt to reify philosophy with code

- Lots of the words philosophers use describing aspects of consciousness tends shows up in CS/AI research
 - Mind, awareness, imagination, reasoning, consciousness, etc.
- (Disclaimer: our backgrounds are CS/AI)

Why attempt to reify philosophy with code

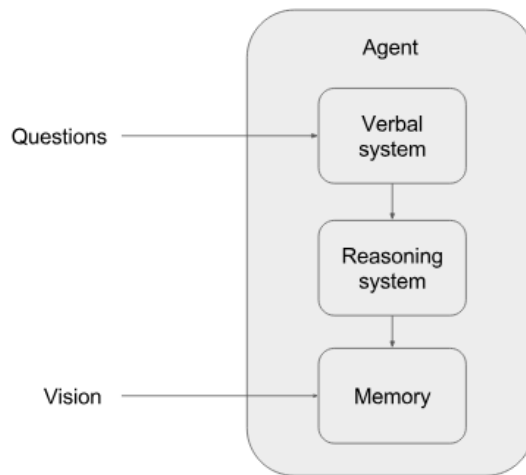
- Lots of the words philosophers use describing aspects of consciousness tends shows up in CS/AI research
 - Mind, awareness, imagination, reasoning, consciousness, etc.
- Our intuition is CS/AI could benefit from a deeper understanding of philosophy
 - But telling people to read more books/papers is not how to make this happen
 - So let's try to do it with code!
- (Disclaimer: our backgrounds are CS/AI)

Why attempt to reify philosophy with code

- Lots of the words philosophers use describing aspects of consciousness tends shows up in CS/AI research
 - Mind, awareness, imagination, reasoning, consciousness, etc.
- Our intuition is CS/AI could benefit from a deeper understanding of philosophy
 - But telling people to read more books/papers is not how to make this happen
 - So let's try to do it with code!
- Possibly benefit philosophy by bringing code-style concreteness
 - (TBD, will let the philosophers in the room speak to this!)
- (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



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.

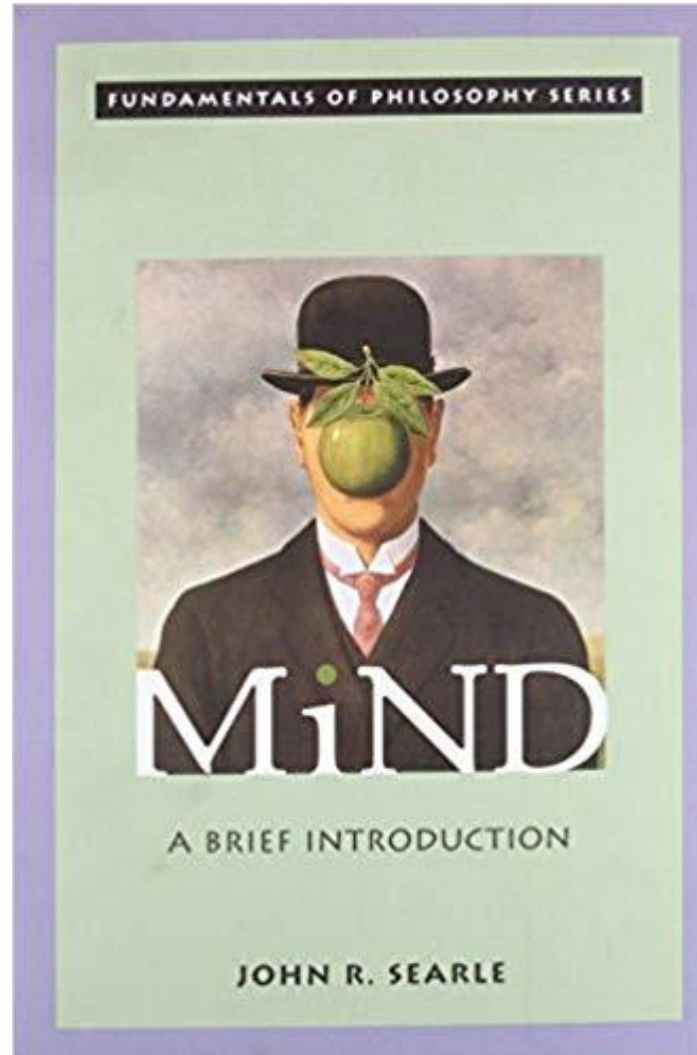
Q: Is it possible for Buck to have an illusion that he is having the experience of redness?
No, that's impossible.

Image from shlegeris.com

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

Reifying philosophy with code

Reifying philosophy with code



Reifying philosophy with code

- Searle's view of the relationship between consciousness and 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

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"

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"
- Let's unpack this with code!

What we're not doing

- Not trying to
 - Propose a cognitive architecture
 - Propose a new AI or machine learning algorithm
 - Claim that the software agent is conscious
 - Convince anyone these are the correct/best/most useful definitions of consciousness or brain states
 - Convince anyone Searle is right or wrong

What we're trying to do

- Trying to 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

- Trying to 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 the system must do
- Design – how will we build the system to meet the requirements
- Implementation – building the system consistent with the design

Requirements: unpacking Searle's view

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

Requirements: unpacking Searle's view

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

Requirements: unpacking Searle's view

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

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

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

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

Requirements: unpacking Searle's view

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

Requirements: unpacking Searle's view

- Searle's view
 - Consciousness is causally reducible to brain states
 - Consciousness is ontologically irreducible to brain states
- ...simpler
 - Conscious mental states are causally reducible to brain states
 - Conscious mental states are ontologically irreducible to brain states

Requirements: unpacking Searle's view

- Searle's view
 - Consciousness is causally reducible to brain states
 - Consciousness is ontologically irreducible to brain states
- ...simpler
 - Conscious mental states are causally reducible to brain states
 - Conscious mental states are ontologically irreducible to brain states
- ...simpler
 - Mental states are causally reducible to brain states
 - Mental states are ontologically irreducible to brain states

Requirements: unpacking Searle's view

- Searle's view
 - Consciousness is causally reducible to brain states
 - Consciousness is ontologically irreducible to brain states
- ...simpler
 - Conscious mental states are causally reducible to brain states
 - Conscious mental states are ontologically irreducible to brain states
- ...simpler
 - Mental states are causally reducible to brain states
 - Mental states are ontologically irreducible to brain states
- ...simpler
 - Internal states are causally reducible to brain states
 - Internal states are ontologically irreducible to brain states

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 causally reducible to brain states
 - Conscious mental states are ontologically irreducible to brain states
- V1
 - Mental states are causally reducible to brain states
 - Mental states are ontologically irreducible to brain states
- V0
 - Internal states are causally reducible to brain states
 - Internal states are ontologically irreducible to brain states

Requirements: unpacking Searle's view

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

Requirements: unpacking Searle's view

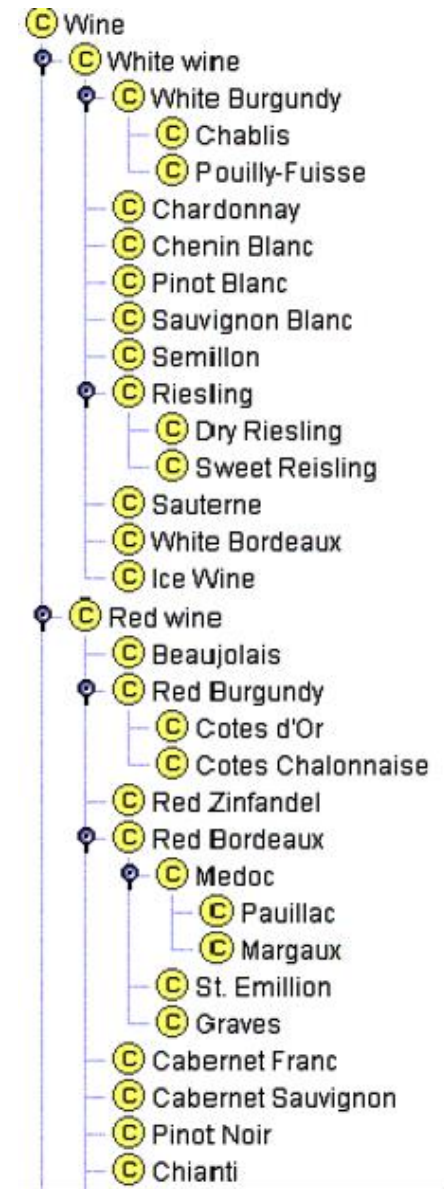
- V0
 - Internal states are causally reducible to brain states
 - Internal states are **ontologically irreducible** to brain states

Requirements: unpacking Searle's view

- V0
 - Internal states are causally reducible to brain states
 - Internal states are **ontologically irreducible** to brain states

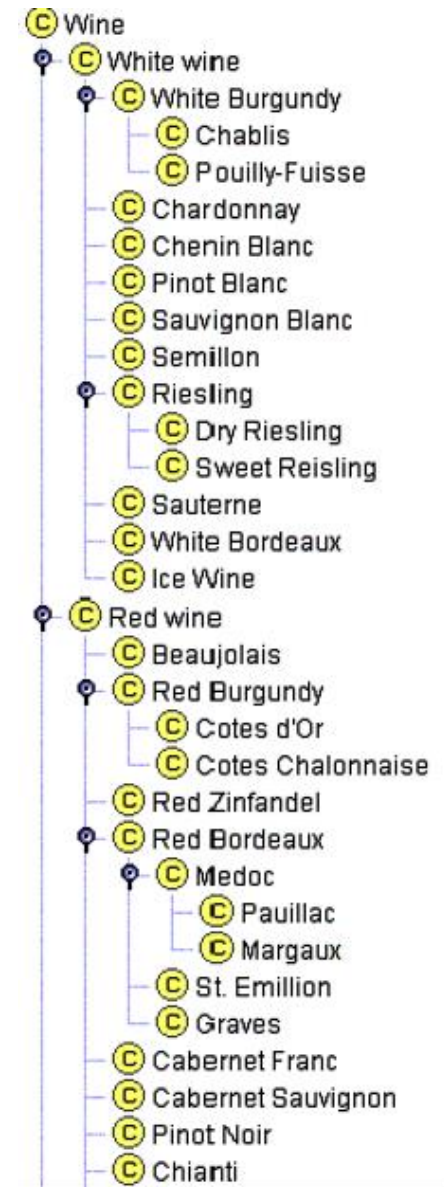
Phenomena of type A are ontologically reducible to phenomena of type B
if and only if A's are nothing but B's

Ontologies in Computer Science



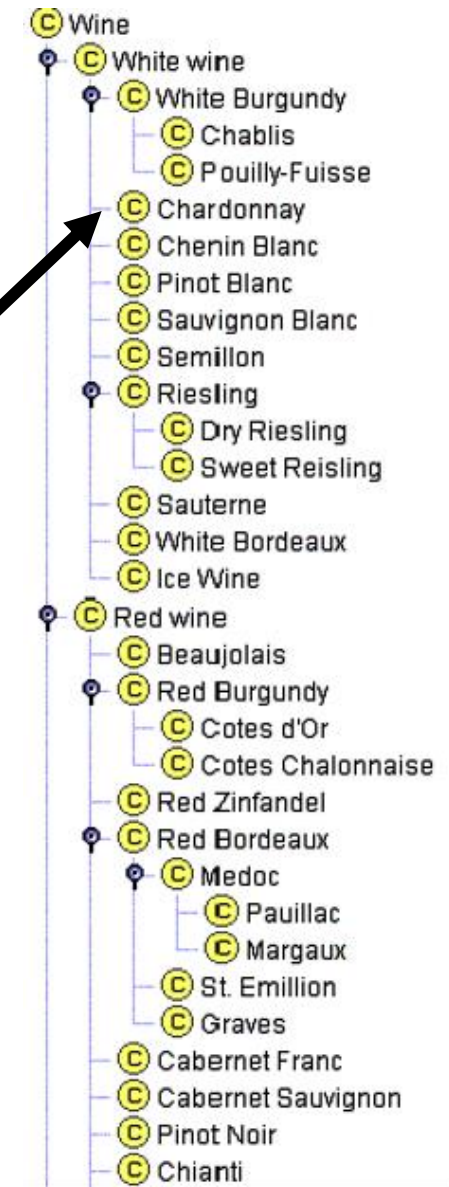
Ontologies in Computer Science

- Class-instance distinction



Ontologies in Computer Science

- Class-instance distinction



(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



- 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

(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



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

(C) A set of wine bottles

(C) Case of wine

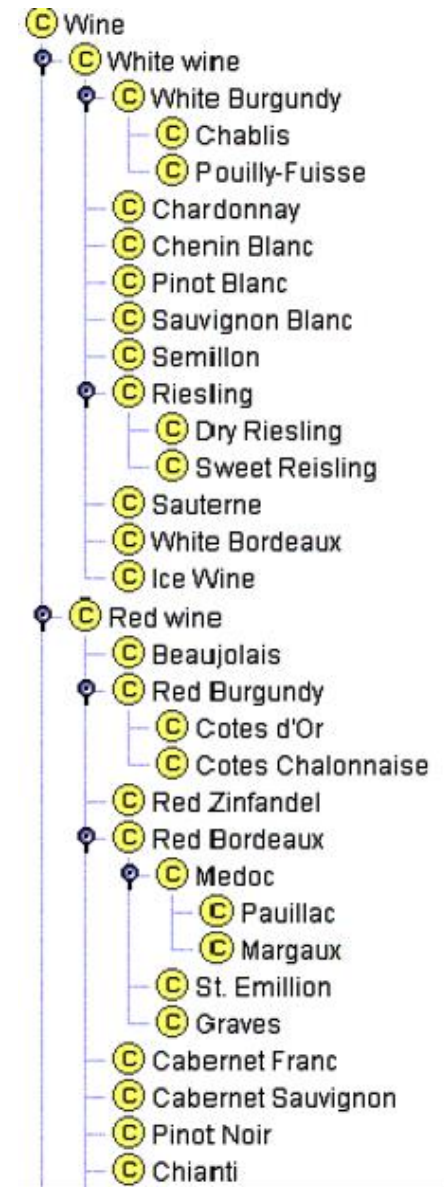
Images from:

<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



(C) A set of wine bottles

(C) Case of wine

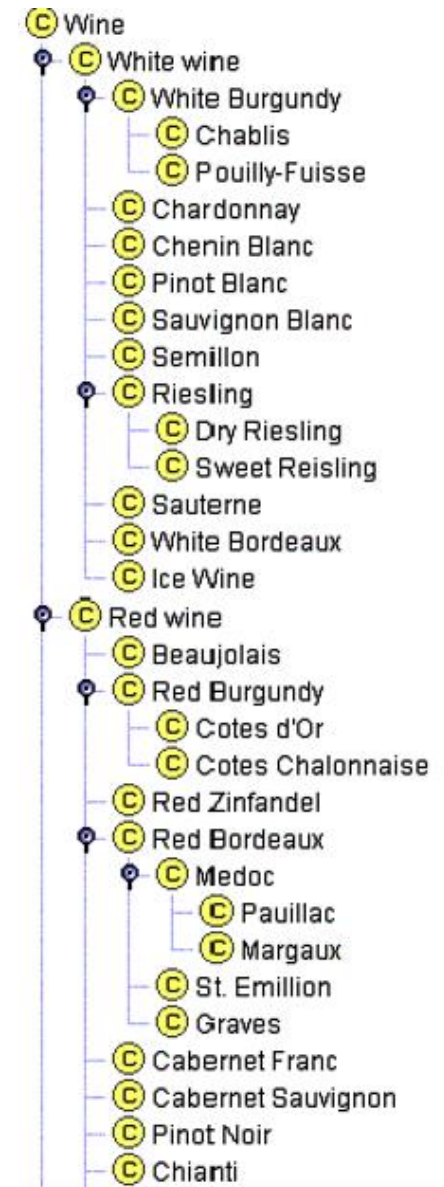
Images from:

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)



(C) A set of wine bottles

(C) Case of wine

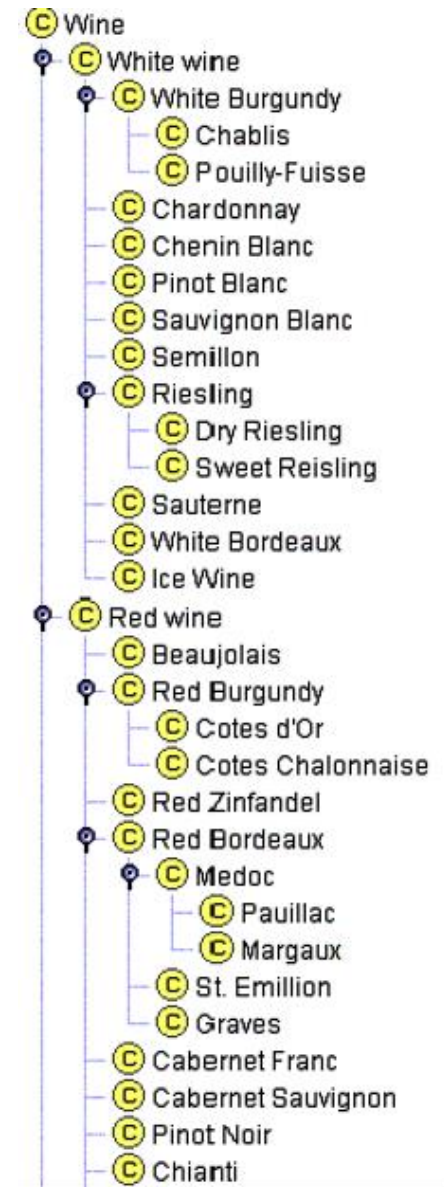
Images from:

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

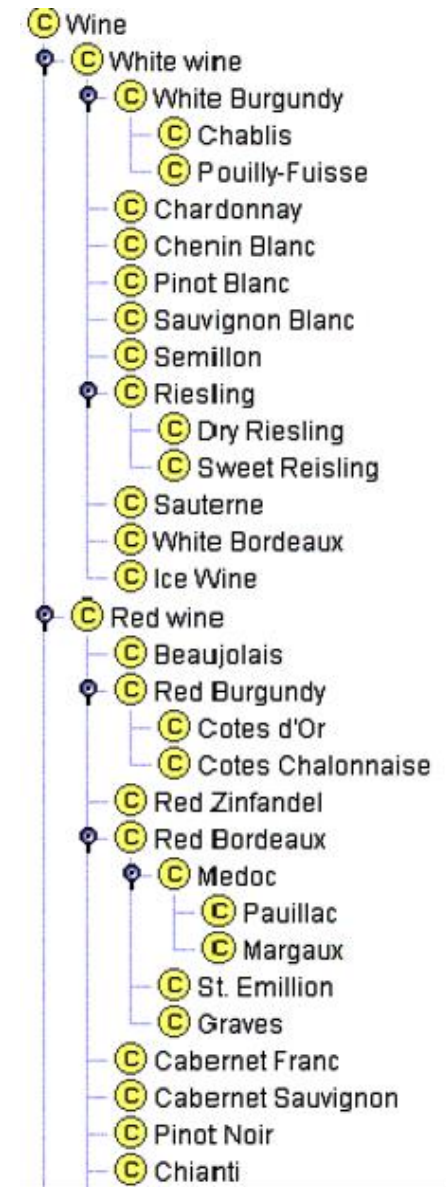
Images from:

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



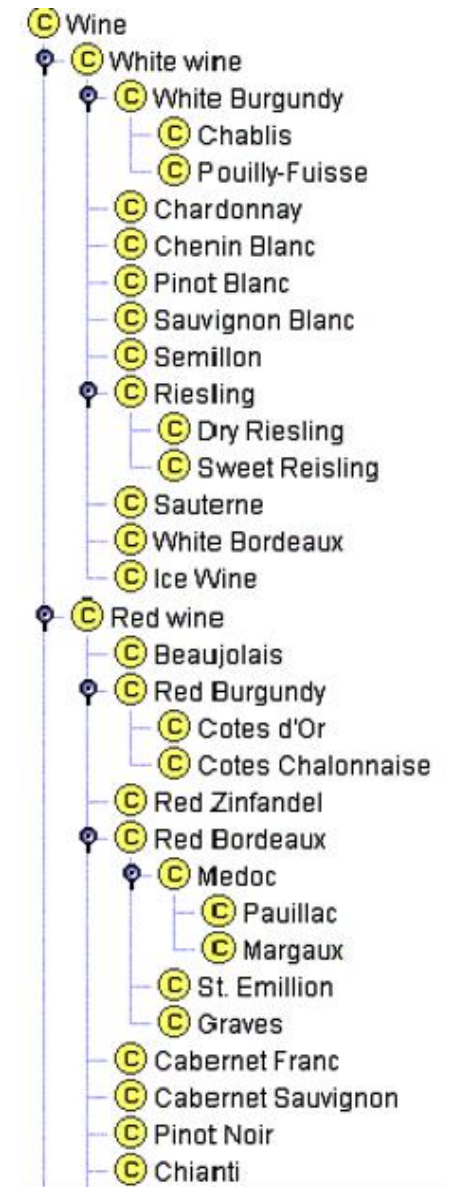
- (C) A set of wine bottles
- (C) Case of wine

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



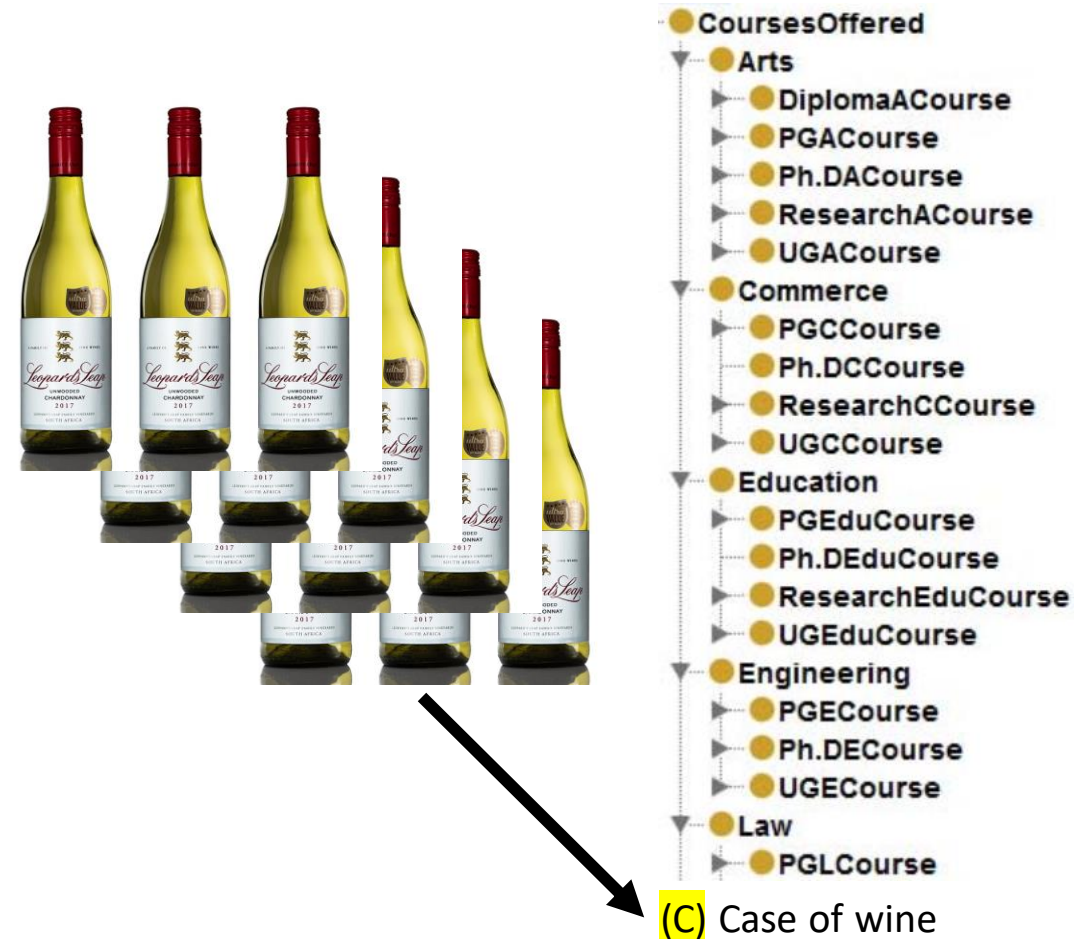
- (C) A set of wine bottles
- (C) Case of wine

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



Requirements: unpacking Searle's view

- V0
 - Internal states are causally 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

Requirements: unpacking Searle's view

- V0
 - Internal states are **causally reducible** to brain states
 - Internal states are ontologically irreducible to brain states

Requirements: unpacking Searle's view

- V0
 - Internal states are **causally 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 causally explained by the behavior of B's
- A's have no causal powers in addition to the powers of B's

Requirements, V0

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

Design, V0

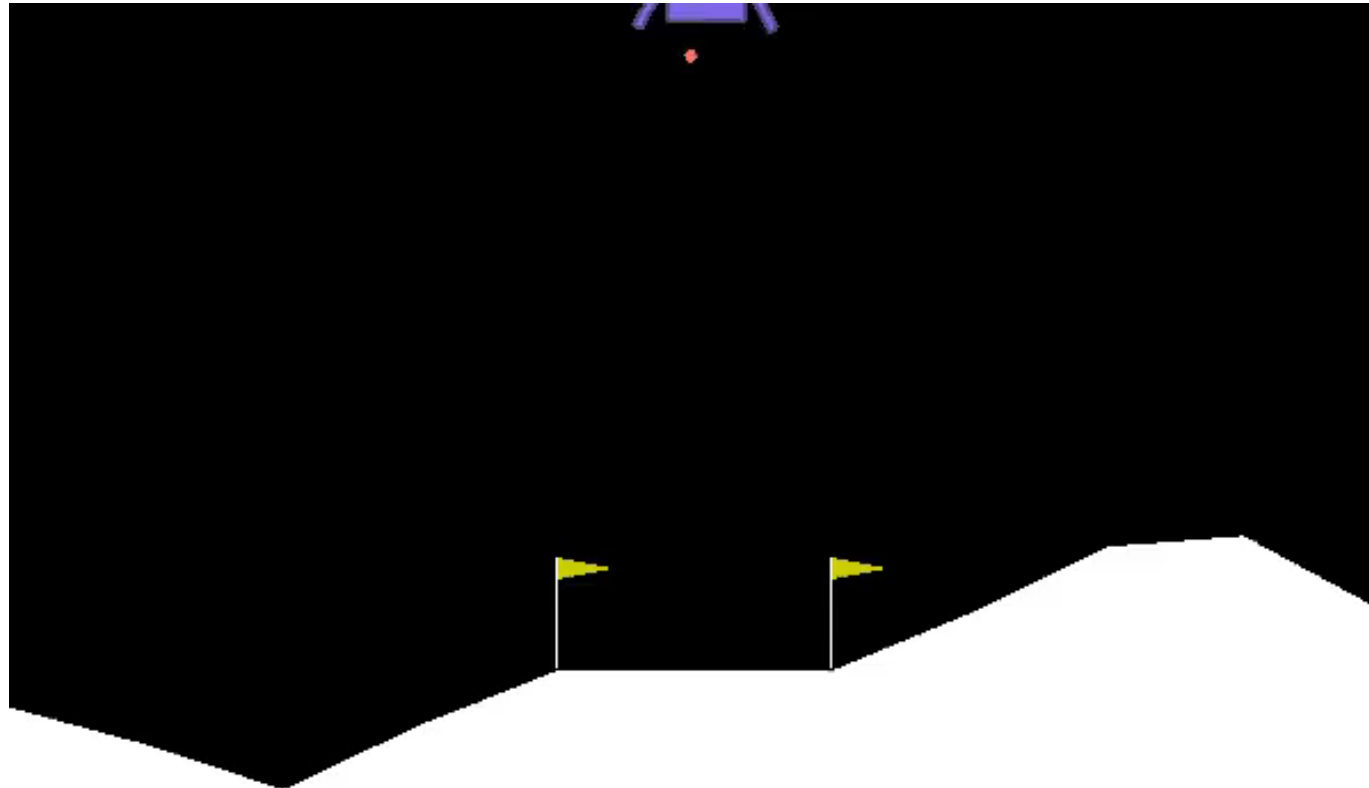
- Design decisions

Design, V0

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

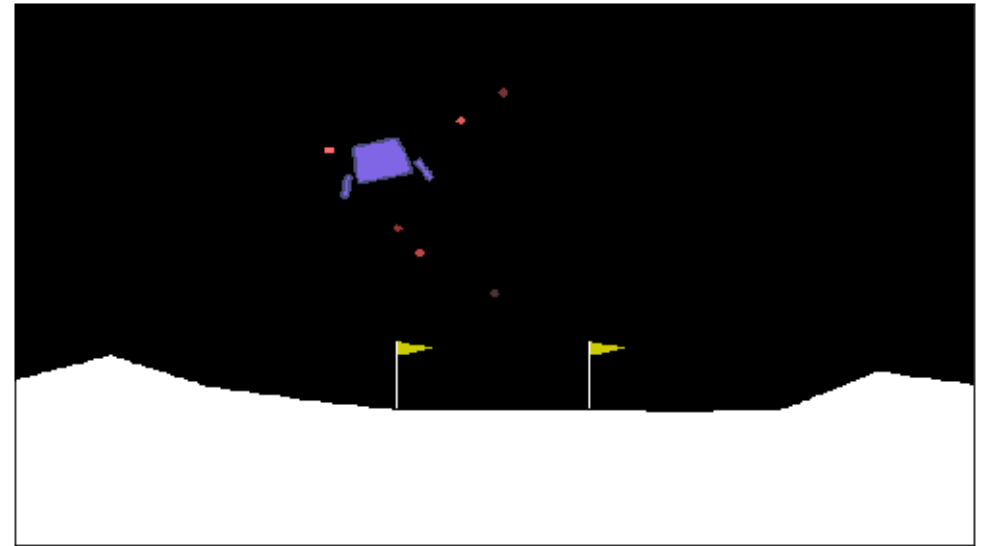
Design, V0

- OpenAI's LunarLander benchmark environment



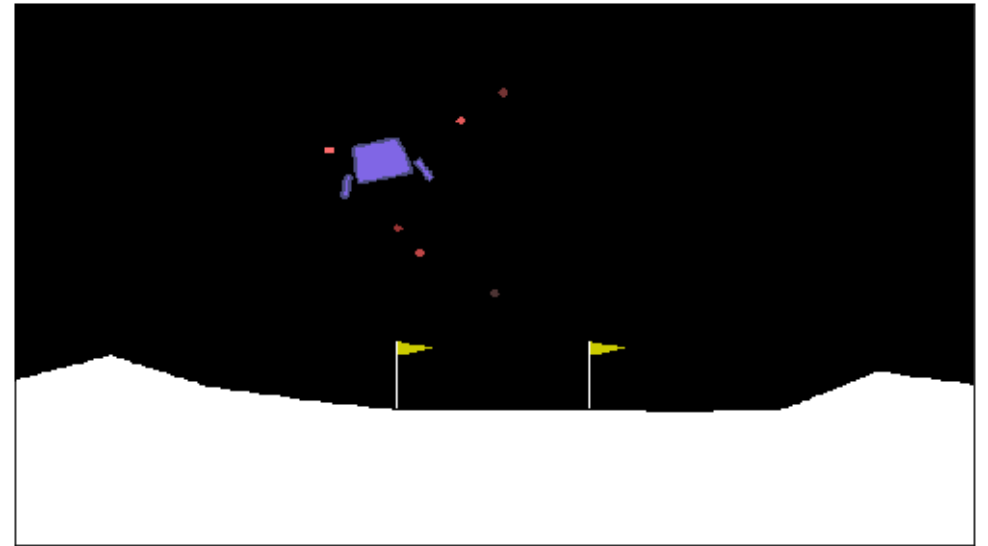
Design, V0

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



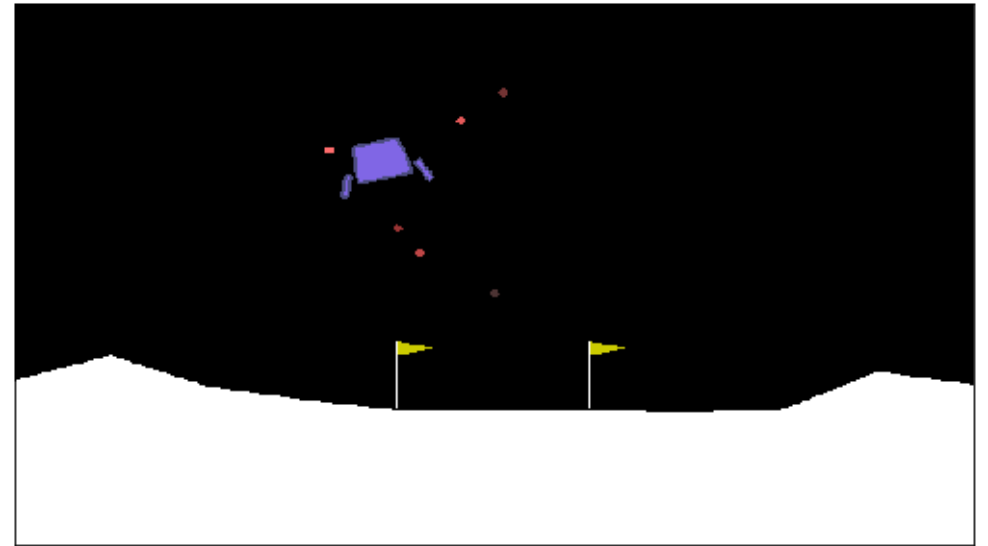
Design, V0

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



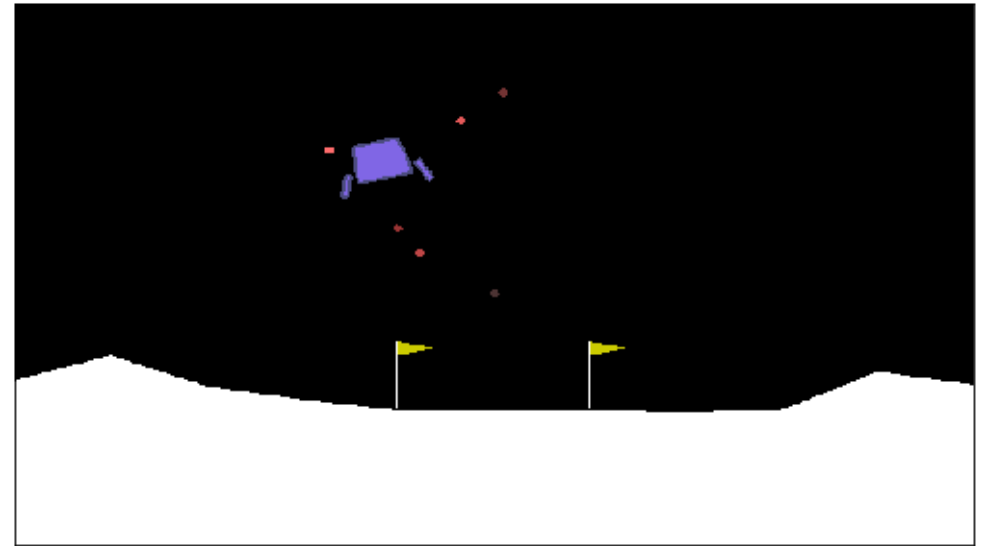
Design, V0

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



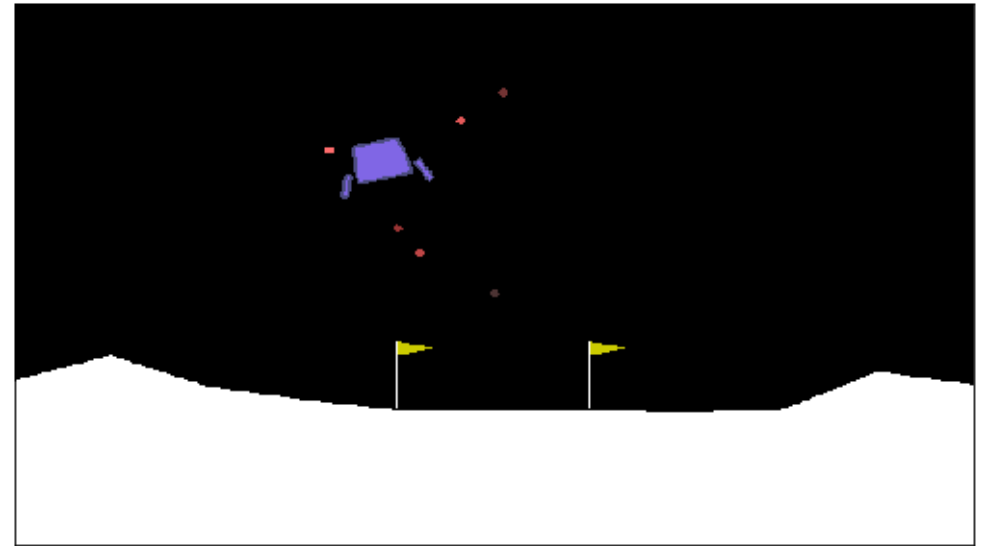
Design, V0

- 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



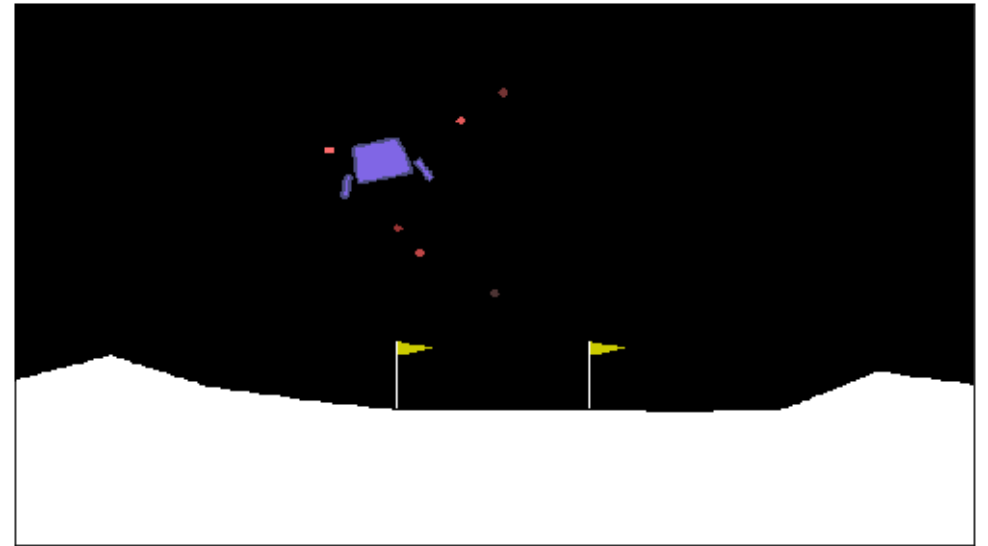
Design, V0

- 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

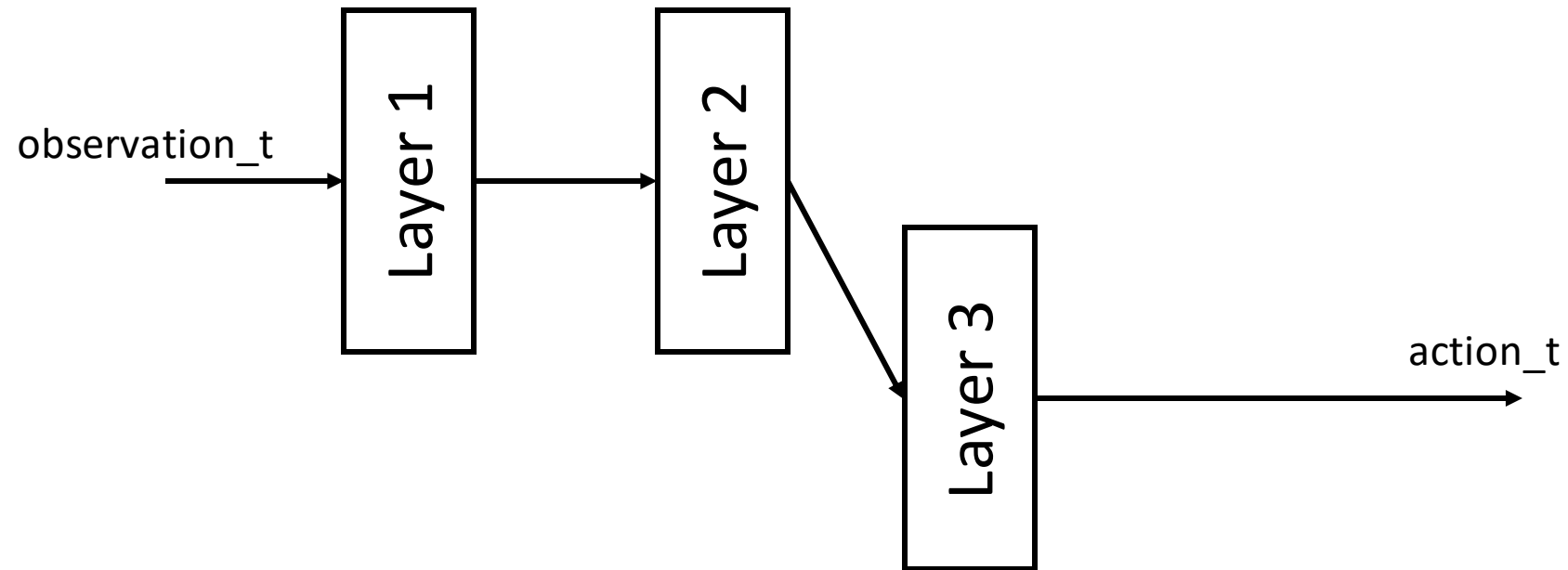


Design, V0

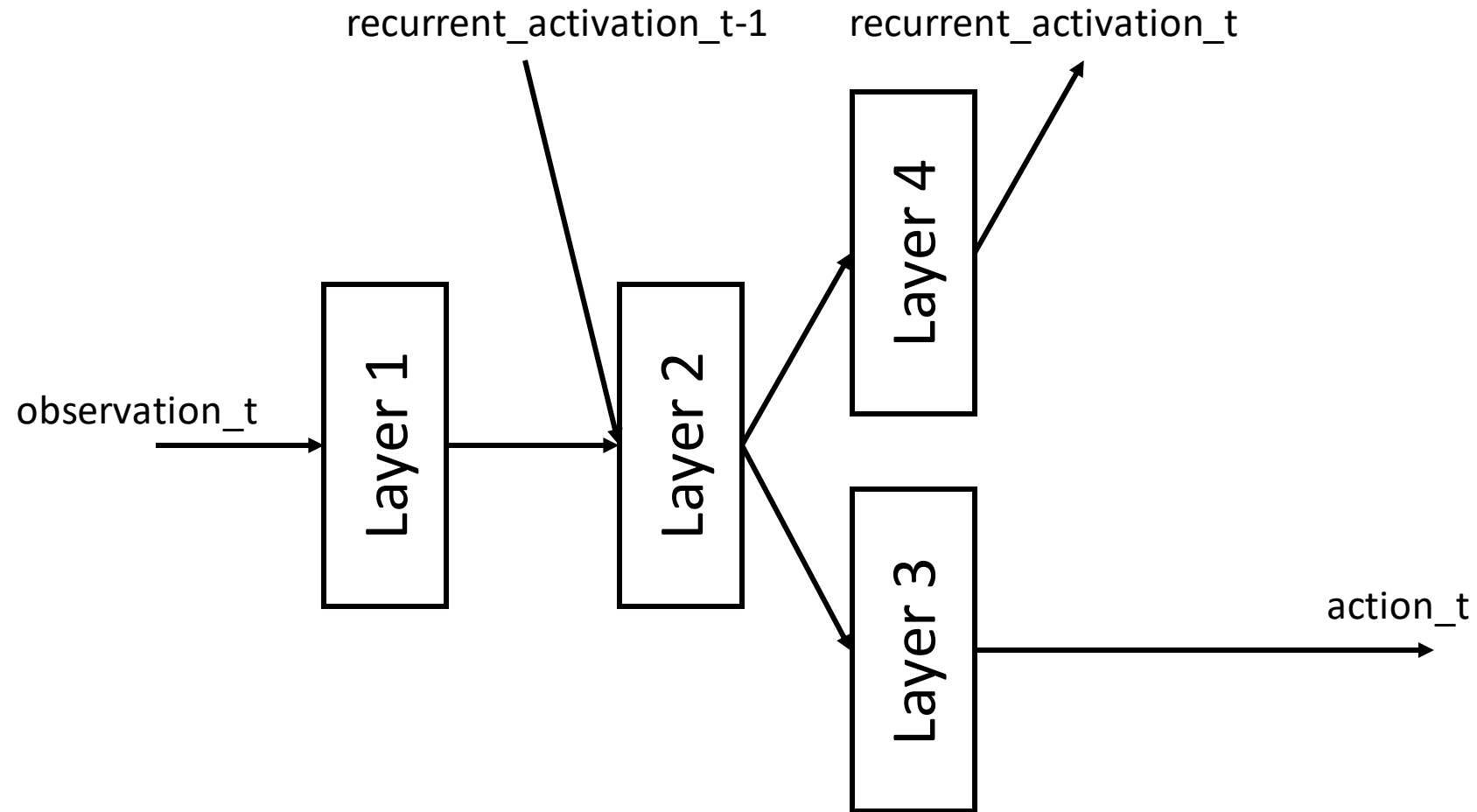
- Design decisions
 - Environment and the agent's "physical" form
 - Internal state of the agent
 - Beliefs about itself relative to semantically important regions
 - Left of the flags, right of the flags, high above the ground, close to the ground, falling too fast
 - Brain state of the agent



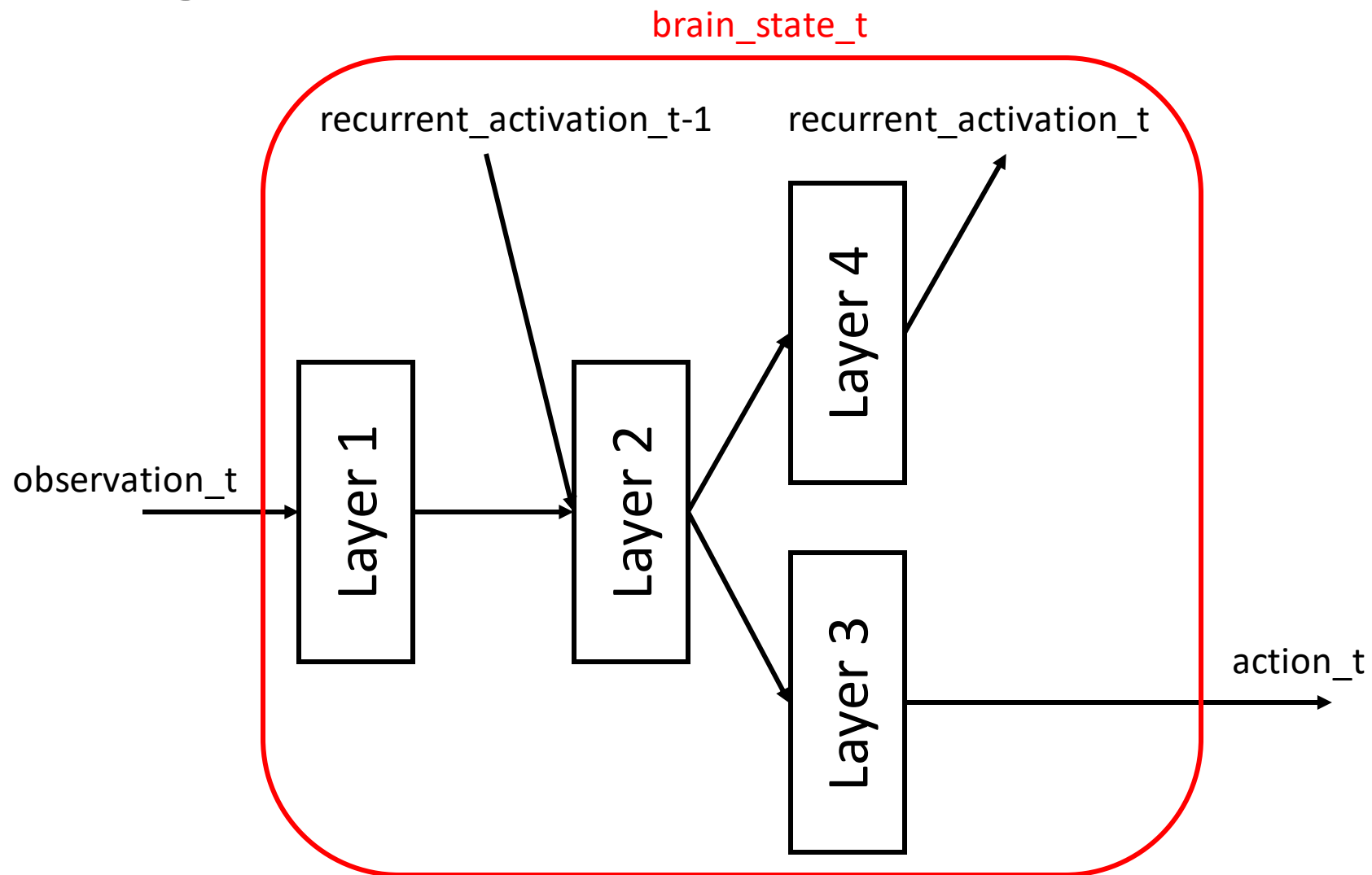
Design, V0



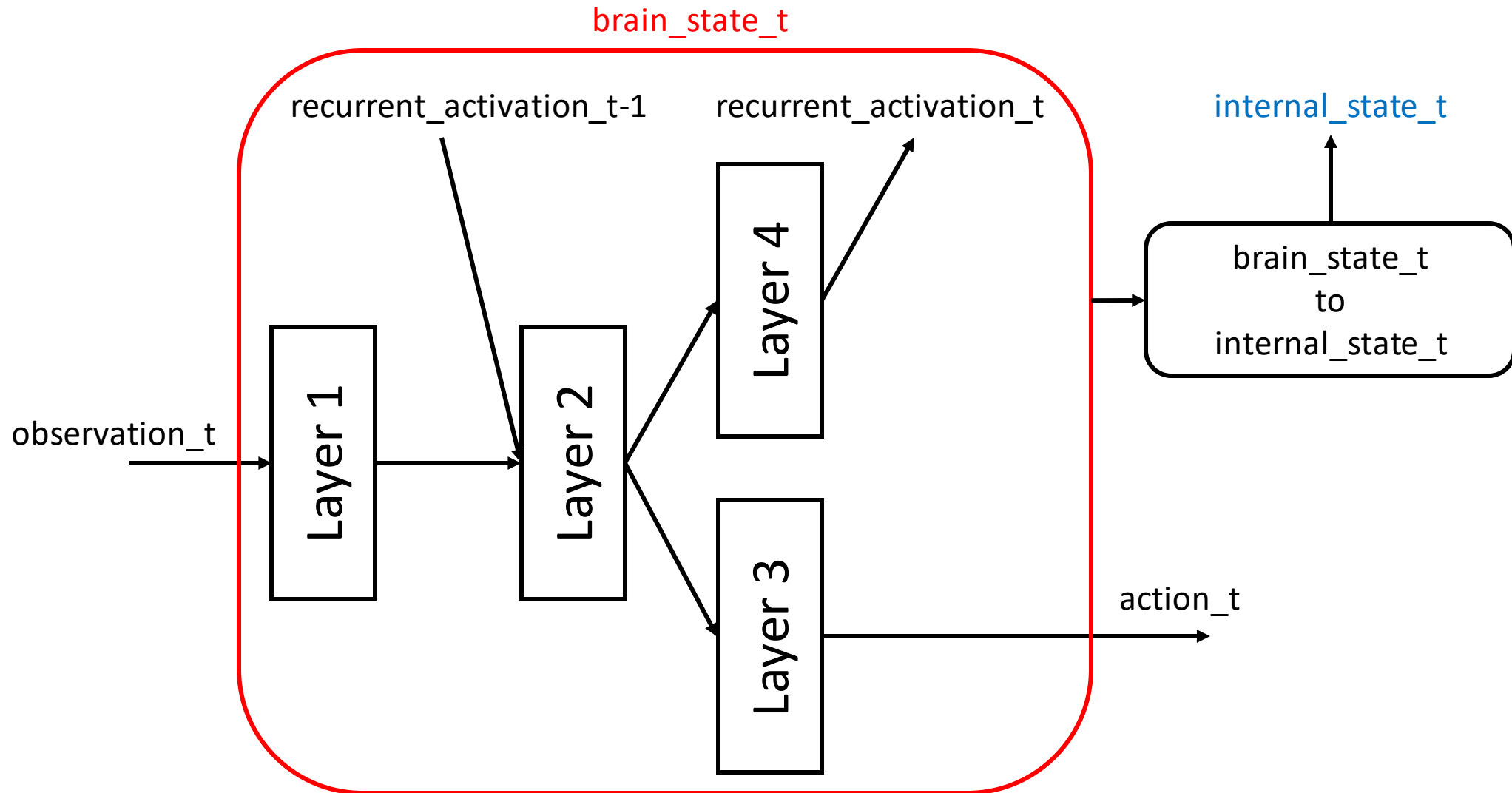
Design, V0



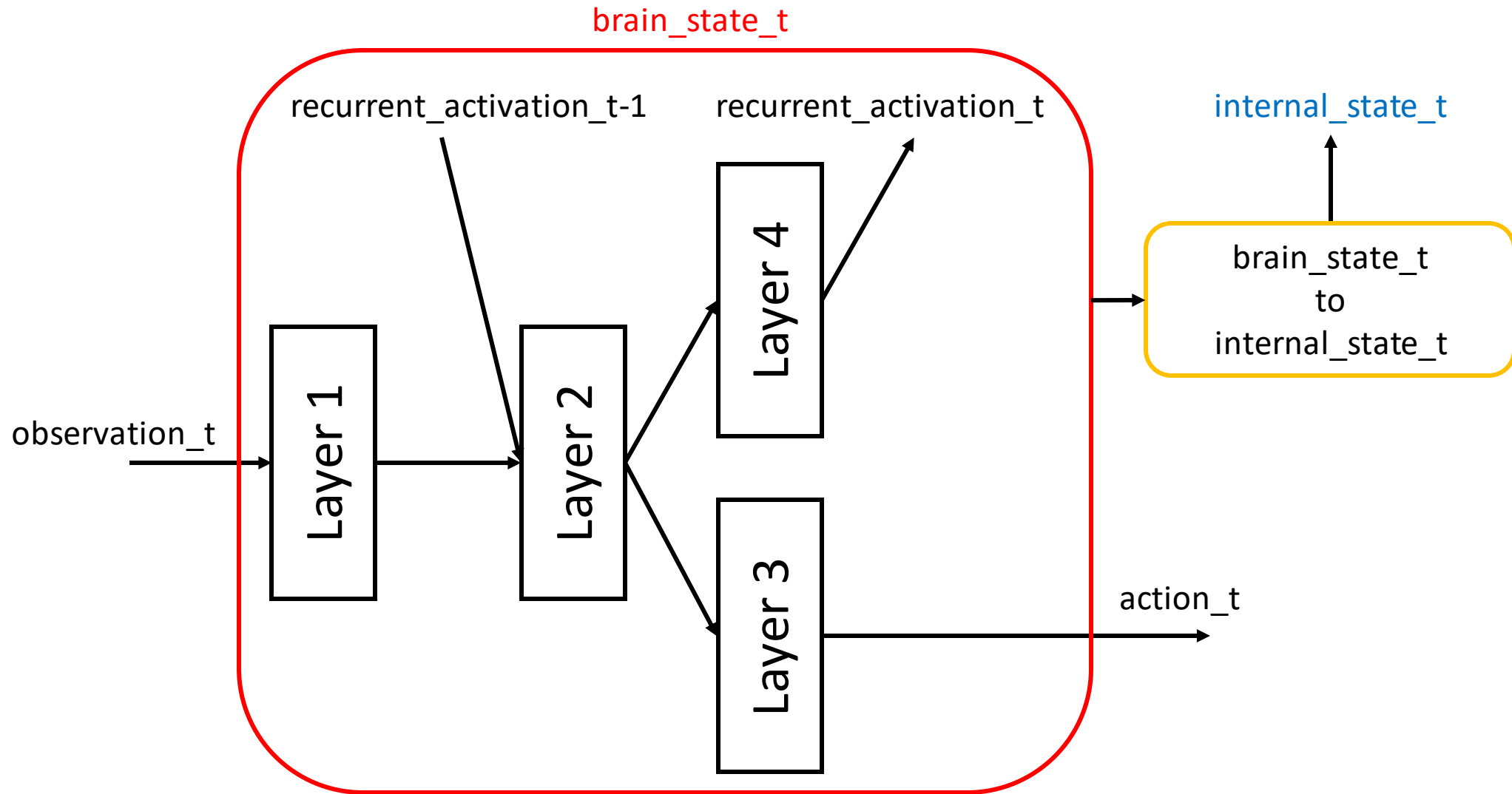
Design, V0



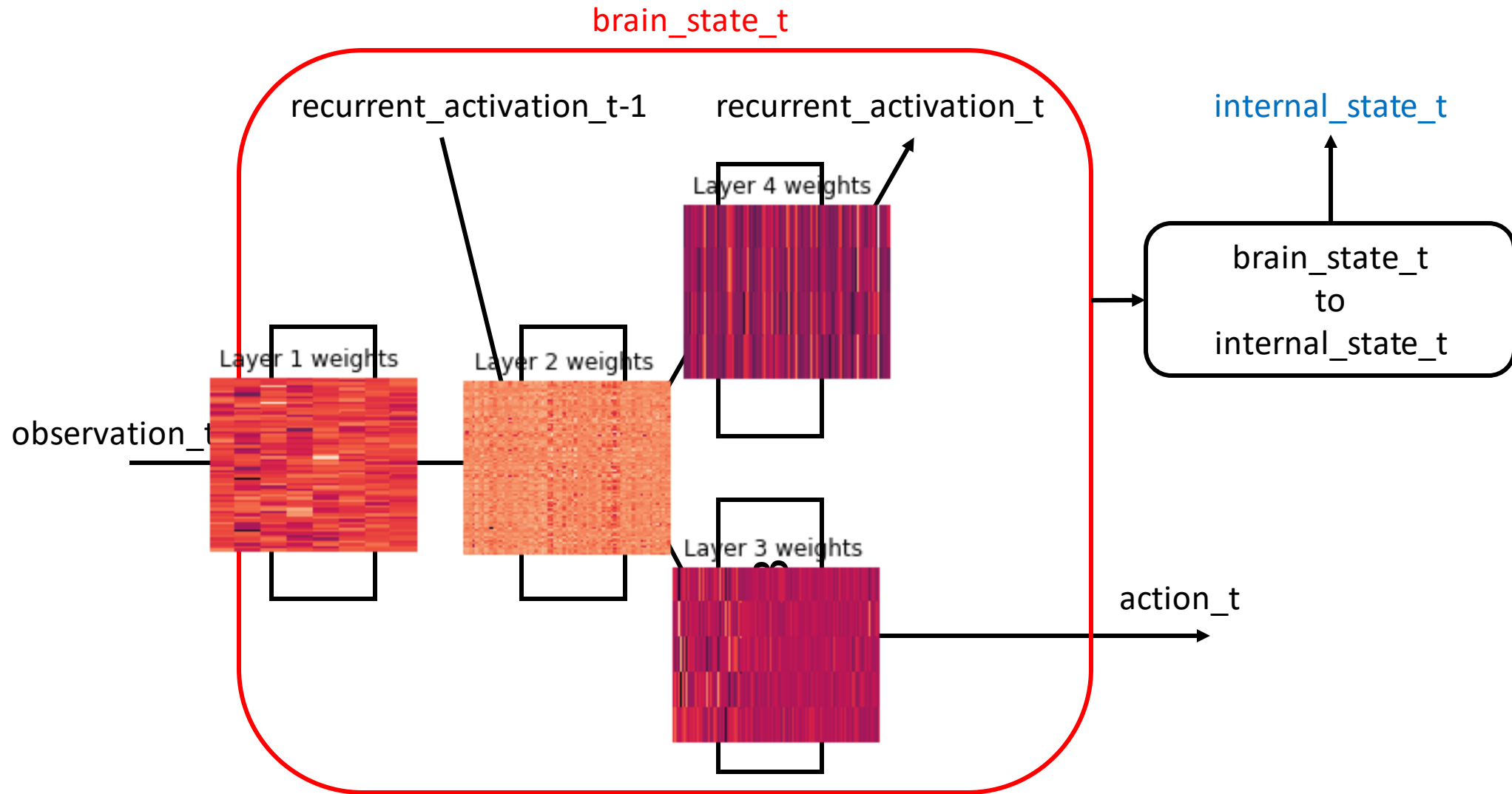
Design, V0



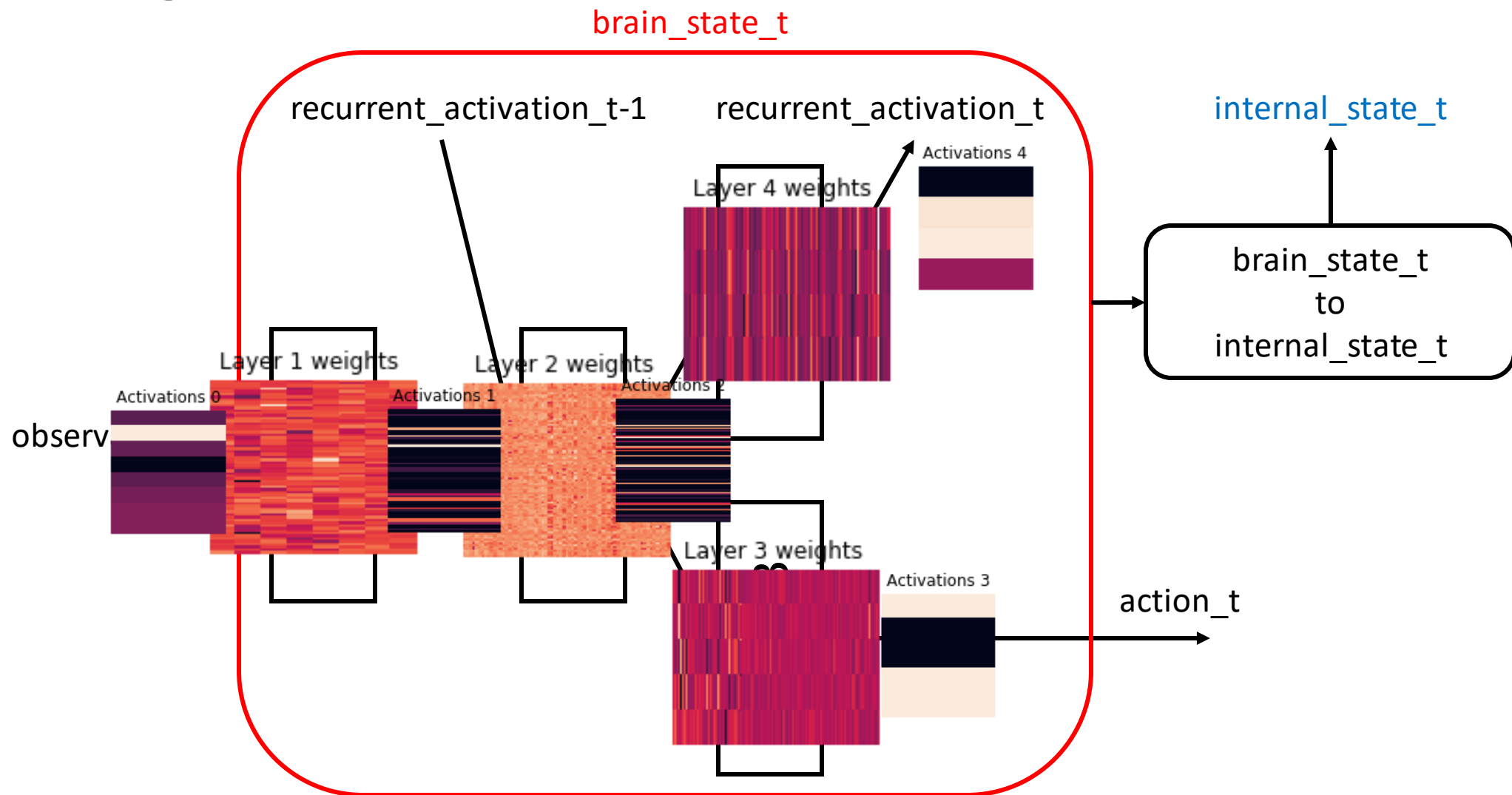
Design, V0



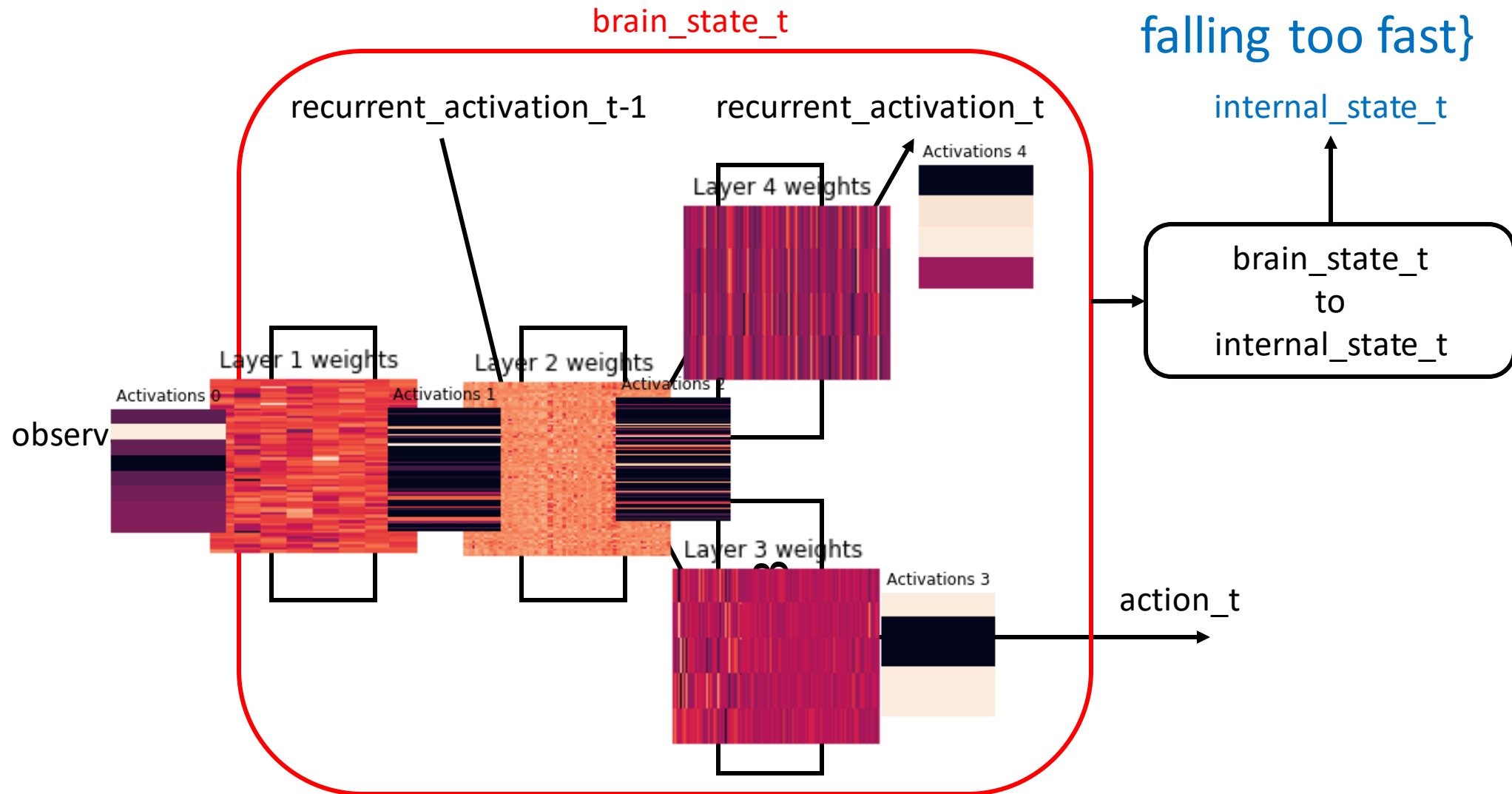
Design, V0



Design, V0



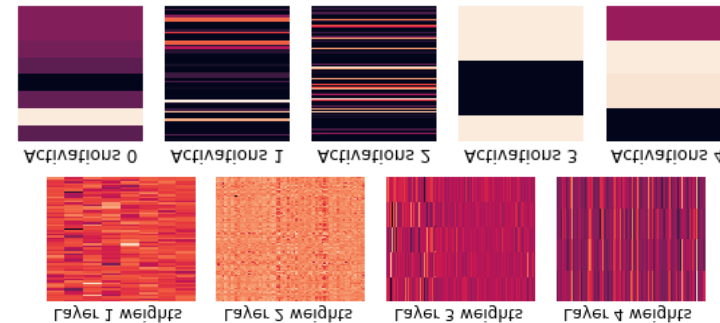
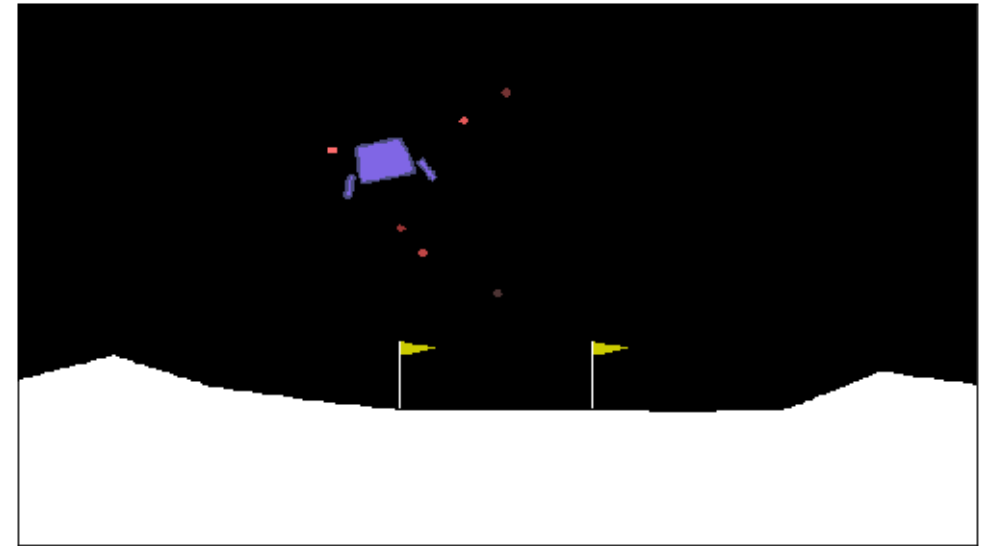
Design, V0



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

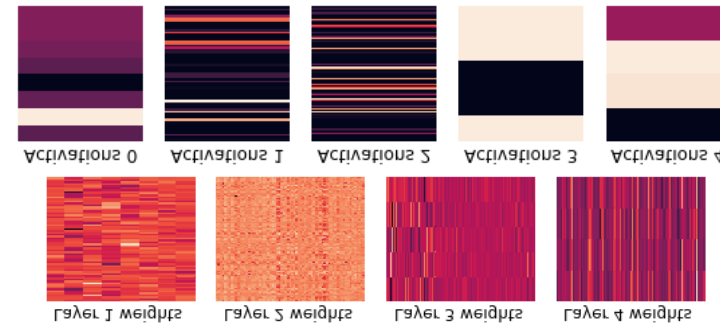
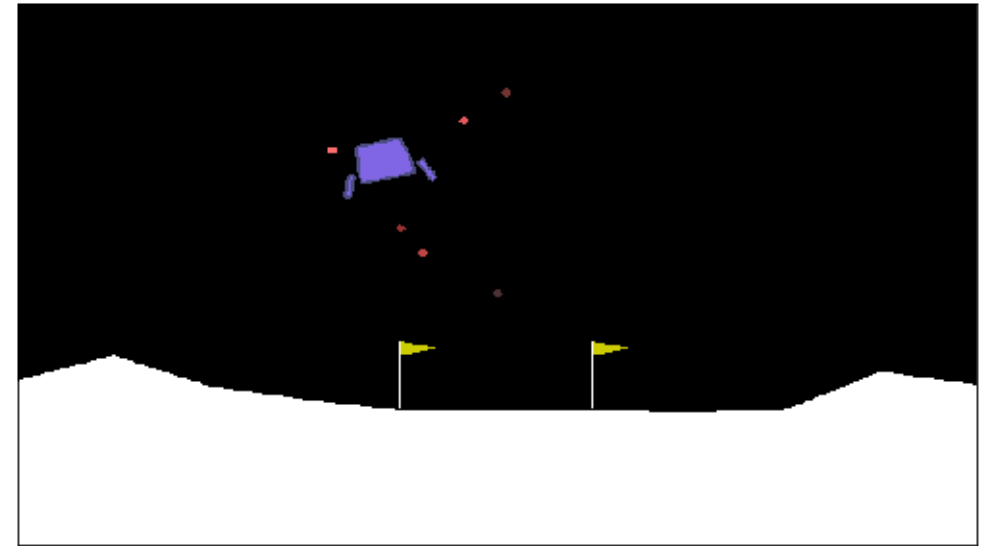
Design, V0

- 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



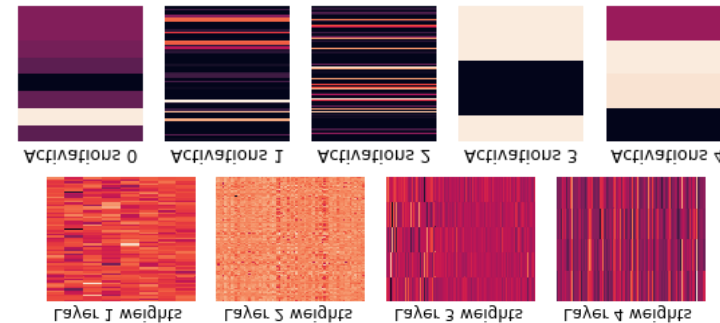
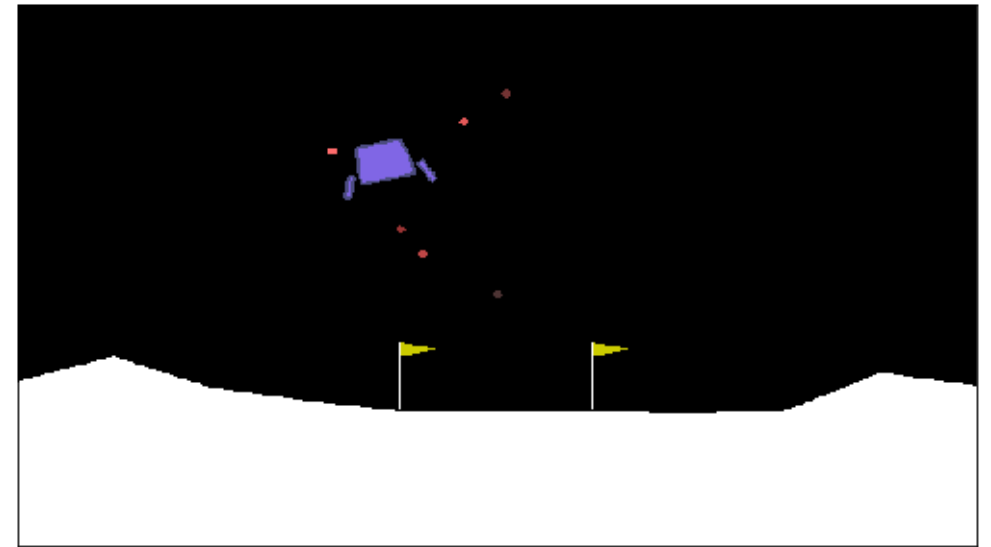
Design, V0

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



Reinforcement learning

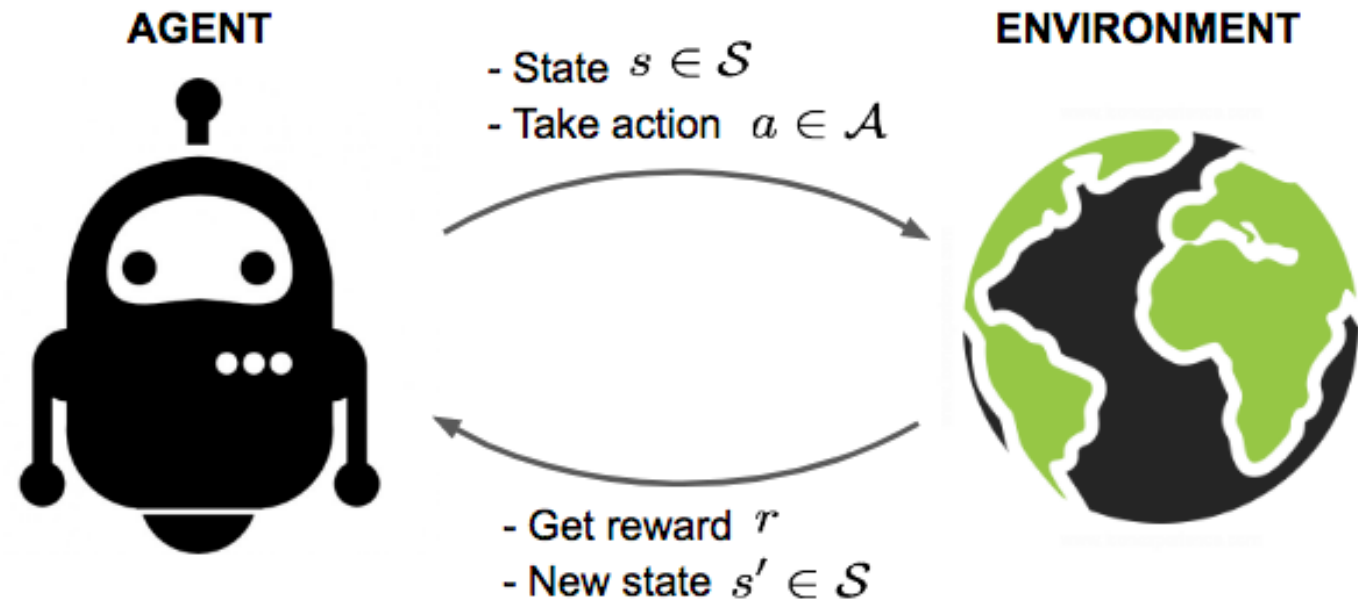


Image from:

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

Implementation, V0

- Jupyter notebook time!
 - <http://localhost:8888/notebooks/notebooks/TSC-2019.ipynb>
 - <https://github.com/Josh-Joseph/tsc-2019/blob/master/notebooks/TSC-2019.ipynb>

Did we satisfy our requirements?

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

Did we satisfy our requirements?

- V0
 - Internal states are causally 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:

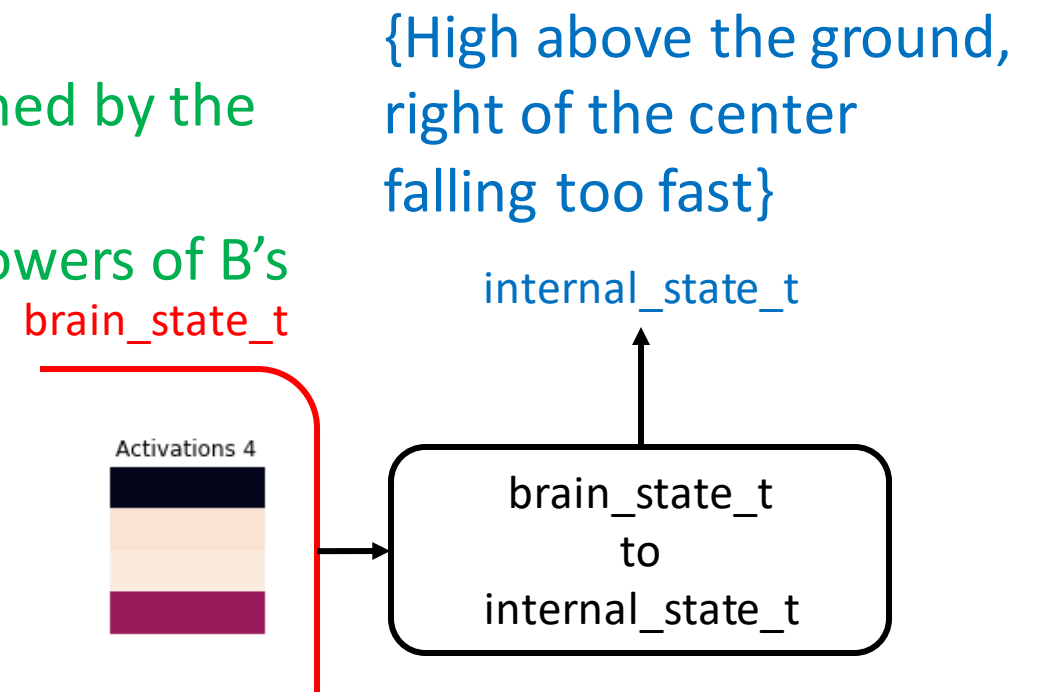
- the behavior of A's are entirely causally explained by the behavior of B's
- A's have no causal powers in addition to the powers of B's

Did we satisfy our requirements?

- V0
 - Internal states are causally 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 causally explained by the behavior of B's
- A's have no causal powers in addition to the powers of B's

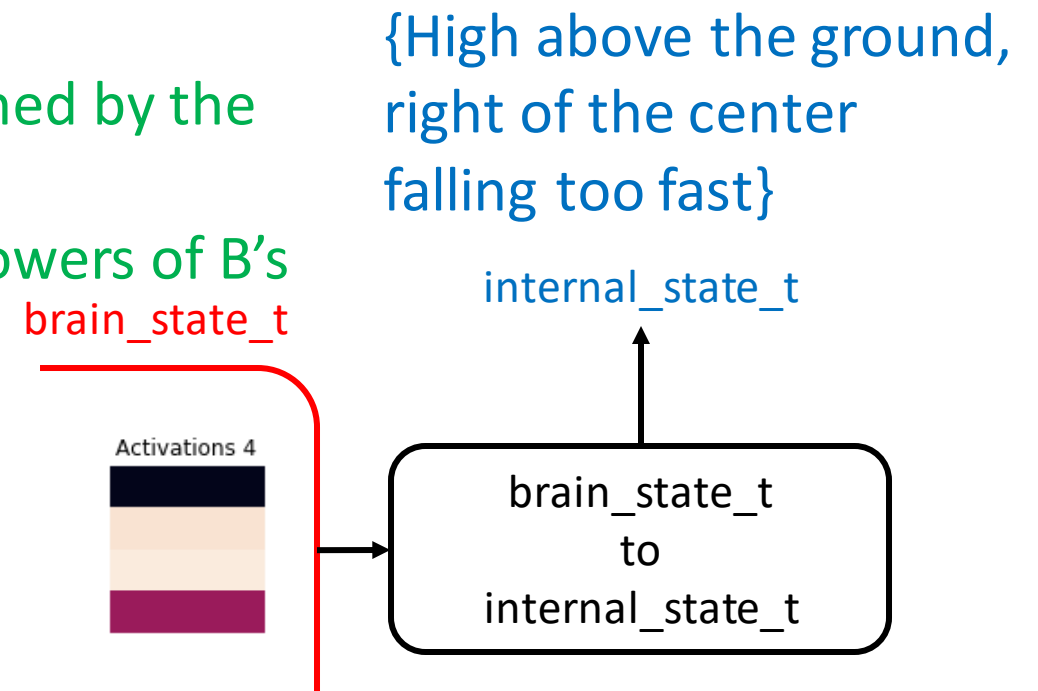


Did we satisfy our requirements?

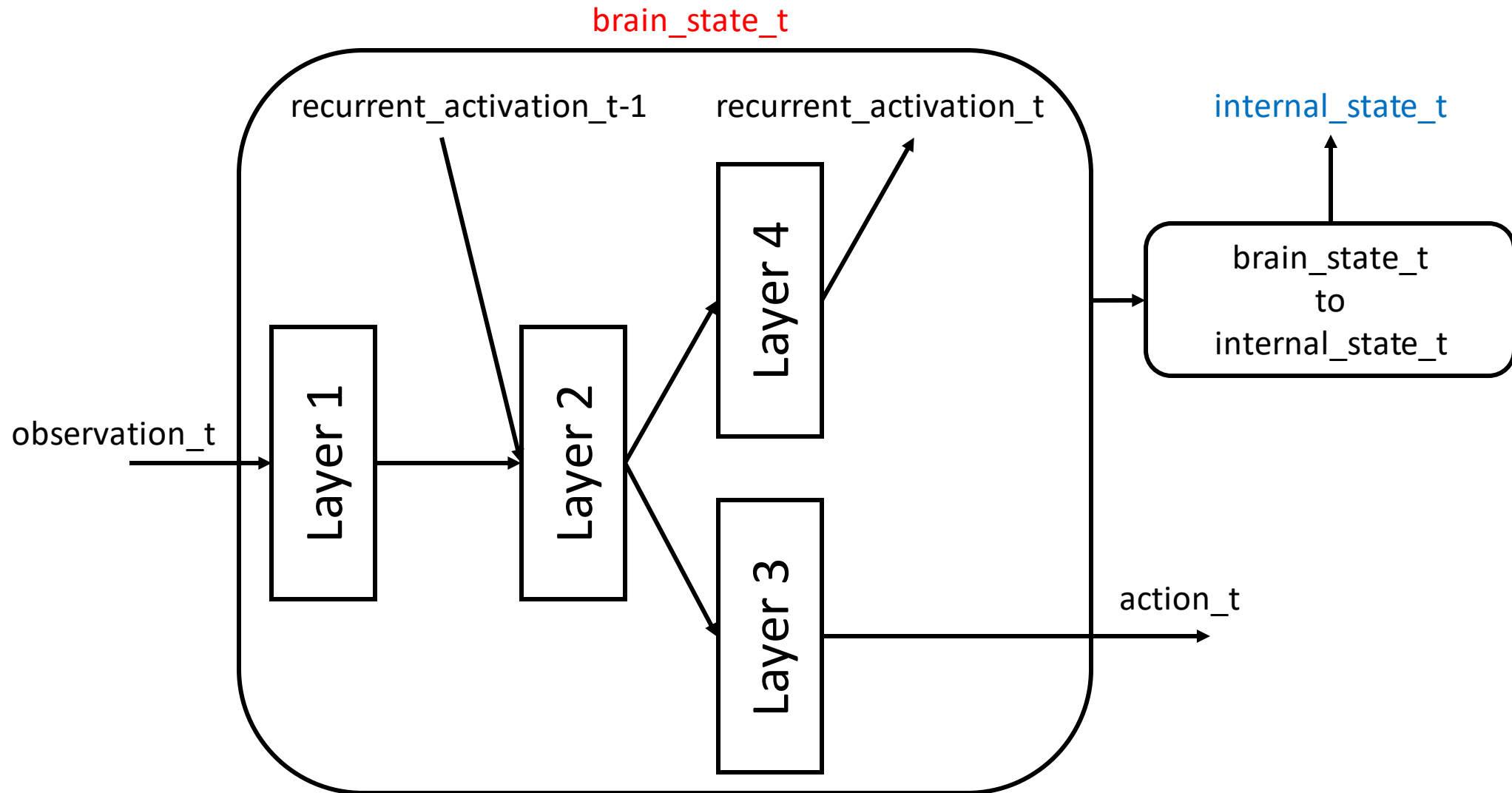
- V0
 - Internal states are causally reducible to brain states
 - Internal states are ontologically irreducible to brain states

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

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



Design, V0



Did we satisfy our requirements?

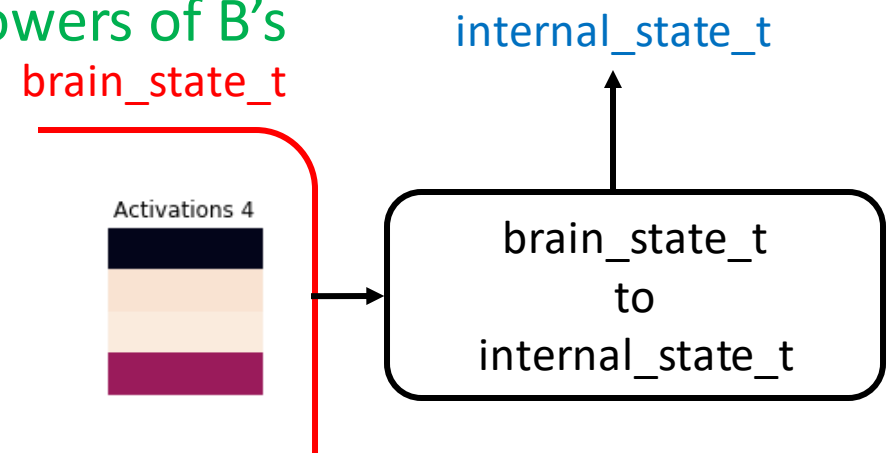
- V0

- ✓ Internal states are causally 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 causally explained by the behavior of B's
- ✓ A's have no causal powers in addition to the powers of B's

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



Did we satisfy our requirements?

- V0

- ✓ Internal states are causally 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

Did we satisfy our requirements?

- V0

- ✓ Internal states are causally 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

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

Did we satisfy our requirements?

- V0

- ✓ Internal states are causally reducible to brain states
 - Internal states are ontologically irreducible to brain states

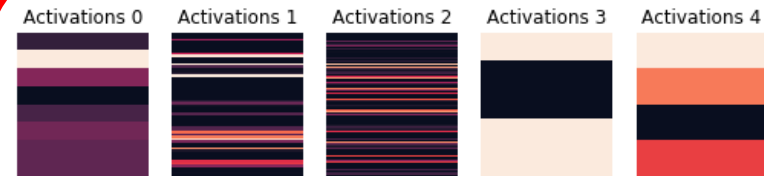
Phenomena of type A are ontologically reducible to phenomena of type B if and only if A's are nothing but B's

Our ontology

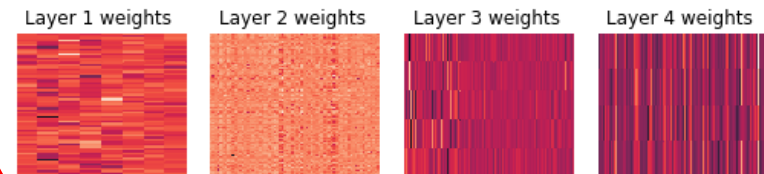
- Layer weights of the neural network
- Connectivity of the neural network
- Activations of the neural network at time t
- The agent's observation at time t
- The agent's action at time t
- The position and velocity of the agent at time t
- A region the agent believes it's in
- Brain state at time t (set of layer weights, activations, and connectivity)
- 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' }

network activations at time t



network layer weights



Did we satisfy our requirements?

- V0

- ✓ Internal states are causally 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

Our ontology

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

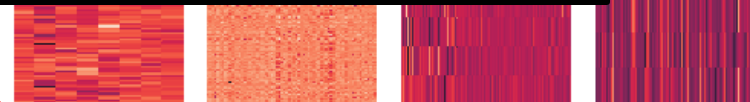
Internal state:
{ 'I_am_high_above_the_ground', 'I_am_to_the_right_of_the_center', 'I_am_falling_too_fast' }

network activations at time t

Activations 0 Activations 1 Activations 2 Activations 3 Activations 4



layer 4 weights



Internal 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 causally 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

Our ontology

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

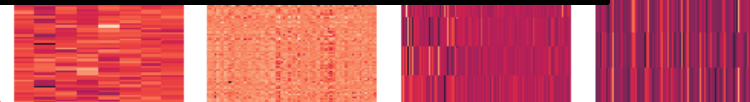
Internal state:
{ 'I_am_high_above_the_ground', 'I_am_to_the_right_of_the_center', 'I_am_falling_too_fast' }

network activations at time t

Activations 0 Activations 1 Activations 2 Activations 3 Activations 4



layer 4 weights



Internal 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 causally 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

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

Is that the “real” ontology though?

- V0



Internal states are causally 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

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
- A region the agent believes it's in
- Brain state at time t
- 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 causally 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

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

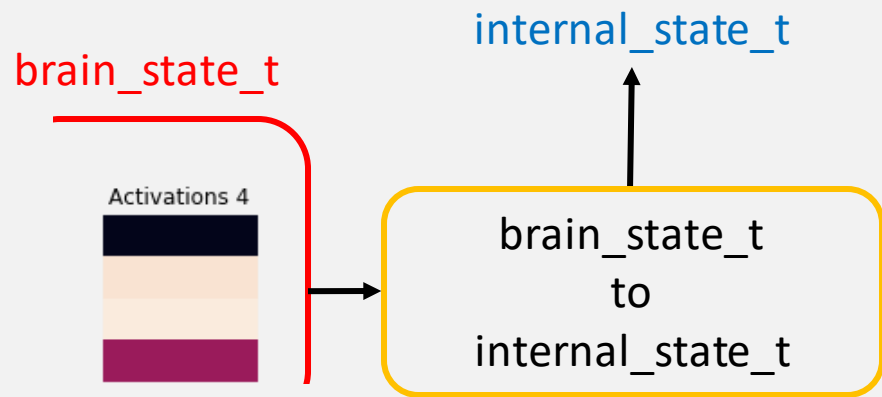
Our ontology

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

- Bits
- Python objects
- Electrons
- Quarks
- ...

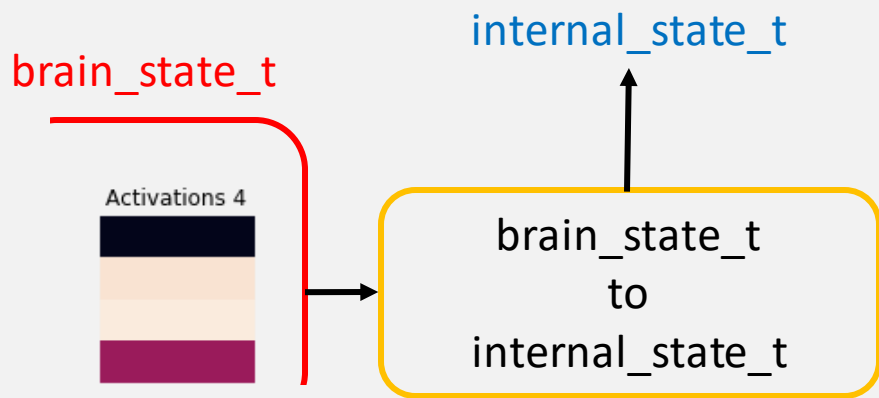
What's the deal with that function?

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



What's the deal with that function?

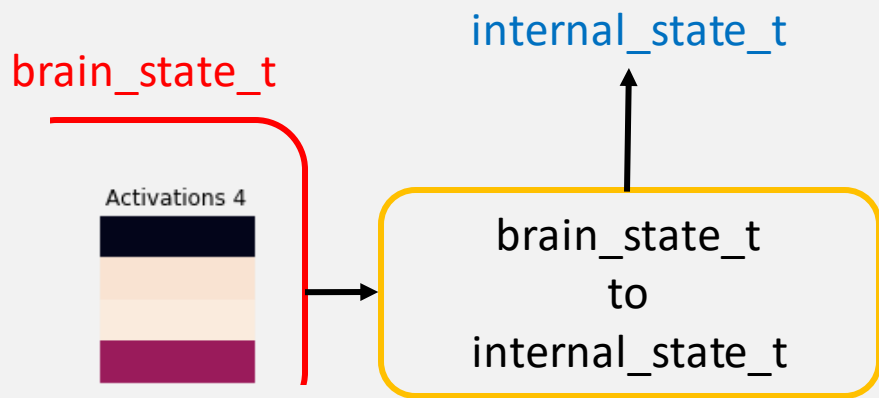
{High above the ground,
right of the center
falling too fast}



```
26 def brain_state_to_internal_state(brain_state):
27     def i_am_high_above_the_ground(observation):
28         return observation[1] > 0.5 # observation[1] accesses y position
29
30     def i_am_low_to_the_ground(observation):
31         return observation[1] <= 0.5 # observation[1] accesses y position
32
33     def i_am_to_the_right_of_the_center(observation):
34         return observation[0] > 0. # observation[0] accesses x position
35
36     def i_am_to_the_left_of_the_center(observation):
37         return observation[0] <= 0. # observation[0] accesses x position
38
39     def i_am_falling_too_fast(observation):
40         return observation[3] < -0.2 # observation[0] accesses y velocity
41
42     regions = [
43         i_am_high_above_the_ground,
44         i_am_low_to_the_ground,
45         i_am_to_the_right_of_the_center,
46         i_am_to_the_left_of_the_center,
47         i_am_falling_too_fast
48     ]
49
50     internal_state = set()
51
52     recurrent_activations = brain_state['activations'][3]
53
54     for activation, region in zip(recurrent_activations, regions):
55
56         if activation > 0.5:
57             internal_state.add(region.__name__)
58
59     return internal_state
```

What's the deal with that function?

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

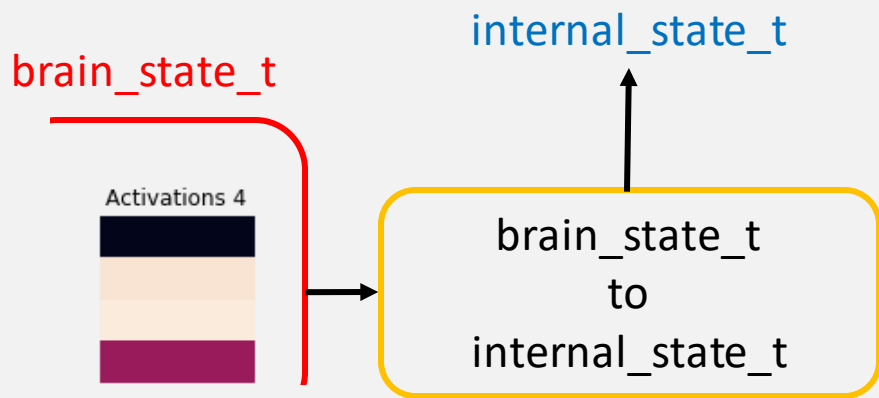


- Is this just some representation of "data flow"?

```
26 def brain_state_to_internal_state(brain_state):
27     def i_am_high_above_the_ground(observation):
28         return observation[1] > 0.5 # observation[1] accesses y position
29
30     def i_am_low_to_the_ground(observation):
31         return observation[1] <= 0.5 # observation[1] accesses y position
32
33     def i_am_to_the_right_of_the_center(observation):
34         return observation[0] > 0. # observation[0] accesses x position
35
36     def i_am_to_the_left_of_the_center(observation):
37         return observation[0] <= 0. # observation[0] accesses x position
38
39     def i_am_falling_too_fast(observation):
40         return observation[3] < -0.2 # observation[0] accesses y velocity
41
42     regions = [
43         i_am_high_above_the_ground,
44         i_am_low_to_the_ground,
45         i_am_to_the_right_of_the_center,
46         i_am_to_the_left_of_the_center,
47         i_am_falling_too_fast
48     ]
49
50     internal_state = set()
51
52     recurrent_activations = brain_state['activations'][3]
53
54     for activation, region in zip(recurrent_activations, regions):
55
56         if activation > 0.5:
57             internal_state.add(region.__name__)
58
59     return internal_state
```

What's the deal with that function?

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

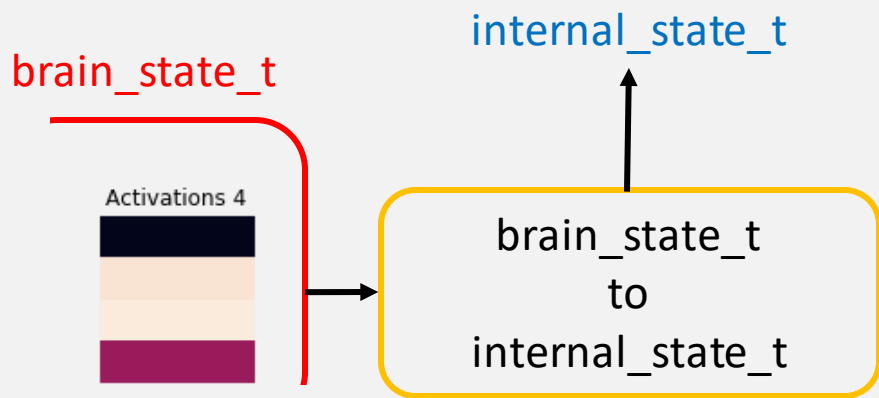


- Is this just some representation of "data flow"?
- Is this something closer to summarization?

```
26 def brain_state_to_internal_state(brain_state):
27     def i_am_high_above_the_ground(observation):
28         return observation[1] > 0.5 # observation[1] accesses y position
29
30     def i_am_low_to_the_ground(observation):
31         return observation[1] <= 0.5 # observation[1] accesses y position
32
33     def i_am_to_the_right_of_the_center(observation):
34         return observation[0] > 0. # observation[0] accesses x position
35
36     def i_am_to_the_left_of_the_center(observation):
37         return observation[0] <= 0. # observation[0] accesses x position
38
39     def i_am_falling_too_fast(observation):
40         return observation[3] < -0.2 # observation[0] accesses y velocity
41
42     regions = [
43         i_am_high_above_the_ground,
44         i_am_low_to_the_ground,
45         i_am_to_the_right_of_the_center,
46         i_am_to_the_left_of_the_center,
47         i_am_falling_too_fast
48     ]
49
50     internal_state = set()
51
52     recurrent_activations = brain_state['activations'][3]
53
54     for activation, region in zip(recurrent_activations, regions):
55
56         if activation > 0.5:
57             internal_state.add(region.__name__)
58
59     return internal_state
```

What's the deal with that function?

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



- Is this just some representation of "data flow"?
- Is this something closer to summarization?
- (or both?)

```
26 def brain_state_to_internal_state(brain_state):
27     def i_am_high_above_the_ground(observation):
28         return observation[1] > 0.5 # observation[1] accesses y position
29
30     def i_am_low_to_the_ground(observation):
31         return observation[1] <= 0.5 # observation[1] accesses y position
32
33     def i_am_to_the_right_of_the_center(observation):
34         return observation[0] > 0. # observation[0] accesses x position
35
36     def i_am_to_the_left_of_the_center(observation):
37         return observation[0] <= 0. # observation[0] accesses x position
38
39     def i_am_falling_too_fast(observation):
40         return observation[3] < -0.2 # observation[0] accesses y velocity
41
42     regions = [
43         i_am_high_above_the_ground,
44         i_am_low_to_the_ground,
45         i_am_to_the_right_of_the_center,
46         i_am_to_the_left_of_the_center,
47         i_am_falling_too_fast
48     ]
49
50     internal_state = set()
51
52     recurrent_activations = brain_state['activations'][3]
53
54     for activation, region in zip(recurrent_activations, regions):
55
56         if activation > 0.5:
57             internal_state.add(region.__name__)
58
59     return internal_state
```


What's the deal with that function?

{High above the ground,
right of the center
f

"[...] consciousness is a state the brain can be in, in the way that liquidity and solidity are states that water can be in."

- *Why I'm Not a Property Dualist*, John Searle

brain_state_t

Activations 4

brain_state_t
to
internal_state_t

- Is this just some representation of "data flow"?
- Is this something closer to summarization?
- (or both?)

```
26 def brain_state_to_internal_state(brain_state):
27     def i_am_high_above_the_ground(observation):
28         return observation[1] > 0.5 # observation[1] accesses y position
29
30     def i_am_low_to_the_ground(observation):
31         return observation[1] < 0.5 # observation[1] accesses y position
32
33     def i_am_to_the_right_of_the_center(observation):
34         return observation[0] > 0.5 # observation[0] accesses x position
35
36     def i_am_to_the_left_of_the_center(observation):
37         return observation[0] < 0.5 # observation[0] accesses x position
38
39     def i_am_falling_too_fast(observation):
40         return observation[3] < -0.2 # observation[0] accesses y velocity
41
42     regions = [
43         i_am_high_above_the_ground,
44         i_am_low_to_the_ground,
45         i_am_to_the_right_of_the_center,
46         i_am_to_the_left_of_the_center,
47         i_am_falling_too_fast
48     ]
49
50     internal_state = set()
51
52     recurrent_activations = brain_state['activations'][3]
53
54     for activation, region in zip(recurrent_activations, regions):
55         if activation > 0.5:
56             internal_state.add(region.__name__)
57
58     return internal_state
```

What's the deal with that function?

{High above the ground,
right of the center

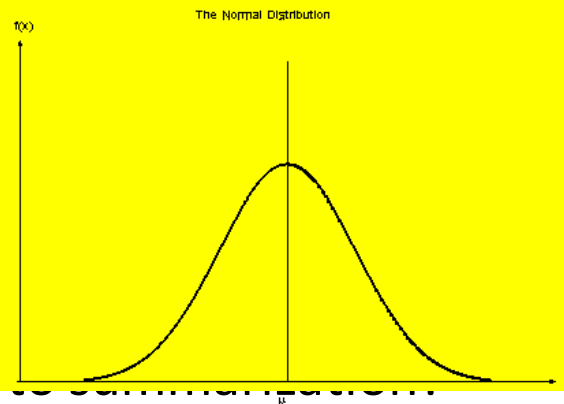
"[...] consciousness is a state the brain can be in, in the way that liquidity and solidity are states that water can be in."

- *Why I'm Not a Property Dualist*, John Searle

Just like a gaussian and its parameters...

brain_state_t

Activations 4



$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum X_i$$

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$$

- Is this just some
- Is this something
- (or both?)

```
26 def brain_state_to_internal_state(brain_state):
27     def i_am_high_above_the_ground(observation):
28         return observation[1] > 0.5 # observation[1] accesses y position
29
30     def i_am_low_to_the_ground(observation):
```

```
56         if activation > 0.5:
57             internal_state.add(region.__name__)
58
59     return internal_state
```

Conclusion

- Software engineer style philosophy reifying seemed to work well
- Created a V0 software agent who's
 - Internal states are causally reducible to brain states
 - Internal states are ontologically irreducible to brain states
- Download and play with the code yourself
 - <https://github.com/Josh-Joseph/tsc-2019>
- Disagree with us?
 - 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?