Twitter Thread by Alberto H

Alberto H
@Alberto_H9

Do machines need emotions to be “really intelligent”? Our best reference for an AGI (human brains) points in that direction.

If so, how on earth would one synthesize or encode “sadness” or “joy” for starters? ■?❤■

Join me and see how far we get! ■

Thx @AlejandroPiad, @svpino

I’ve worked on the topic of emotions + AI for a while now with the goal to “give the Tin Man” a heart.

I haven’t... yet :)

But I did learn a lot of “emotional engineering” along the way that you may find interesting or inspire you new RL approaches.

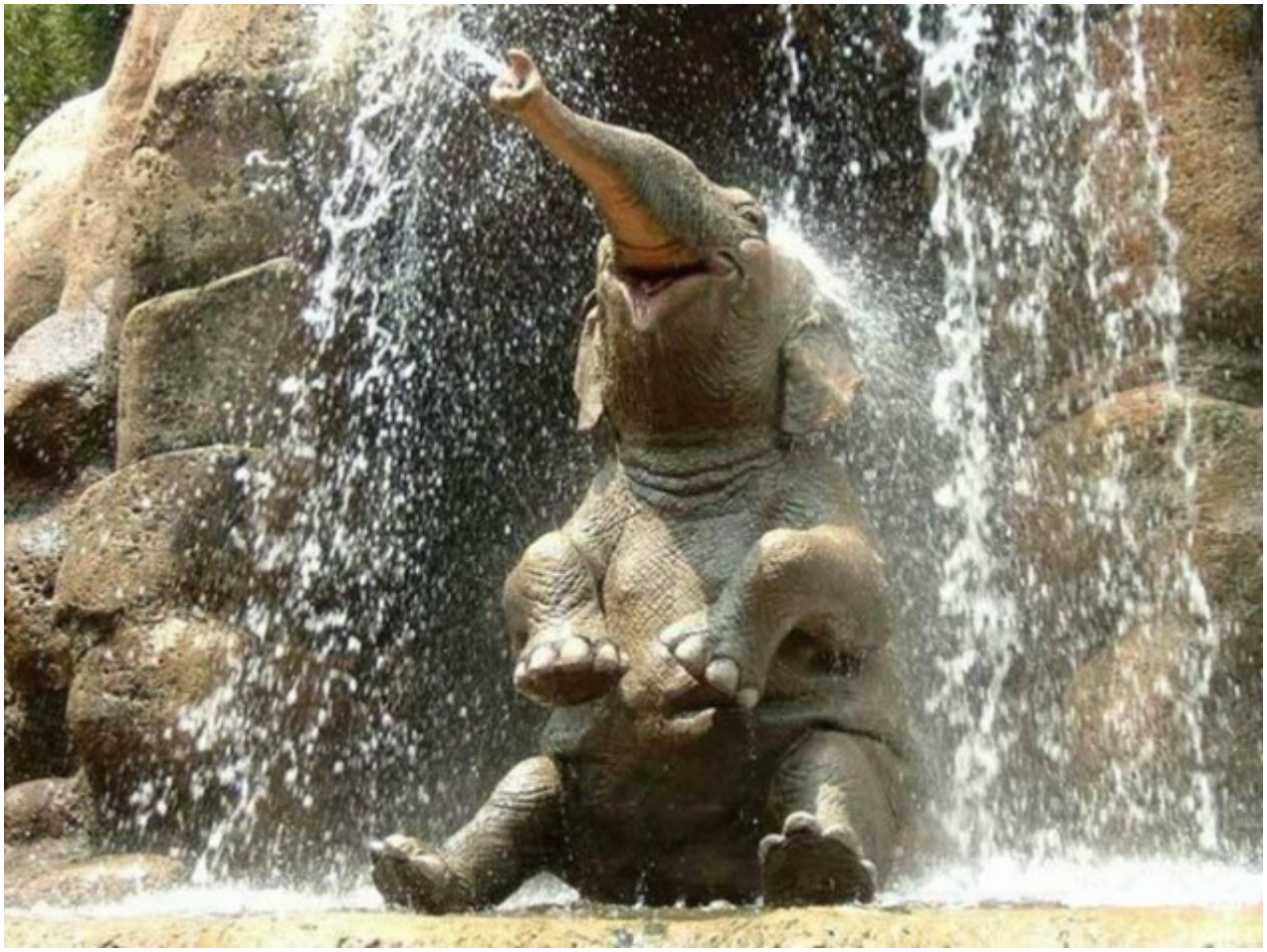
In this thread I'll cover many elusive psycho-bio-neuroscientific concepts from a computational perspective. (Bear in mind I'm a computer scientist, so feedback from people in those fields is welcome!)

Ok, so firstly, HOW are emotions useful for us living things?...

Our behaviours are (very strongly) driven by emotional forces, often prevailing over our more "rational" processes.

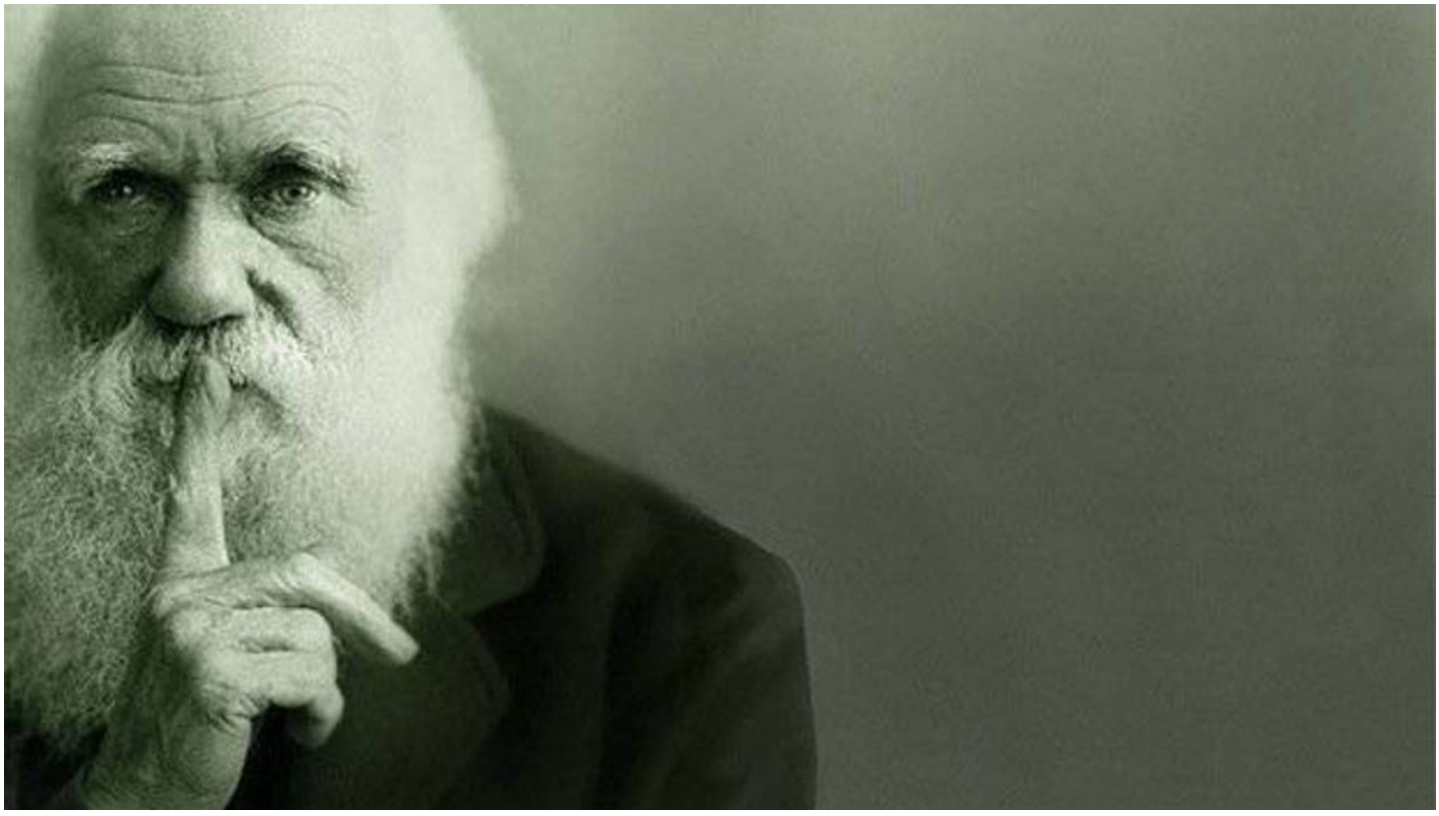
They also spark and feed our learning, focusing it on our most intense emotional episodes ("that hurt, I'll avoid it; that was great, I want more").

Furthermore, when expressed externally, emotions also provide a primordial language that allows inter- and intraspecies communication. (Isn't it fantastic how watching a nature documentary we can tell how animals feel from sounds and body language?)



Btw, animals can display great intelligence too and since Darwin (The Expression of the Emotions in Man and Animals - 1872) they also officially have emotions (who on Earth could have doubted it!?)

Ok, so now, how do you “give a heart” to the machine?



Bad news: there’s no such thing as an IEEE standard for “love” or “nostalgia”. Worse: even the core concepts (emotions, feelings, pain & pleasure, affect, mood…) lack a formal definition all scholars and scientists adhere to.

Where can we engineers start?

I started a while ago with this approach:

1) dive in the domain: biology/psychology/neuroscience have 2 centuries worth of experience
2) treat it like any system: trace information flows step by step from the bottom up: objective reality > senses >...> the subtlest feelings.

How far did I get? Here’s my “pocket summary” of how each emotional concept is seen by most in those fields; if you’re interested in emotions + AI, this could be your first map to the problem.

So it all starts out there: our *senses* provide raw data that we turn into *perception*. While our senses are “factory default”, perceptions can be "trained" (new objects, new symbols...) and have a neutral emotional sign or valence (in principle).

Pain [physical pain] originates in nociceptors all over our body (external and internal), signaling damage.

Its subjective interpretation is intrinsically negative and can’t be trained (though it can be ignored, accepted, etc. by higher mental processes).

Pleasure is generated by hedonic brain circuits as an intrinsically positive appraisal of a state, physical or mental.
Some pleasures are learned (admire a painting), while others are innate (yummy sugar), with blurry lines here. Notice there is NO such thing as “hedoceptors”!

Pain-pleasure constitutes the reward-punishment based system developed during evolution and is the key motivator of behaviour (the only one?). The asymmetric dynamics of this “common currency” are complex, to say the least…

An example: the pain of a loss is ≈2x as powerful as the pleasure of gaining (Kahneman & Tversky, 1979). Most animals take more risks to avoid some loss than the equivalent gain. (Take note, Sutton & Barto!)

We're ready to specify emotions now!

Ok now, *emotions* emerge naturally as innate, automatic appraisals of situations perceived as relevant to our goals.
They are short-lived and can be positive (happiness / enjoyment), negative (anger, fear, sadness, disgust…) or, apparently for some scholars, neutral (surprise).

Emotions, as you will have experienced, trigger clear default behaviours (if not overruled by our rational mind). See a few:

enjoyment -> retain
sadness -> recover
anger -> aggress
fear -> avoid
surprise -> attend

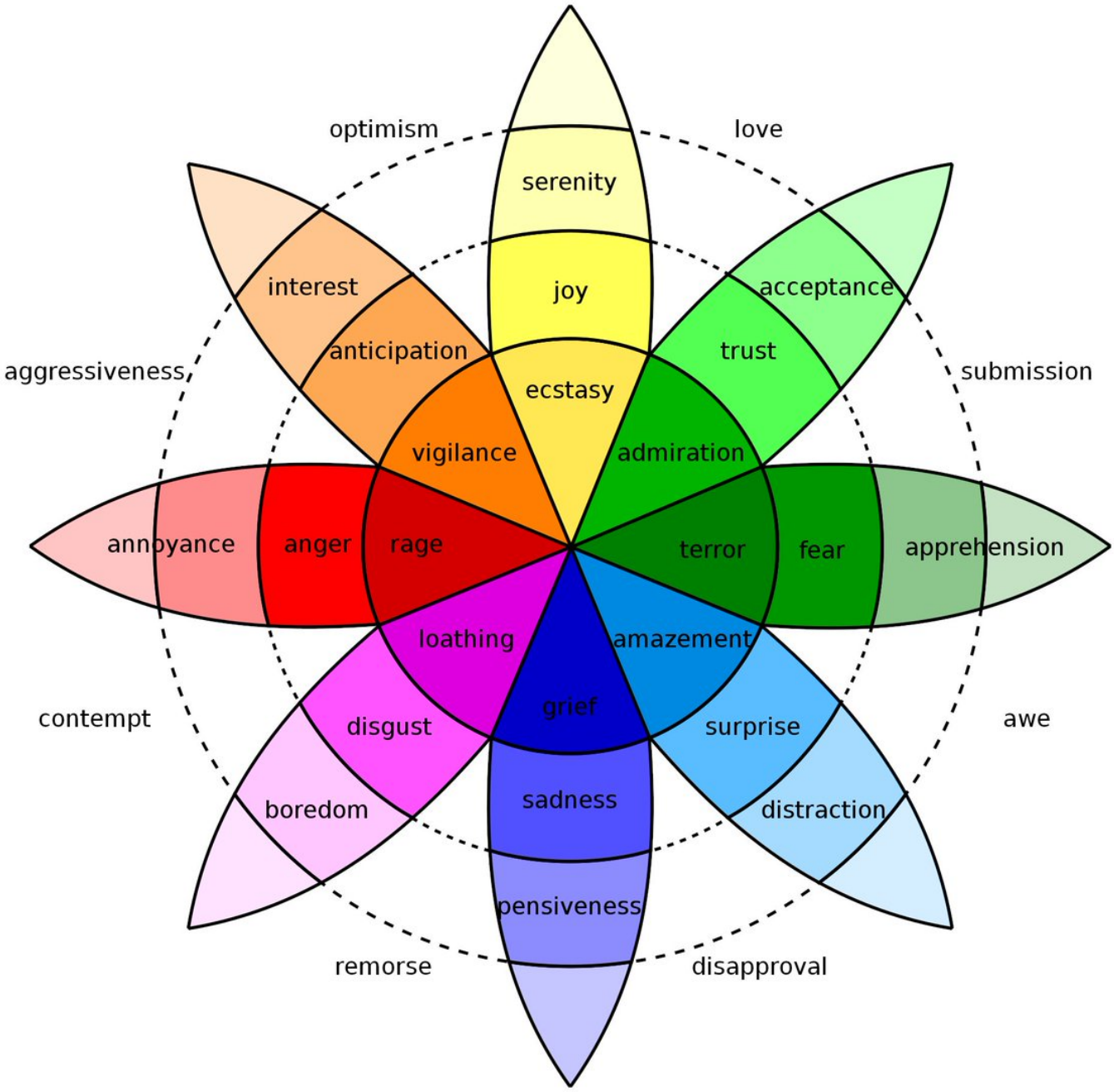
etc.

They seem linked to pain/pleasure states (current and predicted), e.g:
current pleasure -> enjoyment
current pain -> sadness
predicted pain -> fear / anger
predicted (?) -> surprise
...

So how many emotions actually exist? Well, senses & perception seem more straightforward, but when you get to emotions... every scholar brews their theory disregarding others' (reminds me of philosophers?)

Paul Ekman's popular BET (Basic Emotion Theory, 1971) describes 6 basic cross-cultural ones: happiness, sadness, disgust, anger, fear, surprise. But in 1999 he extended the list to 15!

There are many theories to pick from... Plutchick’s wheel of 8 primary emotions x 3 intensities + 8 mixes is very popular and self-explanatory. I love its continuous (non-categorical) nature.



And we finally get to *feelings*, seen as the conscious, subjective experience of emotion. Noticed the step? Your rational mind perceives an irrational emotion turning it into a cognitive object.

Here’s your link from emotion to cognition!

Feelings are still subjective, charged with positive/negative valence, but last longer in your mind.

As cognitive objects, they can be learned or modified by experience (love, phobiae, compassion, disdain, remorse...) The list is *endless* and very culturally dependent.

Philosophic note: many place the edge between innate and cultural right between emotions and feelings respectively. If correct, we can (and should) educate our feelings, but not so easily (or at all) our hard-wired emotions.

Ok, so next-level definitions are less controversial. *Moods* would be a sustained emotion that “colors the perception of the world” (Martin, E. 2007).

An *affect* is a more generic concept encompassing emotions, feelings and moods.

There’s more, but hey, this is Twitter...

So what do we AI researchers do with these computable (no longer esoteric!) concepts? How do they fit in the “machine’s heart”?

I for one believe that the integration of emotions into current approaches to Reinforcement Learning (RL) might be a game-changer.

We living things make good use of them within our mental processes: hey, neuroscience might (once more) be an inspiration.

A few ideas: enrich the perceived state, modulate the reward(s) signal, focused learning, agents’ communication, risk-modulated action selection...

I hope you found these concepts exciting and I encourage you to share your ideas here. Do you think emotions will always be an exclusively biological product? Why? Or why not?

Also, should we even try to give our AIs “a heart”? Why?

I'd love to hear your thoughts. Anyway, my personal guess is we’re still far from hearing from an agent:

“Now I know I've got a heart 'cause it's breaking...”
(Tin Man, The wizard of Oz)