Journal of Marketing Behavior, 2015, 1: 283–291

A Welfarist Approach to Manipulation

Jonathan Baron*

University of Pennsylvania, USA

ABSTRACT

I argue for two modifications in Sunstein's definition of manipulation, designed to make the definition more compatible with a welfarist/utilitarian view of what manipulation is and why we should care about it. I think we need a clearer distinction between good and bad manipulation, and a distinction between good and bad reflection. Good reflection is actively open-minded and serves to achieve the decision maker's goals. Manipulation is designed to prevent such reflection, or prevent the choice that would be made if such reflection occurred, and this is why it is bad, even if it is consistent with reflection that merely bolsters a favored option.

In his article, Sunstein (2015) discusses manipulation from both a deontolotical and welfarist point of view, and attempts to define manipulation in terms of opportunity for reflective deliberation. Here I would like to revise the welfarist definition. I write as a card-carrying utilitarian (a form of welfarist), and a psychologist who is skeptical about the particular version of the two-system theory that Sunstein assumes. I found the many examples to be useful in trying to refine my own thinking, so I will discuss some of them.

In essence, I am led to suggest that we could define manipulation in terms of intentional deception of a certain sort. I also try to draw a clear distinction between "benign manipulation" or even "beneficial manipulation" and "harmful manipulation." Much of the purpose of this discussion is about whether manipulation should be constrained. From a utilitarian point of view, we have no reason to restrain it unless it is harmful. I suggest that the term by itself allows the existence of beneficial manipulation. Manipulation thus has two dimensions, deceit and outcome valence. It should be the combination of harmful outcomes and deceit that bothers us and calls for some sort of control. Here I try to show what is bad about manipulation when it is bad.

*Professor of Psychology, Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104, USA, baron@psych.upenn.edu.

ISSN 2326-568X; DOI 10.1561/107.00000018

The problem of defining a term like "manipulation" need not have a clear solution. We begin with a common language term and try to examine its meaning. But we also want to use it in a technical sense, in law or psychology. The technical sense may require a simpler definition, which does not do justice to all the uses of the term in common language. I try to do some of this simplification here, but not so much as to make the technical term lose its relation to common language completely. (Arguably, this has happened to the term "utility" as used in "utility theory.")

Autonomy and Dignity

The deontological approach to manipulation says that it harms autonomy and dignity. From a utilitarian perspective, autonomy does matter, but not because it is fundamental. Rather, allowing people to make decisions about what affects them serves two purposes. First, as Sunstein notes, people usually know more than others know about which options are best for them. When decisions affect primarily the decision makers, then we should let the decision makers decide when they can achieve their own goals better this way.

In some cases, as Sunstein notes, we should not let people decide. First, obviously, we cannot let them do what is best for them at the expense of greater harm to others. I would not mention this except that the point seems lost on many libertarians, who think, for example, that the law should not compel parents to vaccinate their children against highly contagious diseases.

Second, some people may lack the capacity to make good decisions for themselves, either because of cognitive limitations or ignorance that is not easily remedied with a little information. Children are the obvious example, providing the origin of the term "paternalism" (or "parentalism").

Third, people may lack self-control. Then may be unable to make themselves do what they want to do. People may be better off by their own standards if we make it difficult for them to do things impulsively that they will later regret, such as driving drunk or getting divorced after a single squabble.

In sum, it may be worthwhile to think about manipulation in terms of its benefits and harms, given that its effect on autonomy must be unpacked anyway. For a utilitarian, autonomy violation is a signal that a manipulation should be evaluated, but it is not an ultimate criterion of whether it should be restricted.

As for "dignity," I am not sure what it is. Some examples of dignity violations seem to me to be simply harms, such as making people feel inferior unnecessarily. It adds nothing to say that this is a dignity violation. It is bad anyway. Other examples, such as limiting the wearing of religious symbols in public, or candid photography, seem to me to depend heavily on cultural norms that are variable and labile.

Reflection and Intuition

Sunstein wants to define manipulation in terms of the opportunity for reflection and deliberative choice. This distinction relies on a psychological distinction between two systems, popularized by Kahneman (2011). System 1 is intuitive; System 2 is reflective. Decision making often uses the intuitive system only. Choices are immediate and made without thought. The reflective system sometimes intervenes before the final choice, sometimes overriding the intuitive option. A common psychology test in which this occurs is the Cognitive Reflection Test (CRT; Frederick 2005). One of the three items is: "A bat and a ball cost $1.10 in total. The bat costs $1 more than the ball. How much does the ball cost?" The intuitive response is the result of subtracting $1 from $1.10: 10 cents. A subject who then reflects on this might (or might not) check the answer to see if it meets the conditions, and, finding that it does not, try something else. Kahneman argues that, in general, the intuitive system is affected by various cognitive biases, which are sometimes corrected by the reflective system. Other theorists have loaded up this story with various additional assumptions, such as the claim that emotions largely affect the intuitive system.

Importantly, the intervention of the reflective system is what helps to overcome biases. Sunstein proposes, in essence, that manipulation prevents the engagement of the reflective system, leaving the choice to be determined by uncorrected biases.

The two-system theory, however, is oversimplified. For one thing, it is sometimes difficult to make the distinction between intuition and reflection, and many episodes of reflection do not start out with any intuition at all. Many people do not have an initial intuition even in CRT problems; they treat each problem as if it were a math problem on a test. One student doing the CRT for the first time in my office, did not even consider the intuitive answer but began right away with "Let X equal the price of the ball." On other problems, such as moral dilemmas, many people begin their thinking with two conflicting intuitions (Baron *et al.* 2015).

But the main issue here is the effectiveness of the reflective system. Engagement of the reflective system is a correlate of what matters, but, in my view, it is not the same as what matters. The reflective system presumably leads to decisions that are more consistent with the decision maker's goals. Insofar as the two-system distinction can be made at all, that is arguably true. But the effect is statistical at best. Even in the bat-and-ball problem, many people get the incorrect answer after reflection (Meyer *et al.* 2013).

We may think of reflection (a.k.a. "thinking") as a search process (Baron 2008) in which people search for evidence (reasons, arguments) for or against various options they are considering. They also search for alternative options, and for goals (values, criteria) that might be relevant.

Reflection alone may end up reinforcing intuitive conditions rather than questioning them. A classic demonstration is that of Wason and Evans (1975), who gave subjects a surprisingly difficult problem:

> Each of 4 cards has a letter on one side and a number on the other side. You see one side of each card: B U 3 6. Which cards must you turn over to find out if the following rule is true of all four cards: "If there is a B on one side, then there will be a 3 on the other side."
>
> — (Wason and Evans 1975)

Most subjects gave the answer that seemed intuitive, matching the cards with those in the rule: B and 3. When they were asked to justify this incorrect answer, few changed their minds. Most gave elaborate but nonsensical responses. But when the rule was changed so that it said "then there will not be a three on the other side," they continued to give the matching response of B and 3. This was now correct, and they gave correct justifications when asked to provide them. For both rules, most subjects used their thought to justify their initial choice.

Another type of evidence comes from studies of pre-decisional distortion (e.g., Russo *et al.* 2006). Once people develop a small preference for an option, the strength of that preference, regardless of the option, tends to increase as they think about it, sometimes leading to the choice of an option that would obviously be inferior. Still another relevant result is "belief overkill" (e.g., Baron 2009). Once people form an opinion about some issue, they recruit additional reasons favoring that opinion, so that all arguments point in the same direction. Potentially conflicting reasons are distorted so that they no longer conflict.

An important fact about all these effects is that they do not happen all the time. Haidt (2001) argues that such rationalization and bolstering of intuitive judgments happens most of the time in the moral domain. Such extreme views are surely too extreme (e.g., Bucciarelli *et al.* 2008). Individuals differ widely in all effects of this sort. Some people do not show this sort of habitual bolstering. Most people do it some of the time but not all the time.

Thinking may thus differ in its effectiveness in overcoming intuitive errors or reaching correct conclusions. Ideally, good thinking is "actively open-mided" (AOT; Baron 1993, 2008). It looks for reasons why initial conclusions are incorrect, that is, counter-evidence, for alternative options, and for goals that are not served well by the currently favored option. People differ considerably in AOT, and those who do more of it, or who believe that good thinking is in fact self-critical in this way, perform better on various tasks, including the CRT (Baron *et al.* 2015; Stanovich and West 1998).

In sum, thinking can differ in its extent and its direction. The opportunity for reflective deliberation is opportunity for extent. The direction is not affected by such opportunity. For many people, or for each person in some situations, the direction will be to bolster initially favored options, perhaps those that are intuitive, or perhaps just those that come to mind first or that seem most desirable. For other people, or for each person in some situations, the direction will be more critical, looking for better options before making a final choice. Only in these situations can we be confident that the opportunity for reflection will help people achieve their goals.

Deception

If manipulation is some sort of attempt to influence other people's decisions, how can we characterize that attempt? My proposal here has itself been subjected to very little critical evaluation, so it may need modification. It is that harmful manipulation involves deception with the intent of leading people to choose options that they might regret, or not choose, on the basis of actively open-minded reflection, if they had full information about the nature and intent of the effort to influence their choices. This definition captures the idea that harmful manipulation is hidden, harmful, and intentional.

Intention is relevant because we consider harmful manipulation to be something worth constraining, either through the law (torts, crimes, regulations) or social norms. The threat of punishment, or social pressure, will have less effect on behavior that is accidental. Intention, in this case, means that the manipulating agent expects, and has reason to expect, her manipulation to have the designated effect and would not do it in the absence of this expectation.

Harmless or beneficial manipulation would differ from this kind in one of two ways. First, it may lead people away from options that are better for them but worse for others. Second, it may lead people away from choices that they could be expected to make exactly because they do not engage in AOT, toward options that they would make on better reflection. But it would still be less effective if its nature and intent were revealed, just because people may resist efforts to influence their behavior. It is this hiddenness that makes it fit the everyday concept of being manipulative.

Examples

I now discuss Sunstein's examples in the light of this revised definition.

The use of relative risk information to scare people into do what they ought to do is an example of beneficial manipulation. It would not be manipulation if the communication said, "We are telling you that X triples your risk of some

disease because this is more likely to get you to avoid X than if we told you that your risk increased from 0.00001 to 0.00003." (It might still work. Who knows?) But to say that the relative risk information aims at System 1, or might involve emotion, is beside the point. People might think about this very carefully, depending on what X is.[1]

Use of the loss frame to scare people probably has smaller and more complicated effects (Rothman *et al.* 2006), but it raises the same issues as the relative-risk example. However, in this case people who use AOT could reframe the choice for themselves. Still, the point of this is beneficial, and the communicator's hope is that people do not use AOT.

The use of descriptive or injunctive norms to do good strikes me as another example of beneficial manipulation. Like the example of relative risk, the message gives one sort of information but not another, and the decision maker using AOT might wonder what is not being said. As in the loss-frame case, the communicator might hope that AOT is not invoked, or, if it is, the decision maker might do the right thing anyway.

I count the examples of default rules as beneficial manipulation, if they are based on deception in the sense I described. Only a (Kantian?) deontological view of the sort that holds that any deception is immoral would say that these are wrong or morally "questionable."[2]

Advertising, including much of political campaigning, is interesting because it clearly uses techniques that could be seen as manipulative. The main one, I think, is not so much the association of the product/candidate with pleasant images and sounds. These things have small and transient effects. Rather, it is the use of vivid anecdotes, which can overwhelm pallid statistics even if the anecdote represents one of thousands already counted in the statistics (Nisbett and Ross 1980), such as the Willie Horton story used in the campaign against Michael Dukakis.[3] Is this manipulation? For most adults, I think not, because deception is absent. Everyone knows that advertisers and politicians want you to buy what they have to sell and will tell you whatever will get you to do that. Advertisers are legally not supposed to lie, but politicians can, and do, lie freely, subject only to the criticism of news reporters and other politicians.

Although advertising is not manipulation for sophisticated adults, it is for children, who do not vote, but who have not learned yet to take the other ads

[1]The original communication is deceptive not only in its choice of whether to use relative or absolute risk but also in its failure to provide the base rate (0.0001) along with the relative risk. Thus a truly reflective person might notice the missing information and look for it.

[2]A full utilitarian analysis would take into account the future implications of any sort of deception, as precedents for more harmful lying, if they are misinterpreted, and as undermining somewhat the basis for trust. In the real world such effects are usually very small, a drop in the bucket, given the fact that truth telling is already known not to be universal.

[3]https://en.wikipedia.org/wiki/Willie_Horton.

they see with several grains of salt. Advertising is thus deceptive in this case, and it should be, and is, regulated.

Manipulation with consent is an interesting case, for example, the alcoholic who wants to stop drinking and will welcome any trick that will accomplish this end. By my account, this is manipulation only if the manipulator intends to deceive, that is, if she does not know that the alcoholic knows that the message is designed to influence him. In order for such manipulation to work, the alcoholic may need to deceive himself. After welcoming the manipulation, he must then partially forget that he was the one who chose it. If he fully and rationally took this fact into account, he would not be influenced. I agree with Sunstein that the result is consistent with the alcoholic's best thinking. But, if the manipulator does not know this, from her point of view it is still manipulation of the bad sort. It is a case of moral luck: the manipulator tries to do something bad but fails and does something good instead. Legally, it is "attempted manipulation."

Transparency

A final set of examples challenges my claims that intentional deception is necessary in order to call something manipulation, and that we need to distinguish good and bad manipulation. Sunstein argues that transparency is difficult to define and still appears not to cover all the relevant cases.

He argues that the use of relative risk does not involve a lack of transparency. Yet I have argued that it does, by omission. The alternative frame (absolute risk) is not presented, and neither is the base-rate information from which the recipient might calculate absolute risk. The communicator expects that the recipient will not notice the omission, hence be deceived, and the choice of the relative-risk frame might not have been made otherwise, so it is intentional.

Sunstein argues that "a graphic health warning [...] is perfectly transparent," hence a form of manipulation that does not depend on opacity. I would argue that this is not manipulation, exactly for that reason. It is indeed a form of influence, like advertising, but the intention behind it is perfectly clear. At this point, we are doing subjective semantics. My meaning for "manipulation" may just differ from Sunstein's. But a little distortion of the everyday meaning of the term may be appropriate if we are to make it into a technical concept, for example, one that can be used in the law. (However, my distinction between good and bad manipulation was designed to avoid *excessive* distortion of the everyday meaning.)

What if, Sunstein asks, subliminal advertising (assuming that it had any effect at all) were preceded by an announcement that it would occur? It would still be manipulation. I would argue that such an announcement still does not make it transparent in the relevant sense. Specifically, manipulation requires

that the communicator expects the an actively open-minded thinker to respond differently if the relevant information were revealed. Merely telling people that subliminal ads will be present does not do this.

Another example concerns manipulation of people for their own good, for example, to remedy a self-control problem concerning alcoholism or drug addiction, or for other good purposes such as stopping a kidnapping. I would argue that this is indeed manipulation, but it is good manipulation. I do not see how we can avoid admitting that some manipulation is good, by any account, without distorting the term too much.

Conclusion

I have argued that the definition of manipulation can be improved by making somewhat clearer distinctions between good and bad manipulation and between what good and bad reflection. Good reflection is actively open-minded. It is the sort of reflection that leads to decisions most consistent with all of the decision maker's goals. Thus, preventing this reflection prevents the achievement of goals. Bad reflection is System 2 at its worse, either useless or just bolstering whatever is initially favored. I think that these modifications clarify the welfarist view, helping to explain why we ought to worry about bad manipulation.

I have no further disagreements with Sunstein's argument. The implications of this alternative for questions about whether democratic governments can require, forbid, or otherwise regulate (harmful) manipulation are no different from those of Sunstein's view.

The fact that the term "manipulation" is two-dimensional may lead to confusion. The two dimensions of outcome valence and deceit may be correlated in the real world, or people may think they are. When two dimensions are correlated, people sometimes use one of them to make inferences about the other (Kahneman and Frederick 2002). This attribute-substitution effect may lead people to be suspicious of beneficial manipulation, such as those involved in nudges (Felsen *et al.* 2013; Thaler and Sunstein 2008). For example, people may think of beneficial nudges as violations of autonomy to the extent to which they actually work. This confusion could hold back the adoption of beneficial nudges.

References

Baron, Jonathan (1993), "Why Teach Thinking? — An Essay. (Target Article with Commentary.)," *Applied Psychology: An International Review*, 42, 191–237.

Baron, Jonathan (2008), *Thinking and Deciding*, 4th, New York: Cambridge University Press.

Baron, Jonathan (2009), "Belief Overkill in Political Judgments. (Special issue on Psychological Approaches to Argumentation and Reasoning, Ed. by L. Rips)," *Informal Logic*, 29, 368–78.

Baron, Jonathan, Sydney Scott, Katrina Fincher, and S. Emlen Metz (2015), "Why Does the Cognitive Reflection Test (Sometimes) Predict Utilitarian Moral Judgment (And Other Things)?" *Journal of Applied Research in Memory and Cognition*, special issue on Modeling and Aiding Intuitions in Organizational Decision Making, http://dx.doi.org/10.1016/j.jarmac.2014.09.003.

Bucciarelli, Monica, Sangeet Khemlani, and Philip N. Johnson-Laird (2008), "The Psychology of Moral Reasoning," *Judgment and Decision Making*, 3, 121–39.

Frederick, Shane (2005), "Cognitive Reflection and Decision Making," *Journal of Economic Perspectives*, 19, 24–42.

Haidt, Jonathan (2001), "The Emotional Dog and Its Rational Tail," *Psychological Review*, 108, 814–34.

Kahneman, Daniel and Shane Frederick (2002), *Representativeness Revisited: Attribute Substitution in Intuitive Judgment*, ed. & D. Kahneman T. Gilovich D. Griffin, New York: Cambridge University Press, 49–81.

Meyer, Andrew, Bob Spunt, and Shane Frederick (2013), "The Bat and Ball Problem," Talk presented at the meeting of the Society for Judgment and Decision Making, Toronto, Nov. 16. 2013.

Nisbett, Richard E. and Lee Ross (1980), *Human Inference: Strategies and Shortcomings of Social Judgment*, Englewood Cliffs, New Jersey: Prentice-Hall.

Rothman, Alexander J., Roger D. Bartels, Jhon Wlaschin, and Peter Salovey (2006), "The Strategic Use of Gain- and Loss-framed Messages to Promote Healthy Behavior: How Theory can Inform Practice," *Journal of Communication*, 56, S202–S220.

Russo, J. Edward, Kurt A. Carlson, and Margaret G. Meloy (2006), "Choosing an Inferior Alternative," *Psychological Science*, 17, 899–904.

Stanovich, Keith E. and Richard F. West (1998), "Individual Differences in Rational Thought," *Journal of Experimental Psychology: General*, 127, 161–88.

Sunstein, Cass R. (2015), "Fifty Shades of Manipulation," *Journal of Marketing Behavior*, 1(3-4), 213–44.

Wason, Peter C. and Jonathan St. B. T. Evans (1975), "Dual Processes in Reasoning?" *Cognition*, 3, 141–54.