The Role of Outcome Divergence in Goal-Directed Choice

Prachi Mistry (prachim@uci.edu), Mimi Liljeholm (m.liljeholm@uci.edu)
Department of Cognitive Sciences, UC Irvine
Irvine, CA 92697

Abstract

We assessed the influence of instrumental outcome divergence – the extent to which actions differ in terms of their outcome probability distributions – on behavioral preference in a two-alternative forced choice task. We found that participants preferred a pair of available actions with high divergence to a pair with low divergence. The effect of outcome divergence, dissociated here from that of other motivational and information theoretic factors, potentially reveals the value of flexible control.

Keywords: Instrumental Outcome Divergence; Flexible Control; Goal-Directedness, Choice Preference

Introduction

Goal-directed decisions are supported by a "cognitive map" of state transition probabilities that are flexibly combined with subjective utilities in order to generate action values, the basis of choice (Tolman, 1948; Balleine and Dickinson, 1998; Doya et al., 2002; Daw et al., 2005). Although computationally expensive, the dynamic binding of utilities and probabilities offers adaptive advantage over more automatic, habitual, action selection, which uses cached values based on reinforcement history (e.g., Daw et al., 2005). There are, however, situations in which the processing cost of goal-directed computations does not yield the return of flexible control. Here, we introduce a novel decision variable – *instrumental outcome divergence* – that serves as a measure of flexible control, and assess its influence on behavioral preference.

Consider the scenario illustrated in Figure 1a, which shows two available actions, A1 and A2, where the bars represent the transition probabilities of each action into three potential outcome states, O1, O2 and O3. Here, the goal-directed approach prescribes that the agent retrieves each transition probability, estimates the current subjective utility of each outcome, computes the product of each utility and associated transition probability, sums across the resulting value distribution for each action and, finally, compares the two action values (e.g., Doya et al., 2002; Daw et al., 2005). Of course, given equivalent costs, actions that yield identical outcome states will inevitably have the same value, and thus need not be contrasted further in terms of the utilities of their outcomes. Consequently, the extent to which actions differ in terms of their relationships to future states, that is, the *divergence* of their outcome probability distributions, can be used to prune searches of the cognitive map.

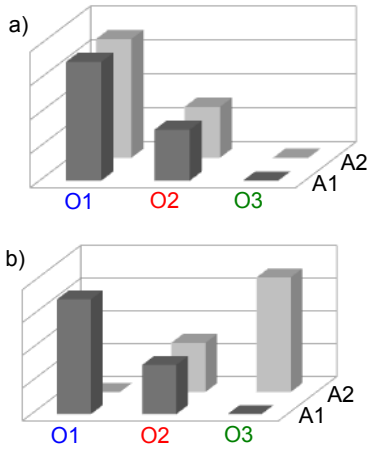

Figure 1: Probability distributions over three possible outcomes, O1, O2 and O3, for two available actions, A1 and A2.

Now consider the scenario in Figure 1b, in which the probability distribution of A2 has been reversed across the three outcomes. Note that, if the utilities of O1 and O3 are the same, the two actions still have the same value. Likewise, the outcome entropies – that is, the uncertainty about which outcome will be obtained given performance of an action – is the same across the two actions. And yet, the two actions clearly differ. To appreciate the significance of this difference, imagine that O1 and O3 represent food and water respectively, and that you just had a large delicious meal but without a drop to drink. Chances are that your desire for O3 is greater than for O1 at that particular moment. However, a few hours later, you may be hungry again and, having had all the water you want, now have a preference for O1. Unlike the scenario illustrated in Figure 1a, that in Figure 1b allows you to produce the currently desired outcome as preferences change, by switching between actions. Thus, instrumental divergence can serve as a measure of agency – the greater the divergence between available actions, the greater the agent's flexible control over the environment.

Scanning human participants with functional magnetic resonance imaging (fMRI) as they performed a simple decision-making task, Liljeholm et al., (Liljeholm et al.,

2013) found that activity in the inferior parietal lobule, an area previously implicated in several aspects of goal-directed processing, including the computation of instrumental contingencies (Seo et al., 2009; Liljeholm et al., 2011) the attribution of intent (den Ouden et al., 2005), and awareness of agency (Chaminade and Decety, 2002; Farrer et al., 2008; Sperduti et al., 2011), scaled with the instrumental divergence of available actions. However, the task used by Liljeholm et al. did not allow them to assess the influence of divergence on behavioral preference. Here, to behaviorally assess the value of control, we used a novel experimental task (see Figure 2 and Methods for details) in which participants chose between pairs of actions with different levels of instrumental divergence. Specifically, participants were required to choose between a high and low divergence action pair at the beginning of each block and, on subsequent trials in that block, choose only between the actions within the selected pair.

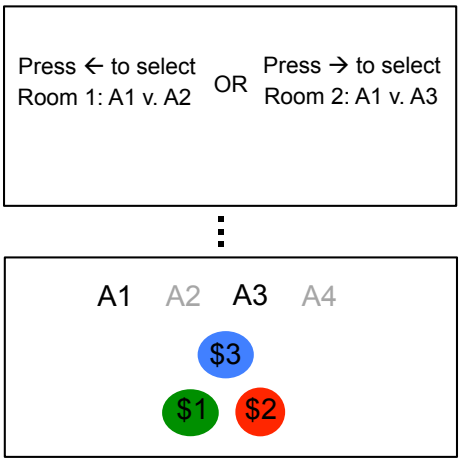

Figure 2: Task illustration showing the choice screen at the beginning of a block (top), and the choice screen on a trial within the block (bottom). On each trial, once an action (e.g., A3) was selected, a feedback screen showed that action inside a selection square, together with the particular token (e.g., red) delivered on that trial.

We predicted that, all other things being equal, participants would prefer action pairs with high divergence, as these yield the highest level of flexible control. Flexible control is particularly important in a dynamic environment, where the subjective utilities of outcome states change frequently. With primary rewards, this is generally the case due to sensory-specific satiety (e.g., Balleine and Dickinson, 1998). Here, to simulate the subjective utility of primary rewards, we used colored tokens as outcomes, and reassigned values to tokens across conditions.

Method

Participants Twenty undergraduates at the University of California, Irvine (12 females; mean age = 20.20 ± 4.81) participated in the study for course credit. All participants gave informed consent and the Institutional Review Board of the University of California, Irvine, approved the study.

Decision Variables We defined the *expected value* of each available action as the sum of the products of its transition probabilities and token utilities: thus, if the values of the green, blue and red tokens are $1, $2 and $3 respectively, and an action produces the three tokens with probabilities of 0.7, 0.0 and 0.3 respectively, the expected value of that action is $1.6. Another important decision variable frequently shown to influence instrumental choice is the variability, or entropy, of outcome states (Erev and Barron, 2005; Weber and Huettel, 2008; Abler et al., 2009), which is greatest when the probability distribution over outcomes is uniform (i.e., all outcomes are equally likely) and smallest when the probability of a particular outcome is 1. We computed the Shannon entropy of the outcome variable X conditional on a particular action Y, defined as:

$$H(X|Y) = \sum_{x\in X, y\in Y} p(x,y)\log\frac{p(y)}{p(x,y)}$$

To rule this variable out as a source of behavioral preference, we kept it constant across all actions throughout the study.

Finally, we formalize instrumental divergence as the Jensen-Shannon (JS) divergence of the outcome probability distributions for the actions in a given pair. A finite and symmetrized version of the Kullback-Leiber divergence, JS divergence specifies the distance between probability distributions M and N as:

$$JSD = \frac{1}{2}\sum_i \ln\left(\frac{M_i}{P_i}\right)M_i + \frac{1}{2}\sum_i \ln\left(\frac{N_i}{P_i}\right)N_i$$

where

$$P = \frac{1}{2}(M+N)$$

Task & Procedure The task is illustrated in Figure 2. At the start of the experiment, participants were instructed that they would assume the role of a gambler in a casino, playing a set of four slot machines (i.e., actions, respectively labeled A1, A2, A3, and A4) that yielded three different colored tokens (blue, green and red), each worth a particular amount of money, with different probabilities. They were further told that, in each of several blocks, they would be required to first select a room in which only two slot-machines were available, and that they could only choose between the two machines in the selected room on subsequent trials in that

block. Finally, participants were instructed that, while the outcome probabilities would remain constant throughout the study, the values of the tokens would change at various times, and these changes might occur after the participant had already committed to a particular pair of machines in a given block. Consequently, although changes in value were explicitly announced, and the current values of tokens were always printed on their surface (to facilitate the computation of expected values), a participant might find themselves in a room in which the values of the two available actions had suddenly been altered.

Two distinct probability distributions over the three possible token outcomes were used: 0.7, 0.0, 0.3 and 0.0, 0.7, 0.3. The assignment of outcome distributions to actions was such that two of the actions (either A1 and A2 or A1 and A3, counterbalanced across subjects) always shared one distribution, while the other two actions shared the other distribution. This yielded a low (zero) outcome divergence for pairs in which the two actions shared the same probability distribution (as in Figure 1a), and a high (0.49) outcome divergence for pairs in which actions had different outcome probability distribution (as in Figure 1b). The unpredictability (i.e., Shannon entropy) of outcomes given a particular action was held constant at 0.61 for all actions. The four actions were combined into six pairs, which were in turn combined into 15 two-alternative choice scenarios (see top screen in Figure 2). For 8 of these scenarios, divergence differed across the two action pairs, while for the remaining 7 scenarios, all decision variables, including divergence, where the same for both available pairs.

We were primarily interested in assessing preference for high- over low-divergence pairs when expected value was held constant across pairs. Consequently, in the majority of blocks, the values assigned to the blue, green and red token respectively were \$2, \$2 and \$1, yielding identical expected values for all actions. However, to simulate a dynamic environment, in a subset of blocks, the token values were changed to \$2, \$1 and \$3 respectively, and in yet another subset they were changed to \$1, \$2 and \$3: For these two subsets of blocks, the expected value of the low-divergence action pair was either higher (\$2.30) or lower (\$1.60) than that of the high-divergence pair (\$1.95), depending on which outcome probability distribution was shared by the two actions in the low-divergence pair.

Changes between the three distinct value assignments described in the previous paragraph (henceforth v_1, v_2 and v_3), although explicitly announced and apparent based on the numbers printed on tokens, were unpredictable in that they always occurred after a participant had already committed to a particular pair of actions in a given block. The order of value assignments was such that four consecutive v_1 blocks were followed by a set of four consecutive v_2 blocks, followed by another set of four v_1 blocks, followed by a set of four v_3 blocks. This entire sequence was repeated once, followed by a final set of four v_1 blocks, yielding 9 sets of 4 blocks for a total of 36 blocks.[1] Each v_2 and v_3 set contained two blocks in which expected value differed in the *same* direction as divergence and two blocks for which expected value differed in the *opposite* direction of divergence. The order of v_2 and v_3 sets was counterbalanced across participants and the order of blocks within each set was random. Finally, each block consisted of 6 trials on which participants choose between the two actions in the selected pair, for a total of 216 trials.

Before starting the gambling task participants were given a practice session in order to learn the probabilities with which each action produced the different colored tokens. To avoid biasing participants towards any particular reward distribution, no values were printed on the tokens in the practice session. During practice, to ensure equal sampling, each action was presented on 10 consecutive trials with only that action being available, and with tokens occurring exactly according to their programmed probabilities (i.e., if the action produced green tokens with a probability of 0.2, the green token would be delivered on exactly 2 of the 10 trials).

Following 10 trials with a given action, participants rated the probability with which that action produced each colored token before proceeding to the next action. Once all actions had been practiced, the four actions were presented in random order and participants again rated the outcome probabilities of each. If a participants' estimate of any given probability deviated by more than 0.2 from the programmed probability, they were returned to the beginning of the practice phase, and this continued until all rated probabilities were within 0.2 points of programmed probabilities. At the end of the study, after the gambling phase, participants again provided estimates of the action-token probabilities.

Importantly, all monetary amounts were fictive, and participants were instructed at the beginning of the experiment that they would not receive any actual money upon completing the study. Nonetheless, given the previously demonstrated correspondence between real and fictive monetary rewards, in both behavioral choice and neural correlates (Bowman and Turnbull, 2003; Bickel et al., 2009; Miyapuram et al., 2012), we predicted that participants would select pairs with the highest expected value whenever expected value differed across pairs, regardless of differences in divergence. We also hypothesized, however, that participants would choose according to divergence whenever expected values were

[1] For completeness, the 7 choice scenarios in which all decision variables were held constant across the two available pairs were randomly distributed throughout the sequence of 36 blocks. These 7 blocks were not analyzed and will not be discussed further.

held constant across pairs, reflecting the postulated value of control.

Results

Participants required on average 1.9 (*SD*=1.2) cycles of practice on the action-token probabilities. Mean probability ratings, obtained right before and right after the gambling phase, are shown in Table 1.

Table 1: Mean probability ratings with standard deviations. Programmed probabilities are shown in the top row. Mean ratings, obtained before and after the gambling task, are averaged across actions and outcomes, yielding three unique outcome probabilities.

	0.7	0.0	0.3
Before	0.70 ± 0.02	0.00 ± 0.01	0.30 ± 0.02
After	0.65 ± 0.17	0.04± 0.12	0.31 ± 0.07

Consistent with our primary hypothesis, we found that, when divergence differed across the two available action pairs while expected value and outcome entropy were held constant, there was a preference for high over low divergence, such that participants selected the high divergence pair 66% (*SD* = 23%) of the time. A planned comparison revealed that this was significantly greater than chance performance, $t(19) = 3.20$, $p < 0.005$.

To assess how preferences for divergence was modulated by expected value, blocks were divided into 3 expected value conditions: In the first condition, expected value was held *constant* across the high and low divergence pair, in the second condition, expected value differed across pairs in the *same direction* as divergence, and in the third condition expected value differed in the *opposite direction* of divergence. We entered the percentage of high divergence choices by each participant into a 3 (expected value) x 2 (order) x 2 (probability) mixed analysis of variance, with "expected value" as a repeated measure and with "order" and "probability", respectively indicating the order of v_2 and v_3 sets and the assignment of probability distributions to actions (see methods), as between-subjects factors.

There was a significant main effect of expected value, $F(2,32)=21.86$, $p<0.001$. There was no significant effect of the order in which changes in value assignments (i.e., v_2 and v_3 sets) occurred within the sequence of blocks, $F(1,16)=1.14$, $p=0.30$, nor any significant effect of which two of the four actions shared a particular outcome probability distribution, $F(1,16)=2.62$, $p=0.13$. There were no significant interactions (smallest $p=0.13$).

Bonferroni adjusted pairwise comparisons revealed that the percentage of high divergence choices was significantly greater when expected value differed in the same direction as divergence, 81% (*SD* = 24 %), than when expected value differed in the opposite direction of divergence, 33% (*SD* = 26%), $p<0.001$. The percentage of high divergence choices was also significantly greater when expected value was held constant across pairs, 66% (*SD* = 23%), than when expected value differed in the opposite direction of divergence, $p<0.001$. Although the percentage of high divergence choices was apparently greater when expected value differed in the same direction as divergence than when expected value was held constant, this difference did not reach significance, $p=0.09$.

Discussion

We assessed the influence of instrumental outcome divergence – the extent to which actions differ in terms of their outcome probability distributions – on behavioral preference in a simple gambling task. In each round of gambling, participants chose between two pairs of actions, knowing that they would be restricted to choosing between actions in the selected pair on subsequent trials in that round. One pair of actions had high outcome divergence while the other pair had zero outcome divergence. We found that, when other decision variables, such as expected value and outcome predictability, were held constant, participant chose the pair with high divergence significantly more often than that with zero divergence.

As noted, actions with high outcome divergence afford an agent flexible control over the environment: a commodity that is particularly valuable when the utilities of states are dynamically changing, as in the current task. We interpret the preference for high divergence demonstrated here as reflecting the intrinsic value of flexible control. Alternatively, however, participant's choices may reflect a previously demonstrated tendency to increase diversity, motivated by a desire to minimize risk in uncertain environments (Hedesstrom et al., 2006; Ayal and Zakay, 2009). Although highly related, in that greater outcome divergence allows for greater diversity, as is the case in the present study, the flexible control afforded by divergence does not necessarily follow from diversity.

To illustrate the distinction between diversity and instrumental divergence, imagine that you are allowed to choose between the two scenarios illustrated in Figure 1a and 1b respectively, but that once you make your selection, a computer algorithm chooses between A1 and A2 with a probability of 0.5. While selecting the high-divergence scenario in Figure 1b would yield the highest diversity, it would not allow you to avoid a particular outcome (e.g., O1) should this outcome suddenly lose its utility. On the other hand, if you were permitted to choose between A1 and A2 yourself, selecting the high-divergence scenario would allow you to completely avoid O1. Alternatively, imagine that the two outcome probability distributions in Figure 1a were uniform, such that all outcomes were equally likely:

this would yield maximum diversity, but zero instrumental divergence. Further work is needed to discriminate between preferences for diversity versus instrumental divergence in goal-directed choice.

On several gambling rounds in the current study, expected value differed across action pairs, in either the same or opposite direction of divergence. Participants' choices were in accordance with expected value on these rounds, such that the percentage of high divergence choices was significantly greater when expected value differed in the same than in the opposite direction. Indeed, the high divergence pair was only selected on 33% of rounds in which expected value differed in the opposite direction. Although this preference for monetary reward over divergence would likely have been even more marked if actual, rather than fictive, monetary amounts had been used, it is also possible that there are relative magnitudes of currency and divergence at which the value of control exceeds that of monetary gain: a breaking point in the trade-off between motivational and information theoretic decision variables.

Model-based reinforcement learning (RL) represents knowledge about action-outcome contingencies as a matrix of state-transition probabilities (e.g., Doya et al., 2002; Daw et al., 2005). In this framework, on each learning trial, leaving one state and arriving in the next contingent on performing a particular action, the agent computes a state prediction-error, which is then used to update transition probabilities. In our previous work (Liljeholm et al., 2013) we computed instrumental outcome divergence based on such trial-by-trial changes in transition probabilities, derived by fitting an RL model to behavioral choices. In the present study, since participants were trained to criterion on outcome probabilities prior to the gambling task, we instead computed divergence based on two predefined outcome probability distributions. Future work may consider a more fine-grained analysis of preference for high divergence under conditions of trial-and-error learning.

Another interesting consideration is the potential role of outcome divergence in stimulus generalization. If two cues signal identical future states, the cost of discriminating between them does not yield a return of improved predictability and, consequently, is likely not worth the effort. Indeed, it is well known that pairing distinct cues with the same outcome enhances subsequent generalization between those cues, a phenomenon known as *acquired equivalence* (Honey and Hall, 1989; Liljeholm and Balleine, 2010). Analogously, in acquired distinctiveness, the pairing of cues with different outcomes decreases subsequent generalization (Bonardi et al., 2005). Thus far, equivalence and distinctiveness effects have, to our knowledge, been limited to cases with two distinct outcome states and deterministic cue-outcome relationships. The use of outcome divergence allows for a potential extension of such effects to the case of multiple probabilistic outcomes.

Finally, outcome divergence may have a modulatory influence on "sense of agency" – a conscious experience of ones capacity to impact the external world commonly measured as a compression of the perceived time interval between voluntary actions and their consequences (Haggard et al., 2002; Haggard and Cole, 2007). Intriguingly, a recent study showed that the degree of temporal compression increased with the number of available actions such that, the greater the number of action alternatives, the smaller the perceived temporal interval (Barlas and Obhi, 2013). A distinct possibility is that not only the number of available actions but also the divergence of their outcome distributions plays a role in this effect. Notably, since schizophrenic individuals have been shown to have a dysregulated sense of agency (Haggard et al., 2003; Voss et al., 2010), the influence of outcome divergence on this measure may prove to be a useful diagnostic tool in the early detection of thought disorders.

In summary, we have introduced a novel decision variable – instrumental outcome divergence – and demonstrated its influence, dissociable from that of other motivational and information theoretic factors, on behavioral preference. Our results complement previous work on the controllability of outcomes (McClure et al., 2001; Haggard et al., 2003; Teodorescu and Erev, 2014) and contribute towards a fuller characterization of goal-directed cognition and action.

Acknowledgements

This work was supported by a start-up fund from the University of California, Irvine to M.L. The authors thank Daniel McNamee for helpful discussion.

References

Abler B, Herrnberger B, Gron G, Spitzer M (2009) From uncertainty to reward: BOLD characteristics differentiate signaling pathways. BMC neuroscience 10:154.

Ayal S, Zakay D (2009) The perceived diversity heuristic: the case of pseudodiversity. Journal of personality and social psychology 96:559-573.

Balleine BW, Dickinson A (1998) Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacology 37:407-419.

Barlas Z, Obhi SS (2013) Freedom, choice, and the sense of agency. Frontiers in human neuroscience 7:514.

Bickel WK, Pitcock JA, Yi R, Angtuaco EJ (2009) Congruence of BOLD response across intertemporal choice conditions: fictive and real money gains and losses. J Neurosci 29:8839-8846.

Bonardi C, Graham S, Hall G, Mitchell C (2005) Acquired distinctiveness and equivalence in human discrimination

learning: evidence for an attentional process. Psychonomic bulletin & review 12:88-92.
Bowman CH, Turnbull OH (2003) Real versus facsimile reinforcers on the Iowa Gambling Task. Brain and cognition 53:207-210.
Chaminade T, Decety J (2002) Leader or follower? Involvement of the inferior parietal lobule in agency. Neuroreport 13:1975-1978.
Daw ND, Niv Y, Dayan P (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nature neuroscience 8:1704-1711.
den Ouden HM, Frith U, Frith C, S.J. B (2005) Thinking about intentions. NeuroImage 28:787-796.
Doya K, Samejima K, Katagiri K, Kawato M (2002) Multiple model-based reinforcement learning. Neural computation 14:1347-1369.
Erev I, Barron G (2005) On adaptation, maximization, and reinforcement learning among cognitive strategies. Psychological review 112:912-931.
Farrer C, Frey SH, Van Horn JD, Tunik E, Turk D, Inati S, Grafton ST (2008) The angular gyrus computes action awareness representations. Cereb Cortex 18:254-261.
Haggard P, Cole J (2007) Intention, attention and the temporal experience of action. Consciousness and cognition 16:211-220.
Haggard P, Clark S, Kalogeras J (2002) Voluntary action and conscious awareness. Nature neuroscience 5:382-385.
Haggard P, Martin F, Taylor-Clarke M, Jeannerod M, Franck N (2003) Awareness of action in schizophrenia. Neuroreport 14:1081-1085.
Hedesstrom TM, Svedsater H, Garling T (2006) Covariation neglect among novice investors. J Exp Psychol-Appl 12:155-165.
Honey RC, Hall G (1989) Acquired equivalence and distinctiveness of cues. Journal of experimental psychology Animal behavior processes 15:338-346.
Liljeholm M, Balleine BW (2010) Extracting functional equivalence from reversing contingencies. Journal of experimental psychology Animal behavior processes 36:165-171.
Liljeholm M, Tricomi E, O'Doherty JP, Balleine BW (2011) Neural correlates of instrumental contingency learning: differential effects of action-reward conjunction and disjunction. J Neurosci 31:2474-2480.
Liljeholm M, Wang S, Zhang J, O'Doherty JP (2013) Neural correlates of the divergence of instrumental probability distributions. J Neurosci 33:12519-12527.
McClure J, Densley L, Liu JH, Allen M (2001) Constraints on equifinality: goals are good explanations only for controllable outcomes. The British journal of social psychology / the British Psychological Society 40:99-115.
Miyapuram KP, Tobler PN, Gregorios-Pippas L, Schultz W (2012) BOLD responses in reward regions to hypothetical and imaginary monetary rewards. Neuroimage 59:1692-1699.
Seo H, Barraclough DJ, Lee D (2009) Lateral intraparietal cortex and reinforcement learning during a mixed-strategy game. J Neurosci 29:7278-7289.
Sperduti M, Delaveau P, Fossati P, Nadel J (2011) Different brain structures related to self- and external-agency attribution: a brief review and meta-analysis. Brain structure & function 216:151-157.
Teodorescu K, Erev I (2014) Learned helplessness and learned prevalence: exploring the causal relations among perceived controllability, reward prevalence, and exploration. Psychological science 25:1861-1869.
Tolman EC (1948) Cognitive maps in rats and men. Psychological review 55:189-208.
Voss M, Moore J, Hauser M, Gallinat J, Heinz A, Haggard P (2010) Altered awareness of action in schizophrenia: a specific deficit in predicting action consequences. Brain : a journal of neurology 133:3104-3112.
Weber BJ, Huettel SA (2008) The neural substrates of probabilistic and intertemporal decision making. Brain research 1234:104-115.