Edmund Rolls: What are Emotions? (Ch.5 of Fellous & Arbib)

[1-page summary]

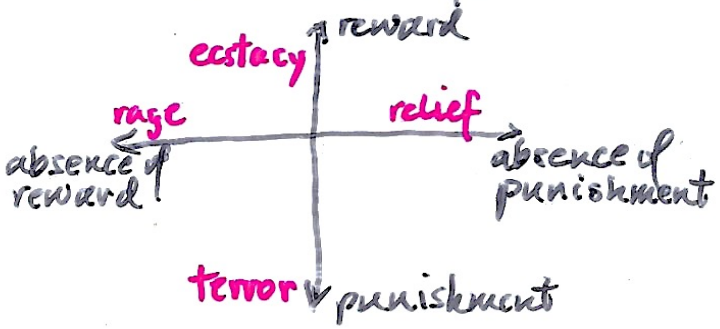
- Rewards & punishments, associated with "stimuli" (perceived situations), allow behavioral adaptation within the organism's lifetime (much more flexible than innate, "hardwired" behavior) — operant conditioning
- 2-stage learning: - association ① betw. stimulus & situations {reward / punishment}
  - operant responses ② {leading to / avoiding} {favorable / unfavorable} situations

① orbitofrontal cortex ② cortex, striatum

- 2 routes to action (apart from reflexes)
  - "implicit" behavioral responses (evaluation only)
    amygdala, orbitofrontal cortex, secondary sensory cortex → striatum ~~etc~~→ action
  - "explicit" behavioral responses (if-then rules, planning)
    additional routes from amygdala to prefrontal planning & language (verbalizable) → action
    conscious (HOT)

attention guidance

- Classification of emotions acc. to "reinforcement contingencies"



Robots

- might usefully have emotions
- but reward/punishment are evolved "selfish gene" mechanisms, not necessarily appropriate for robots
- robots are for humans, exploration, MT ... [!!]

Edmund Rolls: What are Emotions? (Fellous & Arbib Ch.5)
[in more detail]

[Or, Pavlov & Skinner Are Alive & Well in Neuroscience]

Classical Conditioning (Pavlov)

Repeat: neutral stimulus → positive (or -ve) unconditioned stimulus (US)

Cf. recommender systems

conditioned stimulus (CS) → conditioned response → positive (or -ve) US

CS → CR

Operant Conditioning (Thorndike, Skinner)

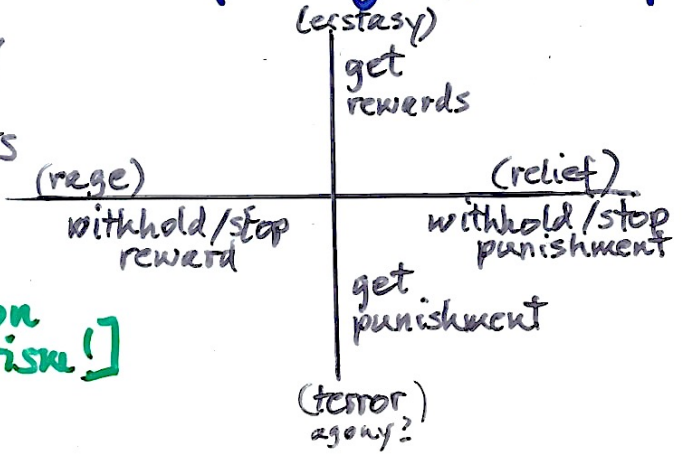
Repeat: choice situation → action1 → punishment (+ve or -ve) } reinforcement
choice situation → action2 → reward (+ve or -ve) } reinforcement

Cf. reinforcement learning sysems

choice situation → action 2 ("operant response")

Rolls: Emotions result from good/bad experiences

Effects of getting / not getting rewards/punishments



[Rolls somewhat neglects cognition here ↔ behaviorism!]

Evolution of reward/punishment system
⇒ enables learning/adaptation in a lifetime!

[So... adaptable robots should presumably also have reward/punishment systems, rather than fixed behavioral rules. But internal KR & thinking are also crucial.]

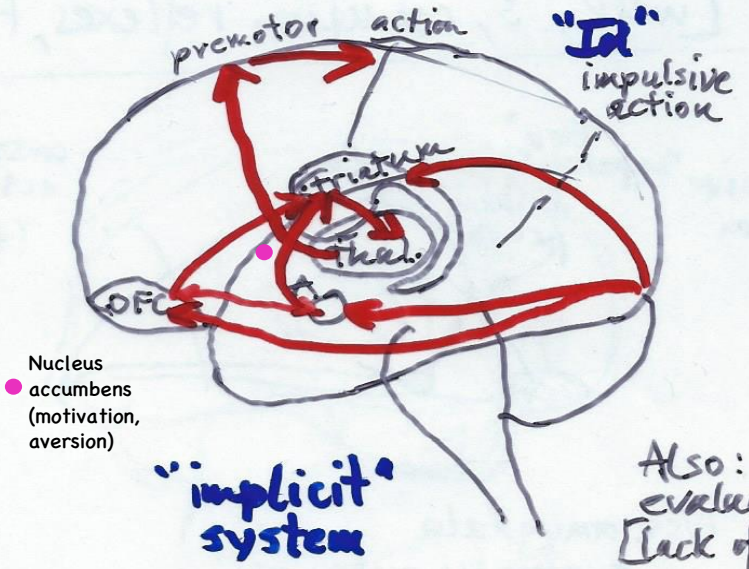
Two stages of learning

1. Situation ⟷ reward/punishment [amygdala, orbitofrontal cortex]
2. situation → **action** → gain reward / avoid punishment

[We also experience anticipatory rewards/punishments]
cf. "conditioned response"

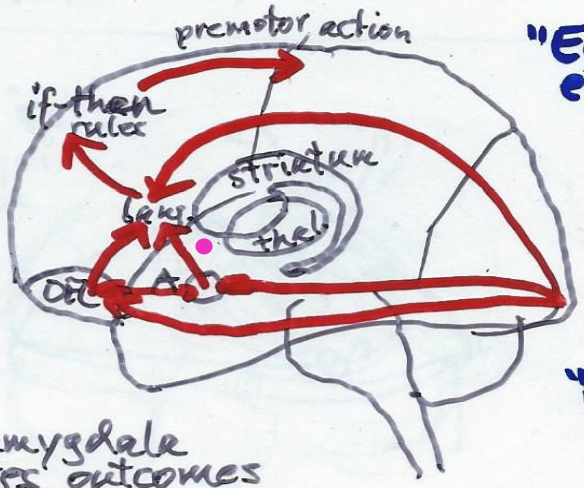
Natural selection "tunes" the reward/punishment system
→ food, sex, sheltered refuge ; fight / flight
→ kin altruism

Two routes to action [well, 3, counting reflexes, Fig. 5.2]



"Id" impulsive action

"implicit" system



"Ego, super-ego" conscious action (HOT!)

"explicit" system

Nucleus accumbens (motivation, aversion)

Also: amygdala evaluates outcomes
[lack of amyg. ⇔ cf Demerol, nitrous oxide]

Implicit system "highlights" aspects of current perceptions to guide the attention of the **explicit system**

[Robots: situations should "suggest" actions, rather than pure goal-directed deliberation]

Functions of emotion

- trigger autonomic responses
- enable flexible responses
- motivate appropriate action
- communicate via facial expression
- support bonding
- bias cognitive evaluation (via mood)
- facilitate memory storage
- promote persistence of motivation
- trigger recall

Concluding comments

- Importance of amygdala, orbitofrontal cortex in emotions
- Computers may not need emotions since MT, heavy hauling, terrain exploration, picture transmission don't need them [!]

I very much believe AI agents (now & in future) need

- rewards & punishments: reinforcement learning!
- emotions, including empathy for sapient, emotional beings