On the rationality of emotion: a dual-system architecture applied to a social game

David C. Moffat
Department of Computing
Glasgow Caledonian University, UK.
D.C.Moffat@gcu.ac.uk

Abstract. The insightful dichotomy between fast and slow thinking, as identified by Kahneman [5], is explored here with a simple model of a rational agent playing the Ultimatum Game.

It is an interesting game to model because it creates a social context between the two players that induces apparently irrational behaviour. One explanation for this is that the players react emotionally to each other in the game; and the emotions are irrational.

Consideration of the model leads to a conclusion that the irrational behaviour patterns can indeed be reproduced by the artificial agent; although the question of whether emotions are truly irrational is not resolved or even addressed here. Another conclusion is that the distinction between fast and slow thinking may not be the most important criterion to distinguish Kahneman's notions of system-1 and system-2. Instead, the related concept of precedence could be prior.

1 Dual systems of cognition — fast and slow

Kahneman has popularised the concept of what he calls system-1 and system-2 thinking, in his excellent book [5]. He is also to be credited for originating many of the key ideas that led the rest of the field in that direction.

To summarise the fundamental dichotomy that Kahneman raises, system-1 thinking is characterised by being *fast*, *intuitive*, *automatic* and *subconscious* or largely *impenetrable* to introspective analysis. On the other hand, system-2 thinking is relatively *slow*, *deliberate*, *logical* and *conscious*, costing *effort* to the thinker.

Into system-2 would go thinking about what to do tomorrow, for example; or solving a puzzle. In AI terms we could associate this kind of thinking with its symbolic, traditional approaches.

System-1 would be for thinking that is more closely following perception, and otherwise more closely coupled to the environment. Kahneman puts emotion into this category as well.

2 The Ultimatum Game

The Ultimatum Game (UG) is an artificial mathematical game that is used in laboratory experiments to probe participants' judgements of fairness in social interactions.

There are two players in the game: the *proposer* and the *responder*, and a sum of money that they have to split between them as follows. The proposer offers a split, which we may express as a percentage of the sum. The responder then chooses to accept the offer, or reject it. Accepting the offer means that both players get their part of the split; but rejecting it means that both get nothing.

For example, if the proposer offers 50% then the responder would surely accept it, and both players would get half the sum. But if the proposer offers much less, say only 4%, then any human responder is likely to reject it. It is easy to see why (if you are also human): the responder is angered by the tiny offer, in which proposer keeps nearly all the money for himself. However, that angry human has behaved irrationally, according to standard economic theory and mathematical game theory. The responder should accept any offer made to him, to maximise his gain in "utility", because even a tiny amount of money is better than nothing.

The fact that people are consistently and robustly "irrational" in this way is what makes the UG such an interesting game for researchers. Is it really true that humans are an inherently irrational species? Is it our emotions that make us irredeemably irrational? Or is there something deeply wrong with standard economic theory?

There are some indications in the literature that it is indeed emotion to be blamed, and probably the emotions of anger or disgust. For example, dosing participants with the oxytocin before they play the UG makes them less likely to reject the offer [9]. As oxytocin is a hormone that fosters affiliative feelings in mammals, (and as we are mammals,) the suggestion is that responders feel more forgiving toward the proposers, and are thus less inclined to punish them.

Let us explore the possibilities of modeling these emotional reactions towards other agents in social situations like the UG. First we construct an abstract architecture of a purely rational agent, in the form of a traditional symbolic-AI planning system. Then we shall add an emotional system to itm and see if it can be made to behave in the "irrational" manner of real humans playing the UG.

3 The Rational Algorithm

1. event perceived
2. maybe replan
   - if no current plan, then
     - maybe construct one from current state and goals
   - else (have current plan), so
     - maybe replan it if the new event was unexpected
   - also generate expectations of any events other than own actions
3. execute next action in the plan
4. repeat from (1)

As an example of a planning algorithm that we could plug into the architecture at line (2) above, we could use any conventional ap-

proach based on the traditional STRIPS representation for actions and events [2, 8]. This would represent the action to reject the offer, say, as having precondition that the proposer has made the offer of a certain percentage *offer(p)*, and postconditions that both players get no money (so *gets(proposer,0)* and *gets(i,0)*, where the agent refers to itself with the personal pronoun *i*).

Without going into more detail of how the agent's planner works, it would arrive at the plan to maximise the profit to the agent itself. This is the intuition that economists have regarding the UG, namely that the rational thing to do is to accept any money offered. We can therefore call that response rational (according to economists' typical views about rationality as maximising utility).

In addition we may assume that the planner deals with the possibilities of other events occurring in the world, that are not its own actions, by making predictions about their likelihood. Without specifying how this might be done, let us say that for our case the agent arrives at the reasonable expectation that the proposer will be "fair" and offer an approximately even split.

- The plan is to wait for the offer, and accept it.
- Expecting an offer around 50%.

With the architecture implied by this algorithm, the agent would perform as follows.

Run through:-

1. event perceived is that I have been offered 20%
2. offer was lower than expected, but still within the plan so continue without replanning
3. next action is thus to accept the offer
4. plan and execution and game terminated: I accepted 20%.

The resulting decision is considered the rational one by the rational actor position in economics. If the agent is offered only 1% or 2% it should accept it, as its aim is to maximise its financial gain. Let us now turn to an emotional variant of this architecture, and see if it might behave otherwise.

4 The Emotional Algorithm

We add in a capacity for (supposedly emotional) *reaction* to the architecture by inserting an extra step, which is (2) below. It occurs before the planner, but could also be after it, and before the plan actions are executed.

The emotional step considers the observed event as potentially relevant to its suite of possible reactions, and reacts accordingly. The reaction rules may be expressed in a similar language to the STRIPS language used above for other planning actions. However the difference is that the reactions rules are not planned; they are triggered, or activated by certain kinds of stimulus events.

An example of an emotional reaction would be for the agent to retaliate when it is hurt by another agent. How it knows that it has been hurt is an interesting problem left on one side here. This is the rule that is exemplified in the execution run below.

1. event perceived
2. maybe react to event
   - if I appraise the event in context as emotionally significant
     - then execute the relevant emotional reaction (in context)
   - maybe break and repeat from (1), to perceive action as new event.
3. maybe replan
   - if no current plan, then
     - maybe construct one from current state and goals
   - else (have current plan), so
     - maybe replan it if the new event was unexpected
4. execute next action in the plan
5. repeat from (1)

Just as with the rational algorithm, the plan is to accept the offer. The planner works in just the same way, even with the emotional component, because in this design, the emotions only occur as reactions to events. In advance of any events (including the offer made by the opponent), then, the same decisions are made as before.

- The plan is to wait for the offer, and accept it.
- Expecting an offer around 50%.

Run through:-

1. event perceived is that I have been offered 20%
2. that is much lower than expectated 50%, so feel pain
   - appraised that action of opponent has hurt me
     - general emotion of "anger" requires retaliation
     - to hurt opponent in context is achieved by rejecting offer
     - therefore reject it
     - and maybe continue to plan, but in this case we have ended.
3. game over, so no replanning
4. and neither is there any need to continue executing the current plan
5. plan and execution and game terminated: I rejected the 20% offer.

The addition of an emotional capacity into the architecture has changed the behaviour to what we would call irrational. The agent itself would have agreed with that assessment, at any time before its own emotional reaction.

Notice that the emotional agent has the same plan as before, and thus the same intentions to accept any offer. But the occurrence of an emotional reaction has upset its plans, presumably to its own consternation afterwards. Later, after punishing the opponent in this way, the agent may repent at leisure: "Oh, but I should have taken the money!"

5 On the reality of cognitive models

We have considered two alternative algorithms, one named rational and the other emotional. The emotional one gives a better account of human behaviour, and in that sense it is a better model. How realistic is it though, and can it be said to be a true model of the cognitive mechanisms inside the human brain?

The matter of models and realism is an interesting issue in the philosophy of science (or the methodology of cognitive science). An influential trichotomy was put forward by David Marr [6], in which he distinguished three levels of analysis which a model could inhabit. The top level is the *computational* level, where models emulate what the natural system (such as a human subject) is doing; how it behaves, and the ultimate (evolutionary) purposes for that behaviour. The middle level is the *algorithmic* one, where the way that the computation is performed is also intended or claimed to be an accurate model of how the natural organism does it. The lowest level is the *implementation* level, where the mechanisms that execute the specified algorithms are also intended to be authentic.

For the human case then, a cognitive model at the implementation level would need to be implemented in some kind of artificial neural network architecture. Artificial intelligence models, and models in cognitive science, are generally pitched at the computational or algorithmic levels. Dennett has described the general methodological approach of the cognitive sciences as a descent down these three levels, from an initially accurate computational model, down through the lower levels by specifying particular algorithms and then mechanisms that in turn should be verified by eventual experiments. This approach toward "reverse engineering" the human mind is what he has called the "intentional stance" [1].

These matters are still debated to this day in cognitive science. See, for example, an interesting discussion by Zednik and Jäkel in 2014 [10].

For an example of a similar sort of argument as the one put forward here, see the interesting account of wishful thinking given by Neumann et al [7]. In that study, the authors propose a model that accounts for some human behaviour by limiting cognitive resources. In other words, they put forward an algorithmic model to explain the phenomenon of wishful thinking. They claim not to have found the unique best algorithmic model, but only an interesting one that would be fruitful for further research. That is the sort of claim that I am making in this paper.

In relation to these levels of analysis then, where do the algorithms here stand? Firstly, they count as computational models, in which the emotional one is found to be superior because it matches human data better. But then: is the emotional algorithm also an accurate model at the algorithmic level of analysis? Not necessarily: that is not the claim in this paper.

The point about the emotional algorithm is that it is a *possible* algorithm that would account for the correct behaviour at computational level. To further validate it as the *only possible* algorithm would require further experimental work, of the type often found in cognitive science. But the fact that it is possible (i.e. consistent with human behaviour) does mean that it excludes claims of alternative algorithms to be uniquely accurate models. In particular, any alternative scheme in which parallel processes for cognition and for emotion (to be crude about it for now) cannot claim to be the best models, if a sequential model like the emotional algorithm presented above can also model behaviour.

Kahneman's dichotomy [5] into system-1 and system-2 types of thinking, that is fast and slow, is a scheme of the above sort. This is what leads me to conclude that the algorithms presented here show that his scheme is not necessarily correct. Rather than speed of thinking processes, in order to explain emotional behaviour as the winner in some cognitive race, we can use the priority or precedence of the two processes, in a sequential algorithm instead. In the emotional algorithm shown earlier, its relative speed had nothing to do with the behaviour patterns shown. Instead, it was that emotional process were simply consulted first, and took precedence over less emotional cognition.

It is not such a significant result as to change research directions in cognitive science; and it does not necessarily invalidate Kahneman's views in any crucial sense. However, it is a curious reminder of how easily we might overstep the mark in our interpretations of mental mechanisms.

This perspective also happens to be consistent with Frijda's notion of *control precedence*, [3], [4]. It was partly because of his term that I have referred to the emotion's precedence; and why I wrote the algorithm out so that the emotion would literally *precede* the later cognitions. What Frijda means by control precedence is not only that emotion takes priority over other cognition; but that it can do so even in the agent's knowledge that it is acting against its own interests. In that sense emotion takes priority over rational preference, as demonstrated in our simple examples earlier.

Some readers may wonder if that is always the case. An example of a scientist giving his research a high priority, although it is only a cognitive goal, might seem to contradict. However, in my personal experience as such a scientist with that high priority goal in life, I can attest to the irritating fact that my own efforts to do research are frequently interrupted and often ruined by emotions of all sorts. While I might say and believe that science is a high priority for me, the evidence is clear that it is not as high as even mundane emotions.

6 Conclusions

The simple architecture outlined here has demonstrated how a component that provides for a kind of *emotional reaction* can issue in behaviour that more realistically resembles human behaviour in the UG experiments. In contrast, the purely "rational" version of the architecture does not behave like a human when it is offered a tiny percentage. Real people reject such unfair offers, possibly because of a sense of unfairness; but in any case because of an emotional reaction.

The emotional version of the model here also rejects the tiny offers, if it has the appropriate rule to do so (which we might call "anger" or "retaliation").

One interesting issue that has been left out here is the matter of how the rule (which is presumably evolved in humans) becomes related to a specific context (like the UG, which can only be learned).

Regarding the matter of rationality, the architecture(s) give an account for why emotion is often seen as irrational, even by the agents that feel them and act upon them. The crux of the matter is that the emotions are unplanned; and that only the agents plans are to be regarded as rational. (Otherwise, why plan them in the first place? The intention to be rational is implicit in the act of planning.)

Regarding Kahneman's dichotomy between system-1 and system-2, it is clear that the planner is system-2 (along with most traditional, symbolic reasoning AI systems). The new entrant here is the emotion subsystem, which falls in the category of system-1 thinking. The emotional reaction shown is not deliberate (it was not planned), but instead rather automatic (when triggered by appropriate events). It is also relatively impenetrable to consciousness or subconscious, although it has a conscious facet in the experience of feeling, for those organisms that can feel their emotions.

The dichotomy between system-1 and system-2 holds up fairly well therefore; but for one surprising exception. In this case (at least) there is no great computational cost in the plans that are constructed, as the plan can only have one action in it. The search algorithm needed to construct the plan is therefore trivial in our example; and so we may reasonably take it that the planning process runs off about as fast as the emotional reaction does, and thus might even direct the agent's next action before the emotion does. But if so, why does the agent react emotionally? The answer is clear from the algorithm: the emotion step occurs earlier in the algorithm's cycle.

This is why the new (emotional) step was introduced at step (2), and not merely added onto the end. If it had been put after the plan execution step in our linear model, then emotions would never occur, as the algorithm would return to repeat at the first step immediately after performing an action (in order to observe its own behaviour). The emotional step takes priority because it literally precedes the other cognitive processes. This is consistent with Frijda's term of

"control precedence" which is one of his defining characteristics of emotion.

We are thus lead to the conclusion, from the architectures here, that the more fundamental distinction between system-1 and system-2 thinking is priority, or control precedence, and not speed as such (despite the title of Kahneman's beautiful book [5]).

7 Acknowledgements

Two anonymous reviewers raised some interesting queries that I have attempted to answer above. One was the nice paradox about the scientist with a high priority goal to do research.

REFERENCES

[1] Daniel Dennett, *The intentional stance*, MIT Press, 1987.
[2] R. Fikes and Nils Nilsson, 'Strips: a new approach to the application of theorem proving to problem solving', *Artificial Intelligence*, (2), 189–208, (1971).
[3] Nico H. Frijda, *The emotions*, CUP Press, 1986.
[4] Nico H. Frijda, *The laws of emotion*, Erlbaum, 2007.
[5] Daniel Kahneman, *Thinking, Fast and Slow*, Macmillan, 2011.
[6] David Marr, *Vision*, Henry Holt Co., 1982.
[7] Rebecca Neumann, Anna N. Rafferty, and Thomas L. Griffiths, 'A bounded rationality account of wishful thinking', in *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, ed., P. Bello et al. Cognitive Science Society, (2014).
[8] Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach (2nd ed.)*, Prentice Hall, 2003.
[9] Paul J. Zak, Angela A. Stanton, and Sheila Ahmadi, 'Oxytocin increases generosity in humans', *PLoS ONE*, **2**(11), e1128, (11 2007).
[10] Carlos Zednik and Frank Jäkel, 'How does bayesian reverse-engineering work?', in *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, ed., P. Bello et al. Cognitive Science Society, (2014).