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
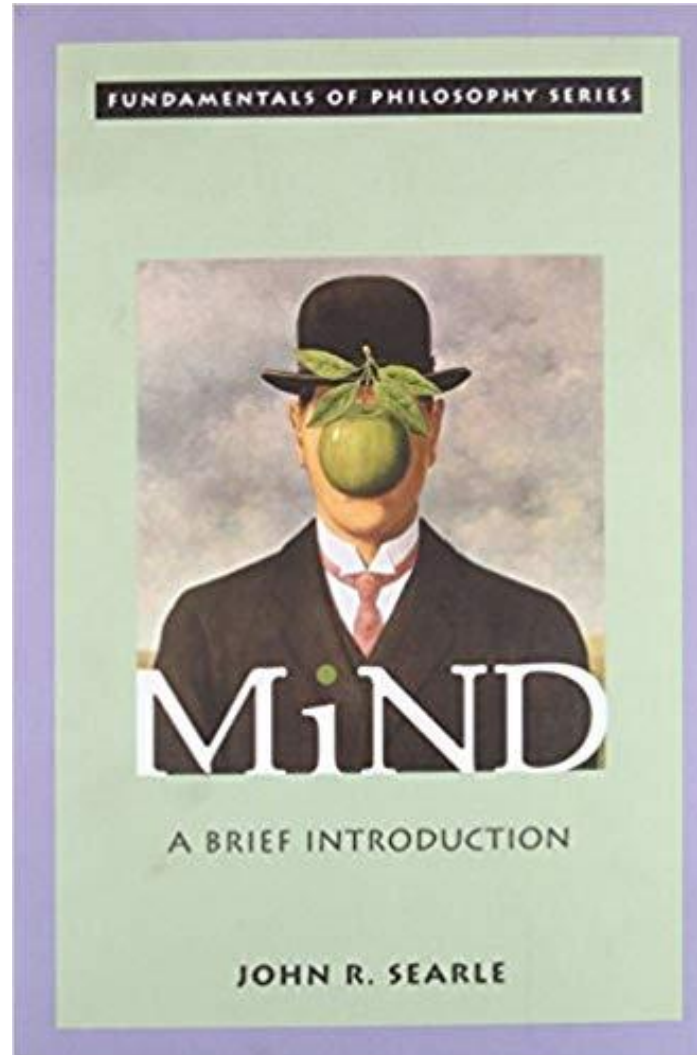
Reifying philosophy with code

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
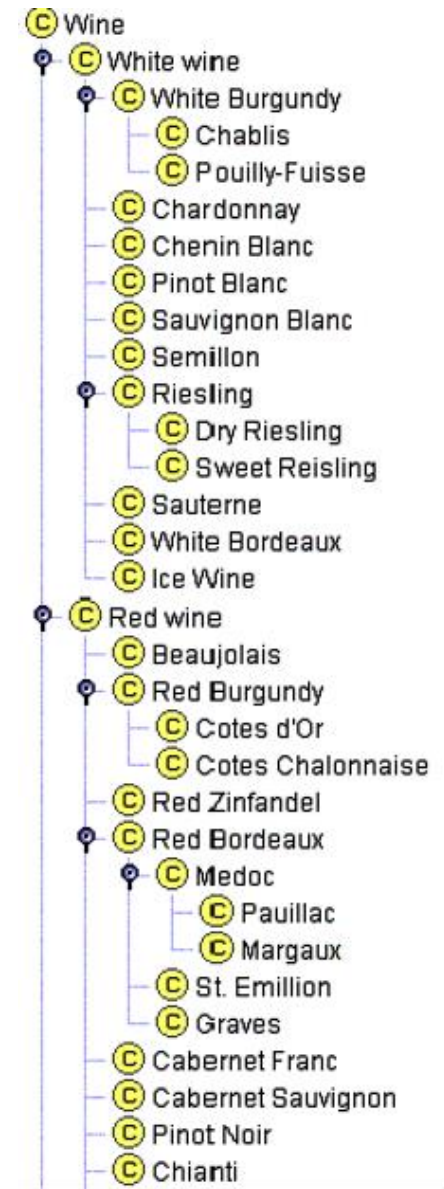
Requirements: unpacking Searle's view
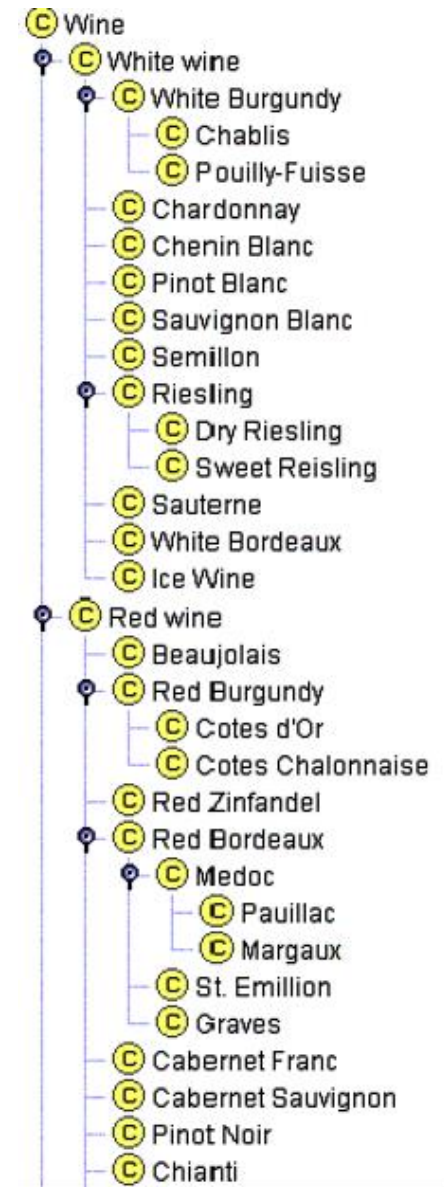
- Searle's view
  - Consciousness is causally reducible to brain states
  - Consciousness is ontologically irreducible to brain states
- V2
  - Conscious mental states are causally reducible to brain states
  - Conscious mental states are ontologically irreducible to brain states
- V1
  - Mental states 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 Wine
  C White wine
    C White Burgundy
      C Chablis
      C Pouilly-Fuisse
    C Chardonnay
    C Chenin Blanc
    C Pinot Blanc
    C Sauvignon Blanc
    C Semillon
    C Riesling
      C Dry Riesling
      C Sweet Reisling
    C Sauterne
    C White Bordeaux
    C Ice Wine
  C Red wine
    C Beaujolais
    C Red Burgundy
      C Cotes d'Or
      C Cotes Chalonnaise
    C Red Zinfandel
    C Red Bordeaux
      C Medoc
        C Pauillac
        C Margaux
      C St. Emillion
      C Graves
    C Cabernet Franc
    C Cabernet Sauvignon
    C Pinot Noir
    C Chianti

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

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

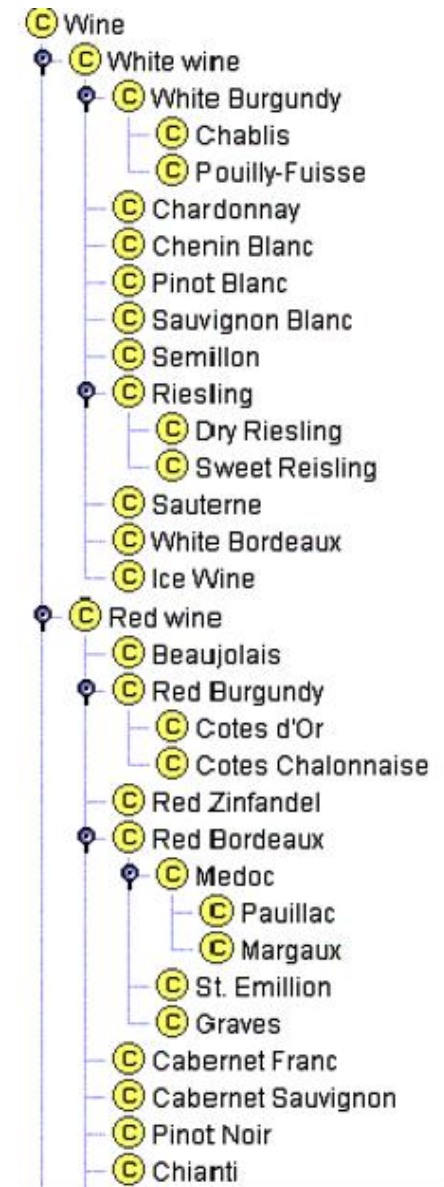# Ontologies in Computer Science

- 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:

Ontologies in Computer Science

- Class-instance distinction

- Type-token distinction

(C) Wine
 (C) White wine
  (C) White Burgundy
   (C) Chablis
   (C) Pouilly-Fuisse
  (C) Chardonnay
  (C) Chenin Blanc
  (C) Pinot Blanc
  (C) Sauvignon Blanc
  (C) Semillon
  (C) Riesling
   (C) Dry Riesling
   (C) Sweet Reisling
  (C) Sauterne
  (C) White Bordeaux
  (C) Ice Wine
 (C) Red wine
  (C) Beaujolais
  (C) Red Burgundy
   (C) Cotes d'Or
   (C) Cotes Chalonnaise
  (C) Red Zinfandel
  (C) Red Bordeaux
   (C) Medoc
    (C) Pauillac
    (C) Margaux
   (C) St. Emillion
   (C) Graves
  (C) Cabernet Franc
  (C) Cabernet Sauvignon
  (C) Pinot Noir
  (C) 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
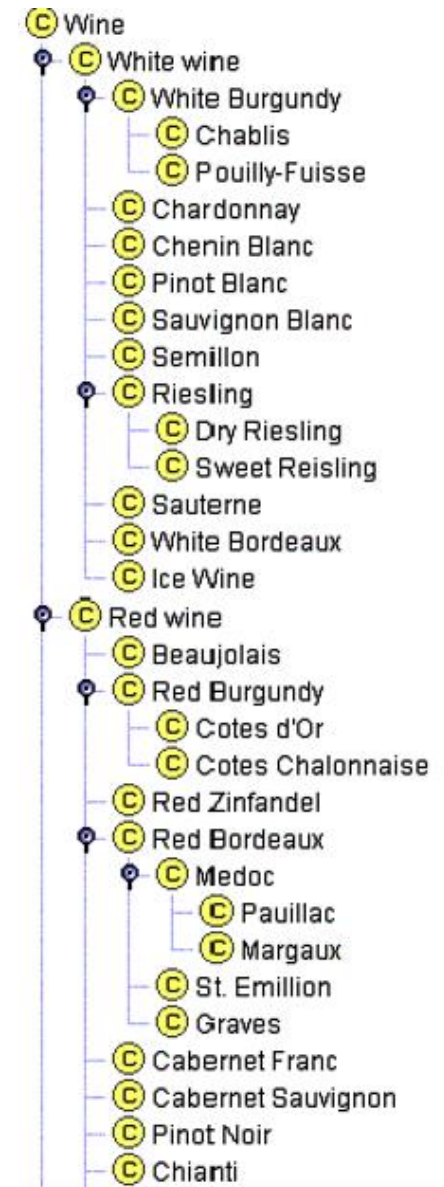   - (the 2003 Toyota Corolla with VIN: 2QFBORHE4KP911561)

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

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

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

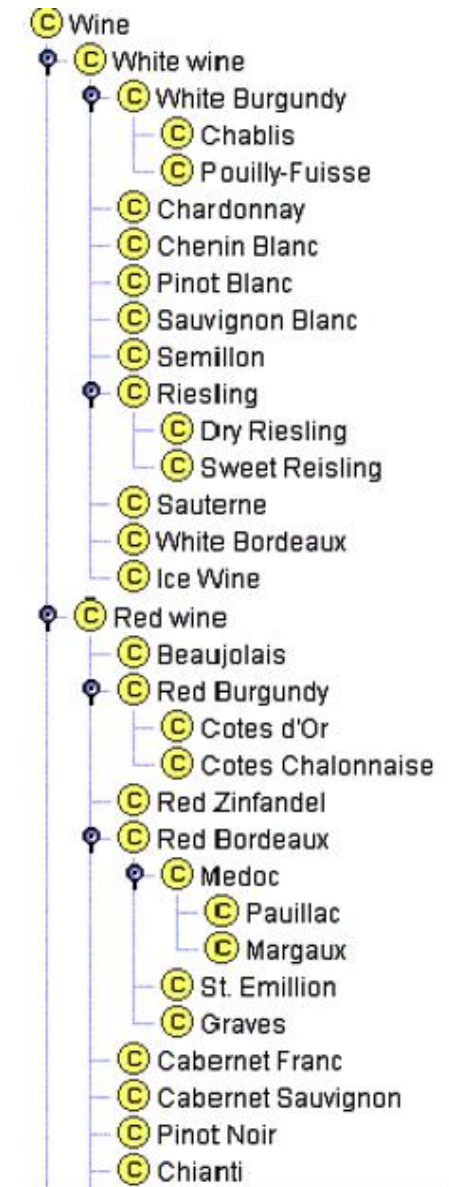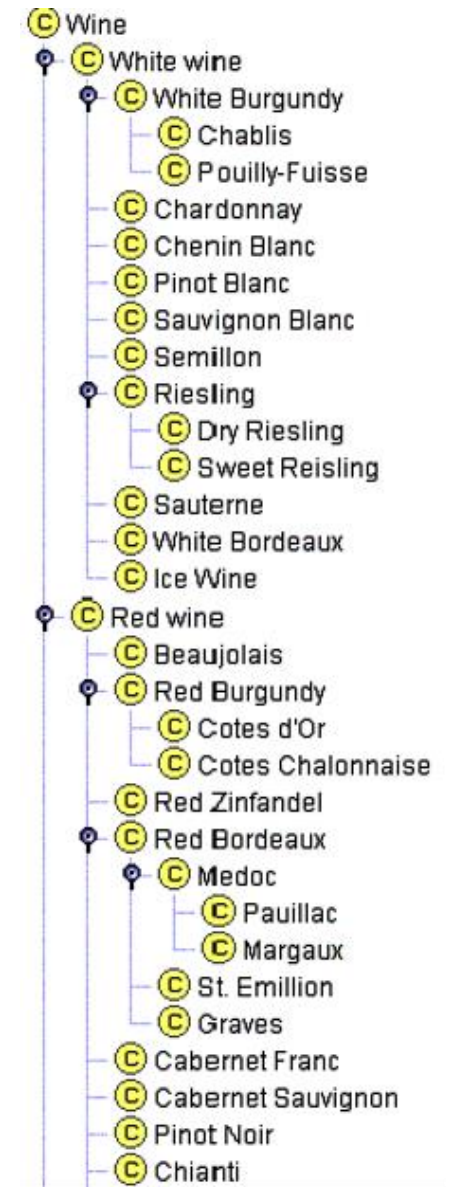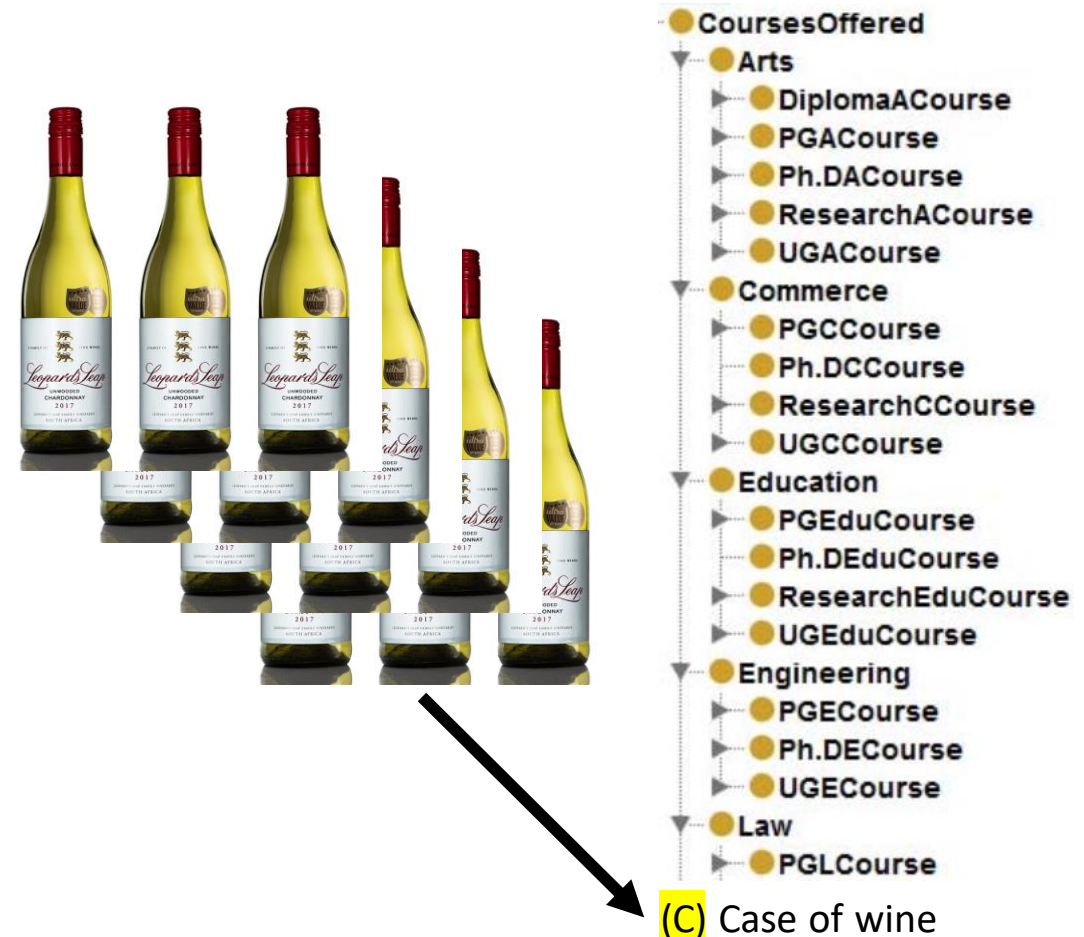# Ontologies in Computer Science

- 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

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

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

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

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

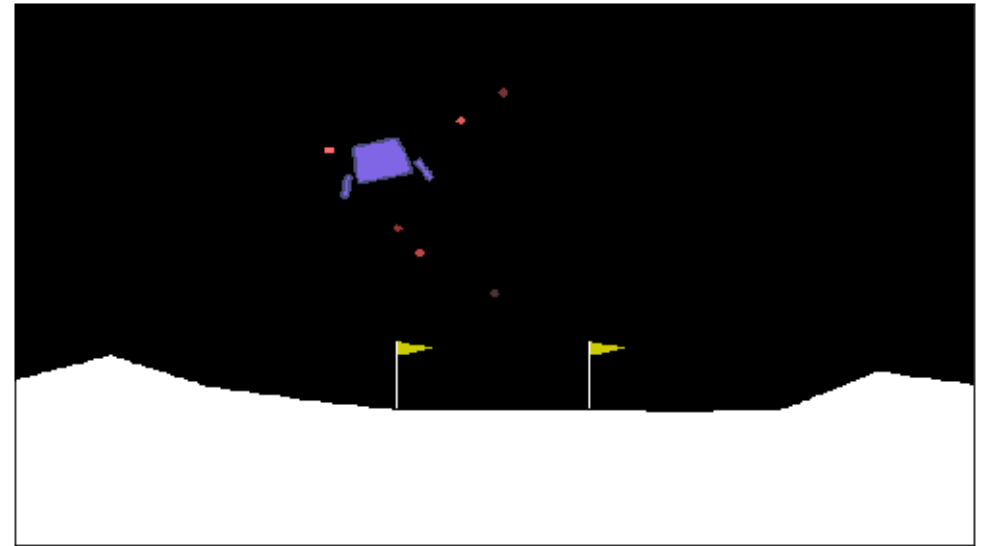- OpenAI's LunarLander benchmark environment
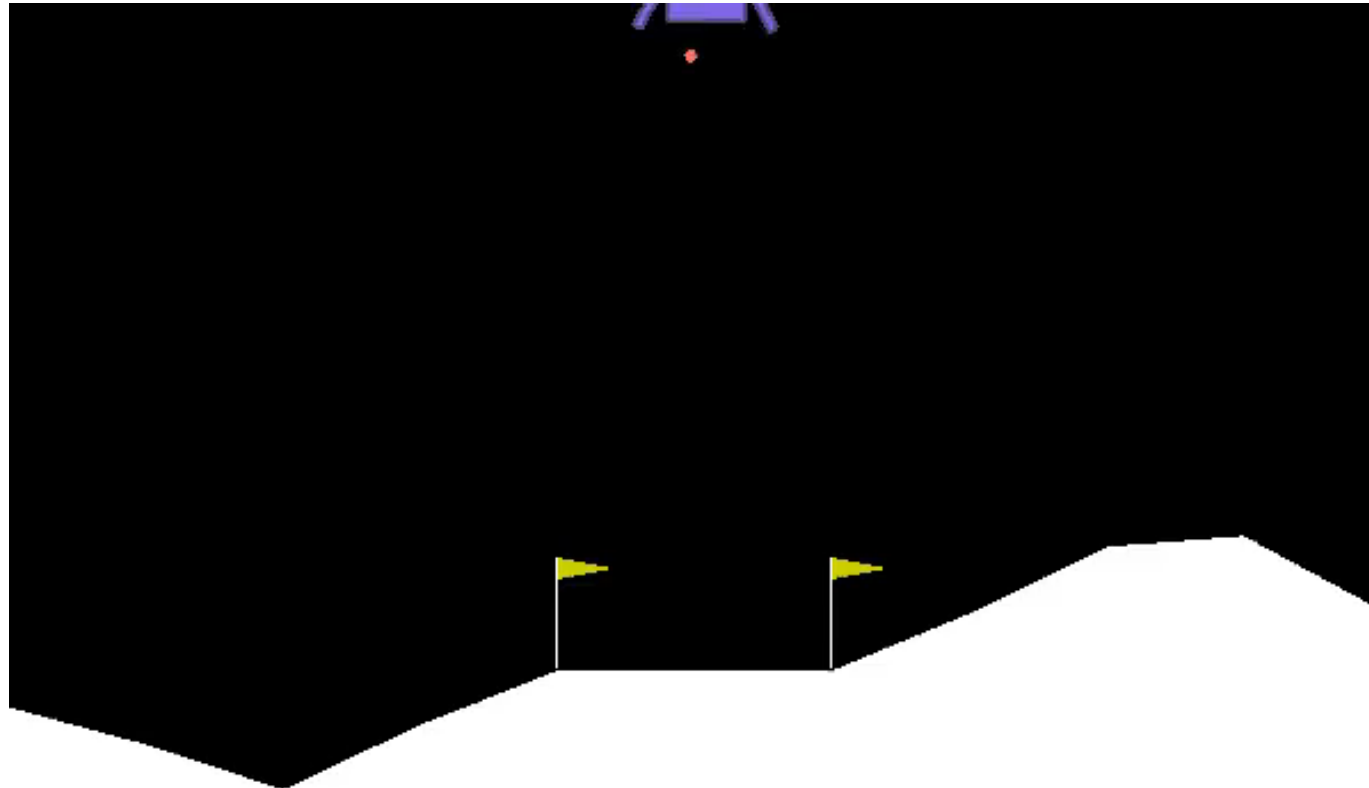
Design, V0

- Design decisions
  - Environment and the agent's "physical" form
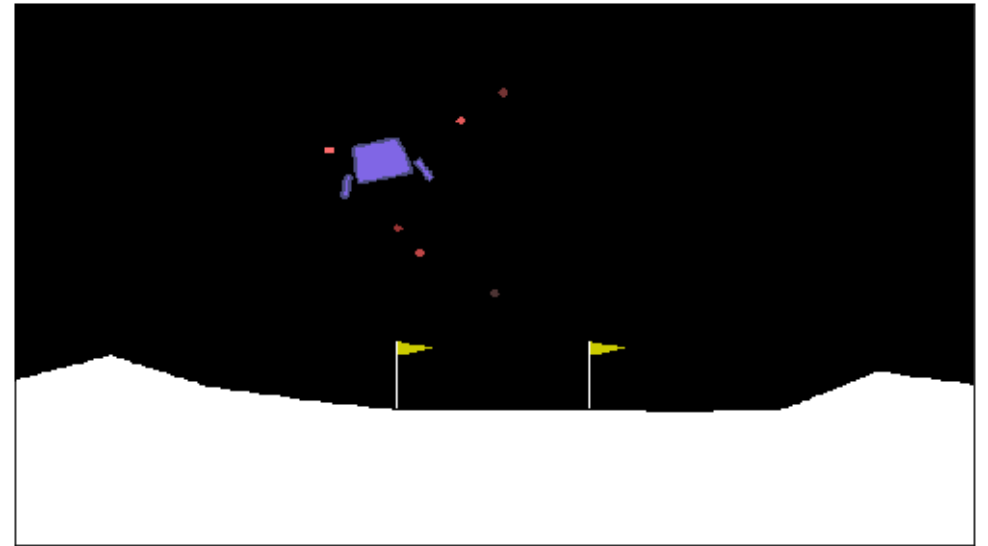
Design, V0

- Design decisions
  - Environment and the agent's "physical" form
  - Internal state of the agent
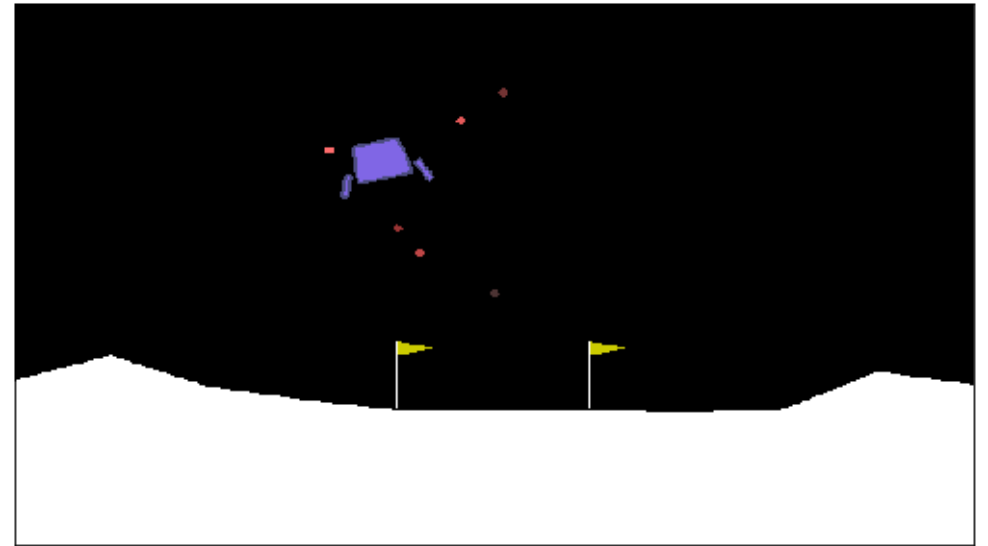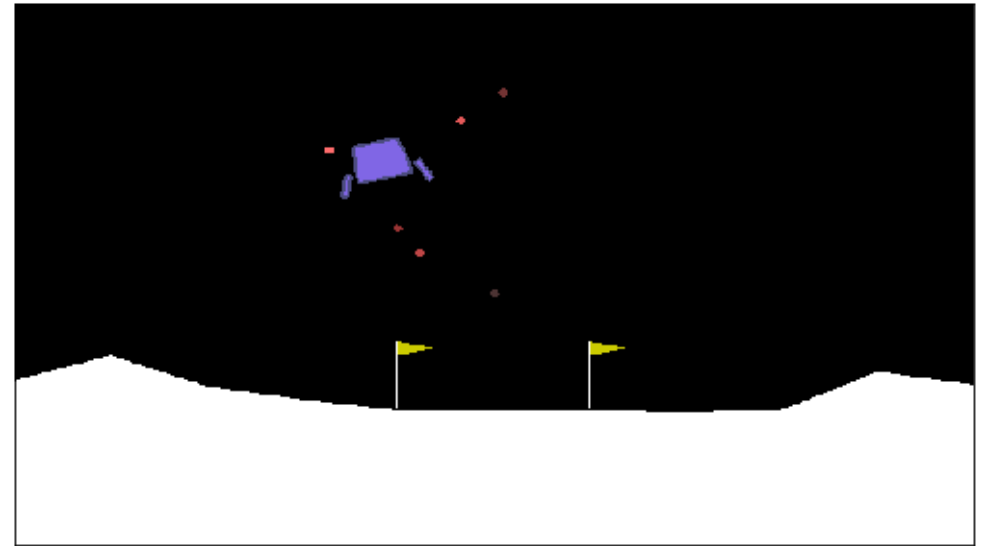
Design, V0

- 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
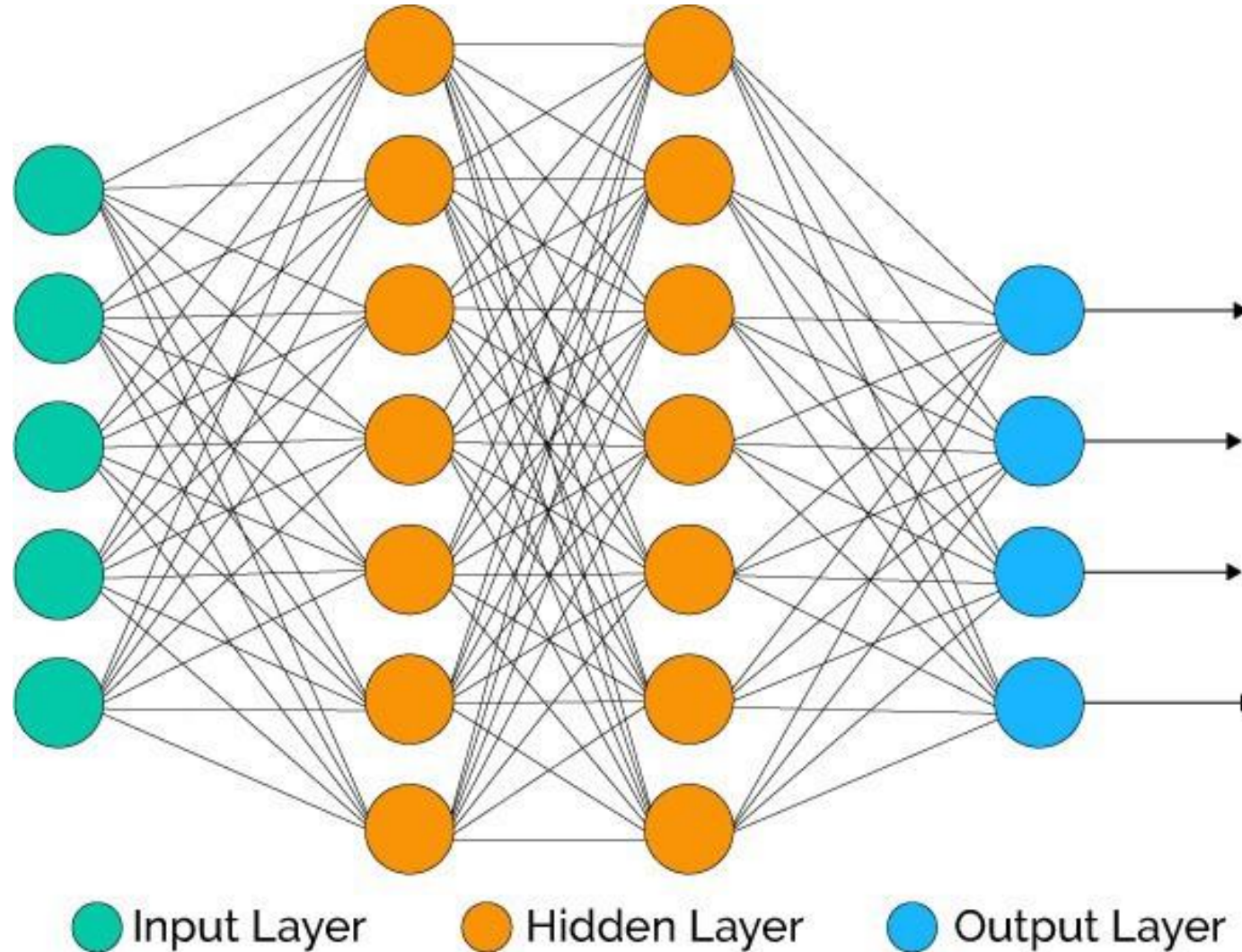
(Artificial) Neural networks

Input Layer   Hidden Layer   Output Layer

(Artificial) Neural networks

For each layer:
- Multiply inputs by weight parameters
- Sum
- Push through non-linear function

Input Layer   Hidden Layer   Output Layer

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

(Artificial) Neural networks

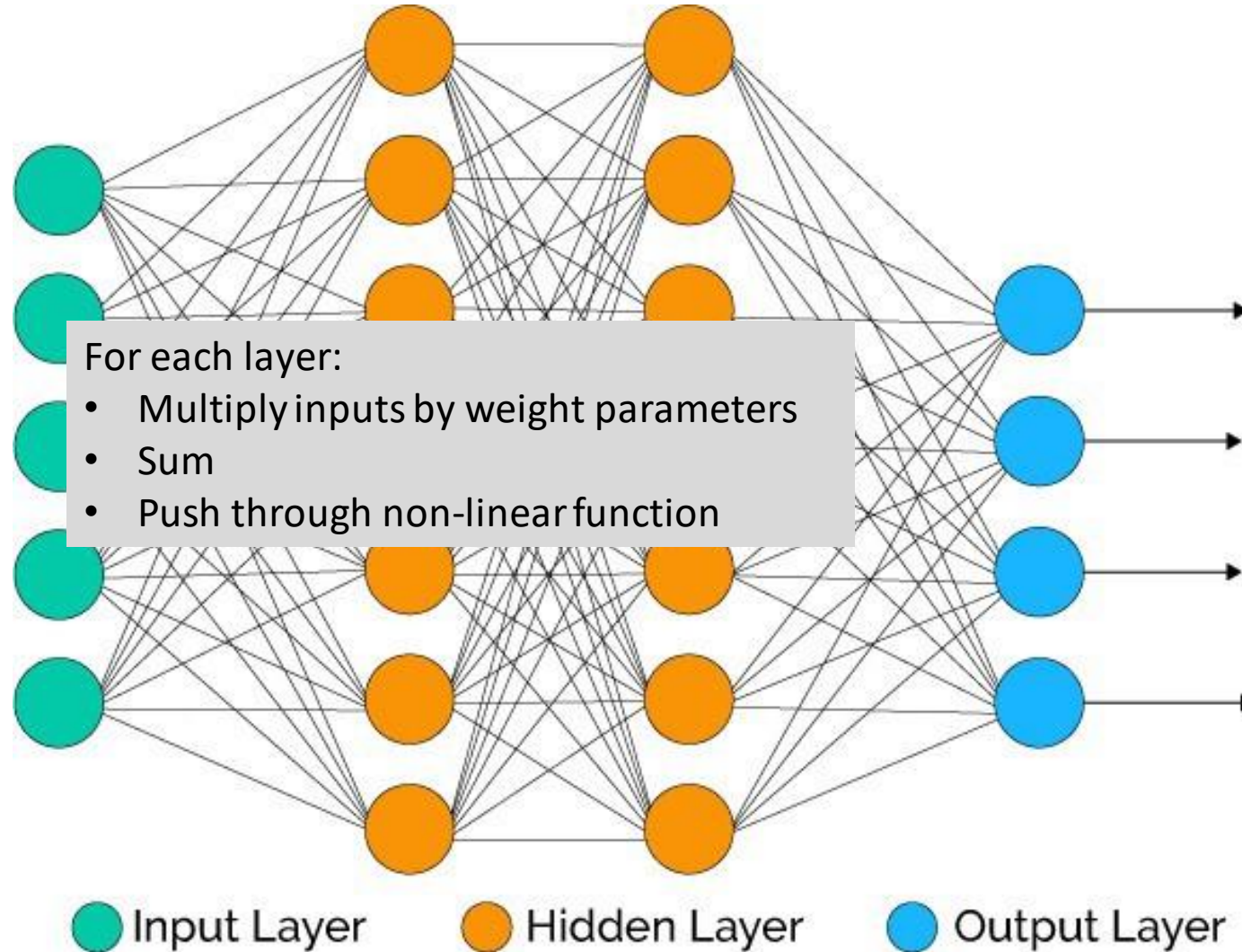Layer 1 weights   Layer 2 weights   Layer 3 weights   Layer 4 weights
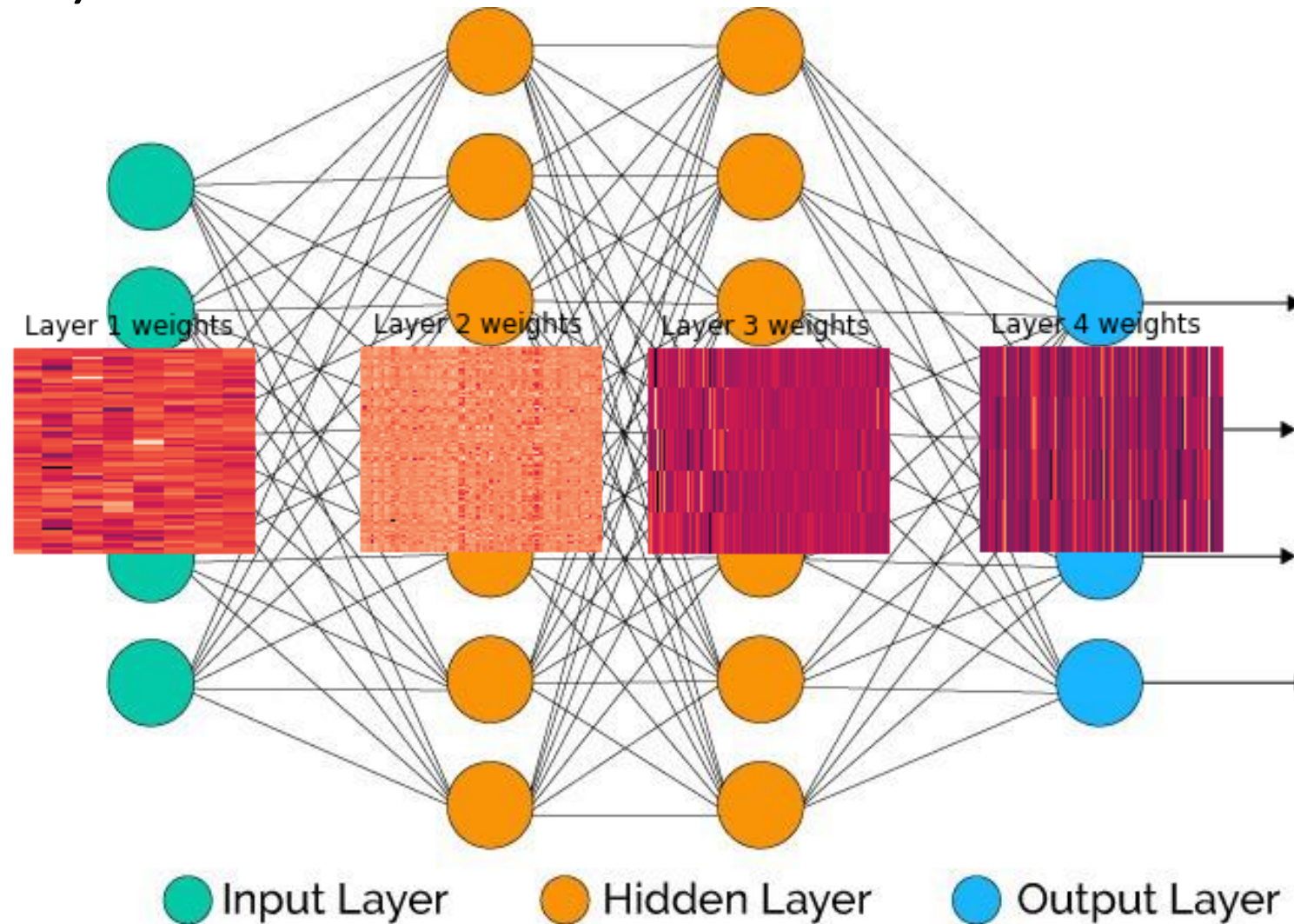
● Input Layer   ● Hidden Layer   ● Output Layer
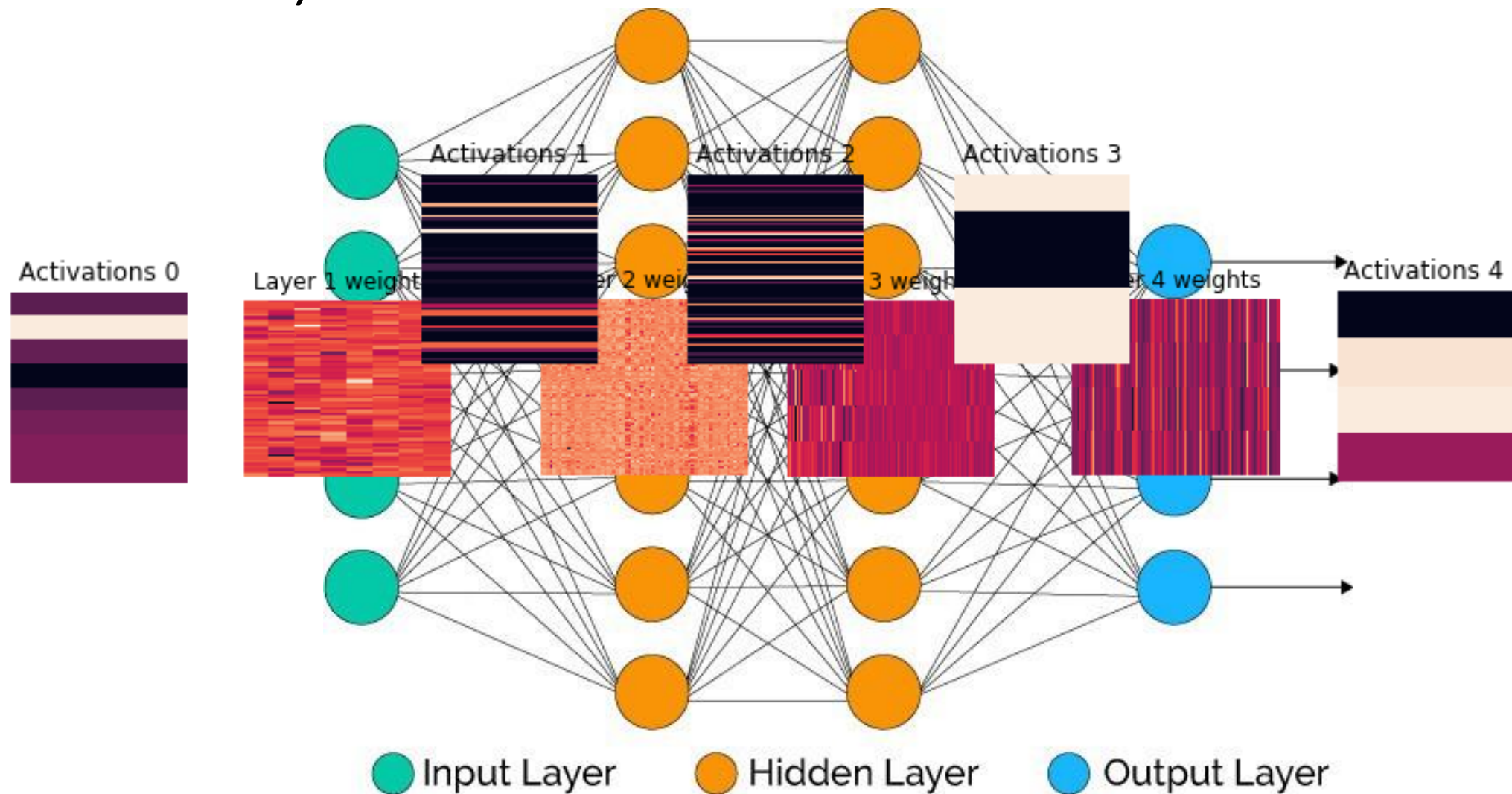
Image from:
https://medium.com/datadriveninvestor/when-not-to-use-neural-networks-89fb50622429

(Artificial) Neural networks

Design, V0

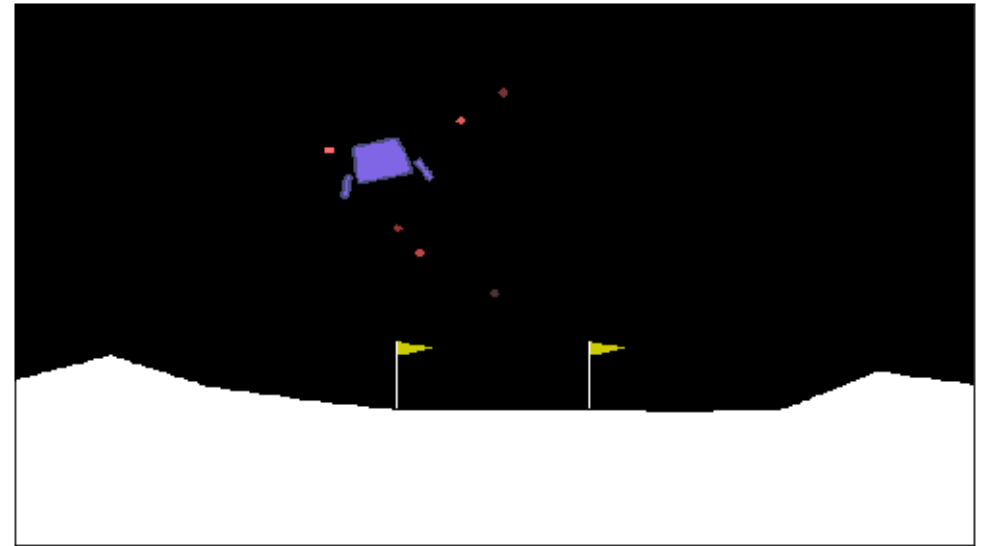
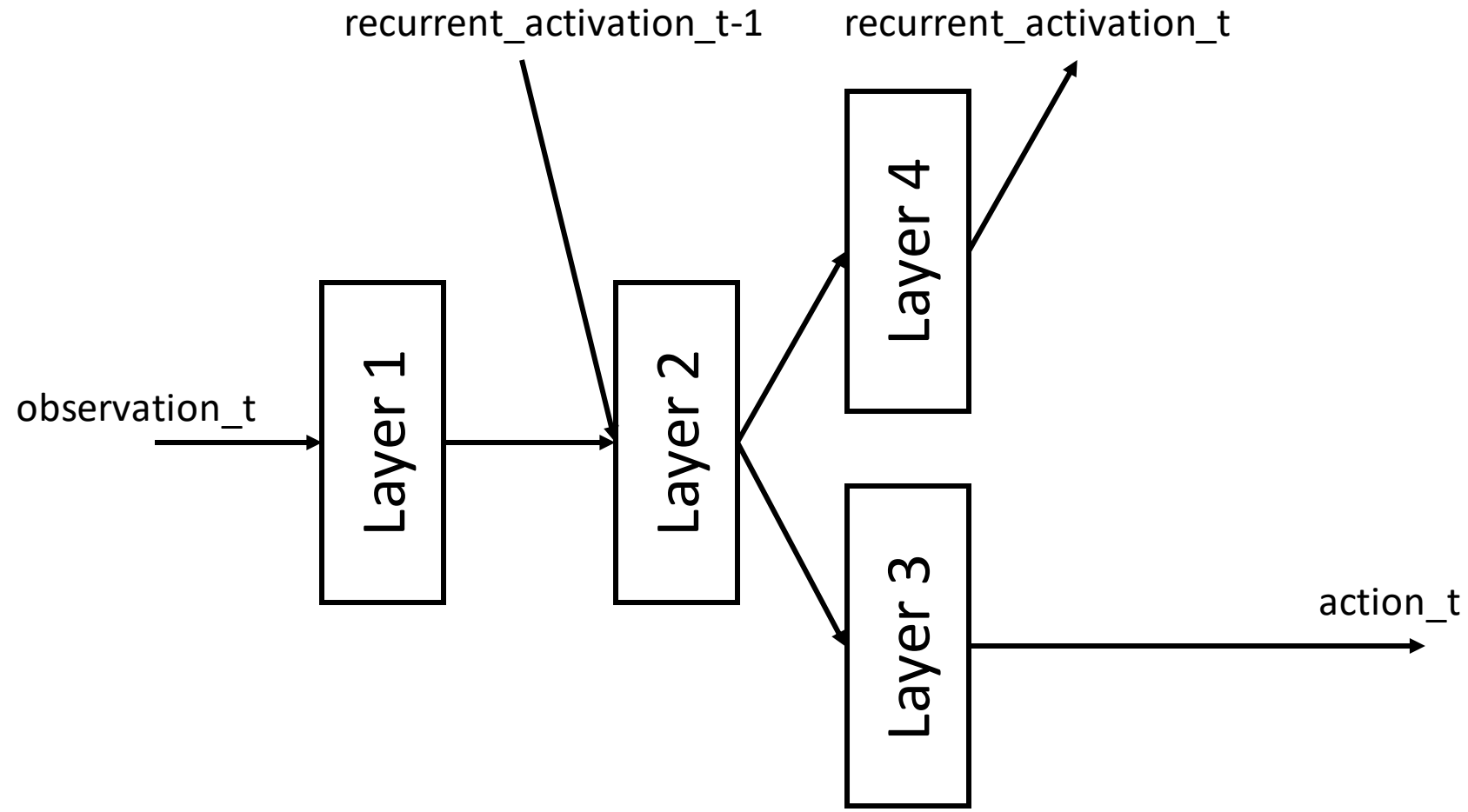- Design decisions
  - Environment and the agent's "physical" form
  - Internal state of the agent
    - Beliefs about itself relative to semantically important regions
      - Left of the flags, right of the flags, high above the ground, close to the ground, falling too fast
  - Brain state of the agent

Design, V0

recurrent_activation_t-1   recurrent_activation_t

observation_t → Layer 1 → Layer 2 → Layer 4 → recurrent_activation_t

Layer 2 → Layer 3 → action_t

Design, V0

Design, V0

Design, V0

Design, V0

brain_state_t

recurrent_activation_t-1

recurrent_activation_t

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

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
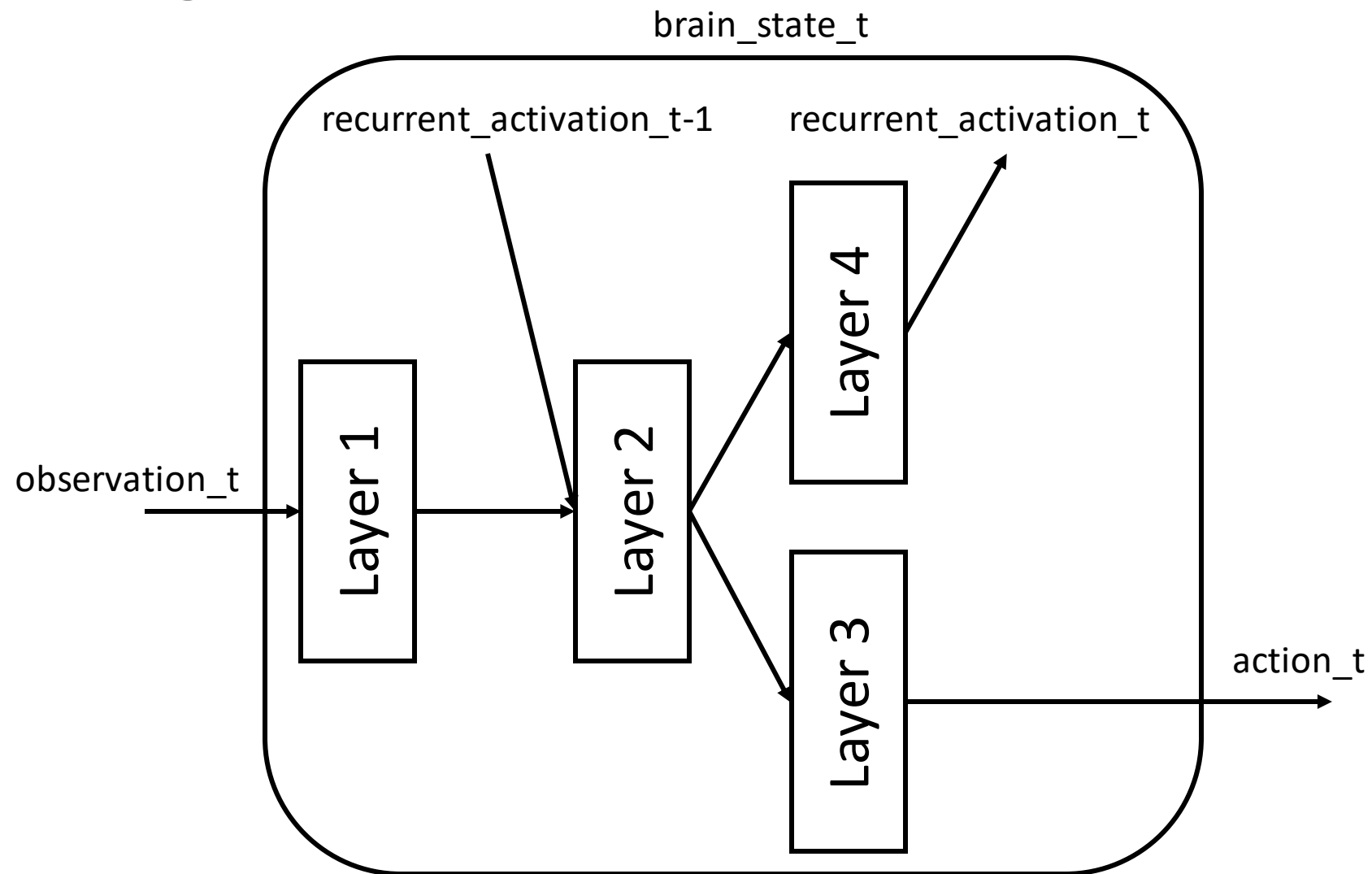
Activations 3

action_t

Design, V0

- 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
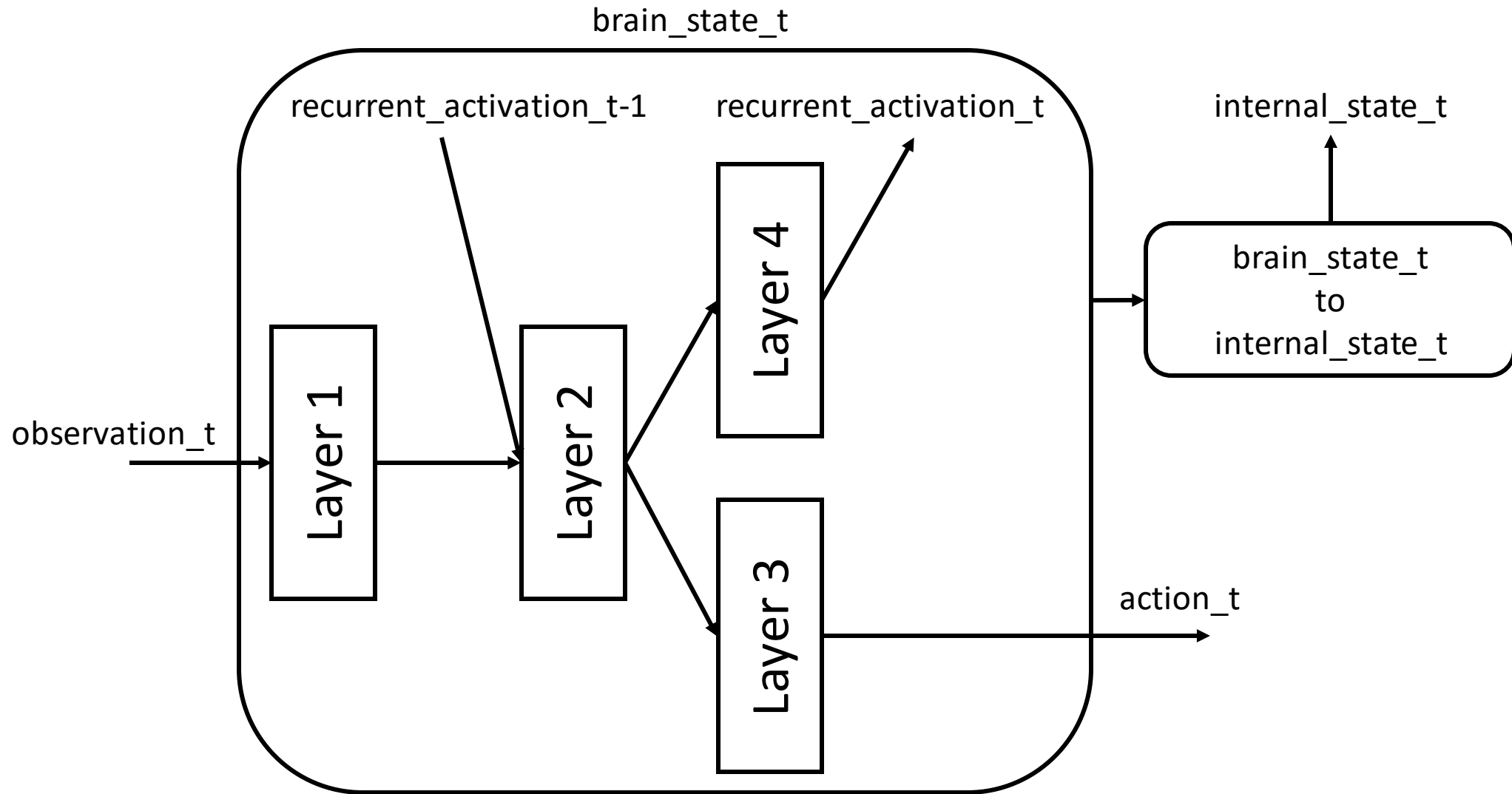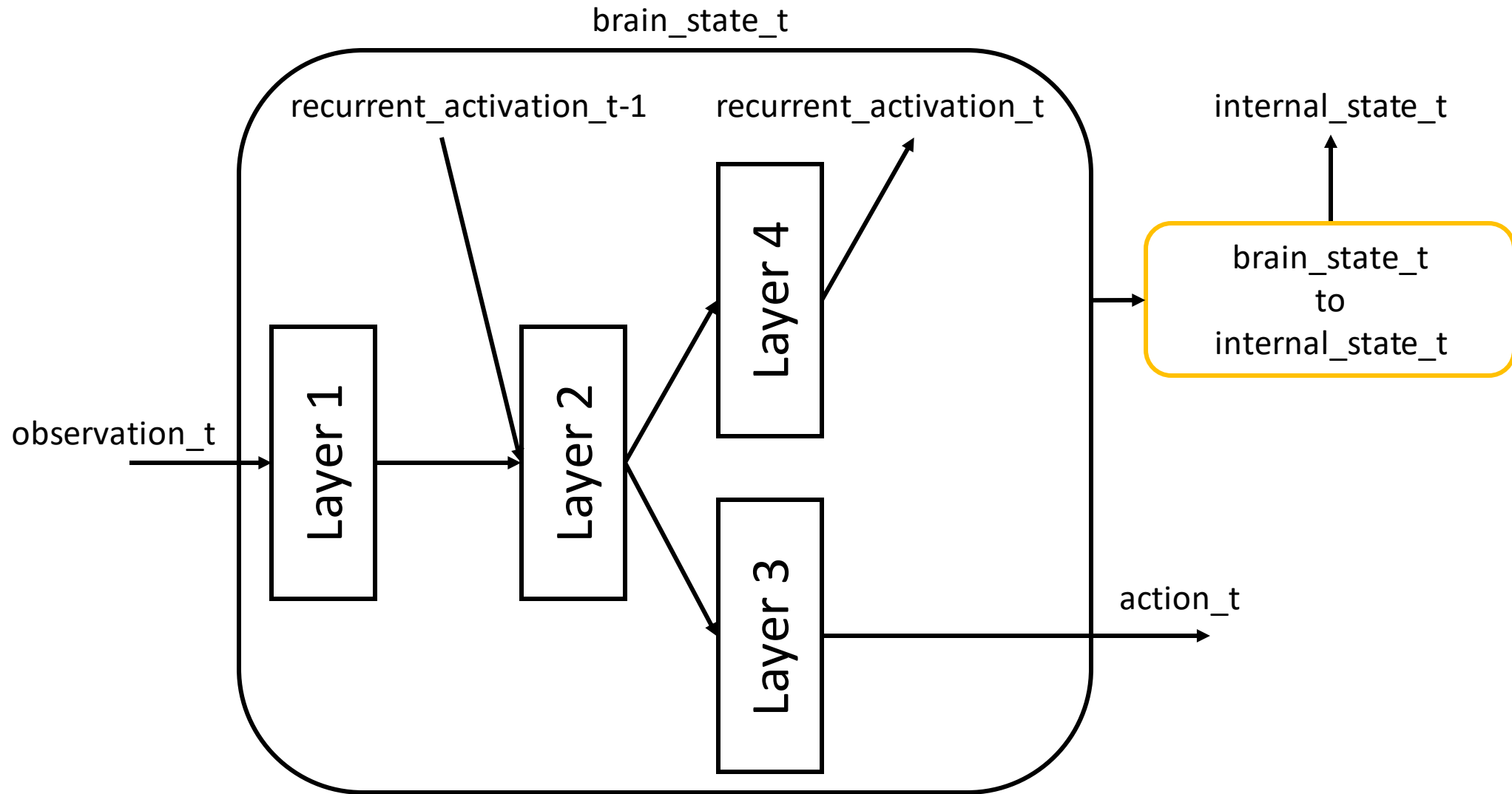 - Brain state of the agent

Design, V0

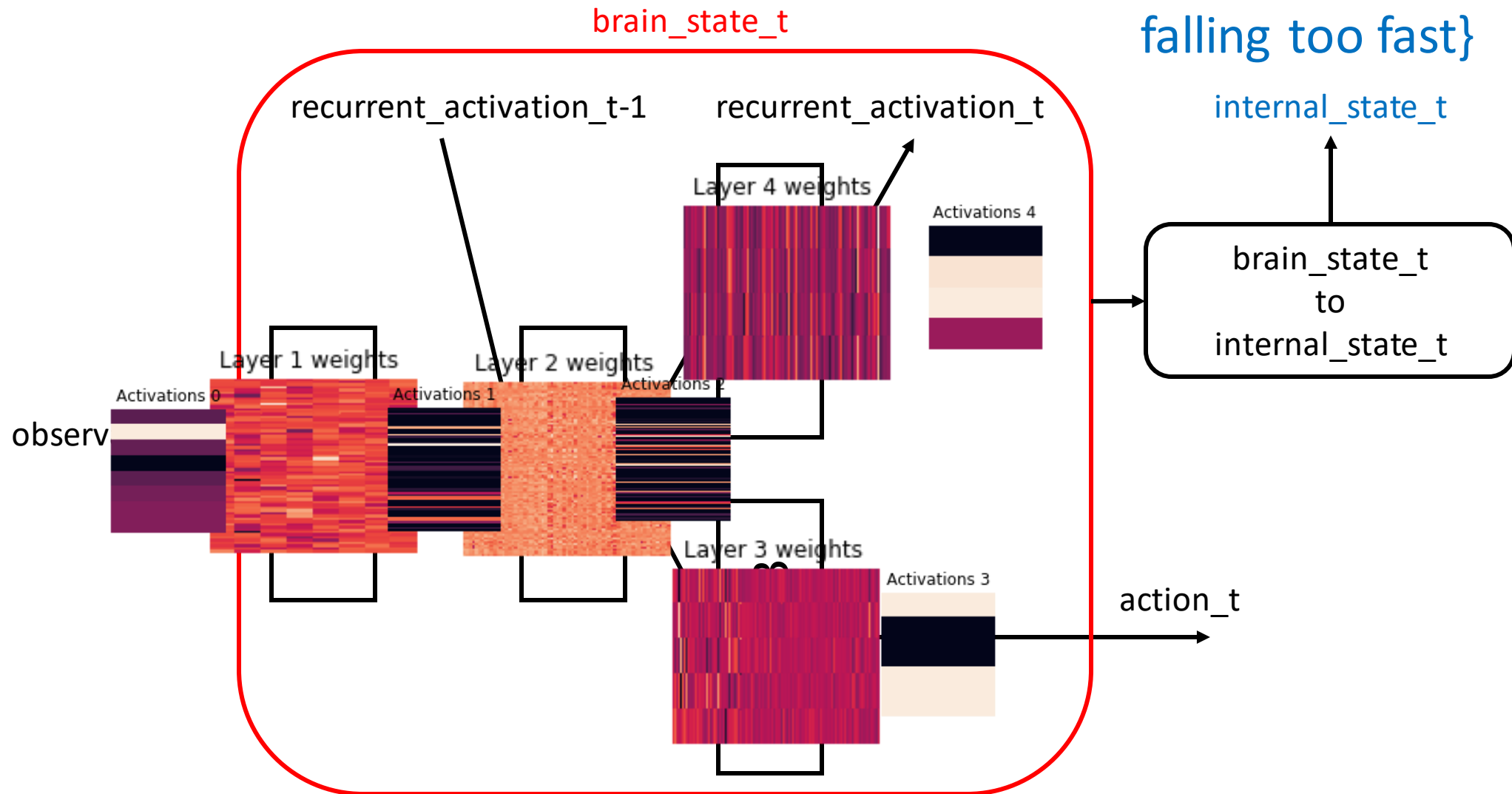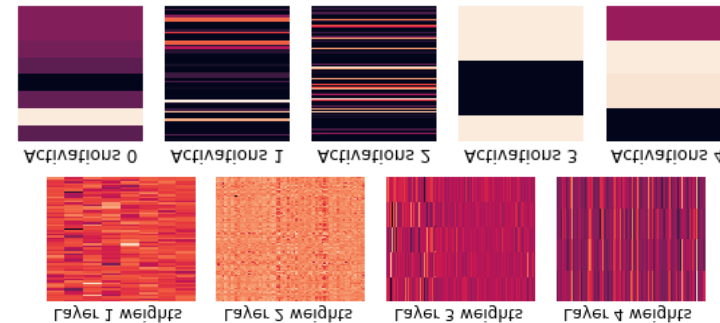- Design decisions
  - Environment and the agent's "physical" form
  - Internal state of the agent
    - Beliefs about itself relative to semantically important regions
      - Left of the flags, right of the flags, high above the ground, close to the ground, falling too fast
  - Brain state of the agent
  - Our ontology
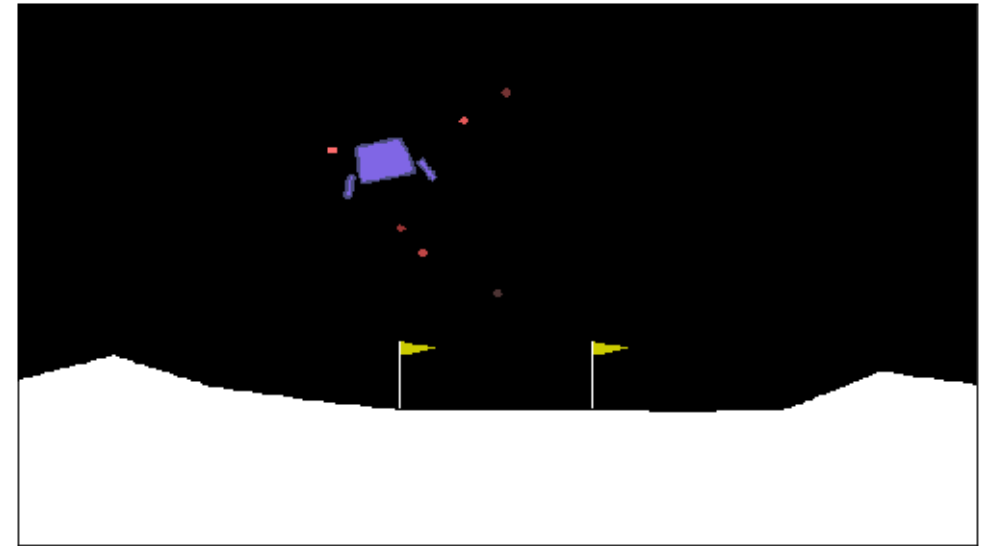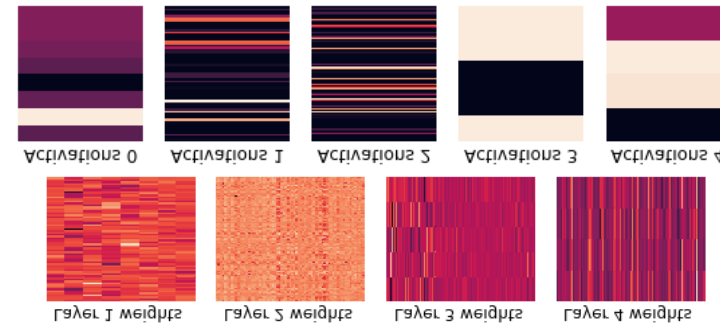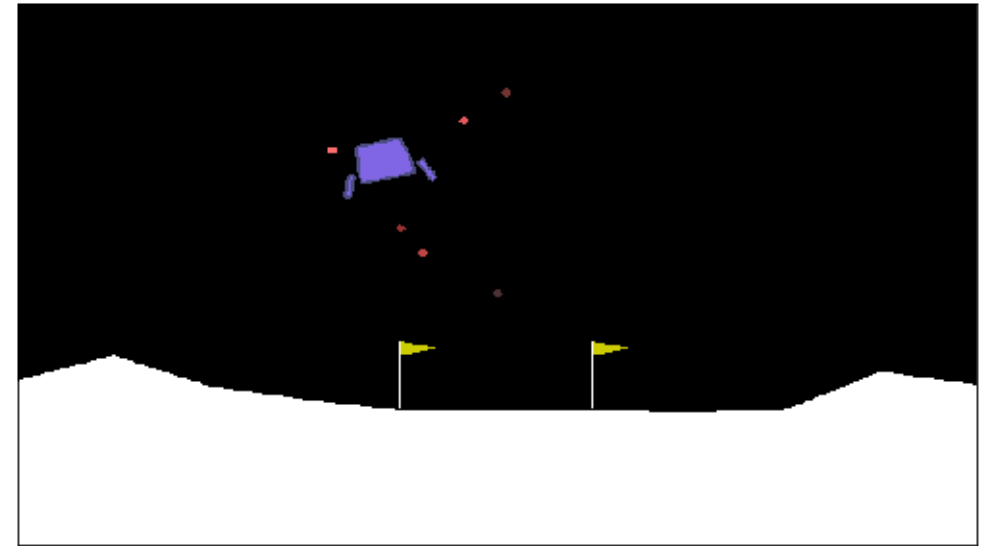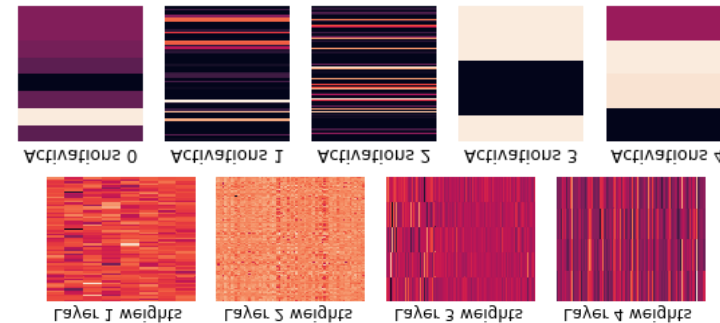
Design, V0

- 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

AGENT

- State $s \in \mathcal{S}$
- Take action $a \in \mathcal{A}$

ENVIRONMENT

- 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
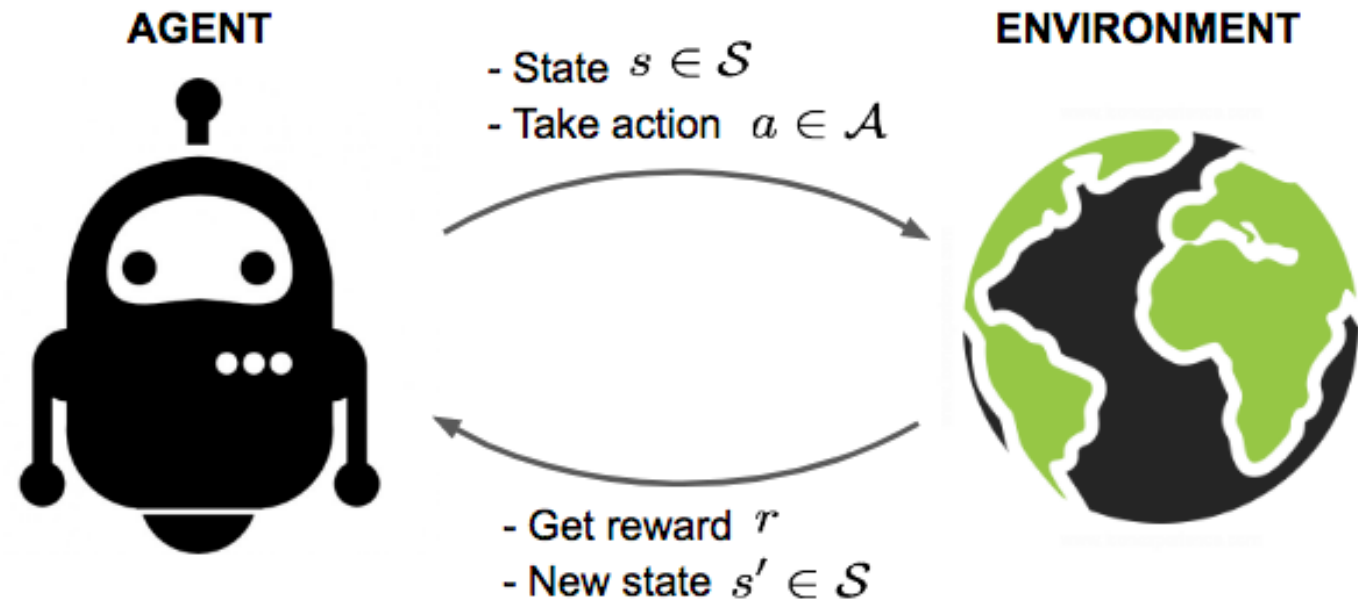
Implementation, V0

- 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 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

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

internal_state_t

brain_state_t

Activations 4

brain_state_t
to
internal_state_t

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}

internal_state_t

brain_state_t

Activations 4

brain_state_t
to
internal_state_t

Design, V0

Did we satisfy our requirements?

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

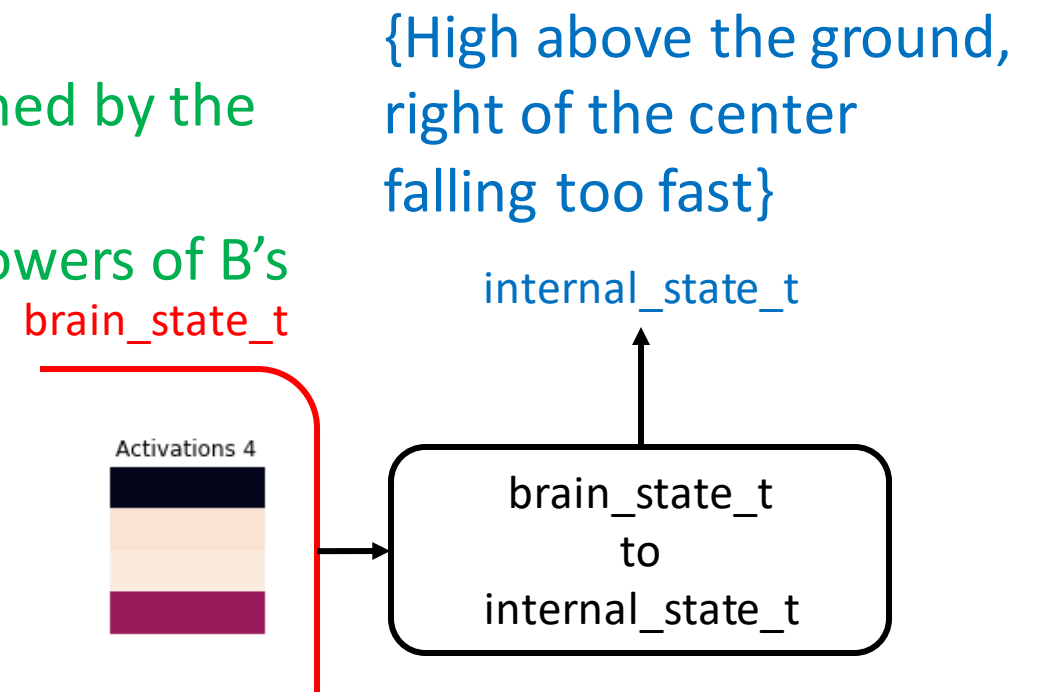Phenomena of type A are causally reducible to phenomena of type B if and only if:

✓ 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}

internal_state_t

brain_state_t

Activations 4
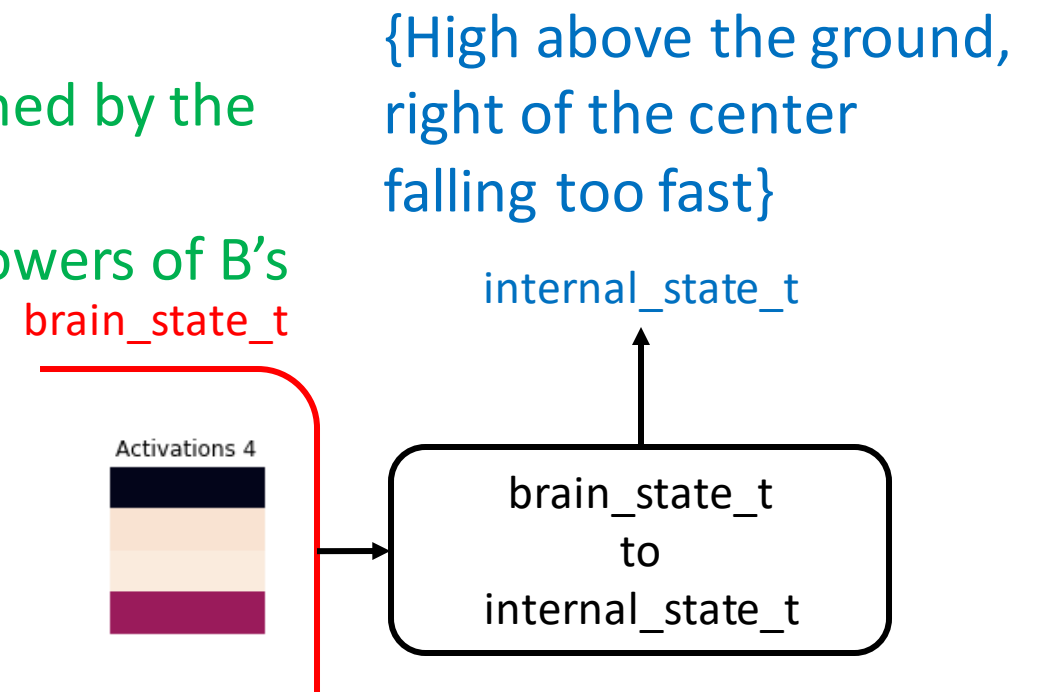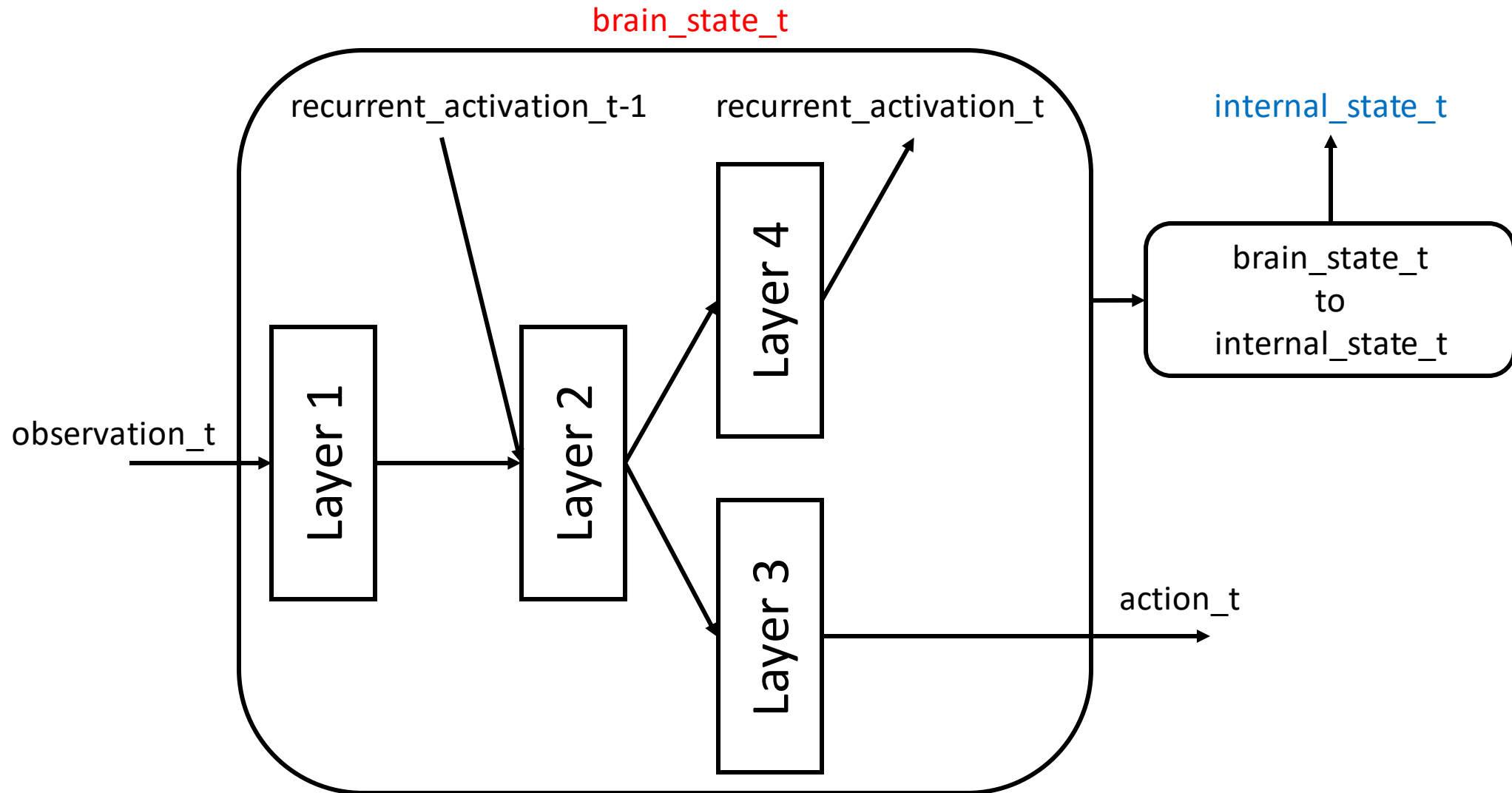
brain_state_t
to
internal_state_t

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
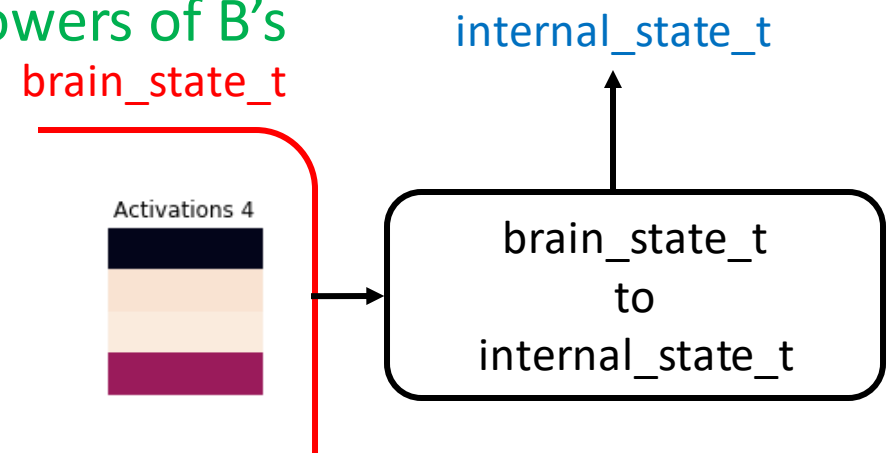
Did we satisfy our requirements?

- 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
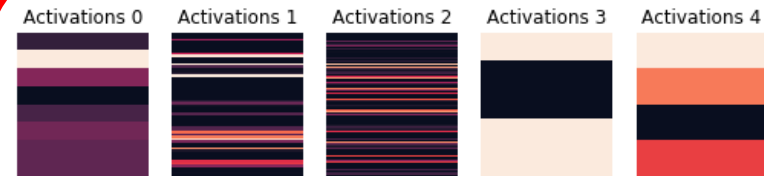   - Internal states are ontologically irreducible to brain states

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

Activations 0 Activations 1 Activations 2 Activations 3 Activations 4

network layer weights

Layer 1 weights Layer 2 weights Layer 3 weights Layer 4 weights

Did we satisfy our requirements?

- V0
  - ✓ Internal states are causally 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...

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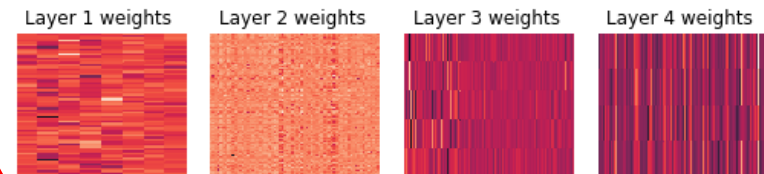
Our ontology
- Layer weights of the neural network
- Connectivity of the neural network
- Activations of the neural network at time t
- The agent's observation at time t
- The agent's action at time t
- The position and velocity of the agent at time t
- A region the agent believes it's in
- Brain state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

network activations at time t

Activations 0   Activations 1   Activations 2   Activations 3   Activations 4

network layer weights

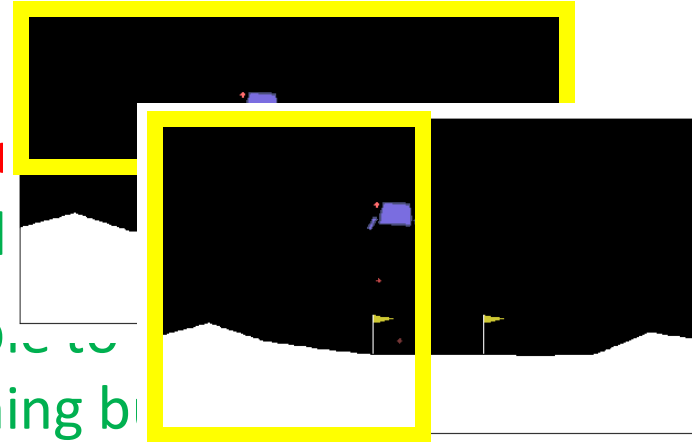Layer 1 weights   Layer 2 weights   Layer 3 weights   Layer 4 weights

Did we satisfy our requirements?

- V0
  - ✓ Internal states are causally reducible to br...
    - Internal states are ontologically irreducibl...

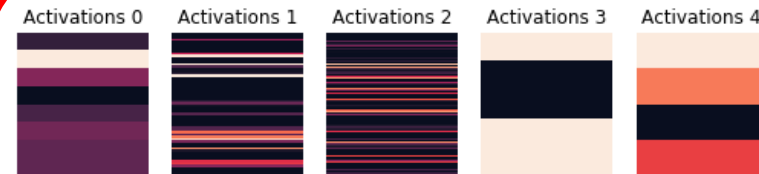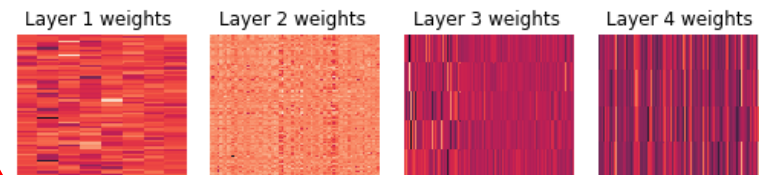Phenomena of type A are ontologically reducib... phenomena of type B if and only if A's are nothing b...

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 o...
- A region the agent believe...
- 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)
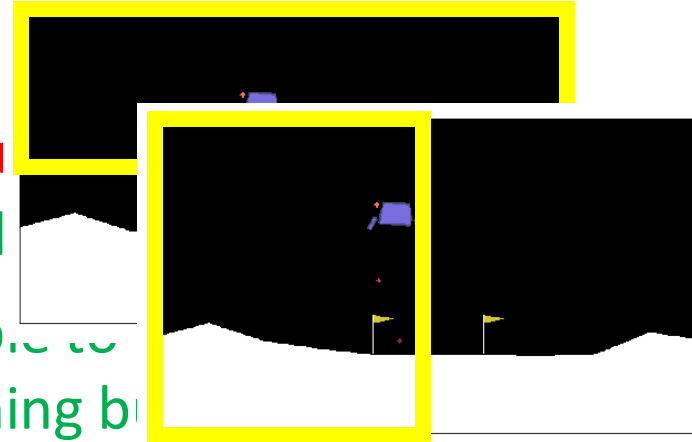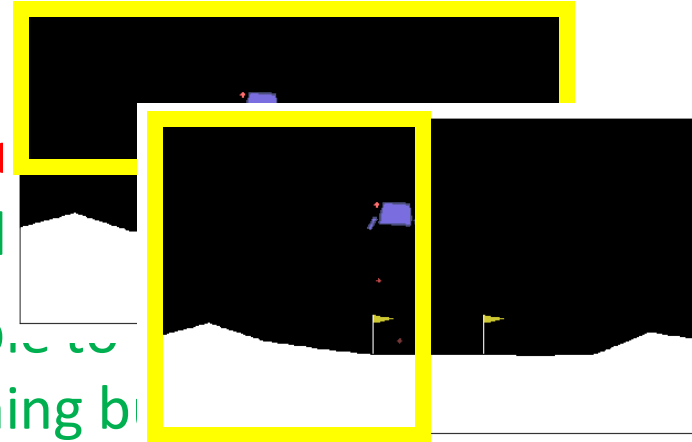
Did we satisfy our requirements?

- V0
  - ✓ Internal states are causally reducible to 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 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 o...
- A region the agent believes...
- 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}

internal_state_t

brain_state_t

Activations 4

brain_state_t
to
internal_state_t
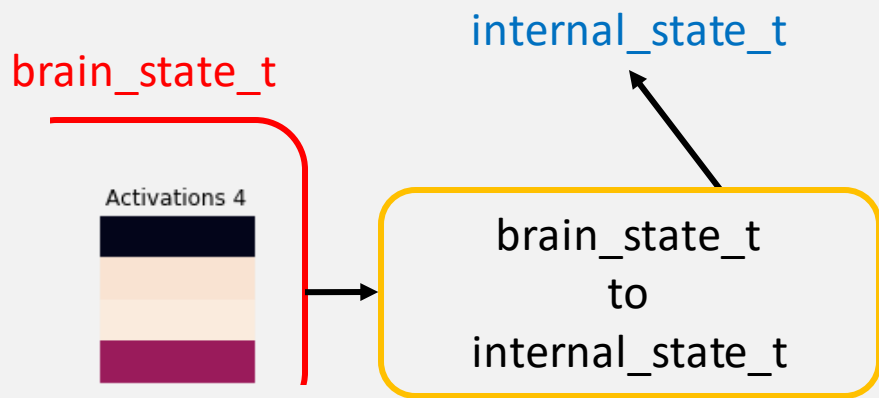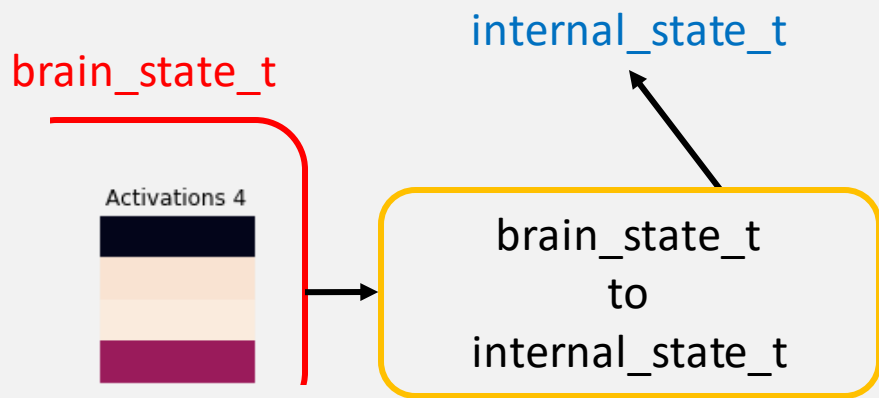
What's the deal with that function?

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

internal_state_t

brain_state_t

Activations 4

brain_state_t
to
internal_state_t

```python
def brain_state_to_internal_state(brain_state):
    def i_am_high_above_the_ground(observation):
        return observation[1] > 0.5  # observation[1] accesses y position

    def i_am_low_to_the_ground(observation):
        return observation[1] <= 0.5  # observation[1] accesses y position

    def i_am_to_the_right_of_the_center(observation):
        return observation[0] > 0.  # observation[0] accesses x position

    def i_am_to_the_left_of_the_center(observation):
        return observation[0] <= 0.  # observation[0] accesses x position

    def i_am_falling_too_fast(observation):
        return observation[3] < -0.2  # observation[0] accesses y velocity

    regions = [
        i_am_high_above_the_ground,
        i_am_low_to_the_ground,
        i_am_to_the_right_of_the_center,
        i_am_to_the_left_of_the_center,
        i_am_falling_too_fast
    ]

    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

Activations 4

brain_state_t
to
internal_state_t

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

```python
def brain_state_to_internal_state(brain_state):
    def i_am_high_above_the_ground(observation):
        return observation[1] > 0.5  # observation[1] accesses y position

    def i_am_low_to_the_ground(observation):
        return observation[1] <= 0.5  # observation[1] accesses y position

    def i_am_to_the_right_of_the_center(observation):
        return observation[0] > 0.  # observation[0] accesses x position

    def i_am_to_the_left_of_the_center(observation):
        return observation[0] <= 0.  # observation[0] accesses x position

    def i_am_falling_too_fast(observation):
        return observation[3] < -0.2  # observation[0] accesses y velocity

    regions = [
        i_am_high_above_the_ground,
        i_am_low_to_the_ground,
        i_am_to_the_right_of_the_center,
        i_am_to_the_left_of_the_center,
        i_am_falling_too_fast
    ]

    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?

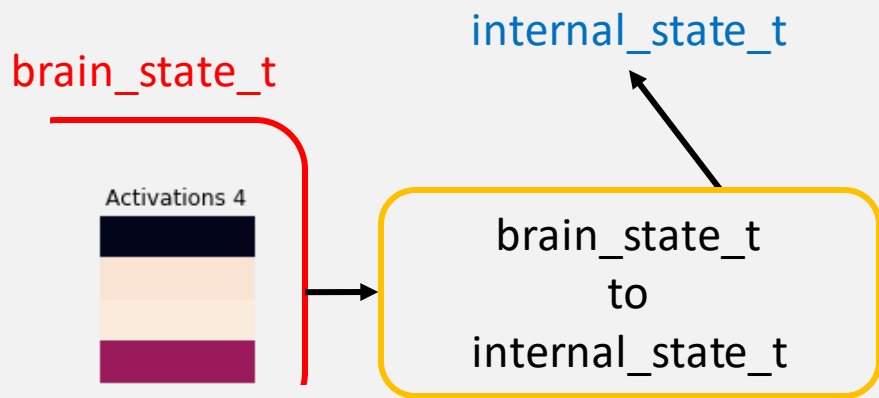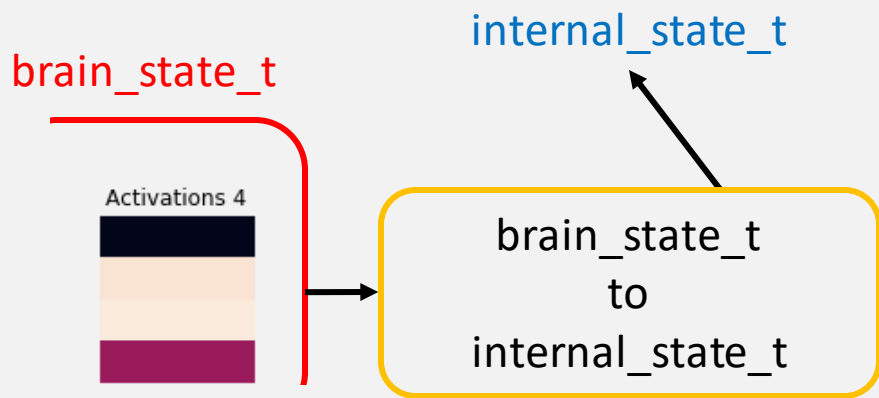{High above the ground, right of the center falling too fast}

internal_state_t

brain_state_t

Activations 4

brain_state_t
to
internal_state_t

```python
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
60
```

- 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}

internal_state_t

brain_state_t

Activations 4

brain_state_t
to
internal_state_t

```python
def brain_state_to_internal_state(brain_state):
    def i_am_high_above_the_ground(observation):
        return observation[1] > 0.5  # observation[1] accesses y position

    def i_am_low_to_the_ground(observation):
        return observation[1] <= 0.5  # observation[1] accesses y position

    def i_am_to_the_right_of_the_center(observation):
        return observation[0] > 0.  # observation[0] accesses x position

    def i_am_to_the_left_of_the_center(observation):
        return observation[0] <= 0.  # observation[0] accesses x position

    def i_am_falling_too_fast(observation):
        return observation[3] < -0.2  # observation[0] accesses y velocity

    regions = [
        i_am_high_above_the_ground,
        i_am_low_to_the_ground,
        i_am_to_the_right_of_the_center,
        i_am_to_the_left_of_the_center,
        i_am_falling_too_fast
    ]

    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?

{High above the ground, right of the center fa

"[…] 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

```python
def brain_state_to_internal_state(brain_state):
    def i_am_high_above_the_ground(observation):
        return observation[1] > 0.5  # observation[1] accesses y position

    def i_am_low_to_the_ground(observation):
        [1] accesses y position

    ):
        0] accesses x position

    ):
        return observation[0] <= 0.  # observation[0] accesses x position

    def i_am_falling_too_fast(observation):
        return observation[3] < -0.2  # observation[0] accesses y velocity

    regions = [
        i_am_high_above_the_ground,
        i_am_low_to_the_ground,
        i_am_to_the_right_of_the_center,
        i_am_to_the_left_of_the_center,
        i_am_falling_too_fast,
    ]

    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?

{High above the ground, right of the center f...
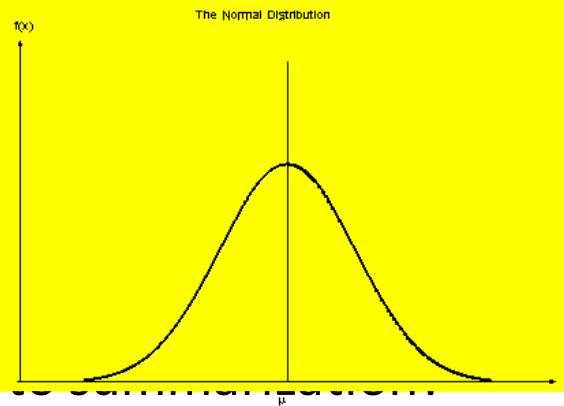
brain_state_t

Activations 4

"[...] 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...

The Normal Distribution

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

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

```python
def brain_state_to_internal_state(brain_state):
    def i_am_high_above_the_ground(observation):
        return observation[1] > 0.5  # observation[1] accesses y position

    def i_am_low_to_the_ground(observation):
        [1] accesses y position

    ):
        0] accesses x position

    ):
        [0] accesses x position

        n[0] accesses y velocity

    ns'][3]

    tions, regions):

        if activation > 0.5:
            internal_state.add(region.__name__)

    return internal_state
```

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

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?