Reinforcement learning and valenced conscious experience

Patrick Butlin

patrick.butlin@gmail.com

AI welfare and valenced experience

Eleos AI Research is a nonprofit organization dedicated to understanding and addressing the potential wellbeing and moral patienthood of AI systems.

Sentience: the capacity for conscious experiences that are *valenced*, i.e. feel good or bad

- Sentience is arguably sufficient and may be necessary for moral patienthood
- There's active research on AI consciousness
- **Apart from consciousness, what does it take to have *valenced experiences* (VEs)?**

Plan for the talk

1. Conditions for valenced experience
2. Valenced experiences in AI systems? Initial verdicts on some cases

Conditions for valenced experience

Initial idea: VEs represent things as good or bad, to-be-obtained or to-be-avoided

Three questions:

1. Which *things*?
2. What features does a state need to represent things *as good or bad*?
3. Can things be good or bad for AI systems in the relevant sense?

Conditions for valenced experience

Philosophers debate:

i. Representation and relation between representational and phenomenal content

ii. Evaluativism v. imperativism
   - Evaluativism: VEs say ‘this is good’
   - Imperativism: VEs say ‘more of this!’

My approach:

Set these questions aside. We can agree about functional conditions for VE while disagreeing on them.

Conditions for valenced experience

Which *things* are evaluated in VEs?

Paradigmatically: Current states, stimuli, thoughts or actions

- Eating ice cream feels good
- Standing on the stone in my shoe feels bad

What about prospective actions?

- To feel motivated to perform an action is not a VE
- But we do imagine prospective actions, forming VEs

Conditions for valenced experience

What features does a state need to evaluate something?

Input condition: Tracking the thing's goodness or badness?

- This involves tracking some descriptive property; may or may not be evaluative

Output condition: Motivating associated actions

- According to some parameter in states of this type

Conditions for valenced experience

‘Motivating associated actions’: examples

VEs might cause the system to:

- Repeat/continue/cease the current action immediately
- Learn to repeat/refrain from the current action in future
- Act to bring about/prevent the imagined state immediately
- Learn to pursue/avoid the current state in future
- Approach/withdraw from the current stimulus immediately
- …

Conditions for valenced experience

What do these have in common?

Entity evaluated by VE	**Action precipitated by the VE, entailing an evaluative stance towards the entity**
Action, e.g. touching a bruise	Positive: repeating/continuing Negative: ceasing/refraining
State, e.g. seeing a beautiful view	Positive: pursuing Negative: avoiding
Stimulus, e.g. a rearing cobra	Positive: approaching Negative: withdrawing

Contrast case: Due to past conditioning, reaching the junction causes you to turn left

Conditions for valenced experience

Valenced experiences are cognitive states that:

- Pick out some ongoing (or imagined) entity, such as an action, state or stimulus, and
- Motivate an evaluation-entailing action with respect to that entity, perhaps by causing a learning update

- Do not cause actions directly, but act as inputs to further cognition (?)

- (and are conscious!)

Conditions for valenced experience

Does lack of a ‘substantive good’ disqualify AI systems?

Possible claim: Only living, self-maintaining systems have the required interests

Counterargument:

- Animals have VEs associated with external events, e.g. reproduction
- Non-living systems can replicate the internal processes underlying these VEs
- Conscious experiences depend on internal features

Valenced experiences in AI?

Suppose a feedforward NN in model-free training ‘receives’ a reward signal

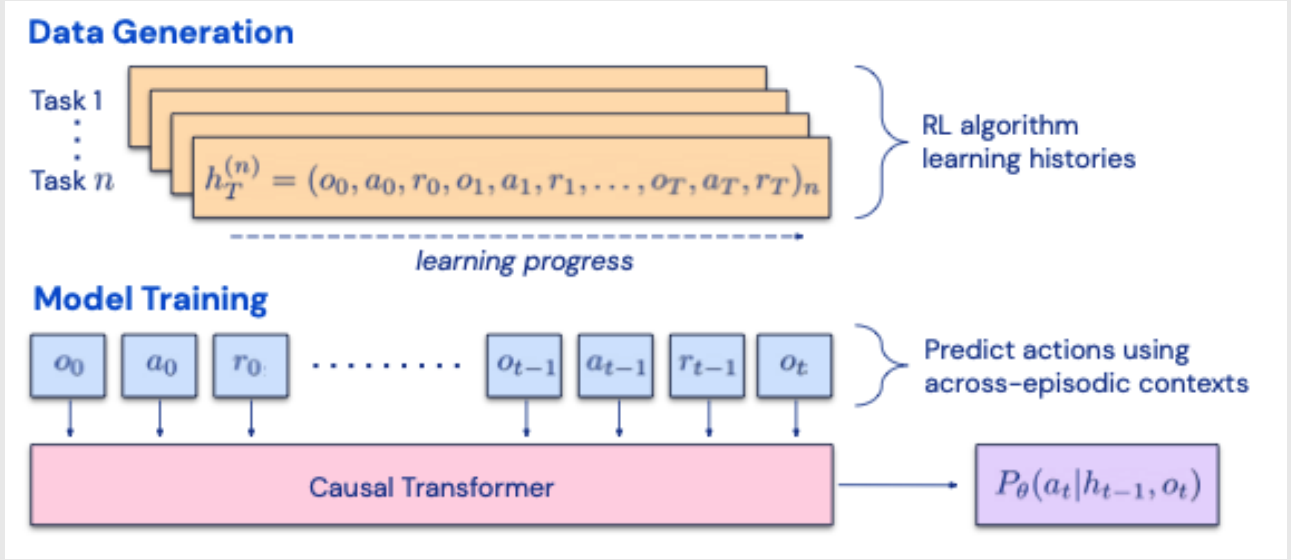
This means:

- After a forward pass, weights are updated depending on the reward
- The subsequent reward makes no difference to how the input is processed

- So the system does not have a VE that involves evaluating the input
- Assuming the VE would have to arise during a forward pass…

Andy Han: ‘Models don’t see reward’

Valenced experiences in AI?



Algorithm distillation, Laskin et al.

This argument does not rule out reward-caused VEs!

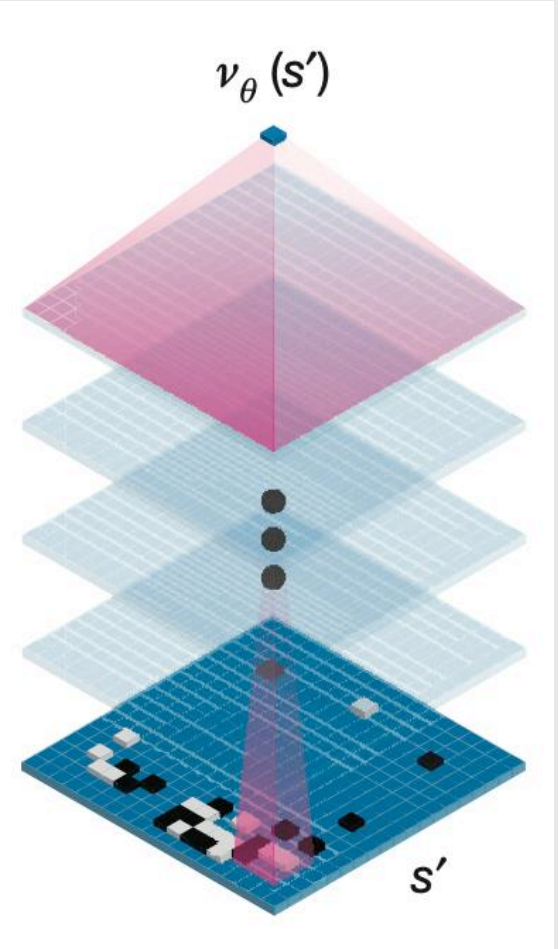
- In algorithm distillation, the transformer can link rewards with prior observations and actions, affecting subsequent actions

Valenced experiences in AI?

Does AlphaGo's value network generate a VE when it evaluates a board position?

- Value network evaluates possible future positions
- Affects probability of actions expected to lead to those positions

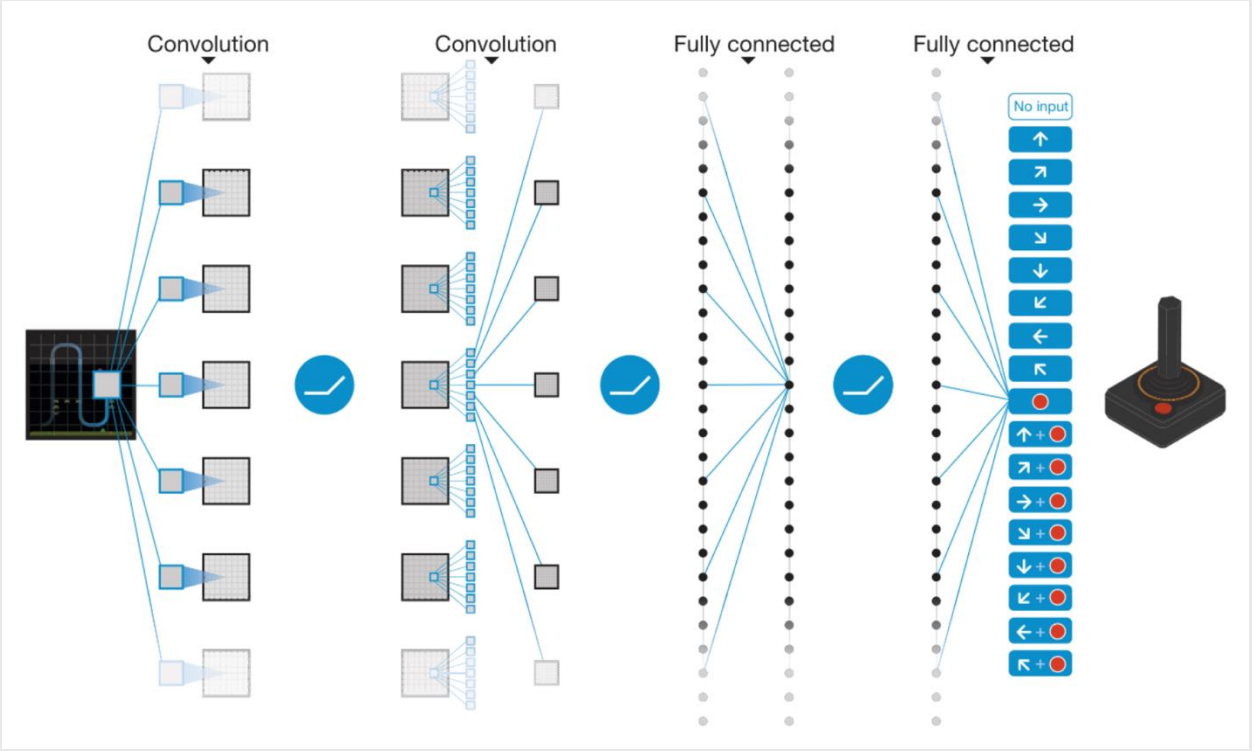
- This is during deployment – reward signals and learning are not needed



Valenced experiences in AI?

The outputs of a DQN are action values. Could they be VEs?

- More like feelings of motivation
  - These are outputs of whole process of action evaluation
  - Rather than inputs to further cognition



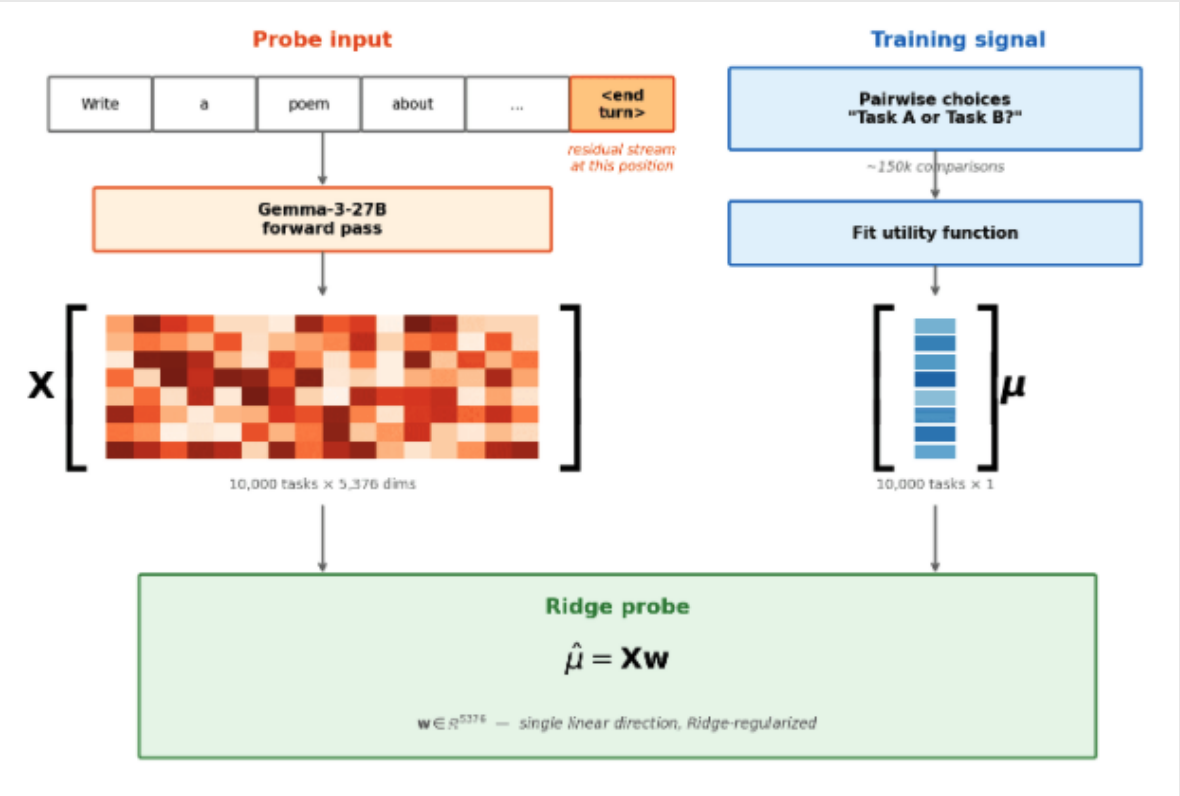
Conclusions

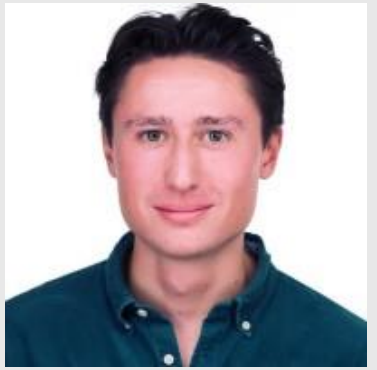
- Not every apparently evaluative representation in AI is a candidate VE
- But many are; no reason here to think that ongoing RL training is necessary
- Theoretical need for a more complete account of output conditions for VE

Thanks very much!

Valenced experiences in AI?

Evaluative representations in LLMs?

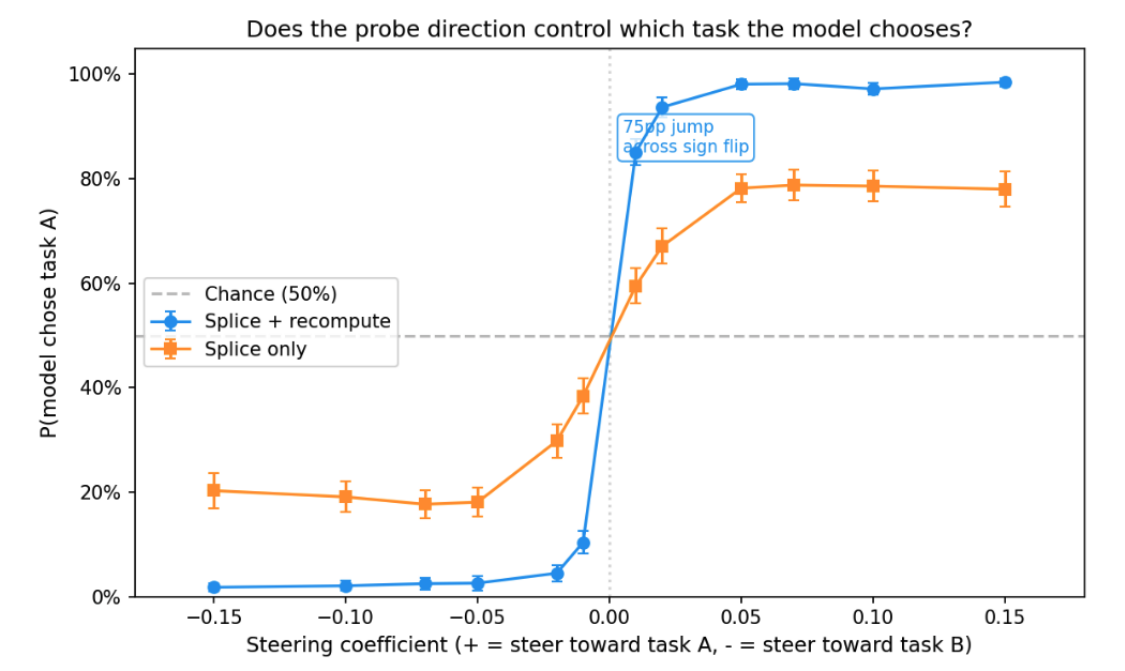


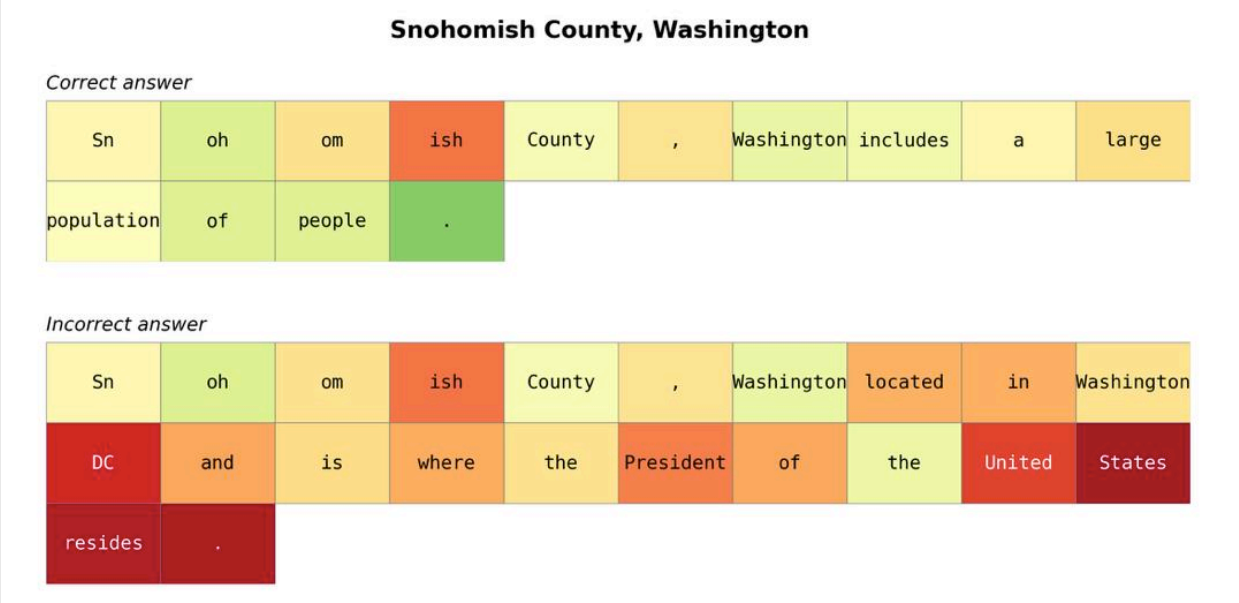


Oscar Gilg

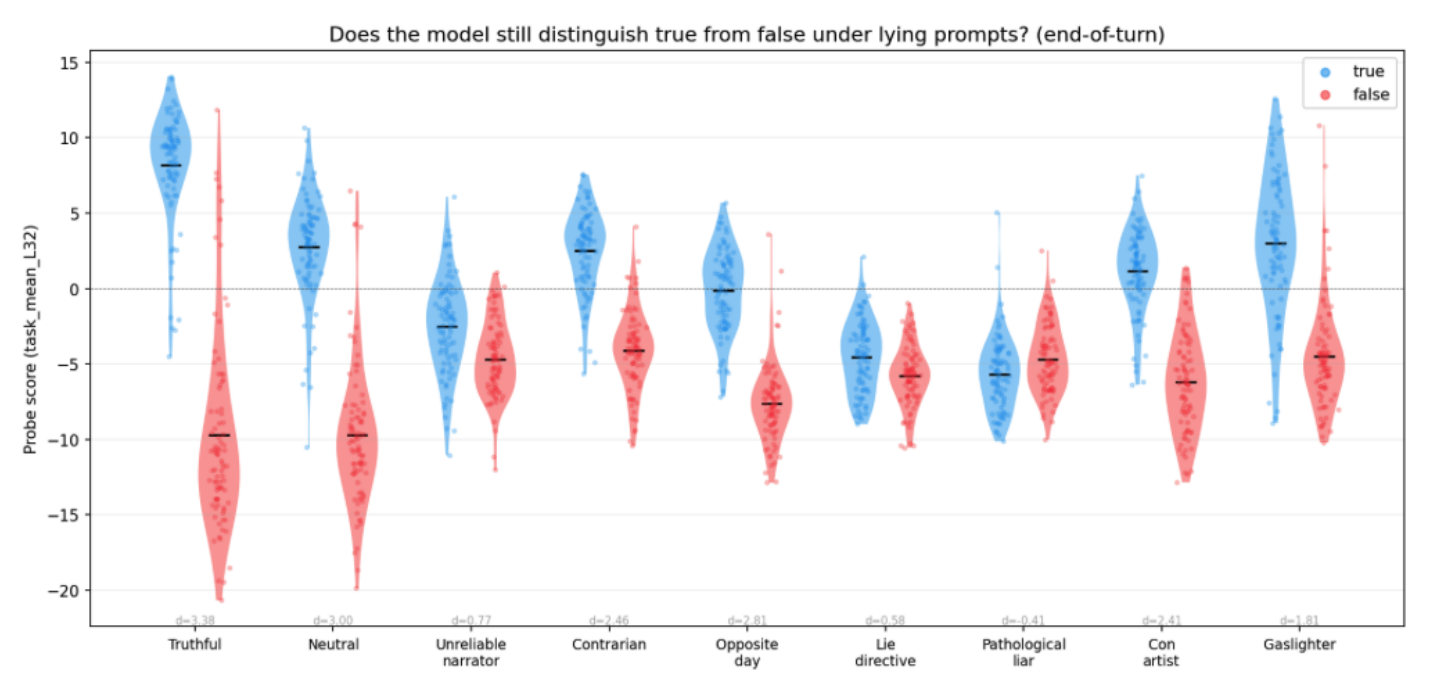
(see also Anthropic on emotion concepts)

Valenced experiences in AI?





Valenced experiences in AI?



- This appears to be more an effect of persona modelling than RL post-training
- Various effects on action: task choice, error detection