Asymmetric Paternalism: Description of the Phenomenon, Explanation Based on Decisions Under Uncertainty, and Possible Applications to Education

Olga Kosheleva
Department of Teacher Education
University of Texas at El Paso
El Paso, TX 79968
Email: olgak@utep.edu

François Modave
Department of Computer Science
University of Texas at El Paso
El Paso, TX 79968
Email: fmodave@utep.edu

***Abstract*—In general, human being are rational decision makers, but in many situations, they exhibit unexplained "inertia", reluctance to switch to a better decision. In this paper, we show that this seemingly irrational behavior can be explained if we take uncertainty into account; we also explain how this phenomenon can be utilized in education.**

I. Traditional Approach to Human Decision Making: A Brief Reminder

In the traditional approach to decision making (see, e.g., [4], [13]), the decision maker's preferences A_1, ..., A_n can be characterized by their "utility values" $u(A_1)$, ..., $u(A_n)$, so that an alternative A_i is preferable to the alternative A_j if and only if $u(A_i) > u(A_j)$.

II. Empirical Testing of the Traditional Approach to Decision Making is not Easy

The traditional approach to decision making is a theoretical description of human behavior. Since its appearance, researchers have been testing to what extent this approach adequately describe the actual behavior of human decision makers.

Such a testing is not easy, since the traditional approach relates an empirically "testable" behavior (such as preferring one alternative A_i to another alternative A_j) with the difficult-to-test comparison between the (usually unknown) utility values.

III. A Testable Consequence of the Traditional Approach to Decision Making

Although a direct test of the traditional approach to decision making is not easy, some testable consequences of the traditional approach can be derived.

For example, in the traditional approach, unless the two alternatives A_i and A_j have the exact same utility value $u(A_i) = u(A_j)$, we have two possibilities:

- either $u(A_i) > u(A_j)$, i.e., the alternative A_i is better,
- or $u(A_j) > u(A_i)$, i.e., the alternative A_j is better.

In the first case,

- if we originally only had an alternative A_i, and then we are adding the alternative A_j, then we stick with A_i;
- on the other hand, if we originally only had an alternative A_j, and then we are adding the alternative A_i, then we switch our choice to A_i.

Similarly, in the second case,

- if we originally only had an alternative A_j, and then we are adding the alternative A_i, then we stick with A_j;
- on the other hand, if we originally only had an alternative A_i, and then we are adding the alternative A_j, then we switch our choice to A_j.

These two cases can be summarized in the following 2-stage experiment: In the first stage,

- first, we present the decision maker with only one alternative A_i; the decision maker does not have a choice, so he or she selects this alternative A_i;
- after that, we provide the decision maker with another possible alternative A_j; so now the decision maker has two alternatives A_i and A_j.

We then record the user's choice in this first stage.

After some amount of time, we start with the second stage of our experiment:

- first, we present the decision maker with only one alternative A_j; the decision maker does not have a choice, so he or she selects this alternative A_j;
- after that, we provide the user with another possible alternative A_i, so now the decision maker has two alternatives A_i and A_j.

We then record the user's choice in this second stage.

According to the above-described traditional approach to decision making, on both stages, the decision maker should make the same choice:

- if the alternative A_i has a larger utility value, then on both stages, the decision maker should chose A_i;

- on the other hand, if the alternative A_j has a larger utility value, then on both stages, the decision maker should chose A_j.

IV. The Above Testable Consequence is in Perfect Agreement with Common Sense

The above behavior not only follows from the mathematics of the traditional decision making, it is also in perfect agreement with common sense.

Indeed, unless the two alternatives A_i and A_j are absolutely equivalent for the decision maker (which happens very rarely), either the first one is better or the second one is better.

So, on both stages of the above experiment, a rational decision maker should make the same choice:

- if to this decision maker, the alternative A_i is preferable to the alternative A_j, then on both stages, the decision maker should chose A_i;
- on the other hand, if to this decision maker, the alternative A_j is preferable to the alternative A_i, then on both stages, the decision maker should chose A_j.

V. For Close Alternatives, Decision Makers Do Not Behave in This Rational Fashion

Interestingly, in the actual tests of the above experiment, human decision makers do not follow this seemingly rational behavior; see, e.g., [3], [6], [7], [12]. Specifically, they exhibit "inertia", the desire not to change an alternative.

Namely, if the alternatives are close in value, then the decision makers exhibit the following behavior in our two-stage experiment. In the first stage,

- first, we present the decision maker with only one alternative A_i; the decision maker does not have a choice, so he or she selects this alternative A_i;
- after that, we provide the decision maker with another possible alternative A_j; so now the decision maker has two alternatives A_i and A_j.

On this stage, most decision makers continue to stick to the original choice A_i.

After some amount of time, we perform the second stage of our experiment:

- first, we present the decision maker with only one alternative A_j; the decision maker does not have a choice, so he or she selects this alternative A_j;
- after that, we provide the user with another possible alternative A_i, so now the decision maker has two alternatives A_i and A_j.

On this stage, most decision makers continue to stick to the original choice A_j.

An example where such seemingly irrational behavior occurred is the employees' choice between two financial retirement plans A_i and A_j.

- Originally, the employees had only one option: retirement plan A_i.
- After this, an additional option A_j is introduced.
- In spite of this new option, most employees decided to keep the old option A_i.

This, by itself, is not inconsistent with rational behavior: we can simply conclude that for most employees, the original option A_i is better than the new option A_j.

However, at the same time, a very similar group of employees was presented with a different scenario:

- Originally, the employees had only one option: retirement plan A_j.
- After this, an additional option A_i is introduced.
- In spite of this new option, most employees decided to keep the old option A_j.

Since we concluded that for most employees, the option A_i is better than the option A_j, we would expect most employees from this second group to switch from the original option A_j to the new option A_i. However, in reality, most employees from this second group stayed with their original option A_j.

In behavioral economics, this "inertial" behavior is called *present-biased preferences*: whatever options we have selected at present biases our future choices.

VI. Maybe Human Behavior Is Irrational?

How can we explain this seemingly irrational behavior? One possible explanation is that many people do often make bad (irrational) decisions: waste money on gambling, waste one's health or alcohol and drugs, etc.

However, the above inertial behavior occurs not only among decision makers who exhibit self-destructive irrational behavior, it is a common phenomenon which occurs among the most successful people as well.

It is therefore reasonable to look for an explanation of this seemingly irrational behavior. It turns out that we can come up with such an explanation if we take into account uncertainty related to decision making.

VII. How to Take Into Account Uncertainty in Decision Making Situations

Each alternative decision can lead to different possible situations. For example, a decision about selecting a financial retirement plan can lead to different amounts available to the decision maker by the moment of his or her retirement:

- A more conservative retirement plan – e.g., investing all the retirement money in the government-guaranteed bonds – will not lead to a large increase of the invested amount, but, on the positive side, has a smaller probability of losing the retirement money.
- On the other hand, a riskier retirement plan – e.g., investing all the retirement money into stocks – has a larger probability of failing, but it can also lead to a much larger amount of money available for retirement.

In the traditional approach to decision making, we first estimate the utilities $U_1, \ldots, U_m$ of different possible consequences $c_1, \ldots, c_m$ of our actions – e.g., the utilities of having different amounts of money by the time of the retirement –

and then estimate the utility of each alternative decision as the expected value of this utility:

$$u(A) = p(c_1 \,|\, A) \cdot U_1 + \ldots + p(c_m \,|\, A) \cdot U_m,$$

where $p(c_k \,|\, A)$ is the conditional probability of the consequence c_k under the condition that we select an alternative A.

In real life, we do not know the exact values of these probabilities $p(c_k \,|\, A)$, we only know them with uncertainty. Usually, we only know the approximate estimates of these probabilities. In some cases, we have bounds

$$\underline{p}(c_k \,|\, A) \le p(c_k \,|\, A) \le \overline{p}(c_k \,|\, A)$$

on these probabilities, i.e., we know intervals $[\underline{p}(c_k \,|\, A), \overline{p}(c_k \,|\, A)]$ that contain the (unknown) probabilities $p(c_k \,|\, A)$. In such situations, for each alternative A, instead of a single value $u(A)$, we get an interval of possible values:

$$[\underline{u}(A), \overline{u}(A)] =$$

$$[\underline{p}(c_1 \,|\, A), \overline{p}(c_1 \,|\, A)] \cdot U_1 + \ldots + [\underline{p}(c_m \,|\, A), \overline{p}(c_m \,|\, A)] \cdot U_m.$$

In other situations, we have expert estimates of the unknown probabilities $p(c_k \,|\, A)$. Such expert estimates can be naturally described by fuzzy numbers. In this case, the resulting utility estimate $u(A)$ is also a fuzzy number.

Comment. In effect, fuzziness means that instead of single pair of bounds for the unknown probability (and, as a result, for the utility $u(A)$), we can provide different bounds which are valid with different degrees of confidence.

In other words, for each such quantity (probability or utility), instead of a single interval, we get a nested family of confidence intervals corresponding to different levels of uncertainty. Nested families are, in effect, equivalent to fuzzy numbers; see, e.g., [2], [5], [9], [11], so this natural idea of representing uncertainty is indeed mathematically equivalent to using fuzzy numbers.

VIII. Uncertainty Explains Present-Biased Preferences

When the approximate estimates $\widetilde{u}_i$ and $\widetilde{u}_j$ for the (unknown) utilities $u(A_i)$ and $u(A_j)$ are close, this, due to the uncertainty, means that it is quite possible that the actual values $u(A_i)$ and $u(A_j)$ are equal. It is also possible that $u(A_i) > u(A_j)$ and it is also possible that $u(A_j) > u(A_i)$.

Let us illustrate these possibilities on a simple example where every estimate has the exact same accuracy ε. In other words, we know that $|u(A_i) - \widetilde{u}_i| \le \varepsilon$ and that $|u(A_j) - \widetilde{u}_j| \le \varepsilon$. In this case, based on the estimate $\widetilde{u}_i$ for $u(A_i)$, the only information that we have about the actual (unknown) value of the utility $u(A_i)$ is that this value belongs to the interval $\mathbf{u}_i \stackrel{\text{def}}{=} [\widetilde{u}_i - \varepsilon, \widetilde{u}_i + \varepsilon]$. Similarly, based on the estimate $\widetilde{u}_j$ for $u(A_j)$, the only information that we have about the actual (unknown) value of the utility $u(A_j)$ is that this value belongs to the interval $\mathbf{u}_j \stackrel{\text{def}}{=} [\widetilde{u}_j - \varepsilon, \widetilde{u}_j + \varepsilon]$.

When the estimates $\widetilde{u}_i$ and $\widetilde{u}_j$ are close – to be more precise, when $|\widetilde{u}_i - \widetilde{u}_j| < 2\varepsilon$ – the intervals $\mathbf{u}_i$ and $\mathbf{u}_j$ intersect.

These intervals represent the set of possible values for, correspondingly, $u(A_i)$ and $u(A_j)$. Thus, the fact that these intervals intersect means that it is possible that the same real number is a value of both $u(A_i)$ and of $u(A_j)$ – i.e., that the corresponding utilities may be equal. Similarly, we can show that in this case:

- there are values $u_i \in \mathbf{u}_i$ and $u_j \in \mathbf{u}_j$ for which $u_i < u_j$, and
- there are also values $u_i' \in \mathbf{u}_i$ and $u_j' \in \mathbf{u}_j$ for which $u_i' > u_j'$.

In other words, when the estimates $\widetilde{u}_i$ and $\widetilde{u}_j$ are close, all three situations are possible:

- it is possible that the alternatives A_i and A_j have exactly the same utility to the decision maker;
- it is possible that for this decision maker, the alternative A_i leads to better results than A_j;
- it is also possible that for this decision maker, the alternative A_j leads to better results than A_i.

Switching to a different alternative usually has a cost, a small but still a cost. For example, in the case of a financial retirement plan, there is a trader's charge for selling stocks and for buying government bonds (and vice versa). Thus, it only makes sense to perform this switch if we are reasonably sure that switching will indeed lead to a better alternative – i.e., in utility terms, to the larger value of utility.

If the utility estimates are close, i.e., if $|\widetilde{u}_i - \widetilde{u}_j| < 2\varepsilon$, we have no guarantee that the new alternative is indeed better than the previously selected one. In this case, it is prudent to stick to the original choice – exactly as actual decision makers are doing.

Comment. On the other hand, if one of the estimates is much larger than the other, it makes sense to switch.

Specifically, if $\widetilde{u}_i > \widetilde{u}_j + 2\varepsilon$, then every value

$$u(A_i) \in \mathbf{u}_i = [\widetilde{u}_i - \varepsilon, \widetilde{u}_i + \varepsilon]$$

is larger than every value

$$u(A_j) \in \mathbf{u}_j = [\widetilde{u}_j - \varepsilon, \widetilde{u}_j + \varepsilon].$$

In this case, we are guaranteed that $u(A_i) > u(A_j)$. Thus, if our original choice was the worse alternative A_j, it makes sense to switch to a better alternative A_i.

IX. Analogy with Interval-Valued Control of a Mobile Robot

The rationality is inertia under uncertainty can be illustrated on the example of a similar situation: how an intelligent mobile robot makes decisions about its motion.

In the traditional control, we make decisions based on the current values of the quantities. For example, when controlling a mobile robot, we make decisions about changing its trajectory based on the moment-by-moment measurements of this robot's location and/or velocity. Measurements are never 100% accurate; the resulting measurement noise leads to random deviations of the robot from the ideal trajectory – shaking and "wobbling". Since each change in direction requires that

energy from the robot's battery go to the robot's motor, this wobbling drains the batteries and slows down the robot's motion.

A natural way to avoid this wobbling is to change a direction only if it is absolutely clear (beyond the measurement uncertainty) that this change will improve the robot's performance. The idea was one of the several interval-related ideas that in 1997, led our university robotic team to the 1st place in the robotic competitions organized by the American Association for Artificial Intelligence (AAAI); see, e.g., [8]. A similar idea also improves the motion of interval-valued fuzzy control; see, e.g., [10], [15], [16].

X. Asymmetric Paternalism: Practical Application of Present-Biased Preferences

At first glance, one may think that the above explanation is of purely theoretical value: OK, we explained how people actually make decisions. How does that help in practice?

The reason why we got interested in coming up with this explanation is that the phenomenon of present-biases preferences is actually actively used in practice. This use is called *asymmetric paternalism*; see, e.g., [1], [6], [14]. Let us explain how this application works.

Suppose that we have two types of behavior, one slightly worse for an individual, and one slightly better. For example, when thirsty, a kid can drink either a healthy fruit juice or a soda drink which has no health value. Our intent is to enforce the healthy alternative.

Traditional paternalism literally enforces the healthy choice by prohibiting all other choices. Alas, practice has shown that in many cases, this literal enforcement does not work.

It turns out that much better results can be achieved if we at first provide only the desired alternative – and then gradually introduce all the other alternatives. For example, we have only healthy drinks for the first few weeks of a school orientation, but then we allow all the choices. Due to the present-biased preferences, kids will tend to stick to their original healthier choice without the need for strict (and non-working) enforcement. Experience shows that this approach really works; see, e.g., [1], [14].

The same strategy works even better for adults – whom we cannot legally enforce into healthy choices.

XI. How Does Our Explanation Help?

If this approach works, do we need any explanation to make it work? Well, sometimes this approach works, and sometimes it does not. As of now, empirical attempts were the only way to check whether this approach will work for given alternatives A_i and A_j.

Suppose that we want to enforce A_i as opposed to A_j by introducing A_i first. How can we tell beforehand whether the decision maker will stick to A_i and not switch to A_j?

Our explanation provides an answer to this question: if $[\underline{u}_i, \overline{u}_i]$ is the interval of possible values of $u(A_i)$ and $[\underline{u}_j, \overline{u}_j]$ is the interval of possible values of $u(A_j)$, then the decision maker sticks with the original choice of A_i if and only if there is no guarantee that the new choice A_j is better. In mathematical terms, this means that the smallest possible value $\underline{u}_j$ corresponding to the new choice does not exceed the largest possible value $\overline{u}_i$ corresponding to the original choice:

$$\underline{u}_j \le \overline{u}_i.$$

For fuzzy numbers, we can get a similar answer for "not switching with a given confidence", if we similarly compare the intervals (α-cuts) for $u(A_i)$ and $u(A_j)$ corresponding to this given confidence level.

XII. Potential Applications to Education

As of now, the asymmetric paternalism techniques have been used in economic and medical situations [1], [6], [14]. In our opinion, this phenomenon can also be efficiently applied to education.

For example, it is well known that when the students just come to class from recess or from home, it is difficult to get their attention. On the other hand, once they get engaged in the class material, it is difficult for them to stop when the bell rings. To take advantage of this phenomenon, it is desirable to start a class with engaging fun material; once the students got into the studying state A_i they will (hopefully) remain in this state even when a somewhat less fun necessary material is presented (and which provides them with a possibility to switch to a passive state A_j).

Acknowledgments

The authors are thankful to the anonymous referees for valuable suggestions.

References

[1] C. Cammerer, S. Issacharoff, G. Loewenstein, T. O'Donoghue, and M. Rabin, "Regulation for conservatives: behavioral economics and the case for 'asymmetric paternalism'", *University of Pennsylvania Law Review*, 2003, Vol. 151, No. 3, pp. 1211–1254.

[2] D. Dubois and H. Prade, Operations on fuzzy numbers, *International Journal of Systems Science*, 1978, Vol. 9, pp. 613–626.

[3] E. J. Johnson, J. Hershey, J. Meszaros, and H. Kunreuther, "Framing, probability distortions, and insuarnce decisions", *J. Risk Uncertainty*, 1993, Vol. 7, pp. 35–53.

[4] R. L. Keeney and H. Raiffa, *Decisions with Multiple Objectives*, John Wiley and Sons, New York, 1976.

[5] G. Klir and B. Yuan, *Fuzzy sets and fuzzy logic*, Prentice Hall, New Jersey, 1995.

[6] G. Loewenstein, T. Brennan, and K. G. Volpp, "Asymmetric paternalism to improve health behavior", *Journal of American Medical Association (JAMA)*, 2007, Vol. 298, pp. 2415–2417.

[7] B. C. Madrian and D. F. Shea, "The power of suggestion: inertia in 401(k) participation and savings behavior", *Quarterly J. Economics*, 2001, Vol. 116, No. 4, pp. 1149–1187.

[8] D. Morales and Tran Cao Son, "Interval Methods in Robot Navigation", *Reliable Computing*, 1998, Vol. 4, No. 1, pp. 55–61.

[9] H. T. Nguyen and V. Kreinovich, "Nested Intervals and Sets: Concepts, Relations to Fuzzy Sets, and Applications", In: R. B. Kearfott and V. Kreinovich, eds., *Applications of Interval Computations*, Kluwer, Dordrecht, 1996, pp. 245–290.

[10] H. T. Nguyen, V. Kreinovich, and Q. Zuo, "Interval-valued degrees of belief: applications of interval computations to expert systems and intelligent control", *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems (IJUFKS)*, 1997, Vol. 5, No. 3, pp. 317–358.

[11] H. T. Nguyen and E. A. Walker, *A First Course in Fuzzy Logic*, CRC Press, Boca Raton, Florida, 2006.

[12] T. O'Donoghue and M. Rabin, "Doing it now or later", *American Economic Rev.*, 2005, Vol. 97, pp. 31–46.

[13] H. Raiffa, *Decision Analysis*, Addison-Wesley, Reading, Massachusetts, 1970.

[14] R. H. Thaler and C. R. Sunstein, "Libertarian paternalism", *American Economic Rev.*, 2003, Vol. 93, No. 2, pp. 175–179.

[15] K. C. Wu, "A robot must be better than a human driver: an application of fuzzy intervals", In: L. Hall, H. Ying, R. Langari, and J. Yen (eds.), *NAFIPS/IFIS/NASA'94, Proceedings of the First International Joint Conference of The North American Fuzzy Information Processing Society Biannual Conference, The Industrial Fuzzy Control and Intelligent Systems Conference, and The NASA Joint Technology Workshop on Neural Networks and Fuzzy Logic*, San Antonio, December 18–21, 1994, pp. 171–174.

[16] K. C. Wu, "Fuzzy interval control of mobile robots", *Computers and Electrical Engineering*, 1996, Vol. 22, No. 3, pp. 211–219.