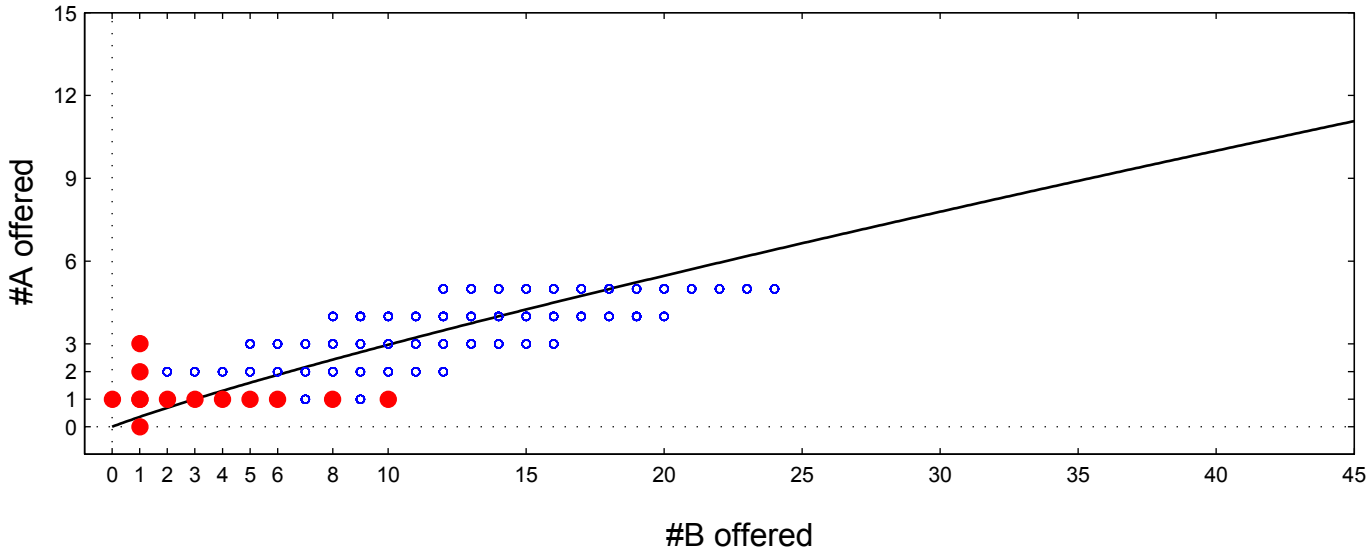

Figure S1. The x-axis and y-axis represent, respectively, the quantities of juices B and A offered in any given trial, and different points in the plane correspond to different offer types. Red dots represent offer types typically employed in our experiments, and the black curve is a hypothetical indifference curve. The indifference curve is non-linear, but close-to-linear within small intervals. In principle, estimating the indifference curve would require testing many different offer types (blues circles). Assuming that the indifference curve is close to linear within small ranges of juice quantity, we can estimate the relative value of the two juices (i.e., the slope of the indifference curve) from a small subset of offer types (red dots).

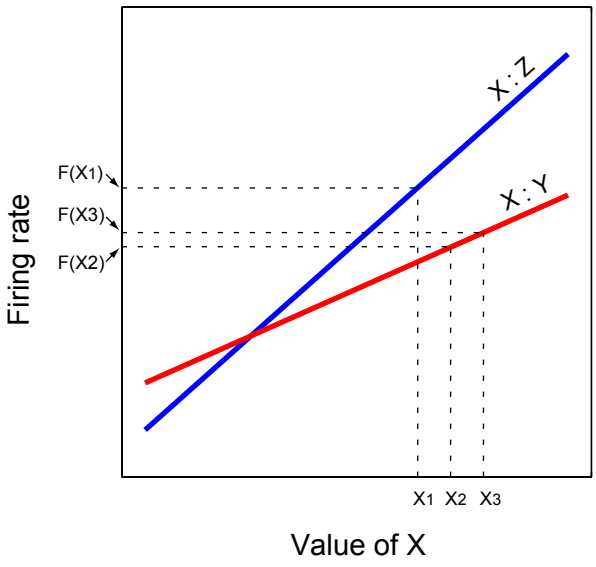

Figure S2. Hypothetical firing rate (y-axis) of a cell encoding the value of good X (x-axis) and such that the response is menu dependent (red when X is offered against Y, blue when X is offered against Z). Transitivity violations (i.e., cases where x-axis orderings are not conserved on the y-axis) are possible, as for X1, X2, X3. However, if the representation is menu invariant (i.e., if red and blue lines coincide), transitivity violations cannot occur. In other words, menu invariance implies transitivity.

The representation of economic value in the orbitofrontal cortex is invariant for changes of menu

By Camillo Padoa-Schioppa and John A. Assad

Supplementary Data

On the assumption of linear indifference curves

The experiments and analysis described here were conducted assuming linear indifference curves. We shall now discuss this assumption in some detail.

Given goods X and Y, the monkey will generally be indifferent between a given quantity q_X of X and another quantity q_Y of Y. For various q_X, the indifference points are described by a curve $q_X = f(q_Y)$. This is called the "indifference curve" (figure S1). If the indifference curve is linear, the function f is a straight line through the origin, and the choice between any quantities q_X and q_Y of X and Y only depends on the ratio q_Y / q_X . Considering three goods X, Y and Z, we will have three indifference curves f, g and h such that $q_X = f(q_Y)$, $q_Y = g(q_Z)$ and $q_X = h(q_Z)$. The behavior of the monkey satisfies indifference transitivity if the following relationship holds true:

$$q_X = h(q_Z) = f(g(q_Z)) \qquad (1)$$

If monkeys choose between three juices A, B and C offered pairwise, estimating the indifference curves f, g and h and their curvatures would in principle require testing many offer types in the planes A:B, B:C and C:A (figure S1). In neurophysiology experiments, this poses a practical problem, because one needs many trials for each offer type to have an accurate estimate of the neuronal firing rate, and monkeys work consistently (i.e., with stable preferences) only for a few hundred trials in each recording session. In designing the experiments, we thus resolved to *assuming* that indifference curves would be linear in the small interval of quantity we tested, and we included in our sessions only offer types such that at least one of the two juices was offered in quantity equal to 1 quantum. Sessions thus lasted 300-600 trials. Under the assumption of linear indifference curves, the relationship of indifference transitivity (Eq.1) reduces to $n_{A:B} * n_{B:C} = n_{A:C}$, where $n_{X:Y}$ is the relative value of juices X and Y (i.e., the slope of the X:Y indifference line) as inferred from the sigmoid fit.

In the language of economics, the relationship between the value the monkey assigns to a juice and the quantity in which that juice is offered is a "utility function." Assuming linear indifference curves is equivalent to assuming that utility functions for different goods are identical to one another up to a scaling factor. Within standard economic theory, where utility functions are defined up to a monotonic transformation, assuming linear indifference curves is the same as assuming linear utility functions. However, from a psychological and physiological point of view, linear indifference curves do not necessarily imply linear value functions. Indeed, values functions for different goods may all have the same, non-linear form (e.g., a particular power law or a log). Here we refer to value functions as psychological constructs.

Indifference curves cannot always be considered linear. For example, a person may prefer 1 chocolate cookie to 1 dollar, and also prefer 1,000 dollars to 1,000 chocolate cookies (i.e., choices may not depend only on the quantity ratio). However, one can in general assume that indifference curves are roughly linear within a small quantity range. Two considerations suggest that assuming linear indifference curves was reasonable in our study. First, relative values remained essentially stable within each session. This suggests that the quantity range over which indifference curves deviated significantly from linearity (i.e., over which the curvatures of different value functions differed appreciably) was at least of the order of the juice volume consumed in one session (hundreds of juice quanta). The maximal juice quantity offered in any given trial (1-10 quanta) was very small compared to such scale. Second, measured relative values did in fact satisfy the relationship $n_{A:B} * n_{B:C} = n_{A:C}$ (figure 2b). In other words, monkeys' choices did not deviate significantly from the prediction based on linear indifference curves.

Variable selection analysis

For this analysis, we consider for each neuron the activity recorded with each juice pair separately, and we submit the neuronal population to the same procedures employed in our previous study[1]. A "trial type" is defined by the offer type and the choice. For example, if the monkey chooses 1A over 3B, the trial type is (3B:1A, 1A). A "response" is the activity of one neuron in one time window as a function of the trial type. Pooling data from different time windows and different juice pairs, a total of 1,660 single-juice-pair responses are significantly modulated by the offer type ($p<0.001$, ANOVA). The variable selection analysis is restricted to this data set.

For each juice pair, we re-name the preferred juice as juice A and the less preferred juice as juice B. We define 19 variables that neuronal responses might

potentially encode: *total value*, *chosen value*, *other value*, *(chosen-other) value*, *(other/chosen) value*, *total number*, *max number*, *chosen number*, *min number*, *other number*, *(max-min) number*, *(chosen-other) number*, *(min/max) number*, *(other/chosen) number*, *offer value A*, *offer value B*, *taste*, *choice value A*, *choice value B*. We also define the "collapsed" variables *offer value* (from *offer value A* and *offer value B*) and *choice value* (from *choice value A* and *choice value B*).[a]

Each response is linearly regressed on each variable. A variable "explains" a response if the regression slope differs significantly from zero ($p<0.05$). To identify variables that best explain the population, we employ two different procedures for variable selection: stepwise and best-subset[1-3]. Replicating our previous findings[1], both procedures identify *offer value*, *chosen value* and *taste* as the most explanatory variables. Collectively, these variables explain 1,295/1,660 (78%) responses (mean $R^2 = 0.65$). A post-hoc analysis indicates that the explanatory power of these variables is significantly higher than that of any challenging alternatives ($p<0.01$), except for *chosen value* versus *total value* ($p=0.15$). However, pooling together the present data set with the previous data set (557 + 931 cells, 1,660 + 1,379 responses), all comparisons are statistically significant, including *chosen value* versus *total value* ($p<10^{-4}$) and *taste* versus *choice value* ($p<10^{-5}$).

On this basis, we classify each single-juice-pair response as encoding one of the variables *offer value A*, *offer value B*, *chosen value*, or *taste*. Responses explained by more than one variable (689/1,295 = 53%) are assigned to the variable with the highest R^2. Responses that do not pass the ANOVA criterion, or that pass it but are not explained by any selected variable, are *unclassified*. The subsequent analysis of classification conflicts is based on this classification.

Analysis of classification conflicts

On the basis of the previous analysis, each single-juice-pair response is either *unclassified*, or it can be explained by at least one of 7 possible variables: *offer value A*, *offer value B*, *offer value C*, *chosen value*, *taste A*, *taste B*, or *taste C*. Given one neuron and one time window, we define an "instance" as a triplet of single-juice-pair responses. Thus there are a total of 3,899 possible instances (557 neurons x 7 time windows). For 760 instances, at least one of the 3 single-juice-pair responses can be explained by at least one variable. Only these 760 instances are included in the analysis of classification conflicts.

We distinguish two possible situations. Cases of classification conflict are instances such that the 3 single-juice-pair responses cannot be explained as encoding the same variable. For example, an instance such that each single-juice-pair response can be explained by only one variable and such that the 3 encoded variables are {*offer value A*, *offer value B*, *offer value A*} presents a classification conflict. Conversely, instances that do not present classification conflicts are cases of no-conflict.

We then proceed to identify cases of classification conflict in our population, with two *caveat*. First, we consider the sign of the encoding. In other words we consider as cases of classification conflict instances in which the cell activity has a positive correlation with the encoded variable when recorded with one juice pair and a negative correlation with the same encoded variable when recorded with another juice pair. Second, we take into consideration the fact that single-juice-pair responses can be explained by more than one variable. In such cases, we identify the instance as a case of classification conflict only if none of the variables can explain the 3 single-juice-pair responses.

Across the population, we observe 157/760 (21%) cases of conflict and 603/760 (79%) cases of no-conflict. To estimate how this measure compares to chance, we perform a bootstrap permutation test. For each instance and for each single-juice-pair response, we re-assign variables that explain the response with a random permutation (across the 7 variables). We thus obtain a new data set, for which we compute the number of classification conflicts. We repeat the operation $N=10^6$ times, and we obtain a random distribution for the number of conflicts expected in the data set by chance. The mean of this random distribution (i.e., the expected number of conflicts) is 320/760 (42%). We then compare the number of conflicts actually present in the data set with the random distribution. This analysis indicates that actual conflicts are significantly fewer than expected by chance ($p<10^{-6}$).

One concern may be whether neuronal responses (in particular, cases of classification conflict) can be explained assuming that the cell activity is determined

[a] In the previous study, variables *offer value A* and *offer value B* were named, respectively, *value A offered* and *value B offered*. Similarly, variables *choice value A* and *choice value B* were named, respectively, *value A chosen* and *value B chosen*. Finally, variable *choice value* was named *value A|B chosen*.

by the preference ranking of the encoded juice as opposed to the identity of the juice. We refer to this hypothesis as neuronal responses "TS-like". One example of TS-like response would be a cell that encodes *offer value A* in A:B trials and C:A trials and *offer value B* in B:C trials. Another example would be a cell that encodes *taste B* in A:B trials and *taste C* in B:C trials and C:A trials. In fact, TS-like responses are very rare in our data set. Across the population, we found only 4/760 (1%) such cases.

It is worth noting that the measure of 21% likely over-estimates classification conflicts, because of a several sources of "noise" in the procedure. For example, we found that imposing a more conservative threshold ($p<0.01$) to establish whether a neuronal response can be explained by a certain variable (which *a priori* should not affect conflicts) results in a lower estimate for classification conflicts (16%). Also, from a statistical point of view, the procedure used to identify conflicts (i.e., analyzing three sub-groups of trials separately, imposing a p-value threshold on each of them, and then comparing the results across sub-groups) is not ideal. A much preferable approach is to combine all trials in a unique analysis, as we do in the analysis of menu dependent encoding (ANCOVA). Nonetheless, the analysis of classification conflicts illustrates two important points. First, even though conflicts are likely over-estimated, they still are many fewer than expected by chance. Second, TS-like responses are very rare.

Statistical power

One possible concern is whether the result showing that OFC responses are typically menu invariant may reflect a poor statistical power due to a limited number of trials. The following analysis suggests that this is not the case.

For each session and each cell, we broke down trials in two groups: those collected in the first half-session and those collected in the second half-session. We thus obtained a "split" population of twice as many "cells," each recorded for half as many trials. If the results were due to poor statistical power, they should differ in this split population compared to the "real" population. In fact, the results are very robust with respect to this manipulation. Referring to the analysis of menu dependent encoding (ANCOVA), significant effects due to the variable are 88% of the total for the real population (table 2) and 89% of the total for the split population. This result suggests that the actual number of trials was adequate.

Appendix: Menu invariance and transitivity

The relationship between transitivity and menu invariance has long been established at the behavioral level for preferences and values[4-6]. Here we provide a heuristic version of that argument, and we show that an analogous relationship holds true for neuronal representations of value.

Behavioral preferences and values

As described in the main text, changes of behavioral context can be conceptualized as follows. Changes of menu are moment-to-moment changes of the options to be chosen between. In contrast, changes of condition are changes on a longer time scale. For example, changes of condition might be induced by changing the range of values offered in the trial block, or may be due to changes of internal motivation (e.g., due to selective satiation). We will restrict this discussion to experiments where changes of behavioral context can be described as changes of menu. In this case, we can assume that preferences remain stable. In other words, indicating with $V_{X:Y}(X)$ the value assigned to good X when it is chosen against good Y, we can assume that for any two goods X and Y, $V_{X:Y}(X)$ remains constant throughout the experiment.

Indicating with $\succ$ the relationship of preference, transitivity is satisfied if for any goods X, Y and Z, conditions $X \succ Y$ and $Y \succ Z$ imply $X \succ Z$. In terms of values, preference transitivity is satisfied if the following implication holds true:

$$V_{X:Y}(X) > V_{Y:X}(Y) \quad \text{and} \quad V_{Y:Z}(Y) > V_{Z:Y}(Z)$$
$$\Rightarrow V_{X:Z}(X) > V_{Z:X}(Z) \qquad (2)$$

That menu invariance implies preference transitivity can be observed as follows. Consider the three inequalities $V_{X:Y}(X) > V_{Y:X}(Y)$, $V_{Y:Z}(Y) > V_{Z:Y}(Z)$ and $V_{X:Z}(X) < V_{Z:X}(Z)$, which collectively violate Eq.2. It is easy to see that the three inequalities cannot all hold true if we ignore the subscripts. In other words, transitivity cannot be violated if values are menu invariant. In contrast, the three inequalities can be satisfied if values depend on the menu, for example if $V_{Y:X}(Y) = V_{Y:Z}(Y)$, $V_{Z:Y}(Z) = V_{Z:X}(Z)$ and $V_{X:Y}(X) < V_{X:Z}(X)$. In summary, transitivity violations imply menu dependent values. Equivalently, menu invariant values guarantee preference transitivity.[b]

[b] The converse is not necessarily true. In principle, preference transitivity could be satisfied even if values are menu dependent.

Neuronal populations and choice

There is one scenario in which the relationship between neuronal menu invariance and preference transitivity would be trivial. Imagine identifying a population of neurons that encode economic value, and further finding that there is a causal relationship between the activity of this neuronal population and behavioral choice (i.e., such that choices are completely determined by the activity of the population). If this is true, and if the population represents value in a menu invariant way, behaviorally measured values must also be menu invariant. Consequently, preferences must be transitive. In other words, assuming a causal relationship between a given neuronal representation of value and economic choice, neuronal menu invariance implies preference transitivity. Unfortunately, whether such causal relationship holds true for value-encoding neurons in OFC is unknown. In the following, we thus establish a link between neuronal menu invariance and transitivity *without* assuming such a causal relationship.

Neuronal representations of value

We will restrict the argument to experiments for which changes of behavioral context can be described as changes of menu (as opposed to changes of condition), and we will assume that behaviorally measured values are menu invariant.

Consider an experiment in which subjects make economic choices between various options. We say that a neuronal population provides a "neuronal representation of value" if the population encodes in each trial the value of each available option. According to this broad definition, neuronal representations of value exist in orbitofrontal cortex (OFC)[1], in the lateral intraparietal area (LIP)[7-9], and possibly in other areas.[c]

Because behaviorally measured values are menu invariant, we can indicate with V(X) (without subscript) the value assigned to good X. As noted in the main text, transitive values establish a common value scale. Consequently, we can use for given goods X, Y and Z the compact notation $V(X) < V(Y) < V(Z)$. (This notation is meaningful if and only if transitivity holds true.) Notably, even if behaviorally measured values are menu invariant, a neuronal representation of value might in principle depend on the menu.

Given one neuron *n* encoding the value of good Z, we indicate with ${}_nF_{X:Y}(Z)$ the firing rate of *n* recorded when the monkey chooses X over Y.[d] We indicate with $F_{X:Y}$ the activity of the neuronal population recorded when the monkey chooses X over Y. The neuronal representation of value is "stable" if for any X and Y, the activity $F_{X:Y}$ recorded in trials in which the monkey chooses X over Y remains constant throughout the experiments.

The distinction between changes of menu and changes of condition described at the behavioral level is also relevant at the neuronal level. For example, a change of condition could be due to a global change in responsiveness (e.g., due to a change in the animal's or to a systemic pharmacological manipulation) that would uniformly increase or reduce the activity of a population of value-encoding neurons while leaving preferences and behaviorally measured values unchanged. Such systemic change, however, would not leave stable the neuronal representation of value. Limiting our discussion to experimental manipulations that affect the menu but not the condition, we can exclude this possible scenario. We can thus assume that the neuronal representation of value is stable.[e]

Transitive values establish a common value scale. Intuitively, a neuronal representation of value "reflects transitivity" if values are encoded in that common scale. More formally, we say that a neuronal representation of value "reflects transitivity" if, for any neuron *n* encoding the value of good X and for any three quantities X_1, X_2, X_3 of X, the following implication holds true:

$$V(X_1) > V(X_2) > V(X_3) \;\Rightarrow$$
$$F_{X:*}(X_1) > F_{X:*}(X_2) > F_{X:*}(X_3) \qquad (3)$$

In other words, a neuronal representation of value reflects transitivity if it maintains value orderings. The

[c] This condition corresponds to the hypothesis that economic choice entails assigning values to the available options. What exactly constitutes an "option" depends on the representation. For OFC, options are goods (juice types), independently of the visuomotor contingencies of choice. For LIP, options are visual stimuli or saccades associated with a desirable outcome. In the following, we interchangeably use "good" and "option", thus implicitly referring to OFC. However, the argument identically applies to the representation of value in LIP.

[d] In the expression ${}_nF_{X:Y}(Z)$, Z is the encoded good, while X and Y are the offered goods. In trials where the monkey is offered the good encoded by neuron *n*, either X or Y is equal to Z. We distinguish between ${}_nF_{X:Y}$ and ${}_nF_{Y:X}$ because the representation may encode variables that depend on the choice.

[e] In the experiments described here, the OFC representation of value can be considered stable.

subscript "X : ∗" indicates that Eq.3 must hold true for any offered good.[f]

To illustrate that a neuronal representation of value reflects transitivity if it is menu invariant, we can proceed as in the previous section. Consider the 3 inequalities $F_{X:Y}(X_1) > F_{X:Y}(X_2)$, $F_{X:Y}(X_2) < F_{X:Z}(X_3)$ and $F_{X:Z}(X_3) > F_{X:Y}(X_1)$, which collectively violate Eq.3 (see figure S2). It is easy to see that the three inequalities cannot all hold true if we ignore the subscripts. In other words, transitivity cannot be violated if values are menu invariant. In contrast, the three inequalities can be satisfied if values depend on the menu, as in the case shown in figure S2. In summary, if the representation violates transitivity it is menu dependent. Equivalently, if a representation of value is menu invariant, it reflects transitivity.[g]

References

1. Padoa-Schioppa, C. & Assad, J.A. Neurons in orbitofrontal cortex encode economic value. *Nature* **441**, 223-6 (2006).
2. Dunn, O.J. & Clark, V. *Applied statistics: analysis of variance and regression*, xii, 445 p. (Wiley, New York, 1987).
3. Glantz, S.A. & Slinker, B.K. *Primer of applied regression & analysis of variance*, xxvii, 949 p. (McGraw-Hill, Medical Pub. Division, New York, 2001).
4. Grace, R.C. Violations of transitivity: Implications for a theory of contextual choice. *J Exp Anal Behav* **60**, 185-201 (1993).
5. Shafir, S. Intransitivity of preferences in honey bees: support for 'comparative' evaluation of foraging options. *Anim Behav* **48**, 55-67 (1994).
6. Tversky, A. & Simonson, I. Context-dependent preferences. *Management Sciences* **39**, 117-185 (1993).
7. Glimcher, P.W., Dorris, M.C. & Bayer, H.M. Physiological utility theory and the neuroeconomics of choice. *Games Econ Behav* **52**, 213-256 (2005).
8. Sugrue, L.P., Corrado, G.S. & Newsome, W.T. Matching behavior and the representation of value in the parietal cortex. *Science* **304**, 1782-7 (2004).
9. Dorris, M.C. & Glimcher, P.W. Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. *Neuron* **44**, 365-78 (2004).

[f] To account for cases in which encoded variable and firing rate are negatively correlated, this definition can be generalized saying that the representation of value reflect transitivity if the relationship $V(X_1) > V(X_2) > V(X_3)$ implies either $F_{X:*}(X_1) > F_{X:*}(X_2) > F_{X:*}(X_3)$ or $F_{X:*}(X_1) < F_{X:*}(X_2) < F_{X:*}(X_3)$.

[g] According to our definition, a neuronal representation of value reflects transitivity if Eq.3 is satisfied. This definition implies that menu invariance is not only sufficient, but also a necessary condition for a neuronal representation of value to reflect transitivity (see figure S2). However, alternative definitions are possible. For example, a neuronal representation of value causally linked to choice could be said to reflect transitivity if it generates transitive preferences. In such case, the representation could reflect transitivity but also be menu dependent. As a consequence, the representation of value in LIP does not necessarily violate transitivity. Similarly, a neuronal representation of value that depends on the behavioral condition does not necessarily violate transitivity.