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

mit
media
lab

MIT
Quest for
Intelligence

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)

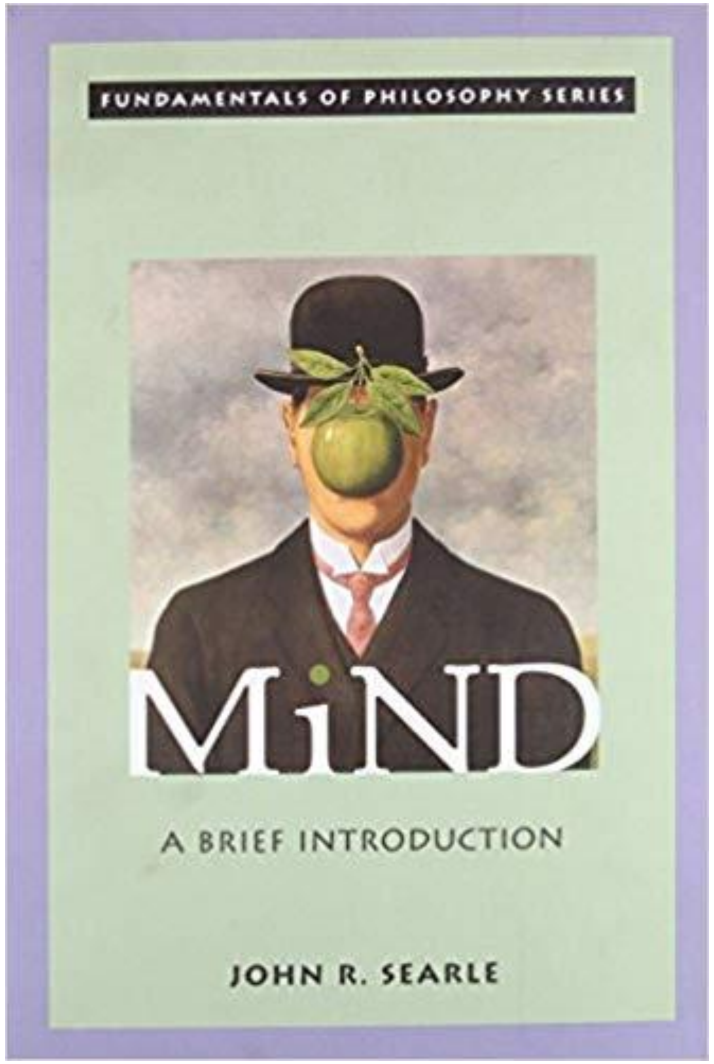
Why attempt to reify philosophy with code

- 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

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
- Not trying to propose a new AI or machine learning algorithm
- Not trying to claim that the software agent is conscious
- Not trying to convince anyone these are the correct/best/most useful definitions of consciousness or brain states
- Not trying to 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 – the built 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
- V2
  - Conscious mental states are casually 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
- 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

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

Requirements: unpacking Searle's view

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

Requirements: unpacking Searle's view

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

Requirements: unpacking Searle's view

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

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

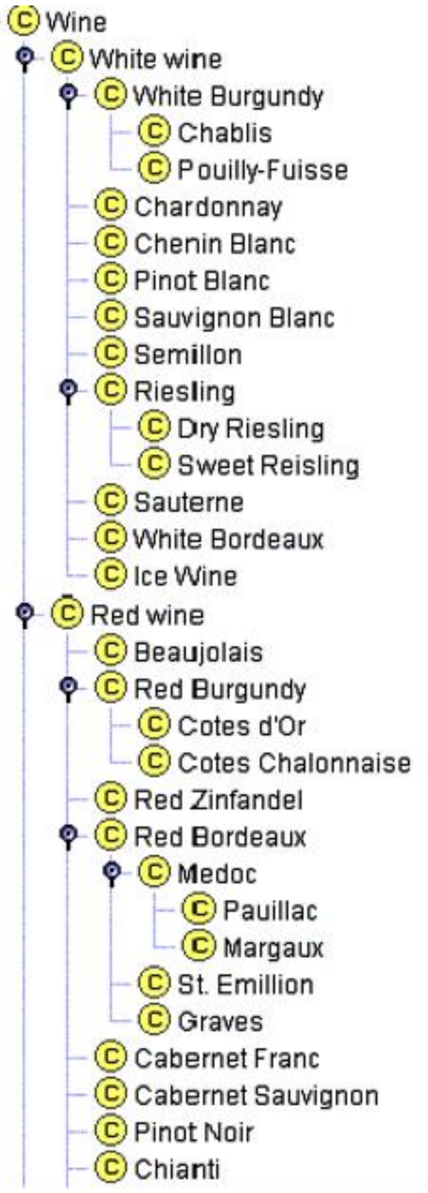
Ontologies in Computer Science

- Class-instance distinction

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

Ontologies in Computer Science

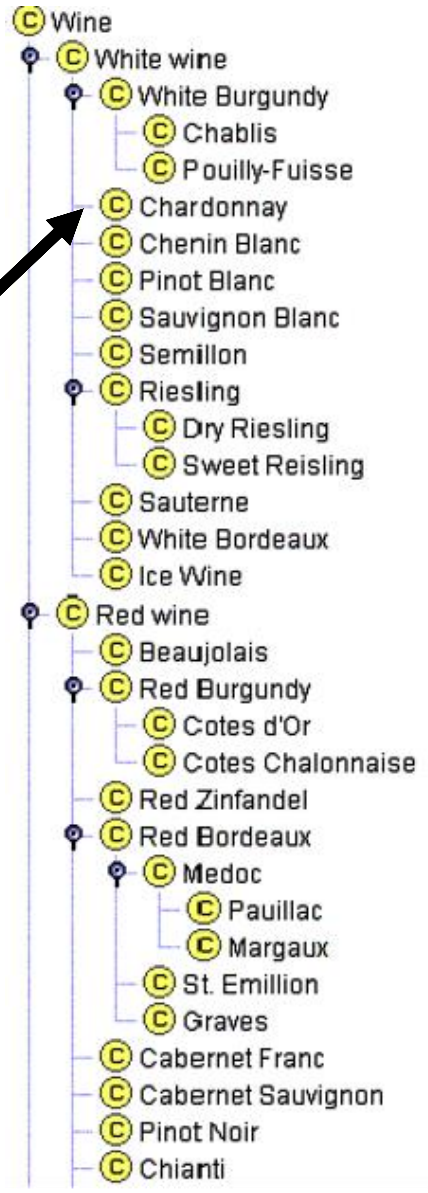
- Class-instance distinction



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

Ontologies in Computer Science

- Class-instance distinction

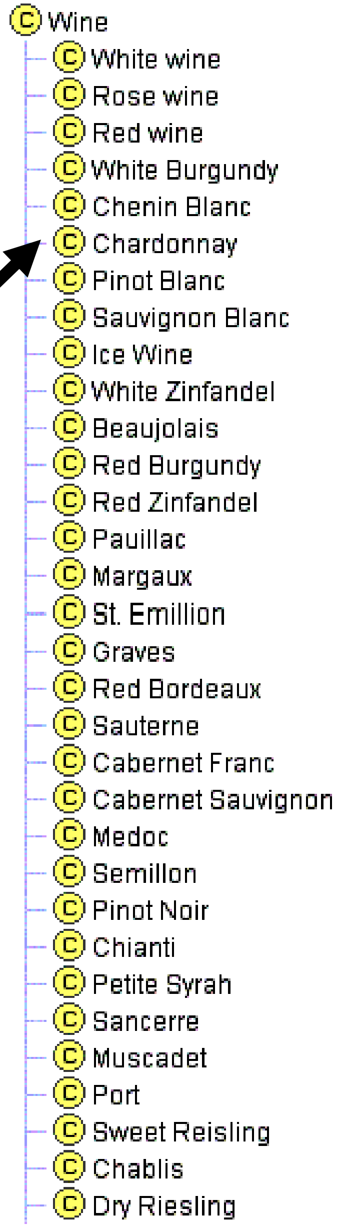


- Wine
  - White wine
    - White Burgundy
      - Chablis
      - Pouilly-Fuisse
    - Chardonnay
    - Chenin Blanc
    - Pinot Blanc
    - Sauvignon Blanc
    - Semillon
    - Riesling
      - Dry Riesling
      - Sweet Reisling
    - Sauterne
    - White Bordeaux
    - Ice Wine
  - Red wine
    -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

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

Ontologies in Computer Science

- Class-instance distinction

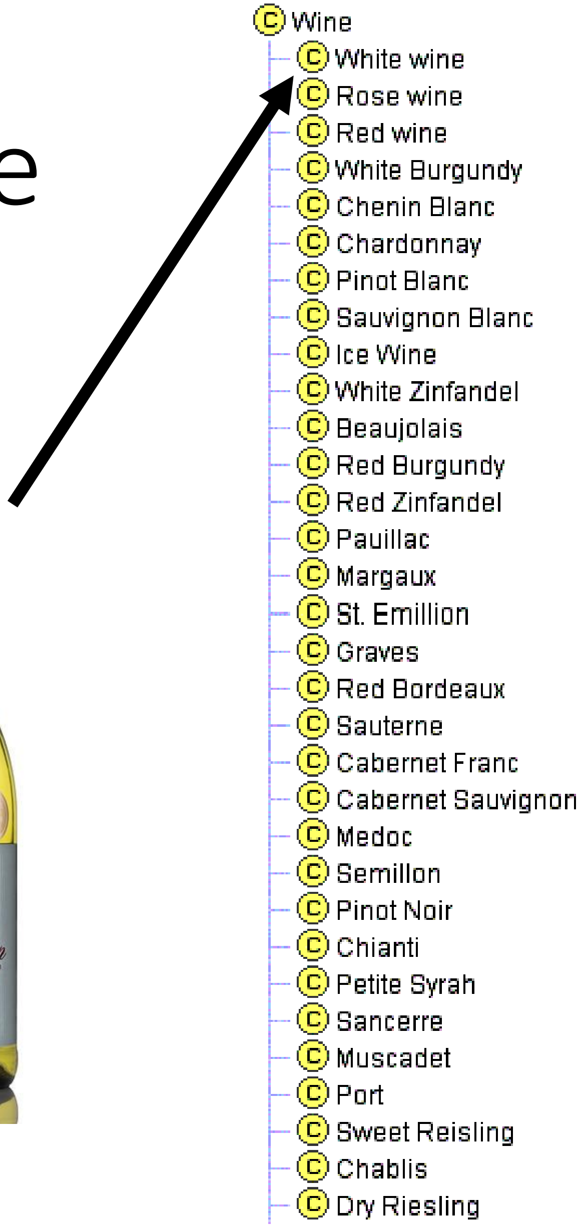


- Wine
  - White wine
  - Rose wine
  - Red wine
  - White Burgundy
  - Chenin Blanc
  - Chardonnay
  - Pinot Blanc
  - Sauvignon Blanc
  - Ice Wine
  - White Zinfandel
  - Beaujolais
  - Red Burgundy
  - Red Zinfandel
  - Pauillac
  - Margaux
  - St. Emillion
  - Graves
  - Red Bordeaux
  - Sauterne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

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

Ontologies in Computer Science

- Class-instance distinction

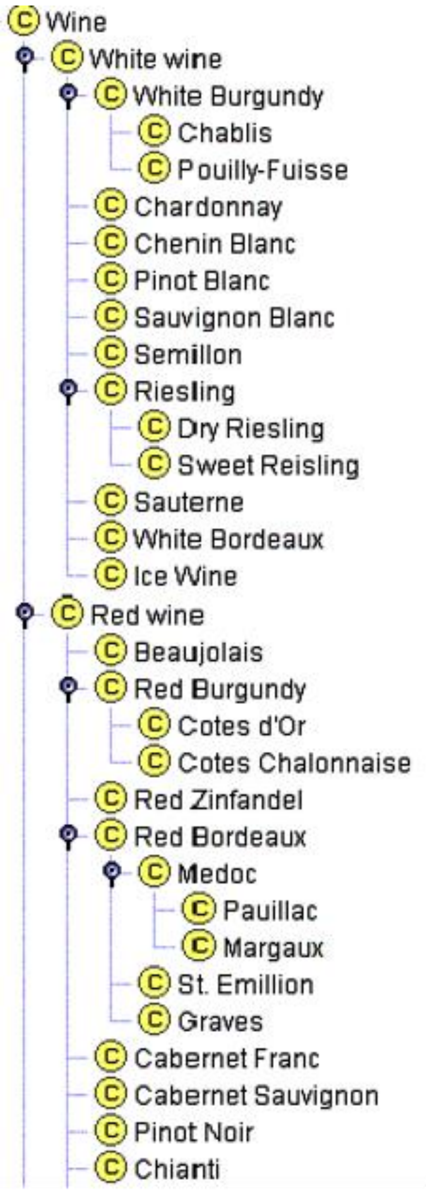


- Wine
  - White wine
  - Rose wine
  - Red wine
  - White Burgundy
  - Chenin Blanc
  - Chardonnay
  - Pinot Blanc
  - Sauvignon Blanc
  - Ice Wine
  - White Zinfandel
  - Beaujolais
  - Red Burgundy
  - Red Zinfandel
  - Pauillac
  - Margaux
  - St. Emillion
  - Graves
  - Red Bordeaux
  - Sauterne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

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

Ontologies in Computer Science

- Class-instance distinction
- Type-token distinction

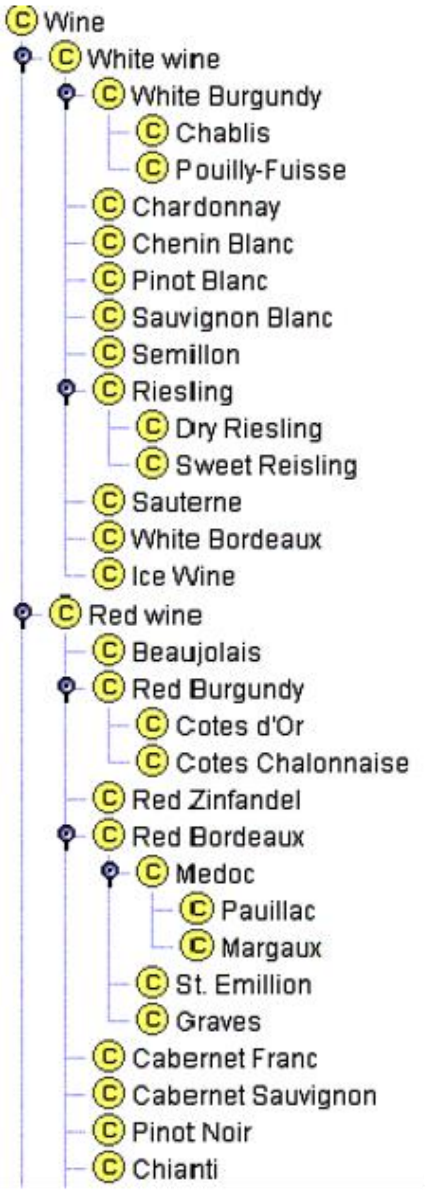


- Wine
  - White wine
    - White Burgundy
      - Chablis
      - Pouilly-Fuisse
    - Chardonnay
    - Chenin Blanc
    - Pinot Blanc
    - Sauvignon Blanc
    - Semillon
    - Riesling
      - Dry Riesling
      - Sweet Reisling
    - Sauterne
    - White Bordeaux
    - Ice Wine
  - Red wine
    -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

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

Ontologies in Computer Science

- Class-instance distinction
- 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)

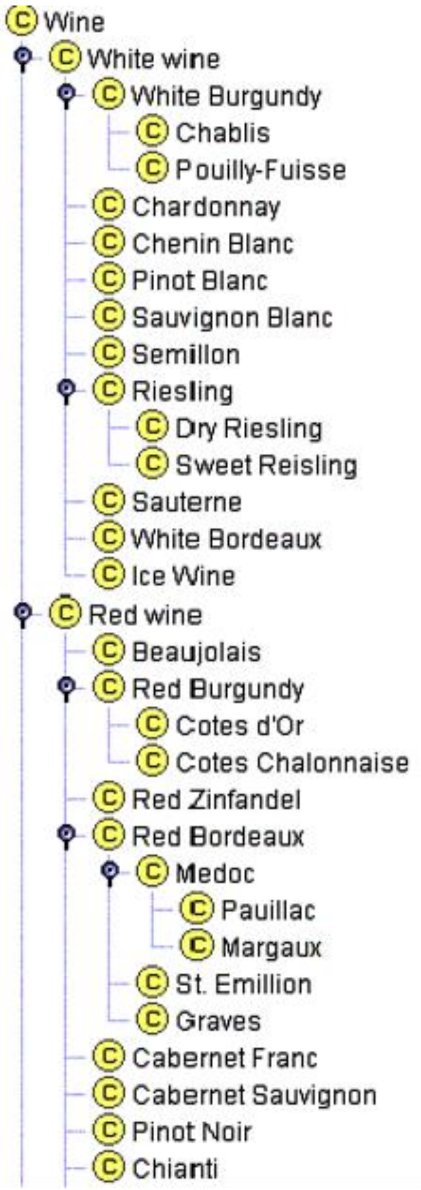


- Wine
  - White wine
    - White Burgundy
      - Chablis
      - Pouilly-Fuisse
    - Chardonnay
    - Chenin Blanc
    - Pinot Blanc
    - Sauvignon Blanc
    - Semillon
    - Riesling
      - Dry Riesling
      - Sweet Reisling
    - Sauterne
    - White Bordeaux
    - Ice Wine
  - Red wine
    -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

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

Ontologies in Computer Science

- Class-instance distinction
- 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



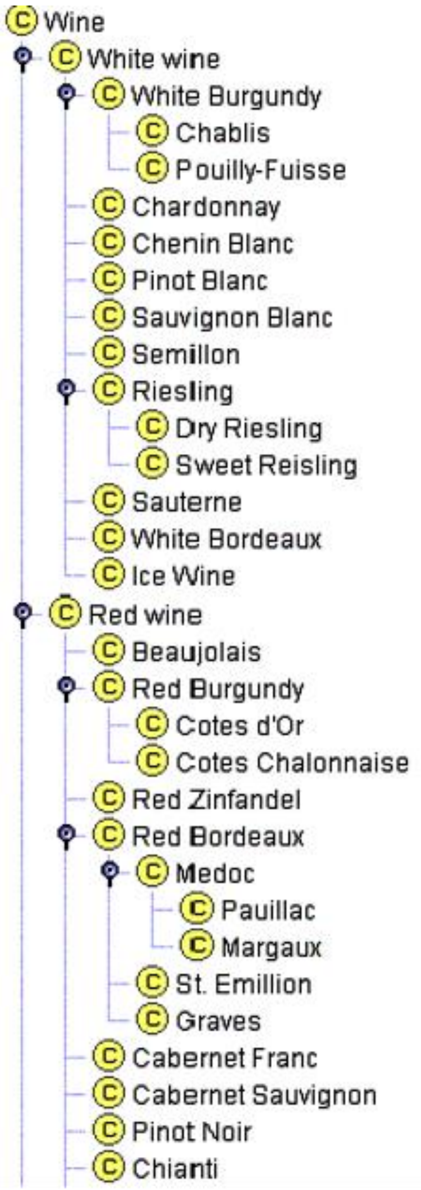
(C) A set of wine bottles

(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



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

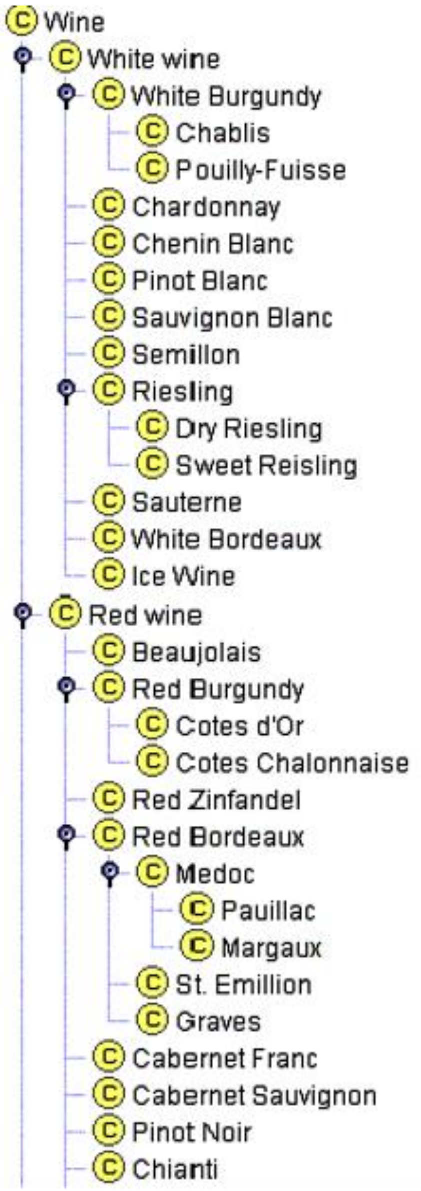
(C) A set of wine bottles

(C) Case of wine

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

Ontologies in Computer Science

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



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

(C) A set of wine bottles

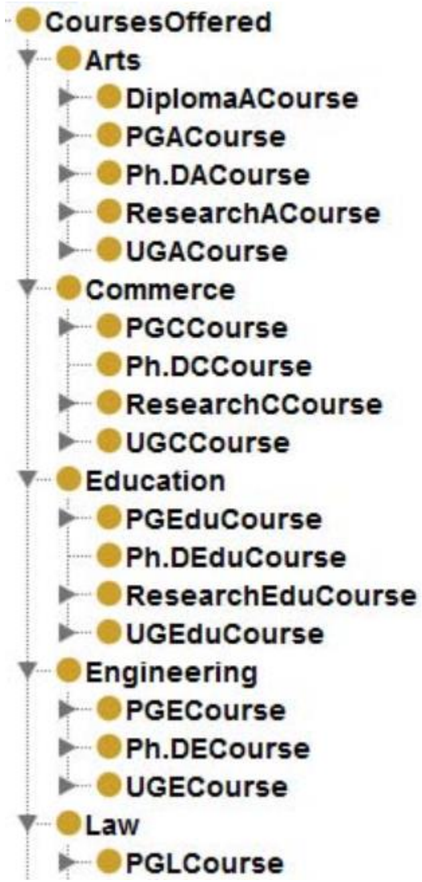
(C) Case of wine

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

Ontologies in Computer Science

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

(C) Case of wine



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

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

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

Requirements: unpacking Searle's view

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

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

Requirements, V0

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

Design, V0

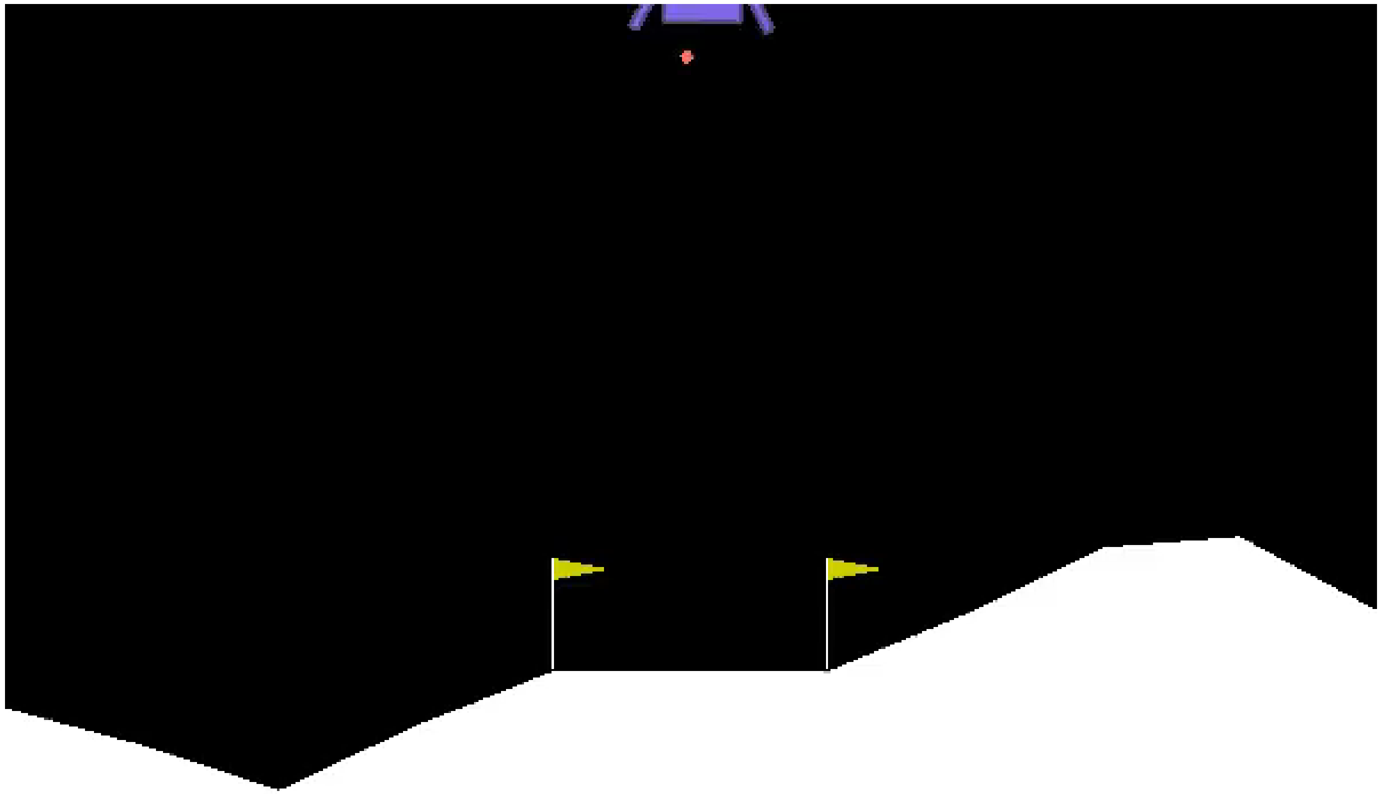
- Design decisions

Design, V0

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

Design, V0

- OpenAI's LunarLander benchmark environment

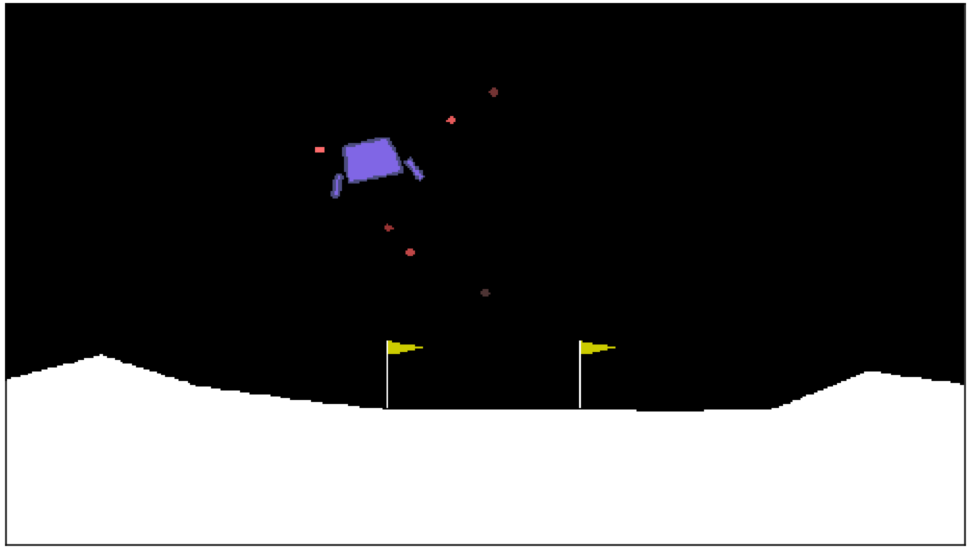


Design, V0

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

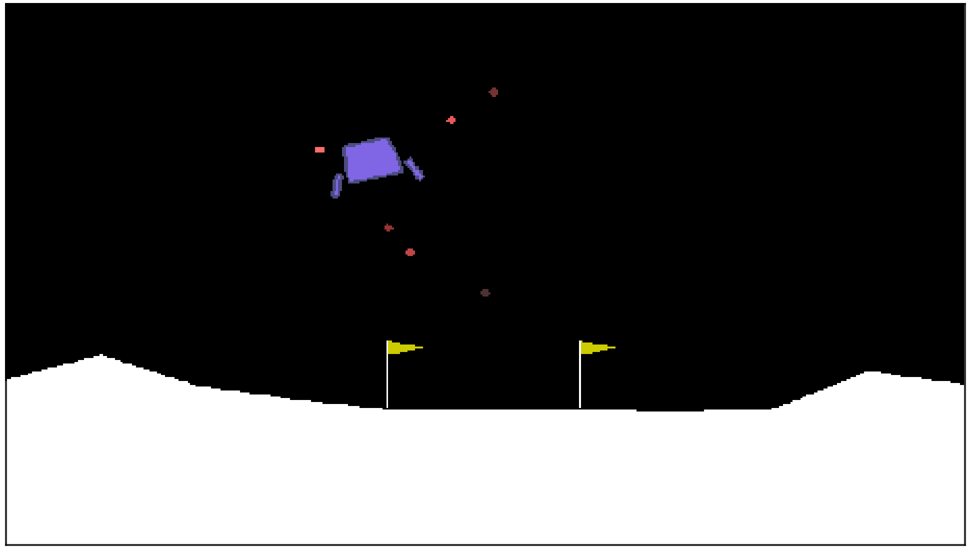
Design, V0

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

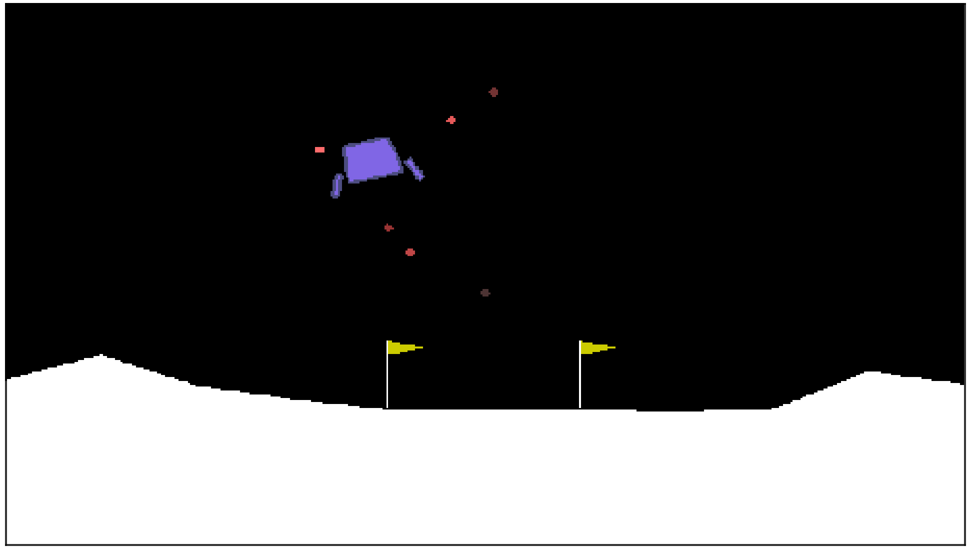


Design, V0

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

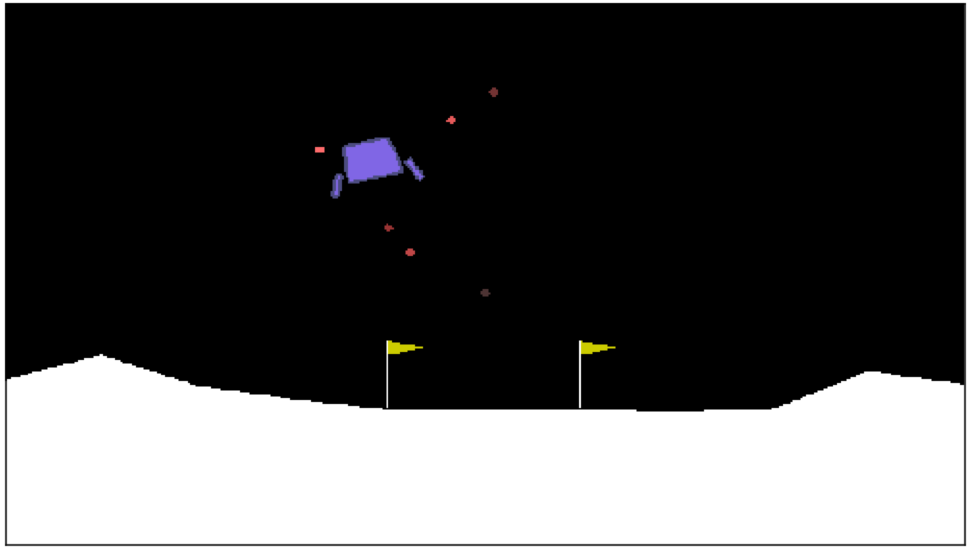


Design, V0



- 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

(Artificial) Neural networks

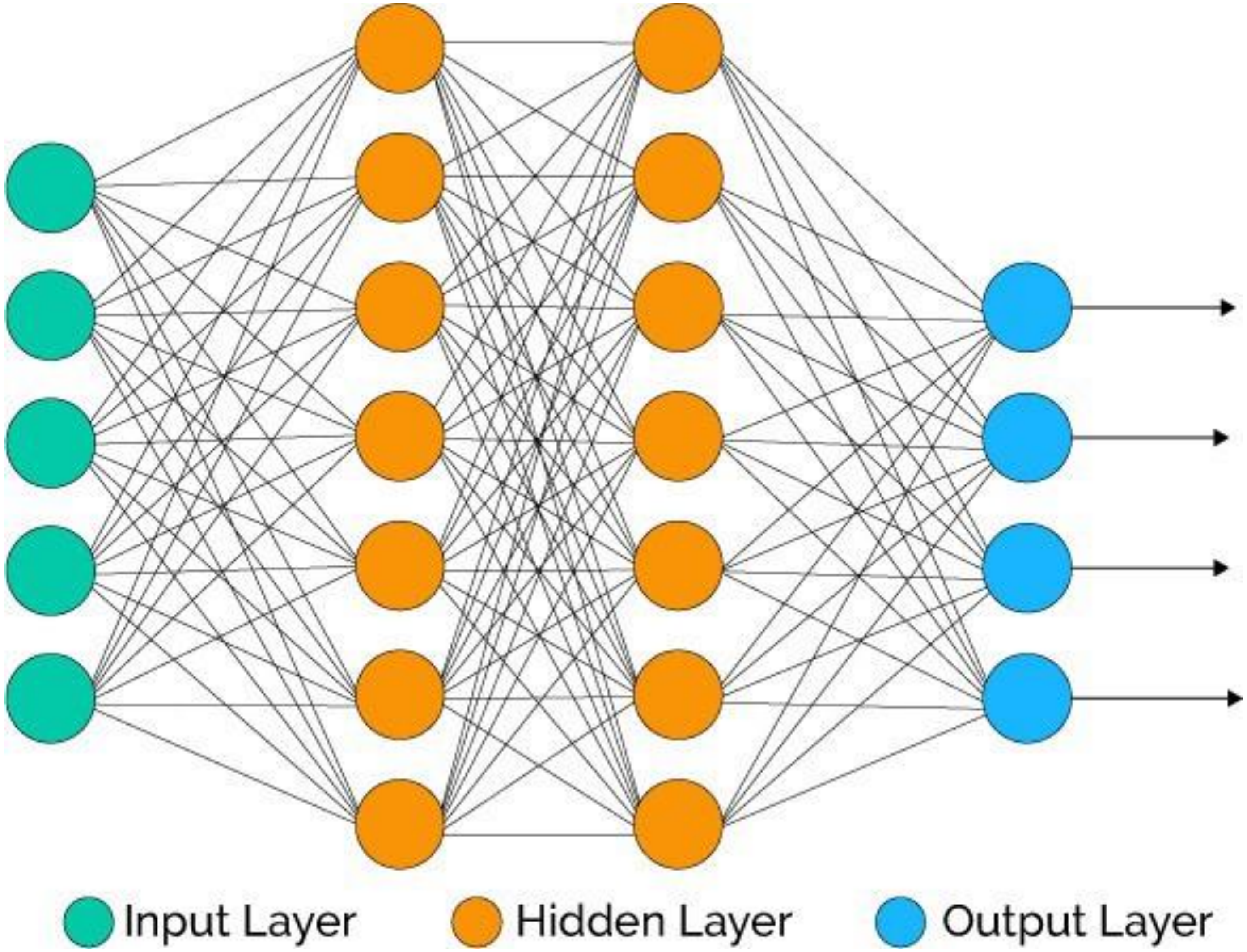
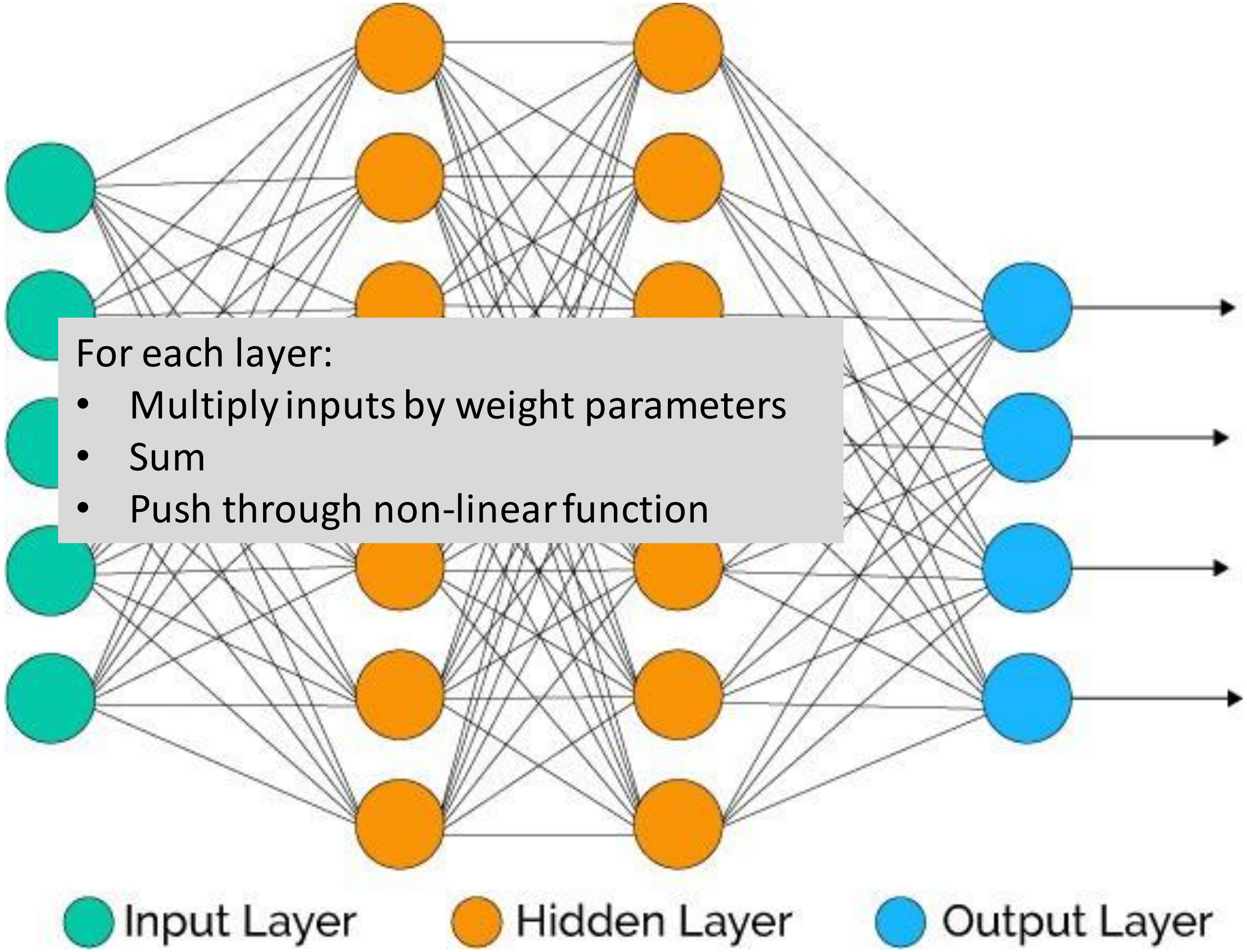


Image from:
https://medium.com/datadriveninvestor/when-not-to-use-neural-networks-89fb50622429

(Artificial) Neural networks



For each layer:

- Multiply inputs by weight parameters
- Sum
- Push through non-linear function

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

(Artificial) Neural networks

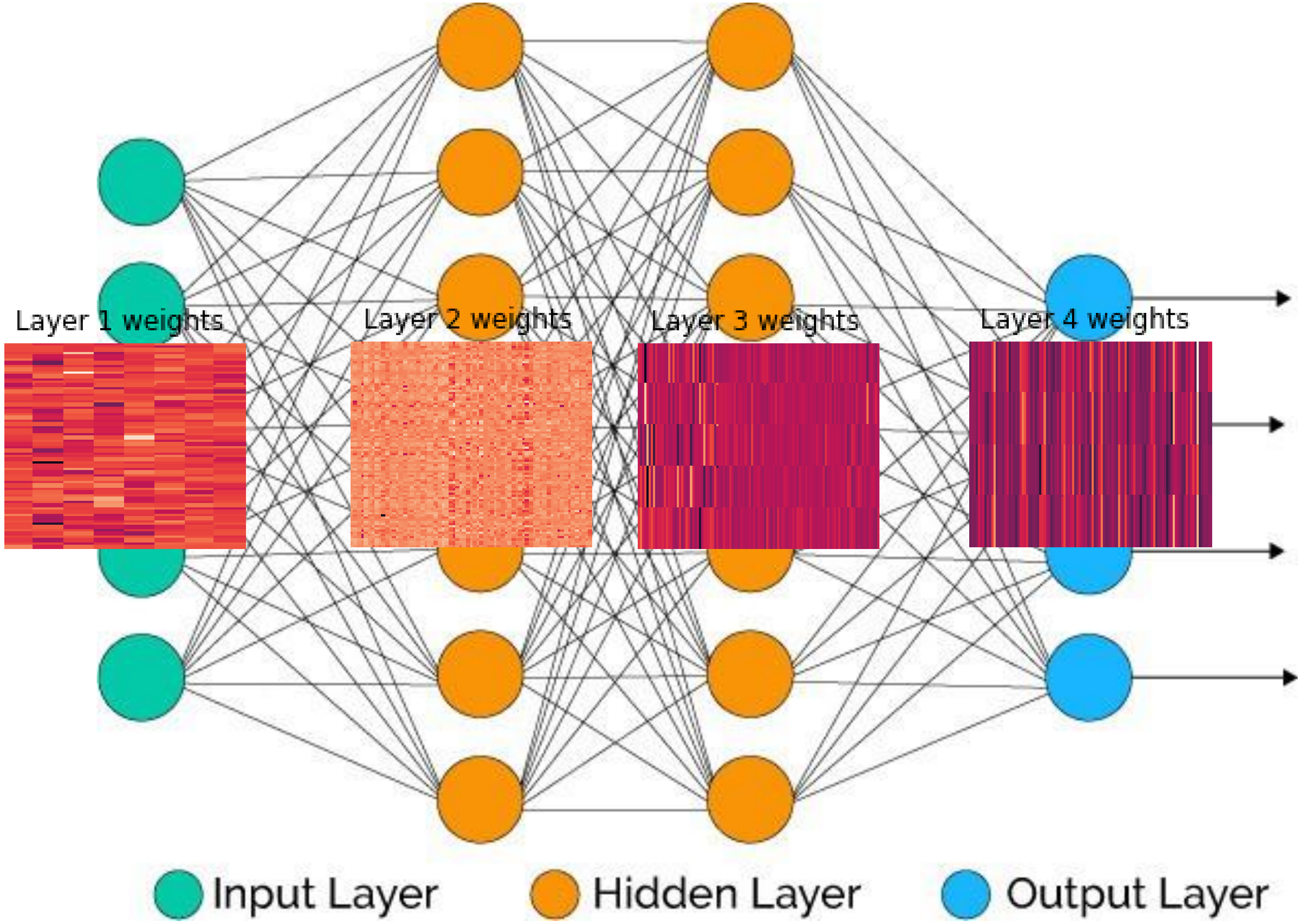


Image from: https://medium.com/datadriveninvestor/when-not-to-use-neural-networks-89fb50622429

(Artificial) Neural networks

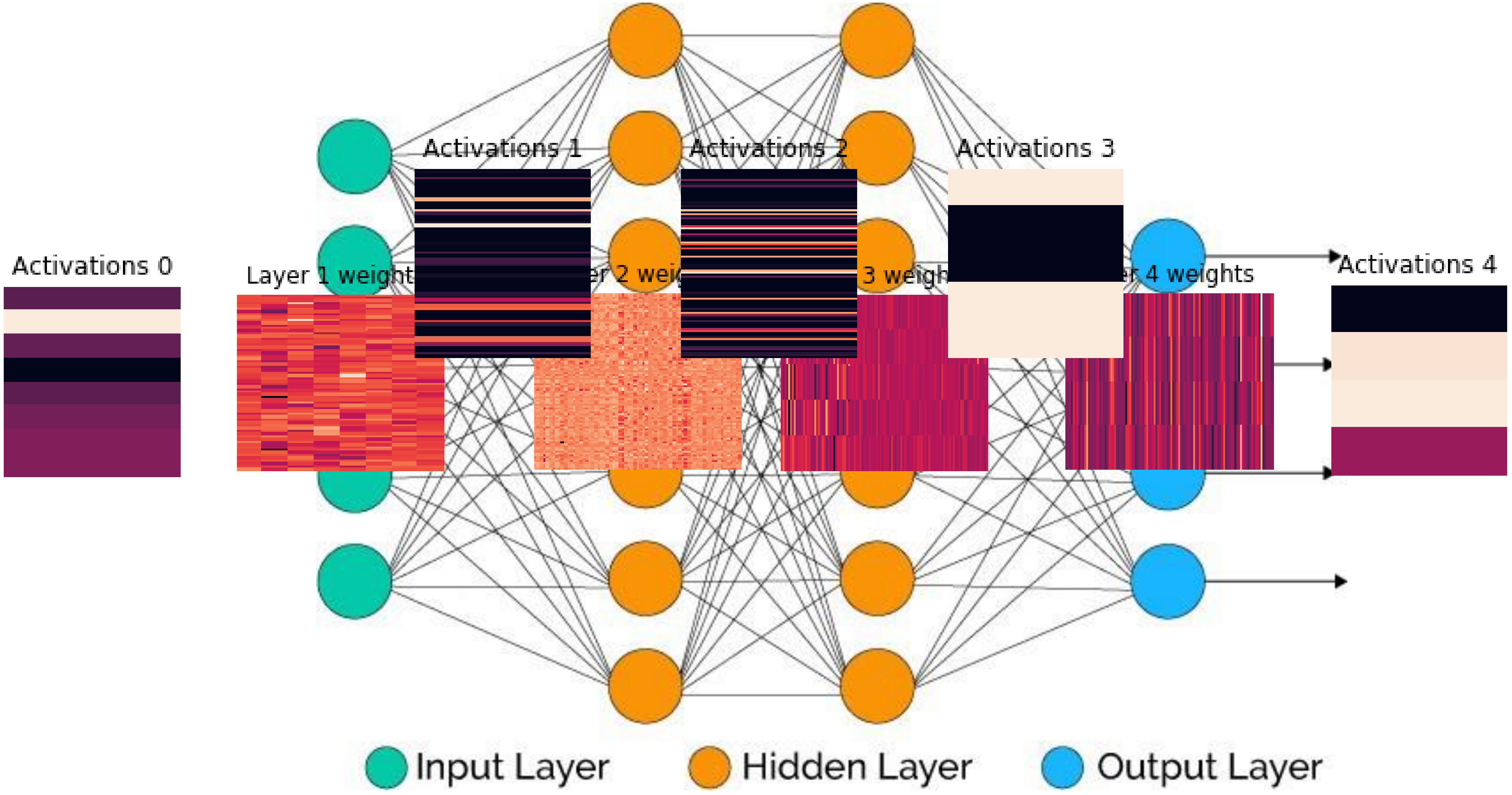
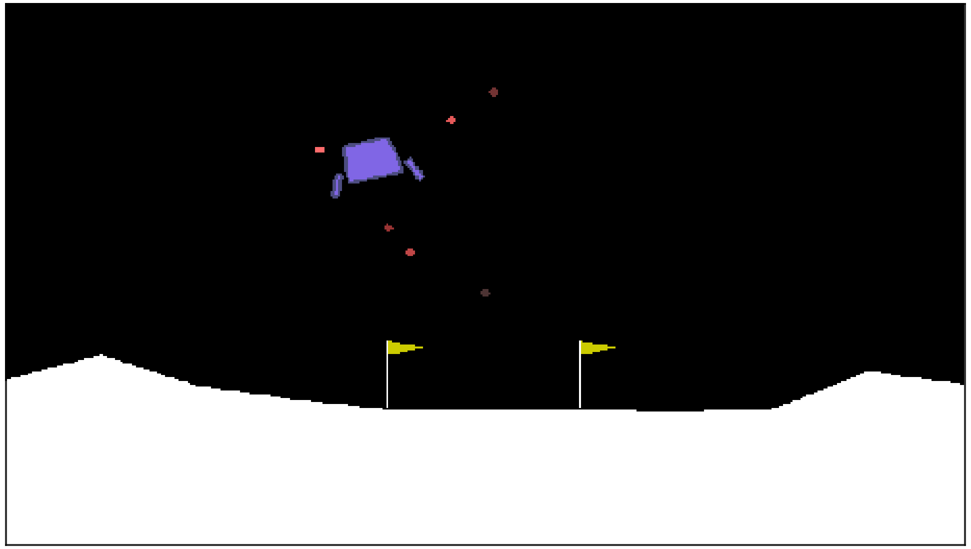


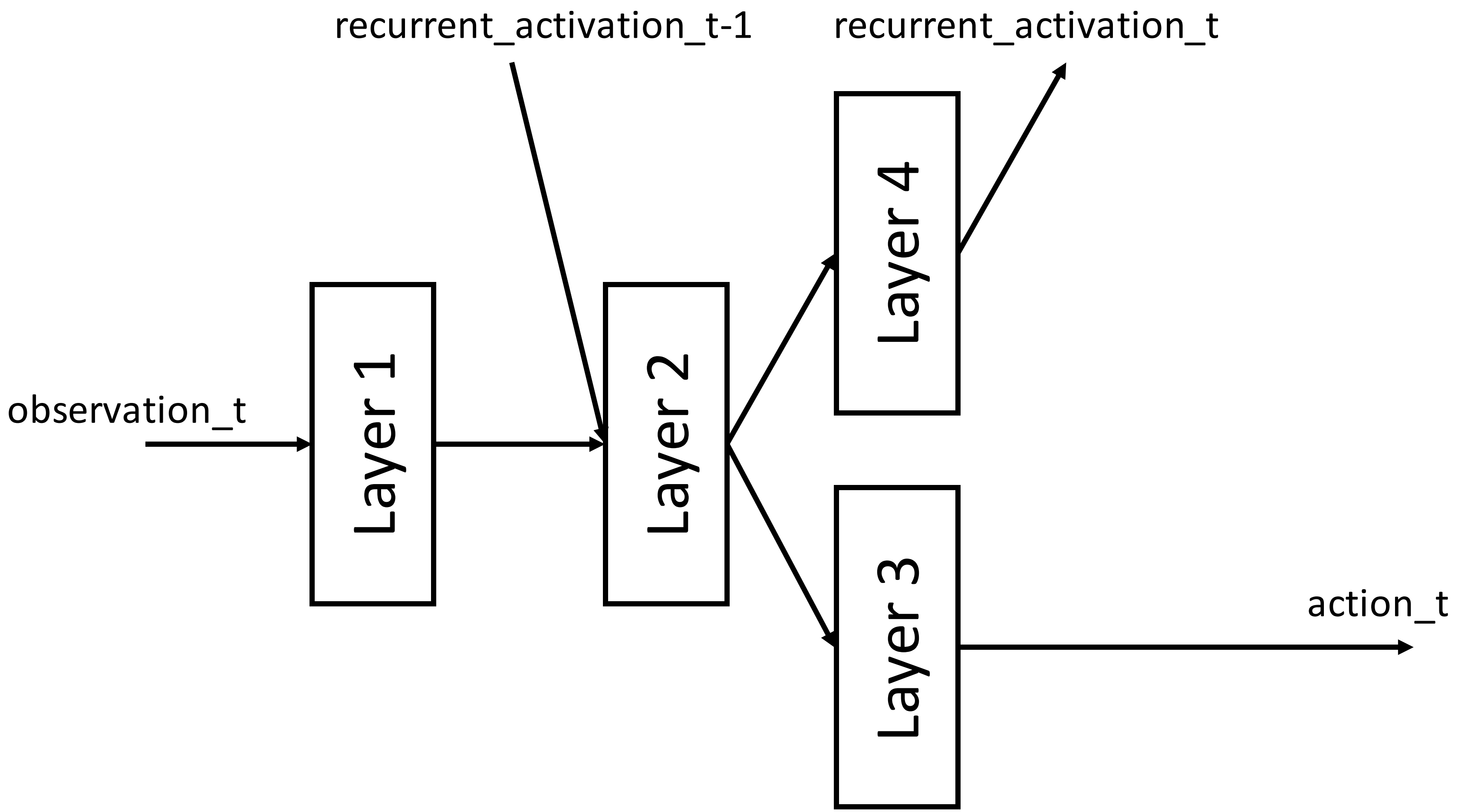
Image from: https://medium.com/datadriveninvestor/when-not-to-use-neural-networks-89fb50622429

Design, V0

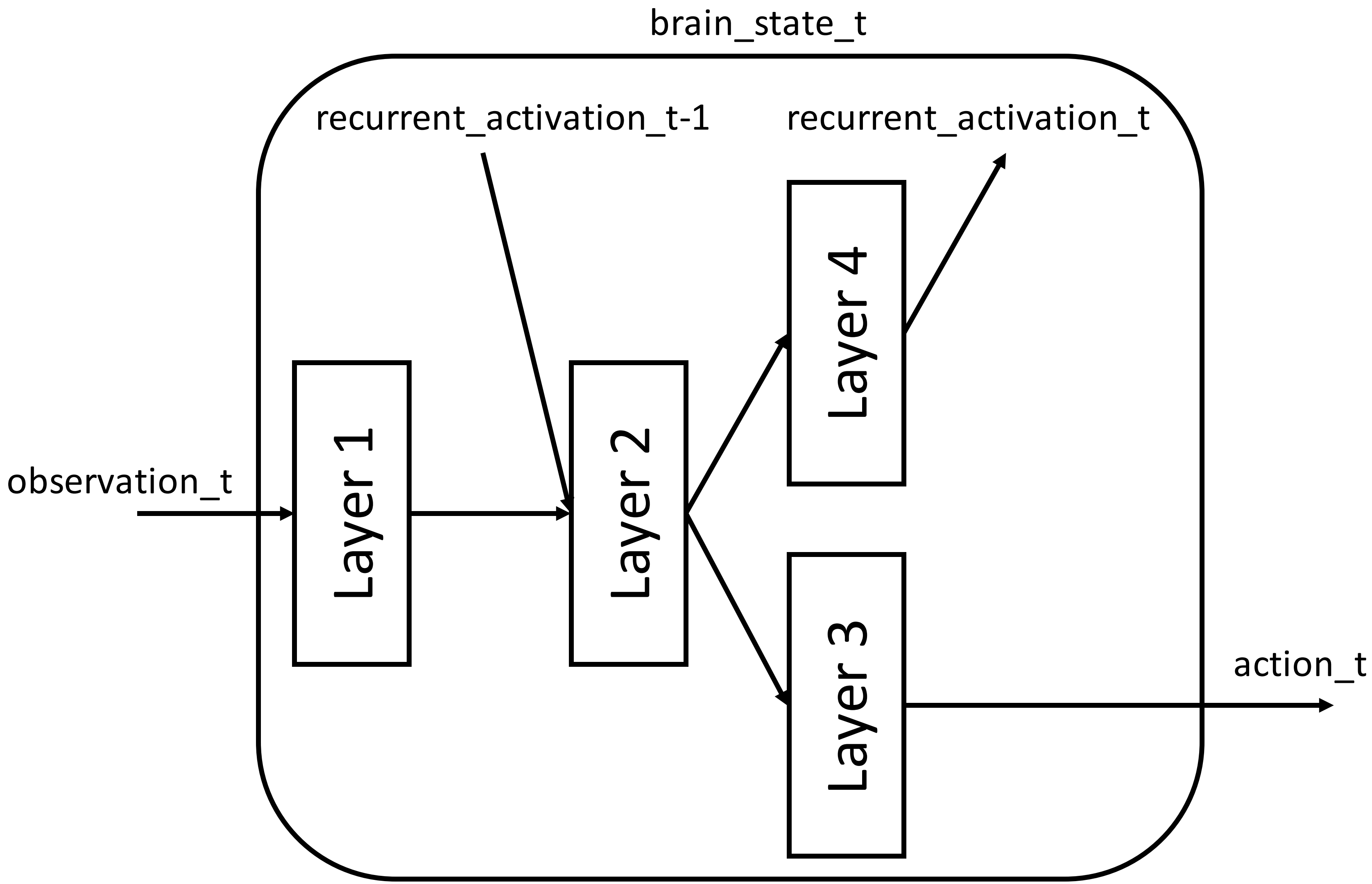


- Design decisions
  - Environment and the agent's "physical" form
  - Internal state of the agent
    - Beliefs about itself relative to semantically important regions
      - Left of the flags, right of the flags, high above the ground, close to the ground, falling too fast
  - Brain state of the agent

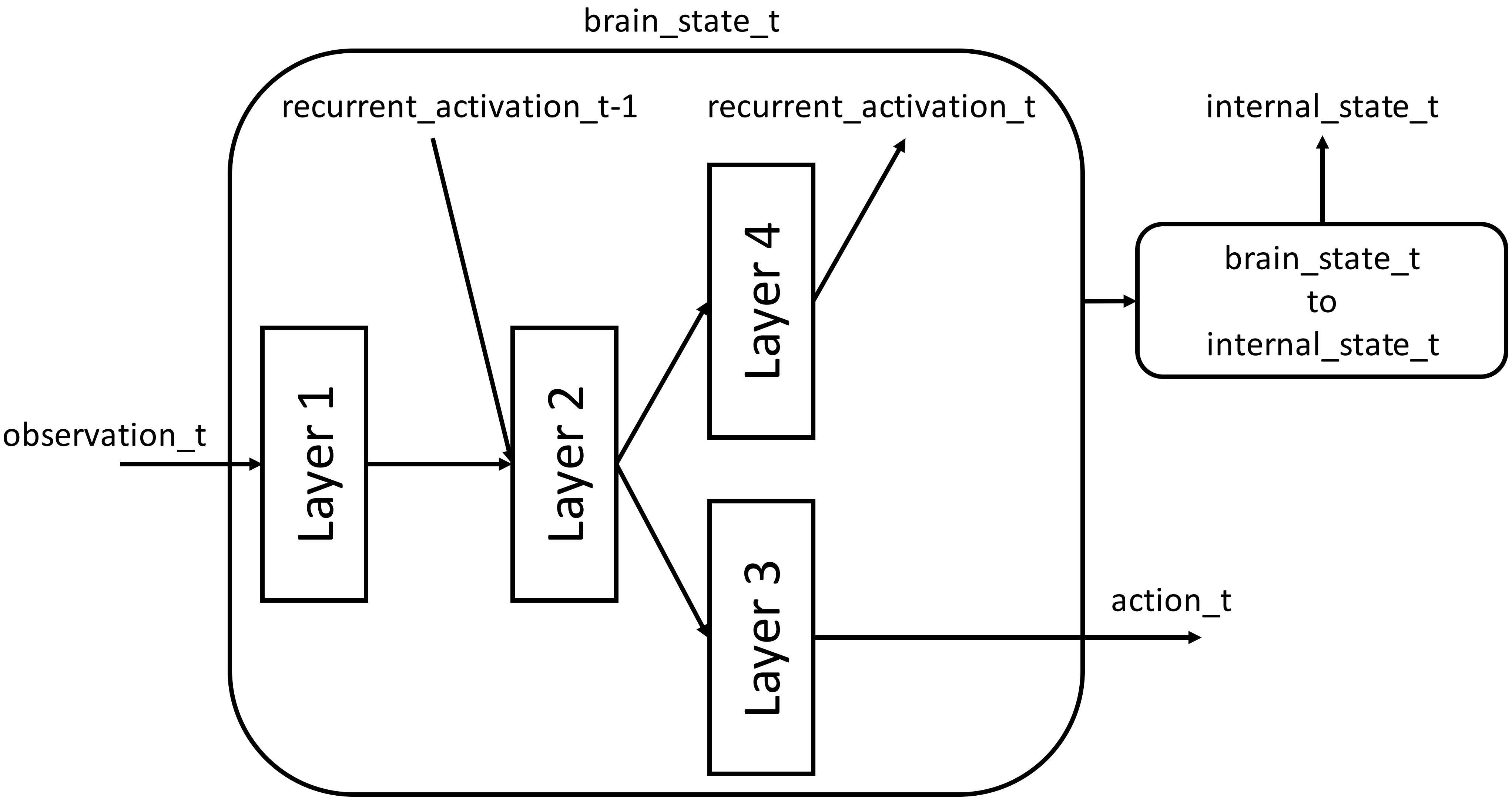
Design, V0


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

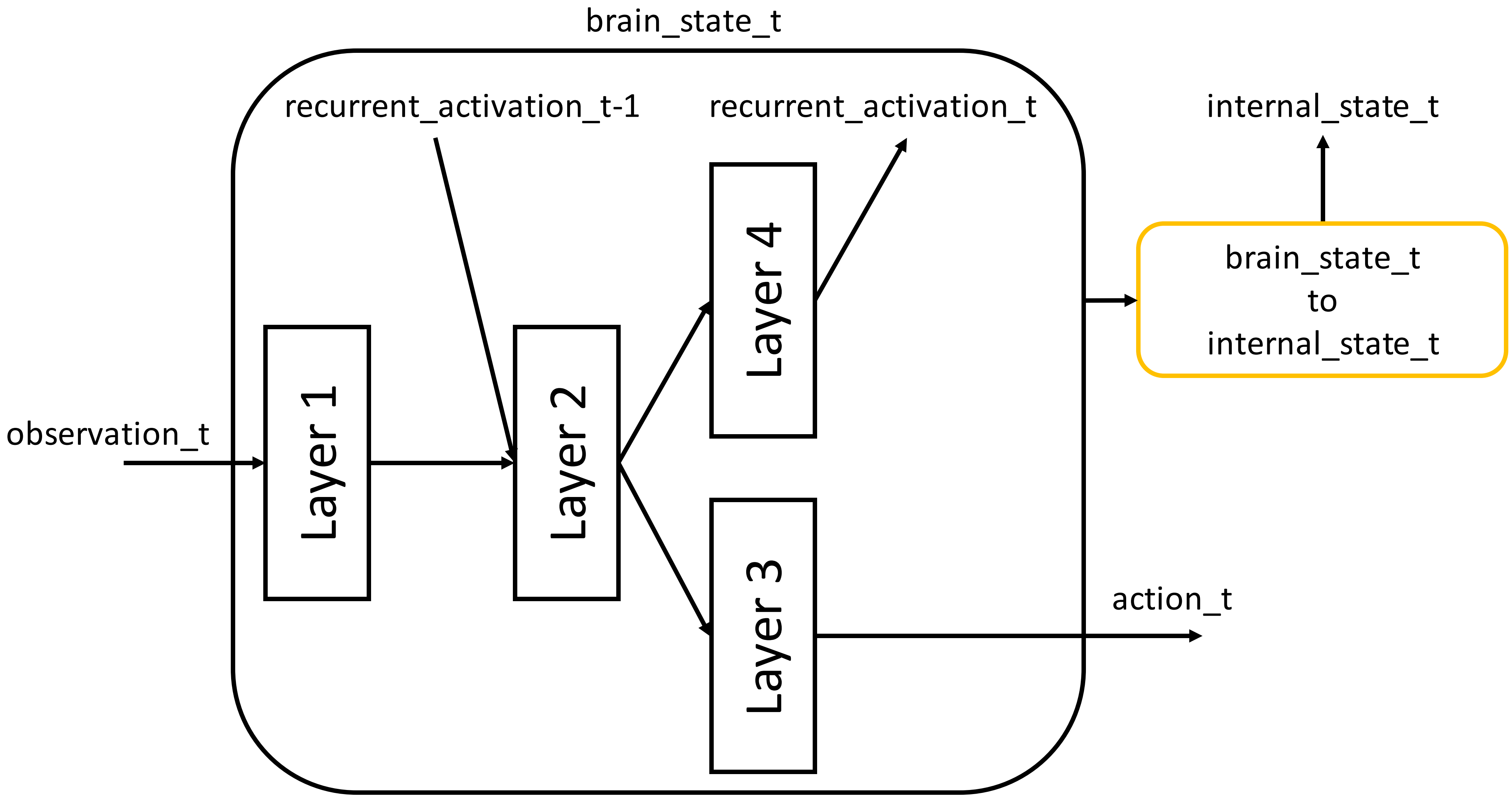
Design, V0


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

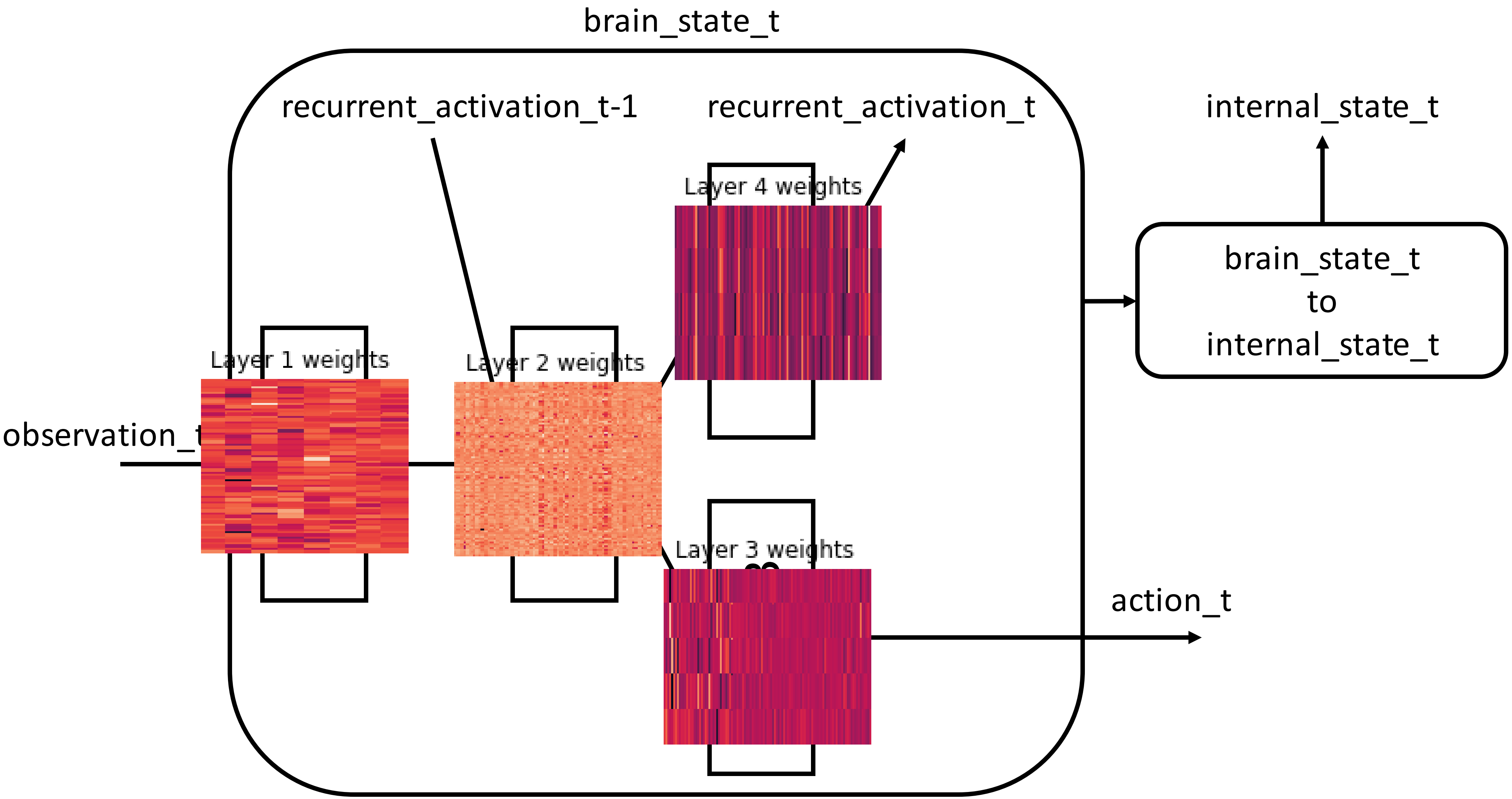
Design, V0


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

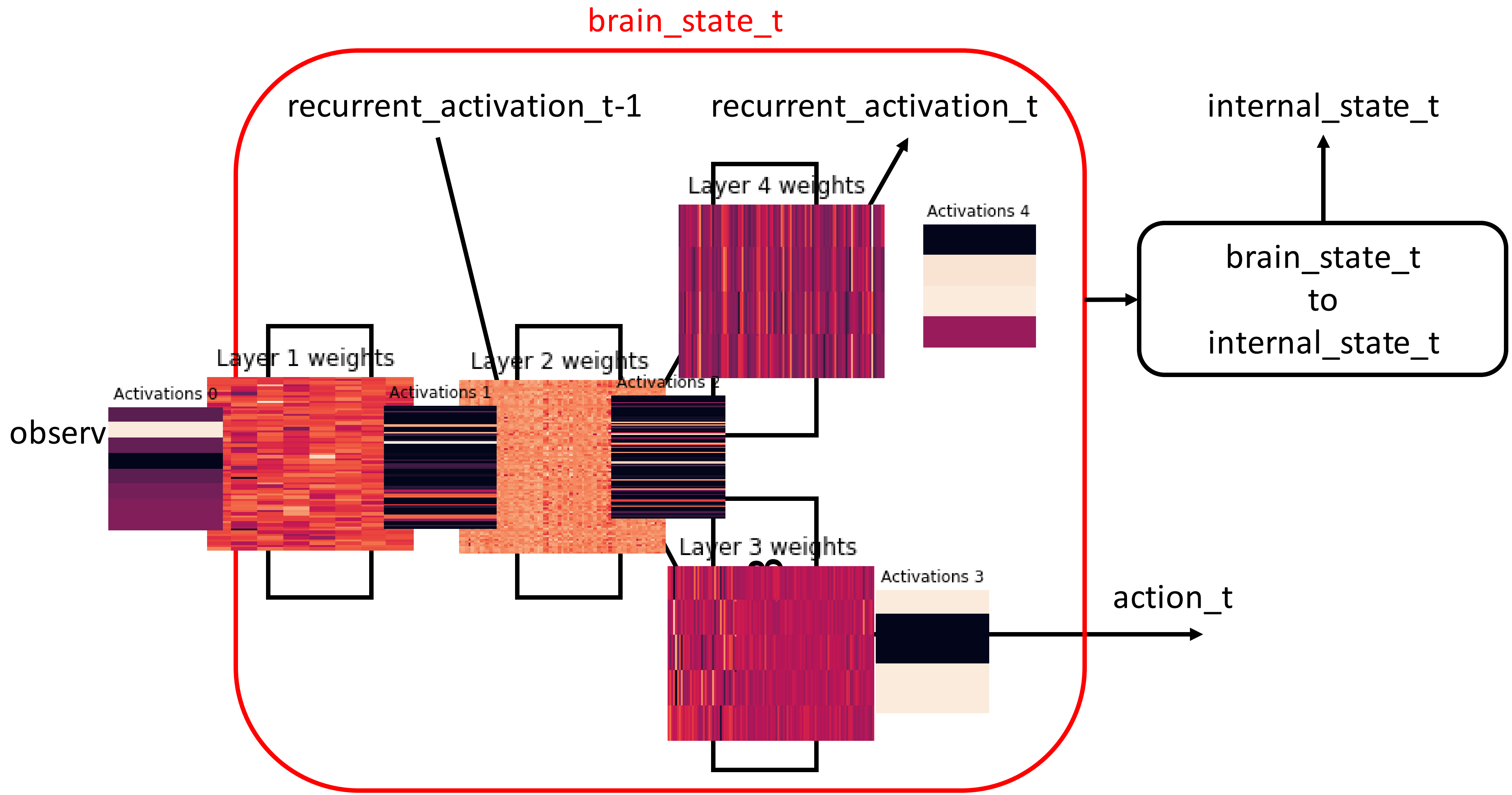
Design, V0


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

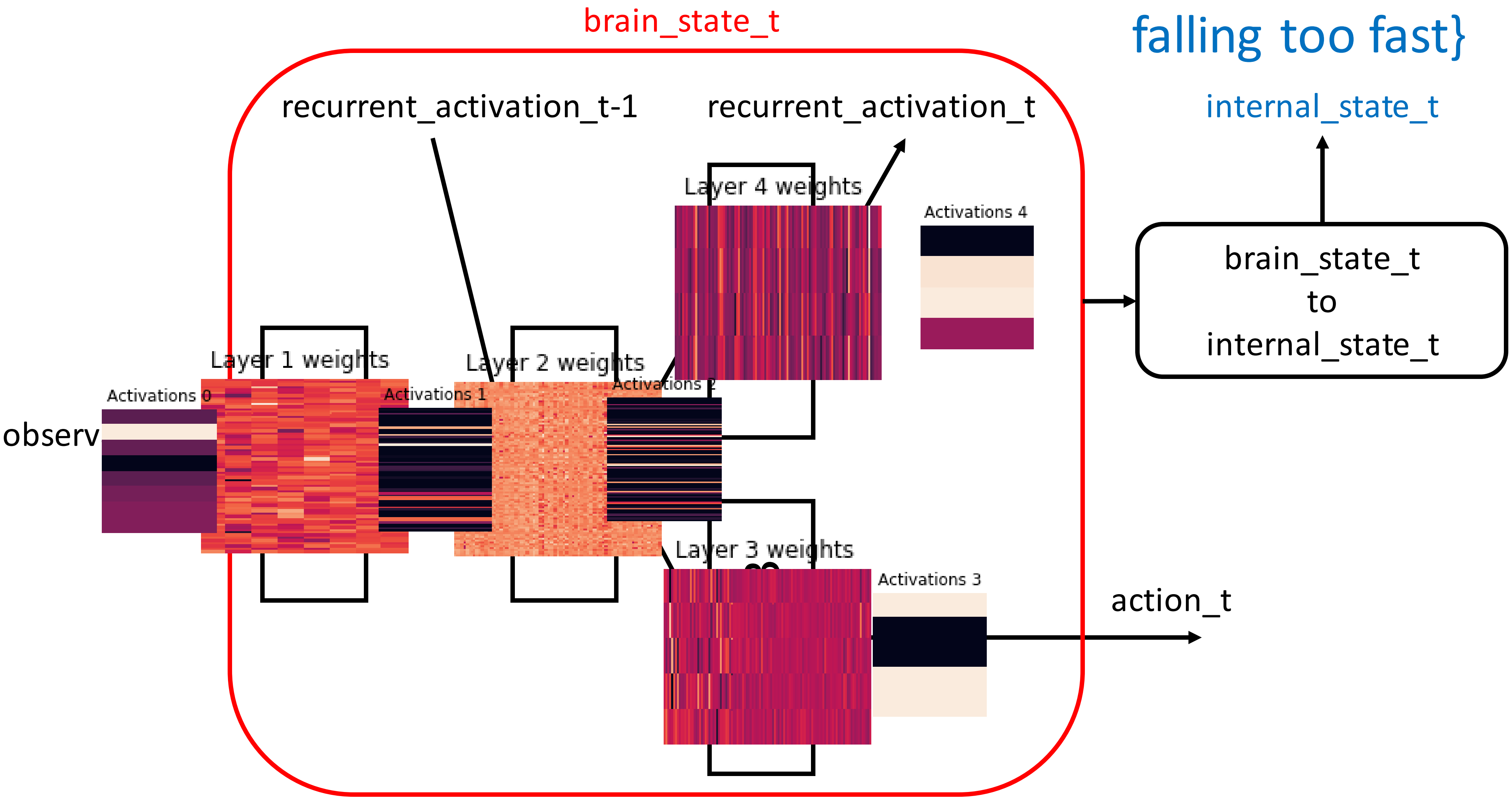
Design, V0



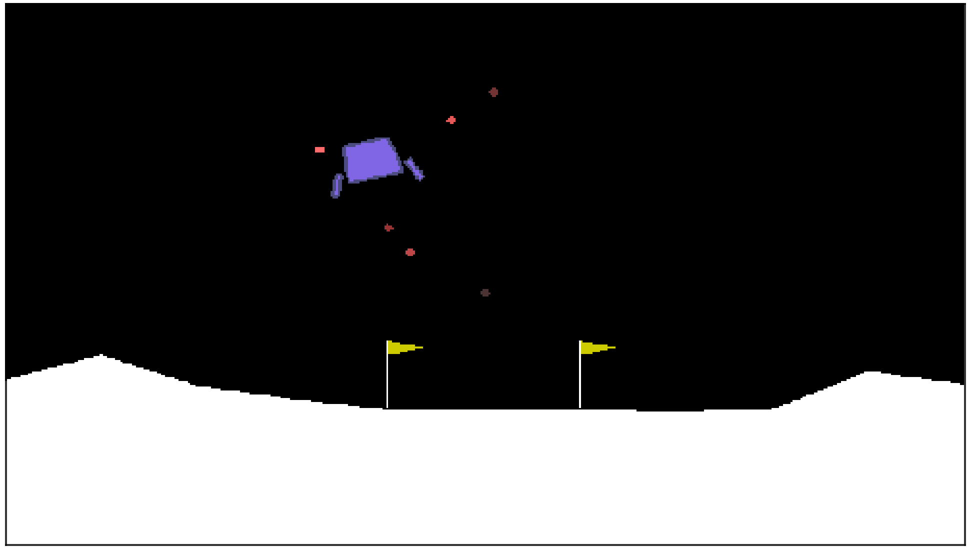
brain_state_t
recurrent_activation_t-1
recurrent_activation_t
internal_state_t
Layer 4 weights
Activations 4
brain_state_t
to
internal_state_t
Layer 1 weights
Layer 2 weights
Activations 0
Activations 1
Activations 2
observ
Layer 3 weights
Activations 3
action_t

Design, V0

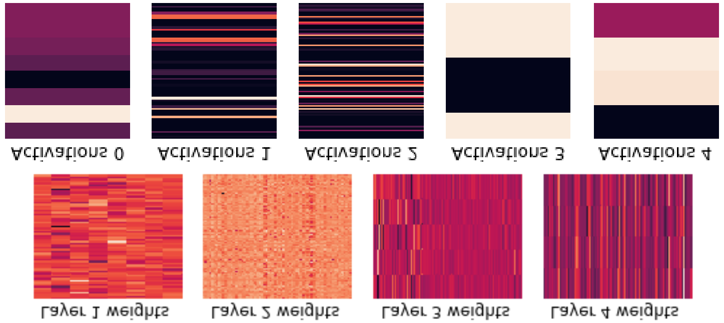
{High above the ground,
right of the center
falling too fast}



Design, V0

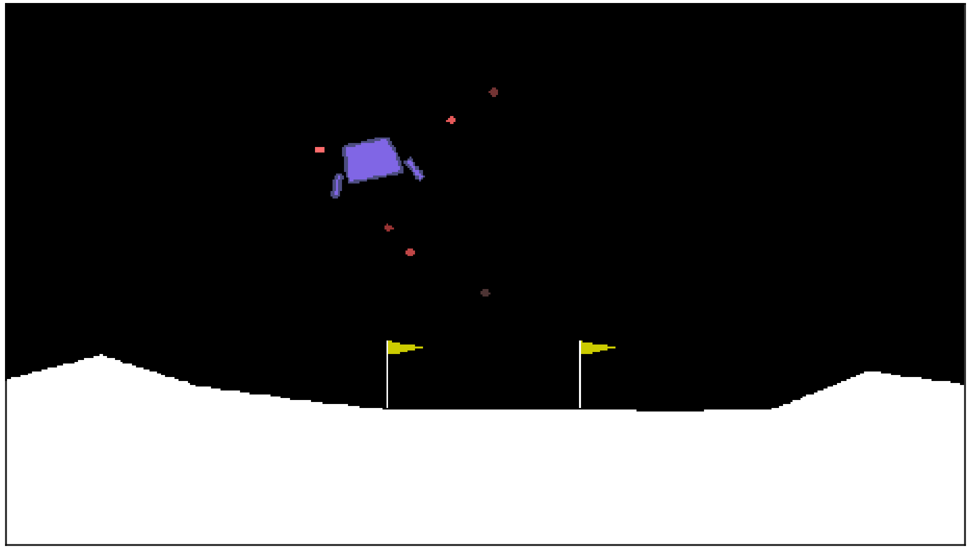


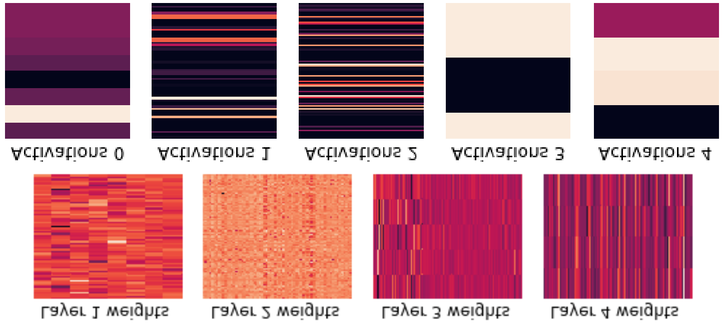
- Design decisions
  - Environment and the agent’s “physical” form
  - Internal state of the agent
    - Beliefs about itself relative to semantically important regions
      - Left of the flags, right of the flags, high above the ground, close to the ground, falling too fast
  - Brain state of the agent



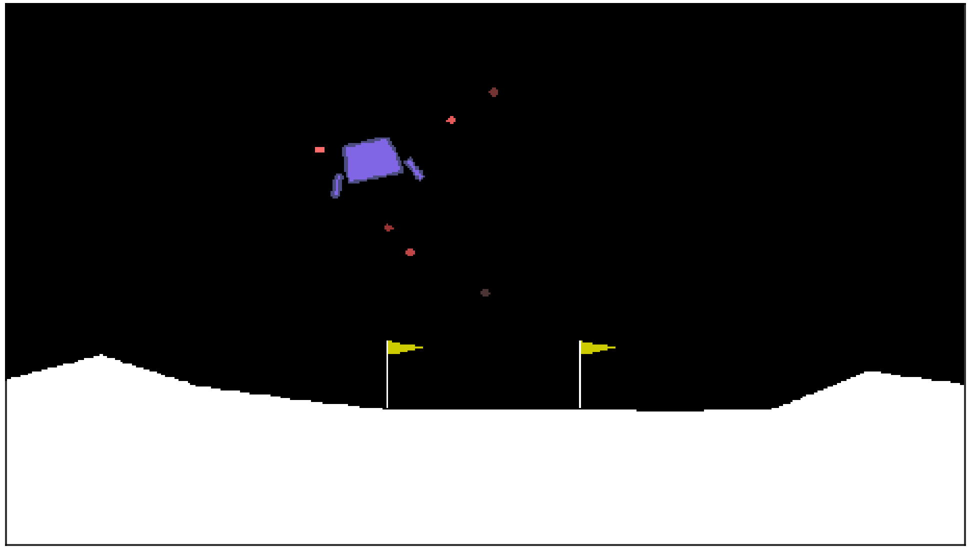
Design, V0

- 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

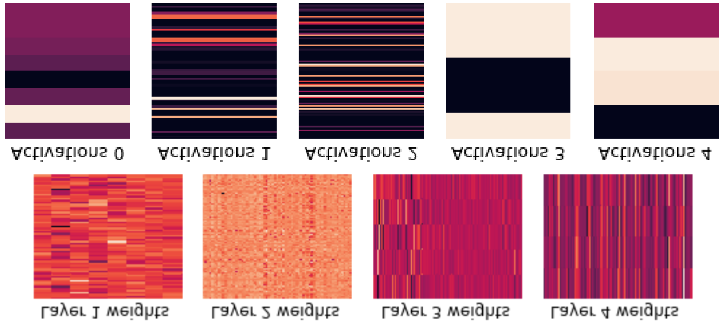




Design, V0



- 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)



Reinforcement learning

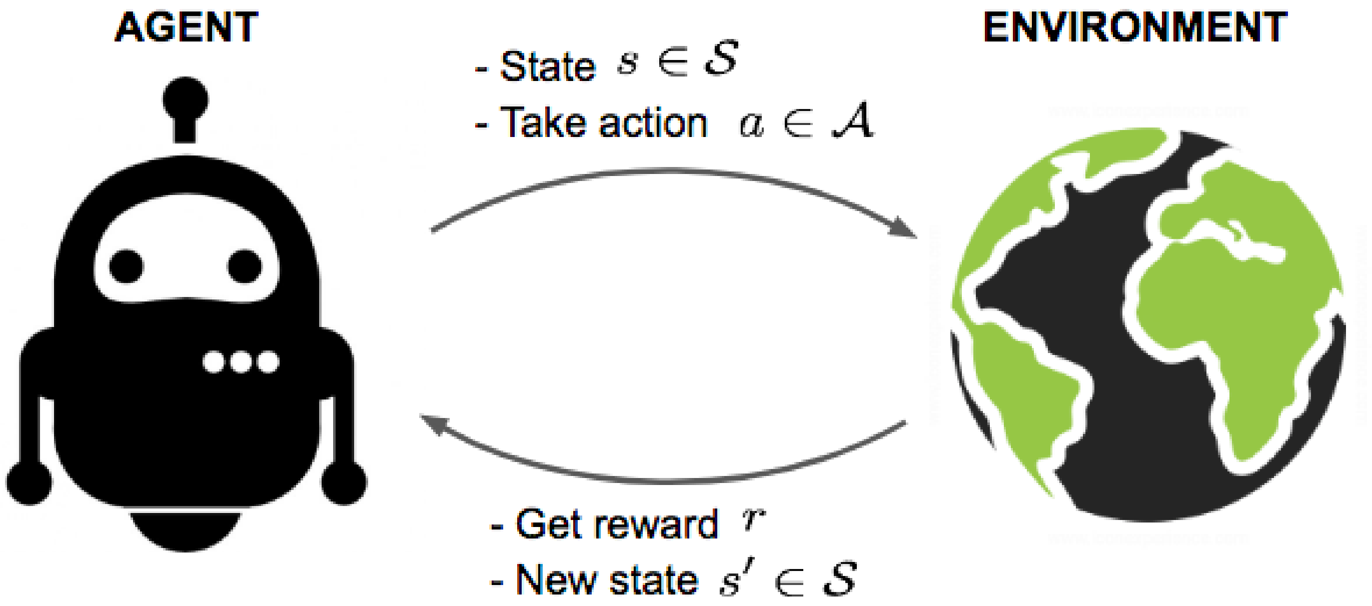

AGENT
ENVIRONMENT
- State $s \in \mathcal{S}$
- Take action $a \in \mathcal{A}$
- Get reward r
- New state $s' \in \mathcal{S}$

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

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 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

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

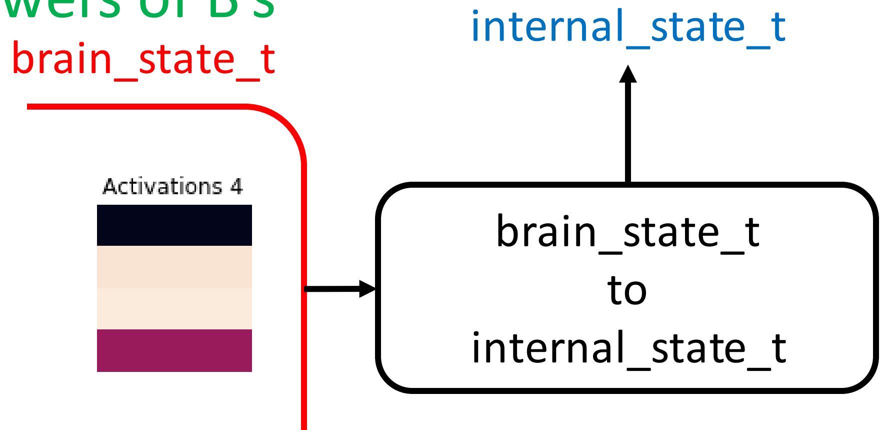
Did we satisfy our requirements?

- 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

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



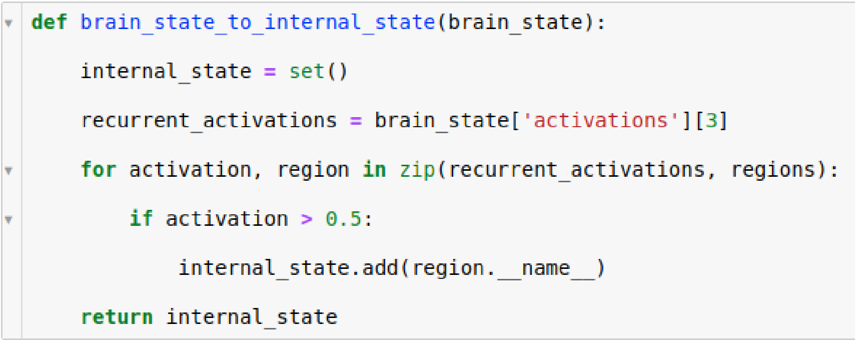
Did we satisfy our requirements?

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

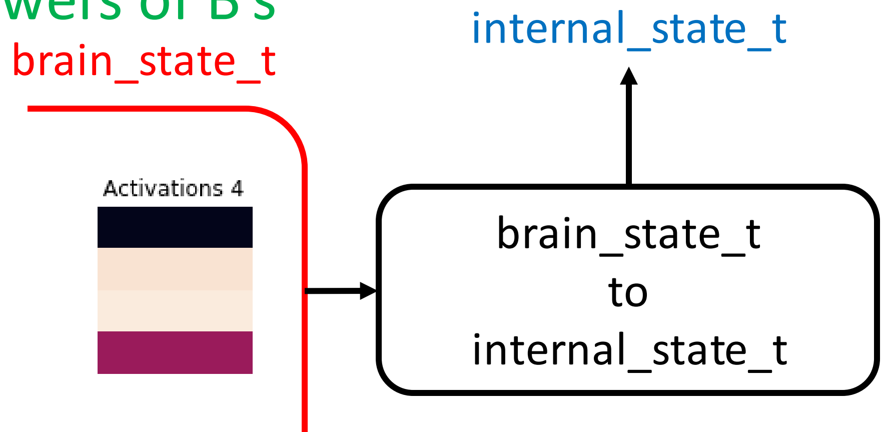
Phenomena of type A are causally reducible to phenomena of type B if and only if:

- 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

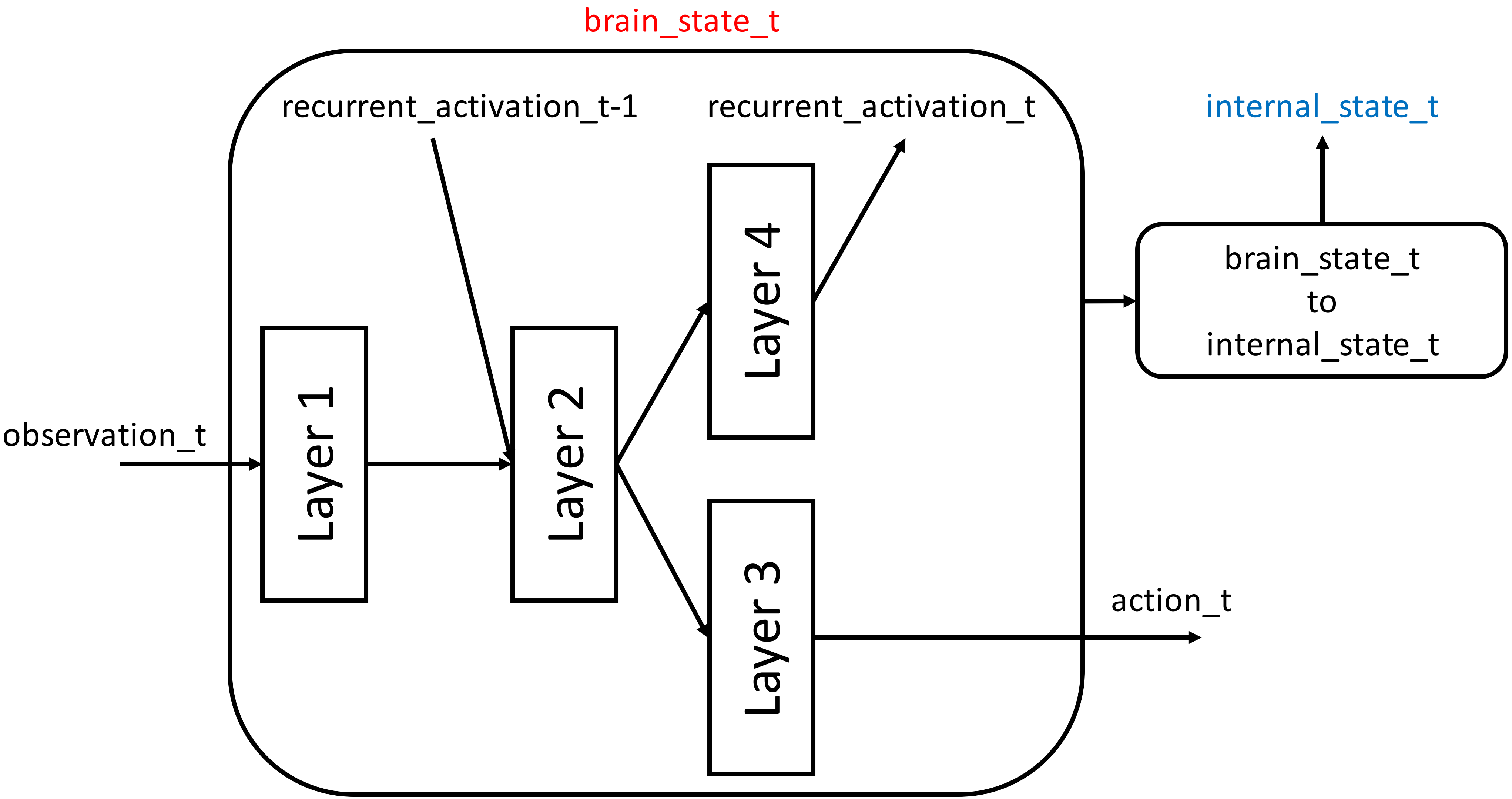
{High above the ground, right of the center falling too fast}



```
def brain_state_to_internal_state(brain_state):
    internal_state = set()
    recurrent_activations = brain_state['activations'][3]
    for activation, region in zip(recurrent_activations, regions):
        if activation > 0.5:
            internal_state.add(region.__name__)
    return internal_state
```



Design, V0


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

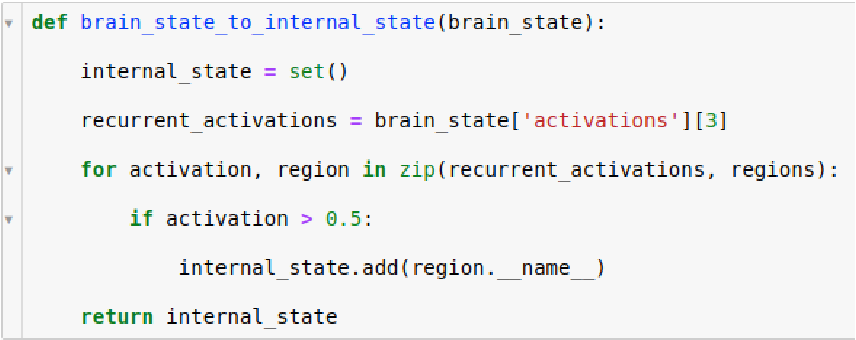
Did we satisfy our requirements?

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

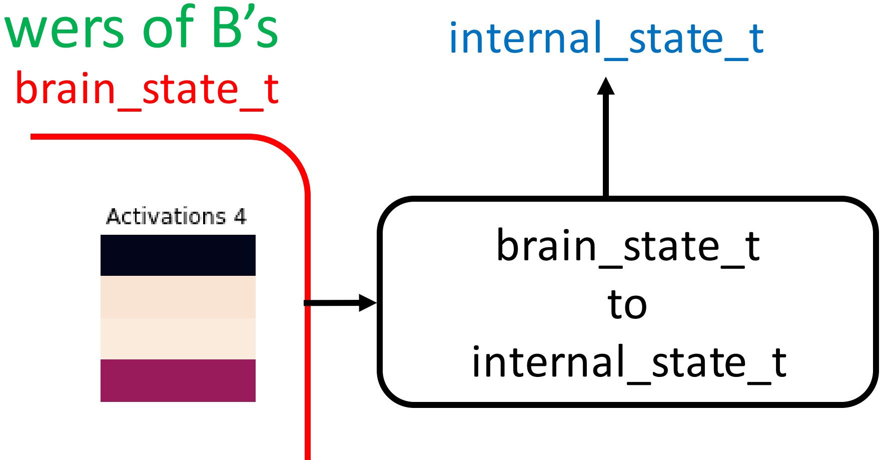
Phenomena of type A are causally reducible to phenomena of type B if and only if:

- 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

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



```
def brain_state_to_internal_state(brain_state):
    internal_state = set()
    recurrent_activations = brain_state['activations'][3]
    for activation, region in zip(recurrent_activations, regions):
        if activation > 0.5:
            internal_state.add(region.__name__)
    return internal_state
```



internal_state_t

brain_state_t

brain_state_t
to
internal_state_t

Did we satisfy our requirements?

- 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

Did we satisfy our requirements?

- 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

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?

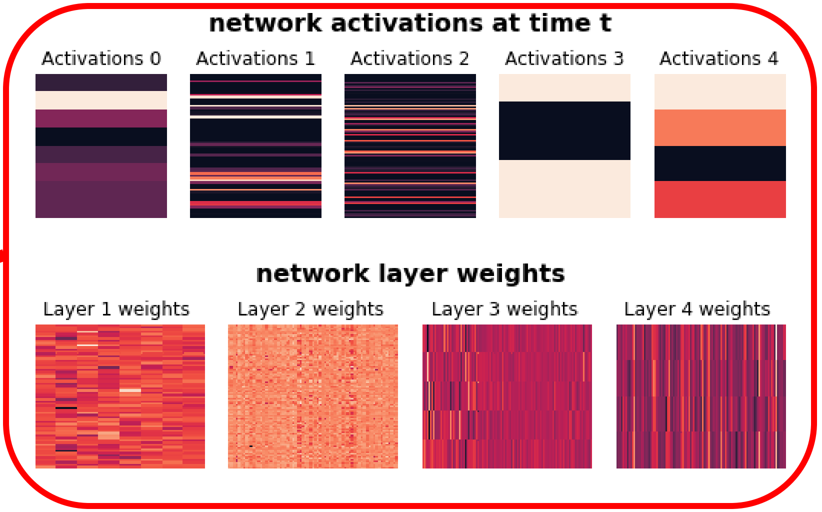
- V0
  - ✓ Internal states are casually reducible to brain states
  - Internal states are ontologically irreducible to brain states

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
- 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)

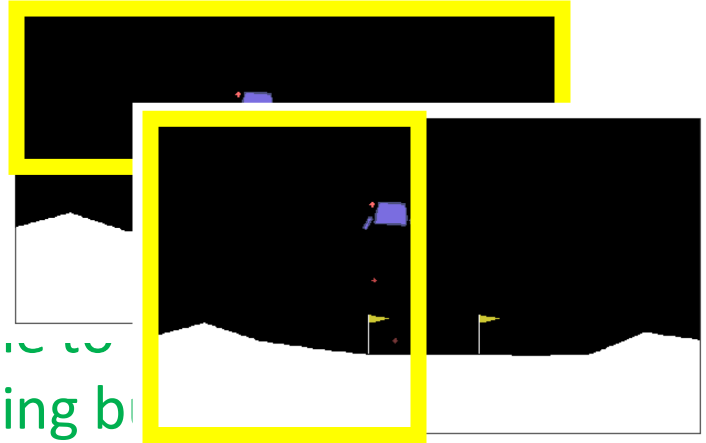
```
Internal state:
{'I_am_high_above_the_ground', 'I_am_to_the_right_of_the_center', 'I_am_falling_too_fast'}
```



Did we satisfy our requirements?

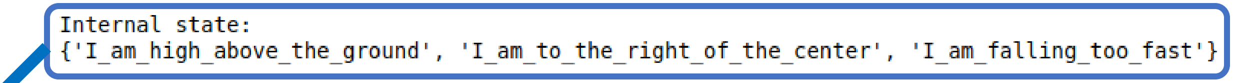
- V0
  - ✓ Internal states are casually reducible to br
  - Internal states are ontologically irreducibl

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

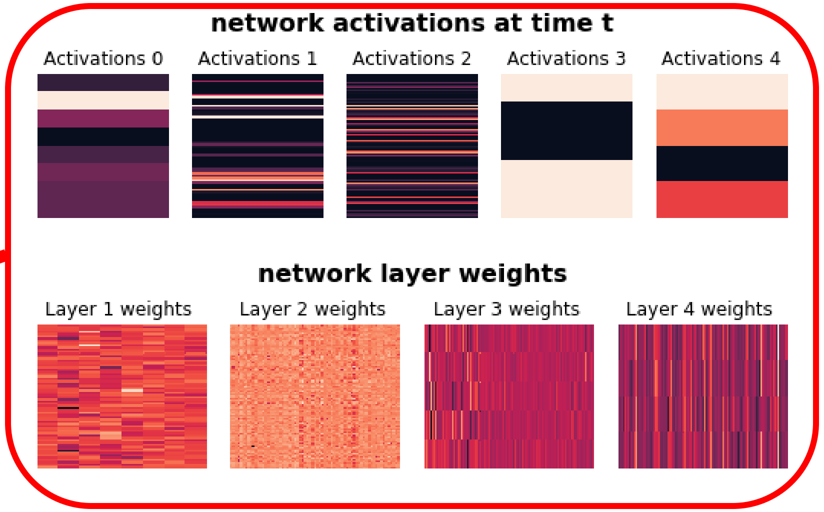


Our ontology

- 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)



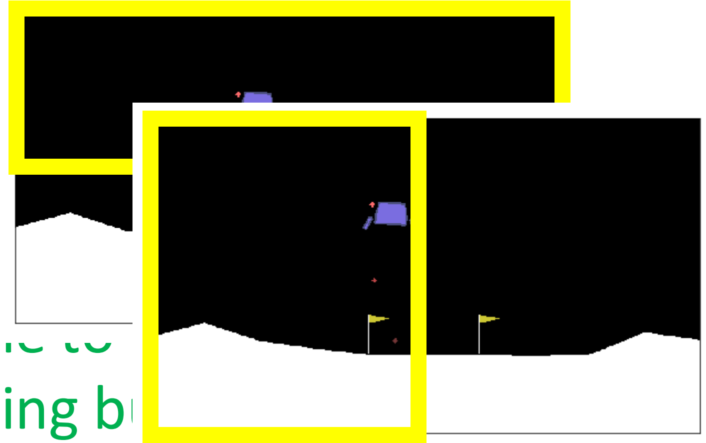
```
Internal state:
{'I_am_high_above_the_ground', 'I_am_to_the_right_of_the_center', 'I_am_falling_too_fast'}
```



Did we satisfy our requirements?

- V0
  - ✓ Internal states are casually reducible to br
  - Internal states are ontologically irreducibl

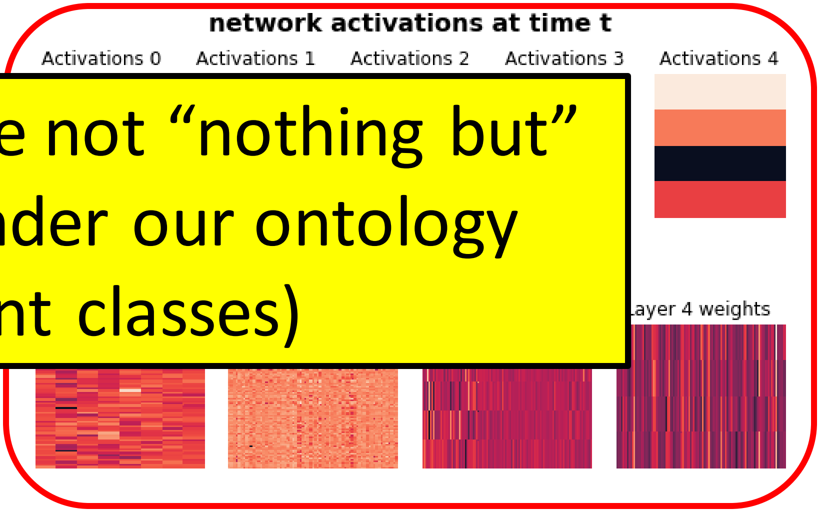
Phenomena of type A are ontologically reducib... to phenomena of type B if and only if A's are nothing bu



Our ontology

- Layer weights of the neural network
- Connectivity of the neural network
- Activations of the neural n
- The agent's observation at
- The agent's action at time
- 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'}
```

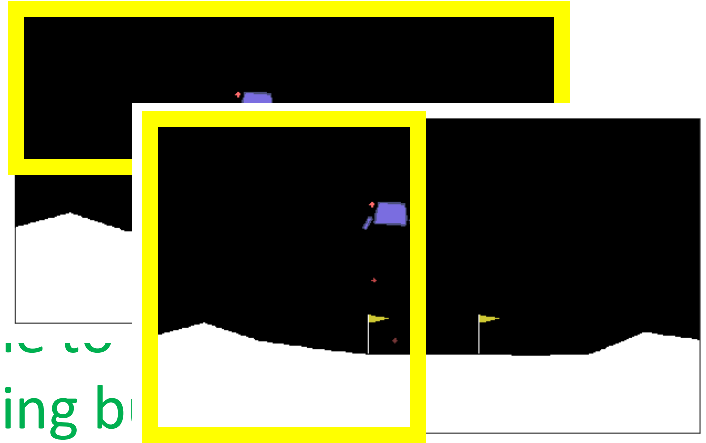


Internal state instances are not "nothing but" brain state instances under our ontology (they are different classes)

Did we satisfy our requirements?

- V0
  - ✓ Internal states are casually reducible to b
  - ✓ Internal states are ontologically irreducibl

Phenomena of type A are ontologically reducib to
phenomena of type B if and only if A's are nothing b

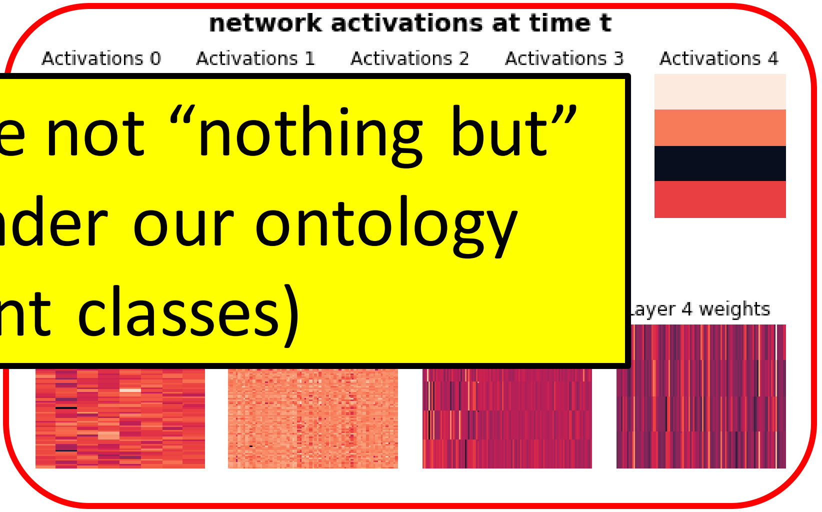


Our ontology

- Layer weights of the neural network
- Connectivity of the neural network
- Activations of the neural n
- The agent's observation at
- The agent's action at time
- The position and velocity c
- 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 instances are not "nothing but" brain state instances under our ontology (they are different classes)

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



Is that the “real” ontology though?

- 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

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

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
- 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

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
- 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

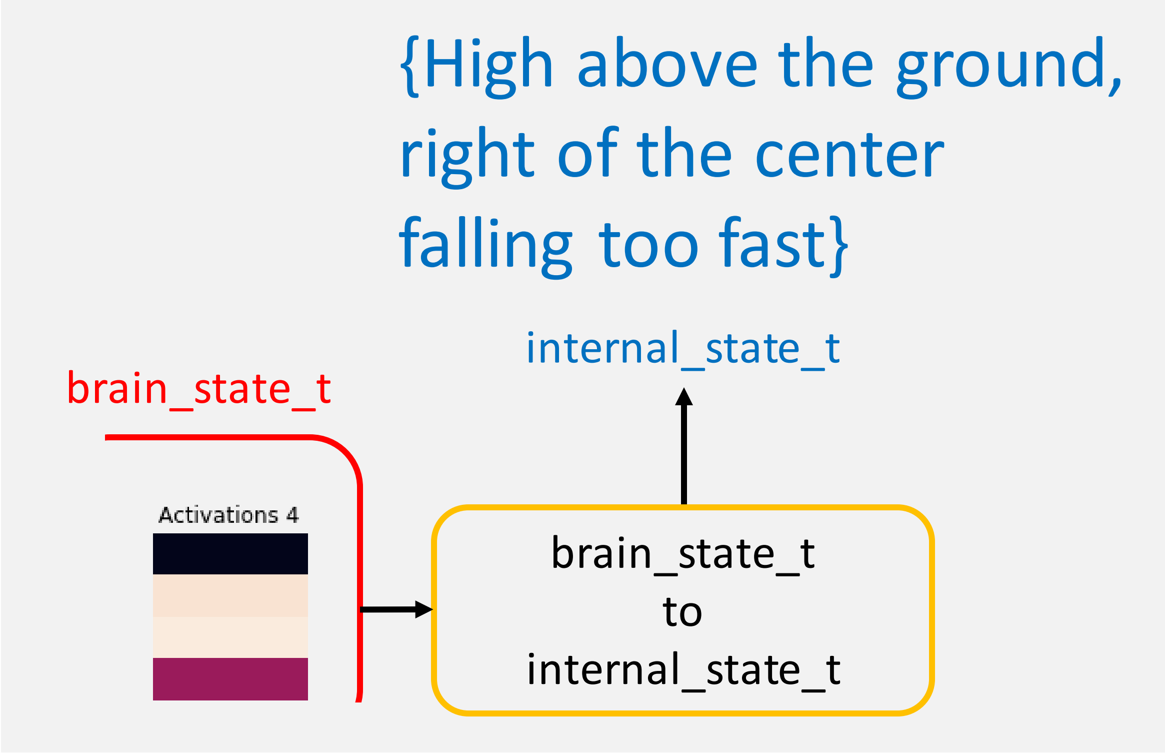
✗ Internal states are ontologically irreducible to brain states

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
- 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
- …

What's the deal with that function?

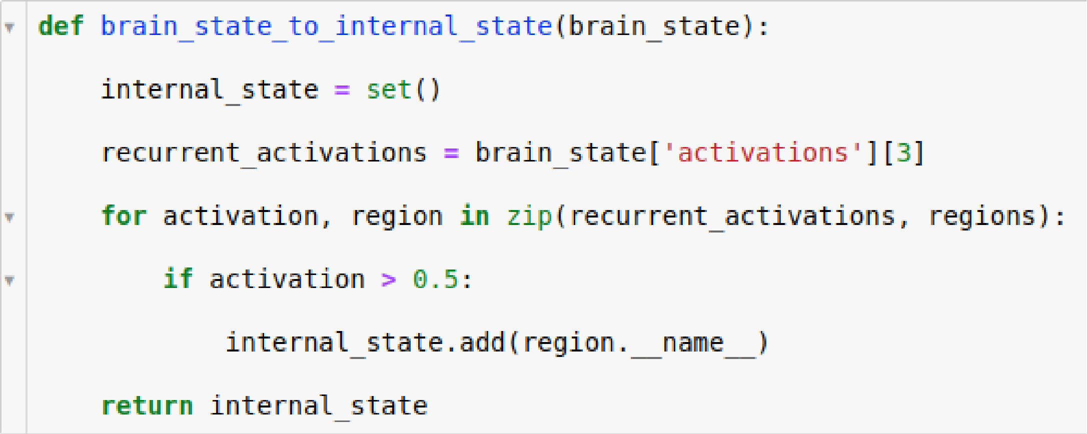


{High above the ground,
right of the center
falling too fast}

internal_state_t

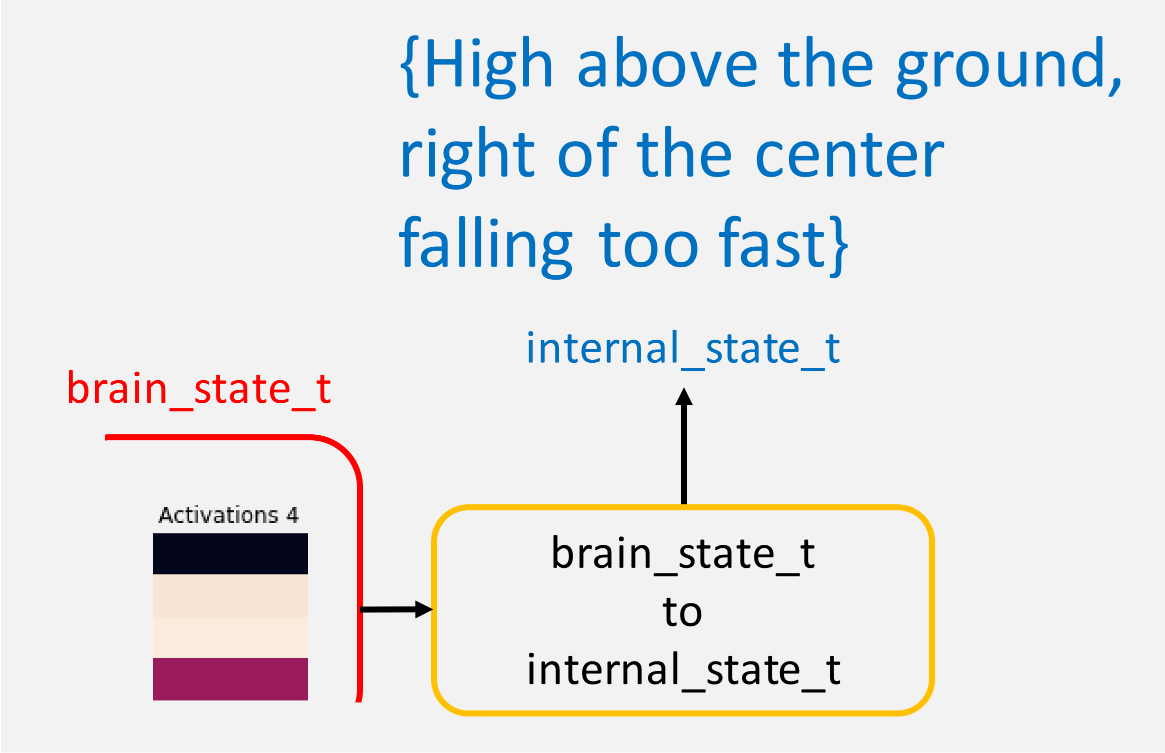
brain_state_t

brain_state_t
to
internal_state_t



```python
def brain_state_to_internal_state(brain_state):
    internal_state = set()
    recurrent_activations = brain_state['activations'][3]
    for activation, region in zip(recurrent_activations, regions):
        if activation > 0.5:
            internal_state.add(region.__name__)
    return internal_state
```

What's the deal with that function?

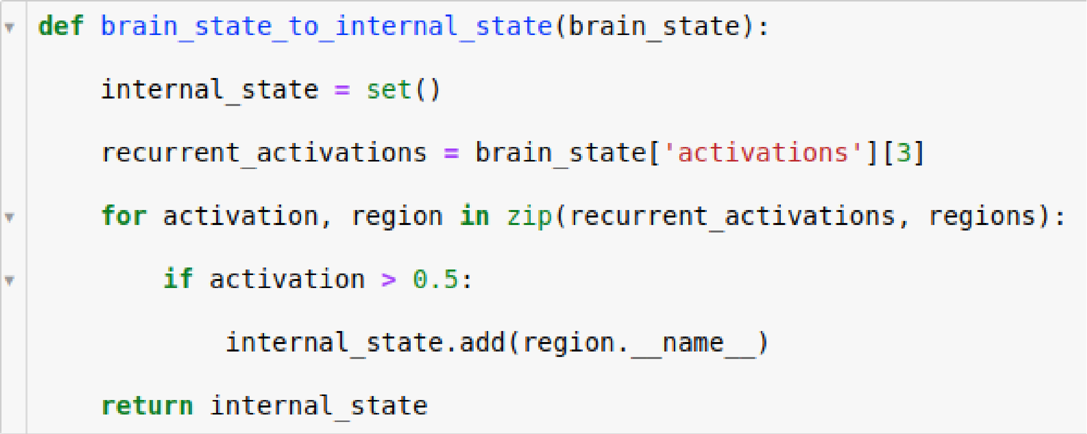


{High above the ground,
right of the center
falling too fast}

internal_state_t

brain_state_t

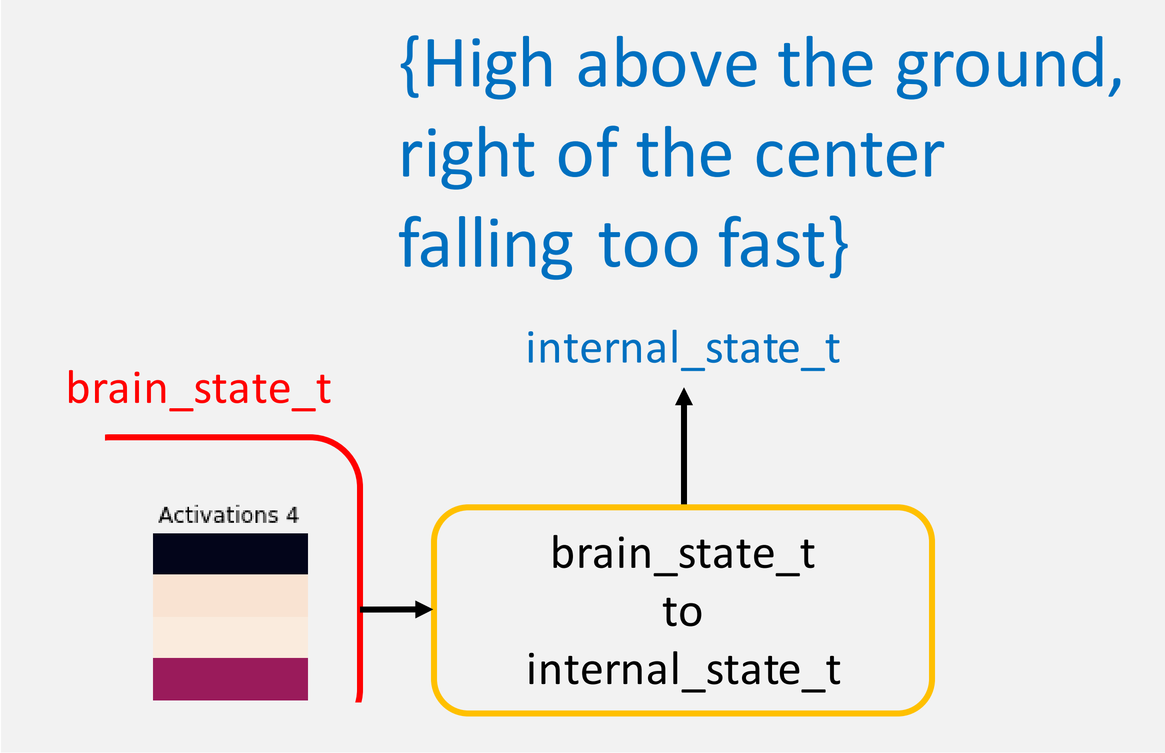
brain_state_t
to
internal_state_t



```
def brain_state_to_internal_state(brain_state):
    internal_state = set()
    recurrent_activations = brain_state['activations'][3]
    for activation, region in zip(recurrent_activations, regions):
        if activation > 0.5:
            internal_state.add(region.__name__)
    return internal_state
```

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

What's the deal with that function?

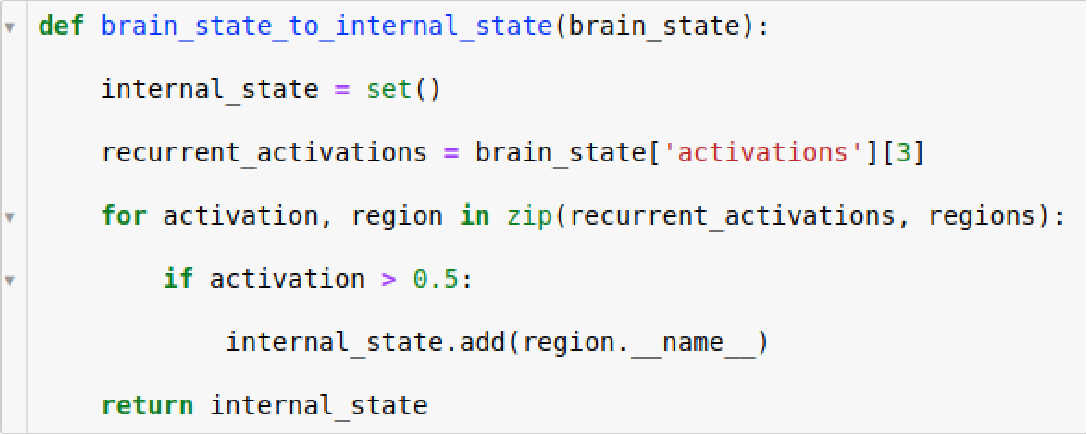


{High above the ground,
right of the center
falling too fast}

internal_state_t

brain_state_t

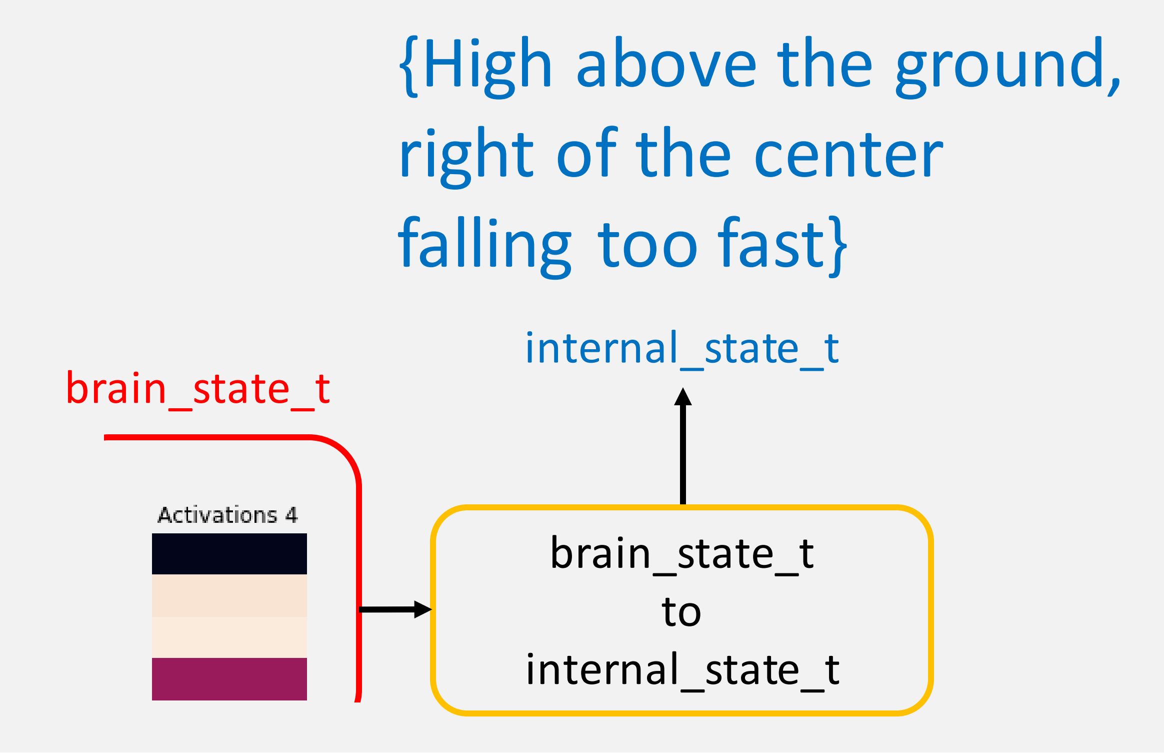
brain_state_t
to
internal_state_t



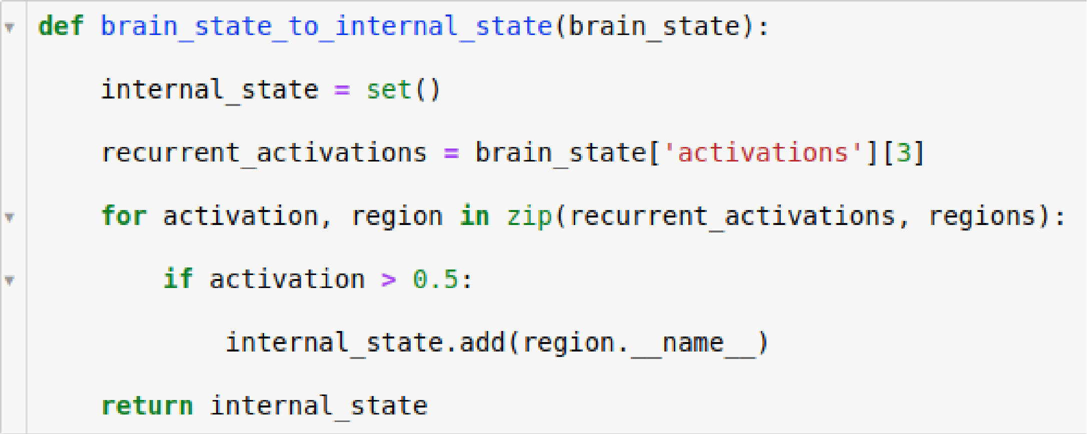
```
def brain_state_to_internal_state(brain_state):
    internal_state = set()
    recurrent_activations = brain_state['activations'][3]
    for activation, region in zip(recurrent_activations, regions):
        if activation > 0.5:
            internal_state.add(region.__name__)
    return internal_state
```

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

What's the deal with that function?



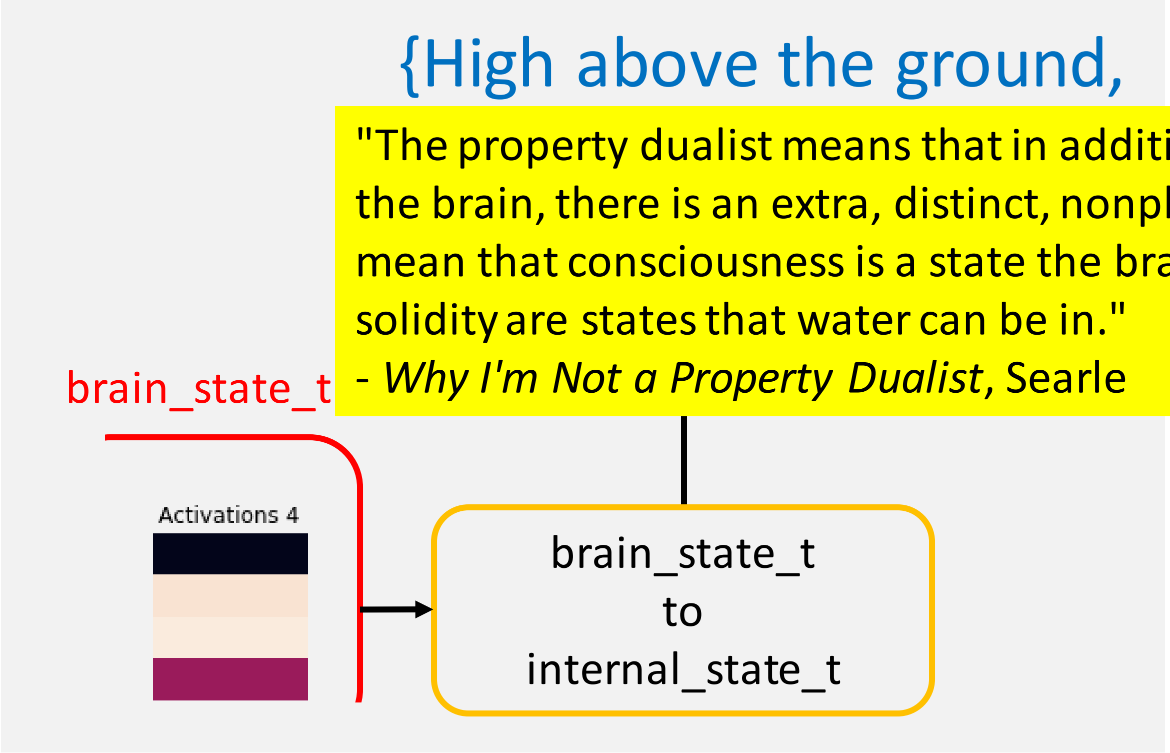
{High above the ground, right of the center falling too fast}



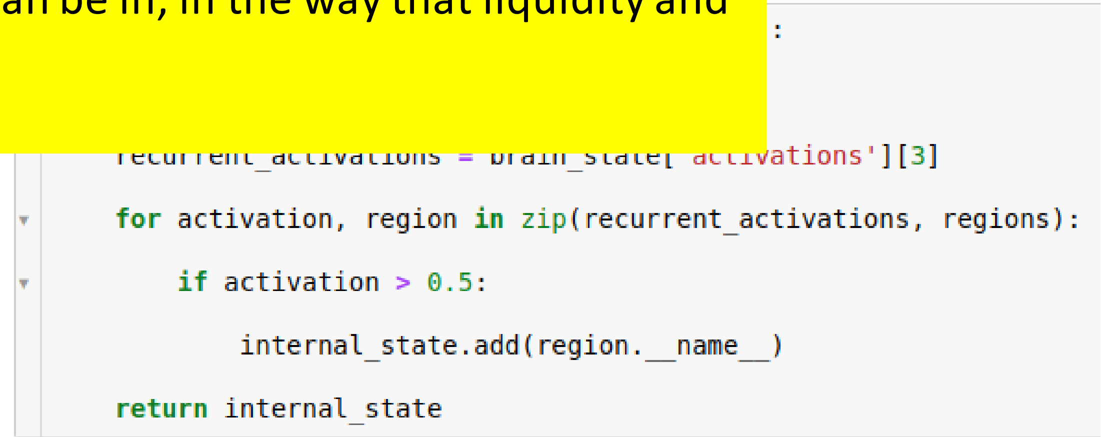
```python
def brain_state_to_internal_state(brain_state):
    internal_state = set()
    recurrent_activations = brain_state['activations'][3]
    for activation, region in zip(recurrent_activations, regions):
        if activation > 0.5:
            internal_state.add(region.__name__)
    return internal_state
```

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

What's the deal with that function?



"The property dualist means that in addition to all the neurobiological features of the brain, there is an extra, distinct, nonphysical feature of the brain; whereas I mean that 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*, Searle



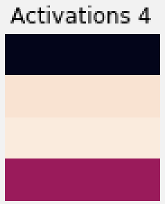
```
:

    recurrent_activations = brain_state['activations'][3]
    for activation, region in zip(recurrent_activations, regions):
        if activation > 0.5:
            internal_state.add(region.__name__)
    return internal_state
```

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

What's the deal with that function?

{High above the ground,

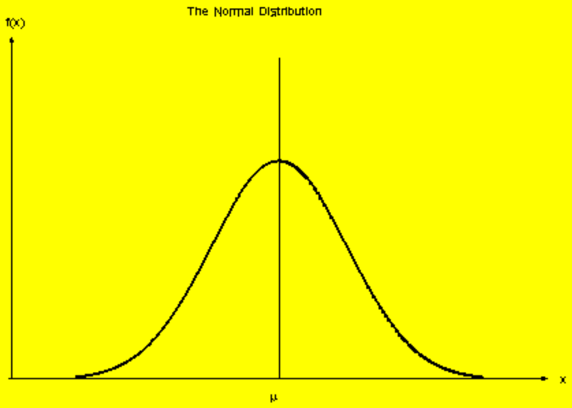
brain_state_t



```
:
ations'][3]
tivations, regions):
)
```

"The property dualist means that in addition to all the neurobiological features of the brain, there is an extra, distinct, nonphysical feature of the brain; whereas I mean that 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*, Searle

Just like a gaussian and its parameters...



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

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

- Is this jus
- Is this so
- (or both?)

Conclusion

- Software engineer style philosophy reifying seemed to work well
- Created a V0 software agent who's
  - Internal states are casually 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 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?