Human decision-making often departs strongly from the predictions of optimal choice. Discuss what cognitive biases and neural limitations cause the human decision maker to depart from optimality. Give at least one example each of cognitive bias and of a neural limitation (in computation e.g. inference, or in mechanism e.g. dynamics).

January 20, 2020

Student ID: 20308798

Module Number: PSGY4061

Word Count (calculated by *Overleaf*, this line excluded): 1482

Decisions and Optimalty (*Introduction*)

In 1944, von Neuman and Morgenstern included in their book *Theory of games and economic behavior* that optimal choice in adversarial games can (in theory) be computed by means of the *minimax algorithm.* For this algorithm to work optimally, one needs to evaluate all leaf nodes within the decision tree; in Chess, the order of the first decision tree (*game tree size*) is 10^{120} (Shannon, 1950), for Go, it is 10^{360} (Tromp, 2016), the number of atoms in the universe is approximately 10^{78} (Eddington, 1956). The intractability of the so-called *real world* problems becomes even more explicit when one takes into account that not only are they laying in a much more complex state space, but also that these states are usually partially observable and transitions from one to the other follow stochastic dynamics (they are Partially Observable Markov Decision Processes (POMDP), as opposed to perfect information and deterministic games like the ones presented above). In addition, the concept of optimalty can also be more obscure and subtle than "winning the game" when one thinks of a living being (Friston, 2010). Taking all this into account, the idea of an *omniscient agent*, which can choose optimally in an environment as complex as that faced by a biological organism, becomes simply infeasible: sub-optimalty is a necessary condition for any plausible biological (or computational) agent learning and deciding in our world.

On Limitations in Decision-Making (*Discussion*)

Cognitive Biases

During the last decades of the 20th century, a series of journal and book publications by Amos Tversky and Daniel Kahneman presented the concept of *cognitive biases* (also called *heuristics*). Their work, such as Prospect Theory (PT, (Kahneman & Tversky, 1979)), is still considered extremely influential in human decision-making, and included models that are valid today. In this last sense, (Ruggeri et al., 2020) tested PT over a population of over 4000 participants from 19 different countries and results were very similar to those obtained in original work.

Prospective Theory follows the pathway started by Herbert A. Simon with the concept of *Bounded Rationality* (1982), in the sense that both are descriptive models; their intention is not to find what an optimal agent should do, but to describe how people do chose in reality (in opposition to normative models). In particular, the fact that some of the Expected Utility Theory axioms (ETU, (Morgenstern & Von Neumann, 1944)) are violated in human behaviour, motivates an extension that is able to acount for such deviations from rationality (an exmample of a violated axiom is *completeness*, given that when presented when the exact same secnario two people might choose differently). In PT, probabilities p_i are replaced with a subjective interpretation of these, called decision weights ($\pi = \pi(p_i)$). Similarly, the utility assigned to an objective outcome x_i is modelled by a quantity $v = v(x_i)$. The maximization of the *prospect* ($\sum \pi(p_i)v(x_i)$), instead of the traditional Expected Utility ($\sum p_i x_i$), provides a framework that excellently predicts people's deviations from optimalty.

Cognitive biases, derived from PT, aim to describe how people deviate from optimalty in different circumstances. An exhaustive review of all cognitive biases exceeds the extension and purpose limitations of this essay, but one example is included for illustration:

Availability Bias This heuristic compares the number of elements in two categories by using only a quick mental sample of them. This cognitive bias can have two sources of error: (1)biased sampling within knowledge (the search structure found more examples of A than B, while I *know* more elements in B than in A), (2)biased knowledge (I might really now more elements in A than B, while there exist in total more elements in B than in A). It should be noted that our choice would only differ from a rational agent in the former, as the later is a case of an incomplete world model, but correct inference process.

Neural Limitations

The study of the brain as a computing machine can provide insight on quantifying how much and in what aspects it differs from optimalty. In the last decades, Reinforcement Learning (RL, (Sutton et al., 1998)) has provided a theoretical framework by which both machines and intelligent animals learn from interacting with the environment (Lee et al., 2012), and in this essay will be used to guide through the different exploration, learning and decision strategies the brain is believed to implement.

RL and value encoding in the brain

State Value function Is the *total discounted reward* (TDR) an agent expects to obtain from a particular state:

$$V(s) = E[R_t|s_t = s] = E[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}|s_t = s] \tag{1}$$

State-Action Value Is the TDR an agent expects to obtain from taking a particular action in a particular state:

$$Q(s, a) = E[R_t|s_t = s, a_t = a] = E[r_{t+1}|s_t = s, a_t = a] + \gamma V(s_{t+1}) \tag{2}$$

In equations (1) and (2), s_t is a state in time t, a is an action, r_{t+1} is the *reward* obtained for transitioning from s_t to s_{t+1}, γ is a *discount factor* and R_t is defined by comparison in (1). The discount factor $\gamma \in [0, 1]$, represents how the reward obtained in the future might be less relevant than an immediate reward, and leads to myopic (focused in near future) and non-myopic (focused in all future states) *policies* (revised in following sections).

The quantities $V(s)$ and $Q(s, a)$ have been found to be surprisingly correlated with neural activity in certain regions of the brain for very simple tasks, where one can easily estimate them computationally. In (Cai et al., 2011) single-neuron recording in the ventral and dorsal striatum of monkeys suggested its encoding of temporally discounted values and state-action values, respectively. Furthermore, there is extensive literature on these quantities being encoded in primates in regions as the dorsolateral Prefrontal Cortex (dlPFC), premotor cortex and supplementary eye field, as well as in the secondary motor cortex and striatum in rodents (Lee et al., 2012).

RL and decision-making in the brain

Policy function Given s, is a decision probability distribution:

$$a \sim \Pi(a|s) \tag{3}$$

For an omniscient agent, the policy would be a Dirac delta centered at the action that has a highest associated action-state value. In the context of a biological agent, both the model of the world and $V(s)$, $Q(s, a)$ values are estimates based on observations: the *explore-exploit dilemma* appears in scene.

One speaks of the exploit-explore dilemma when an agent faces the *meta-decision* of choosing between exploiting what it considers to be the best option and improving the environment and values representations by exploring alternative actions and states. Several mathematical and computational approaches have been proposed in order to guide a policy in this context:

(a) random exploration (ϵ-greedy policy / softmax exploration): the agent chooses *greedily* (the action with highest expected payoff), but with a probability *epsilon* (which in softmax exploration depends on Q(s,a)) it acts randomly.

(b) directed exploration (Thompson sampling / Upper Confidence Bound (UCB): the agent explores with the intention of minimizing uncertainty.

Again, these proposals seem to have their counterparts in the biology of the brain. In (Yu & Dayan, 2005) the relation between Acetylcholine (ACh) and Norepinephrine (NE) concentration is proposed to relate expected and unexpected uncertainty, thus regulating when to revise the approximated model of the world by increasing exploration. Other models, such as that of Aston-Jones and Cohen (2005) and McClure et al. (2006), make use of the tonic and phasic mode of the Locus Coeruleus (LC) described in (Jones et al. 1994, 1997) to model exploratory and exploitative states (respectively).
Recent work has gone a step forward by proposing the PFC as a gating mechanism between stored policy functions that allows for transfer learning (Tsuda et al., 2020).

RL and learning in the brain

Temporal-Difference Error It defines the difference in value observations within a time interval $[t, t+1]$:

$$\delta(t) = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) = r_{t+1} + \underset{a}{argmax}[Q(s_{t+1}, a)] - Q(s_t, a) \tag{4}$$

For an omniscient agent, (4) should be equal to zero. In real life cases, however, this quantity allows to compare between the believed representation of the world and a new observation made, and has thus been proposed to be computed in the brain for learning purposes. In particular, (Montague et al., 1996) and (Schultz et al., 1997) put forward a dopamine mediated temporal-difference learning algorithm based on Hebbian learning, where the temporal derivative of the synaptic weight between two units with respect to time is proportional the temporal-difference error.
In (McClure et al., 2006), the authors built a population model for decision by extending the Pooled Inhibition Model, developed in (Wang, 2002) and extensively reviewed and compared to simpler diffusion models in (Bogacz et al., 2006). Synaptic strengths of the network followed a temporal difference algorithm and learnt to distinguish between a rewarded stimulus and a distractor. The network had troubles, however, in learning new strategies when the rewarded and distractor stimulus were interchanged. By adding a model of the LC that evaluated whether or not to enter in tonic mode, the network was able to detect conflicts in its internal representation of the world, input noise to the system, and then stabilize towards a new learned state, in a very similar manner of the annealing mechanism in machine learning algorithms.

Conclusions

In this essay, suboptimalty has been examined as a necessary condition in human's decision-making processes. In this context, we have revised cognitive biases, that describe how we deviate from optimalty, and RL, a computational approach to model how we generate approximate models of the world and try to maximize a subjective conception of rewards through our actions.

A final note: *if RL is the mathematical (rational) approach to maximize rewards once one accepts computational and observational limitations, and our brain keeps on showing us that it implements RL algorithms, isn't it actually optimal?*

References

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu. Rev. Neurosci.*, *28*, 403–450.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological review*, *113*(4), 700.

Cai, X., Kim, S., & Lee, D. (2011). Heterogeneous coding of temporally discounted values in the dorsal and ventral striatum during intertemporal choice. *Neuron*, *69*(1), 170–182.

Eddington, A. S. (1956). The constants of nature. *JR Newman:"The World of Mathematics*, *2*, 1074–1093.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, *11*(2), 127–138.

Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, *47*(2), 263–291.

Lee, D., Seo, H., & Jung, M. W. (2012). Neural basis of reinforcement learning and decision making. *Annual review of neuroscience*, *35*, 287–308.

McClure, S. M., Gilzenrat, M. S., & Cohen, J. D. (2006). An exploration-exploitation model based on norepinepherine and dopamine activity. In *Advances in neural information processing systems* (pp. 867–874).

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of neuroscience*, *16*(5), 1936–1947.

Morgenstern, O., & Von Neumann, J. (1944). *Theory of games and economic behavior.* Princeton university press.

Ruggeri, K., Alí, S., Berge, M. L., Bertoldo, G., Bjørndal, L. D., Cortijos-Bernabeu, A., ... others (2020). Replicating patterns of prospect theory for decision under risk. *Nature Human Behaviour*, 1–12.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599.

Shannon, C. E. (1950). Xxii. programming a computer for playing chess. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *41*(314), 256–275.

Sutton, R. S., Barto, A. G., et al. (1998). *Introduction to reinforcement learning* (Vol. 135). MIT press Cambridge.

Tromp, J. (2016). The number of legal go positions. In *International conference on computers and games* (pp. 183–190).

Wang, X.-J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. *Neuron*, *36*(5), 955–968.

Yu, A., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46*(4), 681–692.