Behavioral and Brain Sciences

cambridge.org/bbs

Rational framing effects: A multidisciplinary case

José Luis Bermúdez

Samuel Rhea Gammon Professor of Liberal Arts and Professor of Philosophy, Texas A&M University, College Station, TX 77843, USA
jbermudez@tamu.edu

Target Article

Cite this article: Bermúdez JL. (2022) Rational framing effects: A multidisciplinary case. *Behavioral and Brain Sciences* **45**, e220: 1–59. doi:10.1017/S0140525X2200005X

Target Article Accepted: 19 January 2022
Target Article Manuscript Online: 24 January 2022
Commentaries Accepted: 16 April 2022

Keywords:
decentering; decision theory; framing; game theory; perspective taking; rationality; self-control; simulation

What is Open Peer Commentary? What follows on these pages is known as a Treatment, in which a significant and controversial Target Article is published along with Commentaries (p. 16) and an Author's Response (p. 51). See bbsonline.org for more information.

Abstract

Frames and framing make one dimension of a decision problem particularly salient. In the simplest case, frames *prime* responses (as in, e.g., the Asian disease paradigm, where the gain frame primes risk-aversion and the loss frame primes risk-seeking). But in more complicated situations frames can function reflectively, by making salient particular reason-giving aspects of a thing, outcome, or action. For Shakespeare's Macbeth, for example, his feudal commitments are salient in one frame, while downplayed in another in favor of his personal ambition. The role of frames in reasoning can give rise to *rational framing effects*. Macbeth can prefer fulfilling his feudal duty to murdering the king, while also preferring bravely taking the throne to fulfilling his feudal duty, knowing full well that bravely taking the throne just is murdering the king. Such patterns of *quasi-cyclical* preferences can be correct and appropriate from the normative perspective of how one ought to reason. The paper explores three less dramatic types of rational framing effects: (1) Consciously framing and reframing long-term goals and short-term temptations can be important tools for self-control. (2) In the prototypical social interactions modeled by game theory, allowing for rational framing effects solves long-standing problems, such as the equilibrium selection problem and explaining the appeal of non-equilibrium solutions (e.g., the cooperative solution in the Prisoner's Dilemma). (3) Processes for resolving interpersonal conflicts and breaking discursive deadlock, because they involve internalizing multiple and incompatible ways of framing actions and outcomes, in effect create rational framing effects.

1. Against extensionality: A multidisciplinary case for the rationality of (some, but not all) framing effects

Psychologists and behavioral economists have come up with an almost uncountable series of experiments in which people are induced to rate or value the same thing differently depending on how it is framed. You might think: What could be more irrational? And you'd be in good company. Being susceptible to framing effects is a standard example of irrationality in textbooks, and there is a small industry of investment books explaining how framing effects can severely damage your financial health.

But this is a situation where *easy cases make bad law*. We need fundamentally to rethink the almost unquestioned assumption that frame-based reasoning is irrational. Outside the laboratory (and perhaps sometimes in it) there are many situations where it is perfectly rational to be influenced by how things are framed. And many situations, in fact, where being able to frame something in multiple ways is a powerful tool for understanding and making decisions.

Frames and framing factor into decision-making by making one dimension/attribute/value of a decision problem particularly salient. This can take many forms. In the simplest case, frames *prime* responses, as in many of the classic framing experiments (e.g., the Asian disease paradigm, where the gain frame primes risk-aversion and the loss frame primes risk-seeking). But in more complicated situations frames can function reflectively, by making salient particular reason-giving aspects of a thing, outcome, or action. For Shakespeare's Macbeth, for example, his feudal commitments are salient in one frame, while in another they are downplayed in favor of his personal ambition.

This target paper explains how the role of frames in reasoning can give rise to *rational framing effects*, which have the following structure: An agent or decision-makers prefer(s) A to B and B to C, even while knowing full well that A and C are different ways of framing the same action or outcome. Such patterns of *quasi-cyclical* preferences can be correct and appropriate from the normative perspective of how one ought to reason.

The case for rational framing effects has several strands. After reviewing the state of play in section 2, in section 3 I emphasize the role of emotions in decision-making, as well as how complex decision problems are best understood as defined over framed outcomes (as opposed

© The Author(s), 2022. Published by Cambridge University Press

to the neutral, extensional scenarios envisaged by decision theorists). I offer two dramatic examples of quasi-cyclical preferences – Aeschylus's Agamemnon and Shakespeare's Macbeth.

The second half of the paper explores three ways in which rational framing effects can function in practical decision-making. Section 4 shows how consciously framing and reframing long-term goals and short-term temptations can be important tools for self-control. In section 5 we see how, for the prototypical social interactions modeled by game theory, allowing for rational framing effects solves longstanding problems, such as the equilibrium selection problem and the challenge of explaining the appeal of non-equilibrium solutions (such as Cooperation in the Prisoner's Dilemma). Finally, section 6 shows how processes for resolving interpersonal conflicts and breaking discursive deadlock, because they involve internalizing multiple and incompatible ways of framing actions and outcomes, in effect creates rational framing effects.

2. The extensionality requirement: A cornerstone of rationality?

2.1 The extensionality principle

Decision theorists disagree about a lot of things. Almost everything, in fact. But one basic principle receives almost unanimous acceptance. This is the *extensionality principle* (a.k.a. *the invariance principle*), poetically phrased by Shakespeare in *Romeo and Juliet*: "What's in a name? That which we call a rose/By any other name would smell as sweet." How we name things should not affect how we value them. More generally, preferences, values, and decisions should be unaffected by how actions and outcomes are framed.

> The extensionality principle: Preferences, values, and decisions should be unaffected by how outcomes are framed.

It is an immediate consequence of the extensionality principle that framing effects are always irrational.

2.2 Extensionality and framing effects in psychology and behavioral economics/finance

2.2.1 Psychology

The ever-expanding literature on framing effects in psychology goes back to the groundbreaking studies of Tversky and Kahneman (1981), which first presented the Asian disease paradigm. The headline finding from this paradigm was that different frames can prime different attitudes to risk. A positive frame (e.g., talking about survival rates) primes risk aversion, while a negative frame (e.g., talking about mortality rates) primes risk-seeking behavior. Subjects prefer a certain outcome to a risky one in the positive/survival frame, but preferences reverse in the negative/mortality frame.[1]

Research on framing effects in psychology has identified a wide range of valence-consistent framing effects, in which risk is not a factor and so the effect is driven purely by valence. These include: the fat content of ground meat (Levin & Gaeth, 1988); condom use (Linville, Fischer, & Fischoff, 1993); evaluating basketball players (Levin, Schneider, & Gaeth, 1998); contract negotiation (Neale & Bazerman, 1985); and social dilemmas (Brewer & Kramer, 1986). In all these cases subjects consistently respond to the same thing differently as a function of how it is framed (preferring meat labeled as 25% fat to meat labeled as 75% lean, and even finding that it tastes better).[2]

2.2.2 Behavioral economics and behavioral finance

Framing effects have been extensively investigated in behavioral finance (for an overview see Bermúdez, 2020a, Ch. 3). Multiple studies have confirmed that investors tend to be risk-averse for gains and risk-seeking for losses. A particular focus of research is how this bias, if such it is, plays out across temporally extended investment behavior through varieties of *mental accounting*, as in Johnson and Thaler's theory of *hedonic editing*, which explores how people keep track of losses and gains across multiple gambles/investment decisions (Thaler, 1999; Thaler & Johnson, 1991).

The disposition effect is the tendency to hold losing investments and sell those doing well (Shefrin & Statman, 1985) and implicates multiple types of framing – focusing on losses/gains rather than absolute levels of wealth, for example, and also framing losses narrowly rather than across an entire portfolio (Barberis & Huang, 2006; Barberis & Xiong, 2009). The theory of *myopic loss aversion* that has been proposed to explain the *equity risk premium* (the fact that equities are priced much higher relative to bonds than would be predicated by plausible measures of risk aversion – Mehra, 2008; Mehra & Prescott, 1985) is a framing explanation. Myopic loss aversion is the tendency to evaluate losses and gains in terms of particular (and relatively short) timeframes, for example, 12 months. This is a framing effect, because the same absolute levels might well be a gain rather than a loss relative to a longer timescale.

2.3 The consensus view

Tversky and Kahneman themselves drew a stark conclusion from their experiments on framing effects and breaches of the extensionality principle:

> Because framing effects and the associated failures of invariance are ubiquitous, no adequate descriptive theory can ignore these phenomena. On the other hand, because invariance (or extensionality) is normatively indispensable, no adequate prescriptive theory should permit its violation. Consequently, the dream of constructing a theory that is acceptable both descriptively and normatively appears unrealizable (Tversky & Kahneman, 1986, p. S272)

Despite some challenges and objections (see notes 2 and 3), it is fair to say that those words, written in 1986, still represent a consensus among psychologists of reasoning, and in the cognitive sciences more broadly.

In economics and finance, the irrationality of framing effects is rarely questioned. Experimental paradigms typically result in

José Luis Bermúdez is Professor of Philosophy and Samuel Rhea Gammon Professor of Liberal Arts at Texas A&M University. He is the author of eight single-authored books at the intersection of philosophy, psychology, and the theory of decision, and over 100 journal articles and book chapters. His books include *The Paradox of Self-Consciousness* and, most recently, *Frame It Again: New Tools for Rational Thought*. His work has been supported by grants and fellowships from the National Endowment for the Humanities, the National Science Foundation, the American Association of Learned Societies, and the British Academy, among others.

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

subjects choosing options in which they receive less money (or lose more), and, from the perspective of economics/finance, that is the height of irrationality. Investing effects such as the disposition effect are much discussed precisely because they are frequently charged with destroying wealth. Myopic loss aversion seems, on the face of it, to be irrational. Certainly, the importance of avoiding framing effects is a central theme in popular investing books, such as *The Little Book of Behavioral Investing: How Not to Be Your Own Worst Enemy* (Montier, 2010).

2.4 Beyond the rationality wars: A radical proposal

The framing experiments played a key role in the "rationality wars" – an interdisciplinary debate about whether the extensive literature on cognitive biases and widespread fallacies in reasoning shows that human beings are in some sense intrinsically irrational.[3] The rationality wars seem to have ended in the manner of the battles reported by Xenophon and Thucydides, with both sides raising victory monuments, despite devastating and roughly equal casualties on both sides.

In one important respect, though, both sides of the rationality wars missed a fundamental point. Even those who contested the claim that human beings are intrinsically irrational, never cast doubt on the intrinsic irrationality of framing effects. This paper shifts the terms of engagement. The focus on the canonical experiments has blinded us to the existence of *rational framing effects*. These framing effects are fundamentally different kind from those we have been looking at up to now.

In the classic risky choice and valence-consistent framing experiments, subjects typically fail to recognize that the two proffered outcomes are equivalent, and so that they are in the grip of a framing effect. One indication is that subjects typically back-pedal when the framing effect is pointed out to them, and presenting both frames simultaneously works as a debiasing effect, as in Bernstein, Chapman, and Elstein (1999).[4]

The rational framing effects I will be discussing are very different. They all involve subjects *knowingly and consciously* valuing outcomes or actions differently as a function of how they are framed. This situation might be represented as follows, where "o" represents the outcome: The agent prefers $F_1(o)$ to A, for some A, but also prefers A to $F_2(o)$, where F_1 and F_2 are known to be different ways of framing the same outcome.

The next section motivates the general idea that there can be rational preferences with the structure just described – and hence that there can be rational framing effects.

3. Motivating the alternative: Rational framing effects?

3.1 Quasi-cyclicality versus cyclicality

Previous discussions of framing effects have failed to make an important distinction.

There are good reasons to think that it is irrational to have preferences that are *cyclical*. A decision-maker has cyclical preferences when, for example, she simultaneously prefers A to B, B to C, and C to A. A decision-maker with cyclical preferences will never be able to decide to do what she prefers most (assuming that transitivity holds). For each of A, B, or C, there will always be something she prefers to it.[5]

In contrast, this paper focuses on decision problems where the decision-maker is aware that there is a single outcome framed in different ways and nonetheless insists on evaluating it differently in the two frames. As suggested earlier, this situation might be represented as follows, where "o" represents the outcome: She prefers $F_1(o)$ to A, for some A, but also prefers A to $F_2(o)$, where F_1 and F_2 are known to be different ways of framing the same outcome. I term this pattern of preferences *quasi-cyclical*. Figure 1 illustrates the difference between cyclical and quasi-cyclical preferences.

You might object: How can these be different? Surely, quasi-cyclicality just collapses into cyclicality? If I know that $F_1(o)$ and $F_2(o)$ are different ways of framing the same outcome, then surely that outcome itself is all that matters for my preferences and choice – how it is described or framed should be irrelevant.

Against this I suggest that the objects of preference are framed outcomes. There is no such thing as making choices over a purely extensional opportunity set, independent of any way of describing or framing the things in it. We cannot help but see the objects of choice as framed, or described, or conceptualized in certain ways. As a matter of fact, we often do ignore these framings and in effect choose and reason as if we were choosing and deliberating about a purely extensional opportunity set. But some of the time we do not, and instead allow framings and descriptions to influence our choices. When we do that, we are not choosing between outcomes, viewed completely neutrally, nor between frames or descriptions, viewed as cognitive or linguistic entities. Rather we choose between outcomes framed in a certain way. This is how quasi-cyclical preferences arise. (For more on the objects of preference, in the specific context of how it can be rational to have quasi-cyclical preferences, see sect. 3.4).

Before looking at specific examples of quasi-cyclical preferences in sections 4 through 6, I review general theoretical reasons for my way of thinking about preferences and, by extension, for the idea that there can be rational, quasi-cyclical preferences (and hence, rational framing effects).

3.2 Complexity, emotion, and framing: Three working hypotheses

Classical decision theory explicitly adopts a Humean model of rational decision-making, treating reason as a slave of the passions. In the theory of expected utility, a preference order is taken as given (reconstructible, in the ideal case, from patterns of suitably consistent choices), and the decision-maker's task is to maximize utility relative to that preference order. Decision theory is silent on where those preferences might come from. And a rational decision-maker's preferences are constrained only by considerations of consistency, as given, for example, by axioms such as transitivity and substitution (Bermúdez, 2009, Ch. 2; Harsanyi, 1977; Jeffrey, 1983).

Emotions and reason cannot be as insulated from each other as this instrumentalist picture suggests, however. Numerous studies have shown, for example, that integral emotions (i.e., those directly relevant to the task at hand) play an important role in modulating decision-making (see Loewenstein & Lerner [2003] for a review, as well as the influential neuropsychological data reported in Damasio [1994] and Bechara, Damasio, Damasio, & Lee [1999]). There is strong evidence also that incidental emotions influence decision-making (e.g., Keltner & Lerner, 2010; Lerner & Keltner, 2000).

Particularly important for framing and quasi-cyclical preferences is that emotions are multi-dimensional phenomena. Current research on emotions has largely moved away from the traditional idea (e.g., Russell, 1980) that emotions influence

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

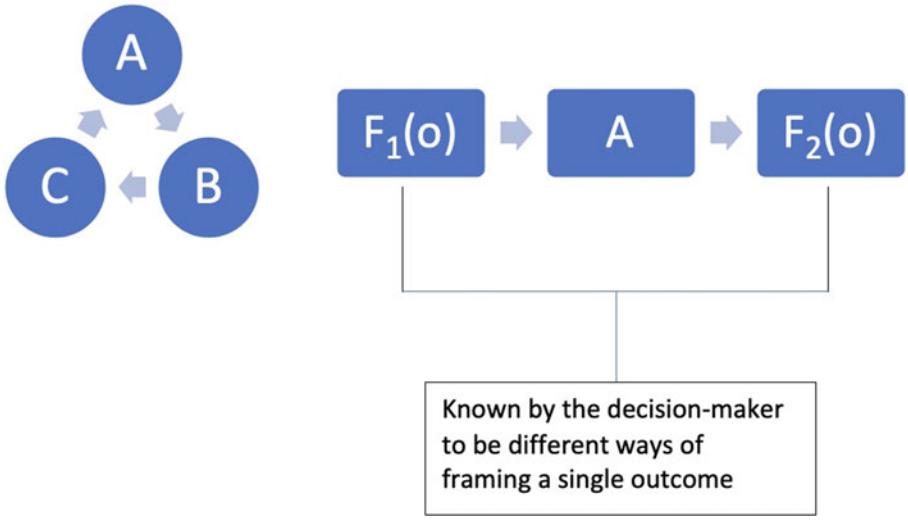


Figure 1. Diagram on the left shows the structure of cyclical preferences. Assuming the transitivity of preference, it has the counter-intuitive consequence that everything is preferred to itself, and also that everything has something that is preferred to it. On the right is an illustration of quasi-cyclical preferences. Here there is no circle, even though the decision-maker knows that $F_1(o)$ and $F_2(o)$ are different ways of framing o.

decision-making and behavior through one or both of arousal and valence. Instead, contemporary theories identify several different dimensions along which emotions vary. According to the Emotion-Imbued Choice model, for example, there are six dimensions: certainty, pleasantness, attentional activity, anticipated effort, individual control, and others' responsibility (Lerner, Li, Valdesolo, & Kassam, 2015). Adolphs and Anderson have a different list, which includes scalability, valence, persistence, generalization, global coordination, automaticity, and social communication (Adolphs & Anderson, 2018). This multi-dimensional perspective opens up the possibility that different framings of a decision problem can elicit different dimensions of emotions, and so engage the motivational system in different ways.

Relatedly, decision theorists and philosophers have proposed thinking about decision-making in multi-dimensional terms. Keeney and Raiffa (1976) develop the idea that decision-making involves multiple different criteria with inevitable trade-offs between them. The field of multi-attribute utility analysis (or multi-criteria decision-making) proposes tools for solving this type of problem.

A more dramatic development of this basic idea has emerged in moral philosophy, where several authors have argued that many decisions, including those traditionally known as moral dilemmas, involve incommensurable values that cannot be compared on a single scale or ordering. The notion of *transformative experiences* developed by Paul explores a similar idea, where the incomparability lies between value systems before and after some transformative event (Paul, 2014).

These very different but converging ideas suggest a working hypothesis for how frames factor into decision-making.

(H1) Frames and framing factor into decision-making by one dimension/attribute/value of the decision problem highly salient, which influences how the subject engages emotionally and affectively.

H1 in turn suggests two further hypotheses specifically about quasi-cyclical preferences. The first has to do with when they will arise, namely, in situations with multiple dimensions/attributes/values pulling in different directions:

(H2) Quasi-cyclical preferences are likely to be found in decision problems that are sufficiently complex and multi-faceted that they cannot be subsumed under a single dimension/attribute/value. Different frames engage different affective and emotional responses, which the decision-maker cannot resolve either by ignoring all the frames except one or by subsuming them into a larger frame.

If H2 correctly characterizes how quasi-cyclical preferences can arise, then that suggests when it will be rational (i.e., correct and appropriate from a normative perspective) to have such preferences:

(H3) Framing effects and quasi-cyclical preferences can be rational in circumstances where it is rational to have a complex and multi-faceted response to a complex and multi-faceted situation.

The next sub-section brings these three working hypotheses to life with two dramatic literary examples of quasi-cyclical preferences. In section 3.4 I motivate H3 in more detail.

3.3 Quasi-cyclical preferences: Agamemnon and Macbeth

3.3.1 Agamemnon at Aulis

The chorus in Aeschylus's *Agamemnon* tells the story of Agamemnon's sacrifice of his daughter Iphigenia. Agamemnon is leading the Greek fleet against Troy to avenge the abduction of Helen by Paris. The fleet is becalmed at Aulis when two eagles swoop down to kill and eat a pregnant hare. It is a portent, which the prophet Calchas interprets as reflecting the displeasure of the goddess Artemis at the prospect of innocents being killed at Troy. The lack of wind has the same source. The only solution (although it is not clear why!) is for Agamemnon to sacrifice to the goddess his own daughter Iphigenia.

The chorus recalls Agamemnon's struggle:

> And I can still hear the older warlord saying,
> "Obey, obey, or a heavy doom will crush me! –
> Oh, but doom will crush me
> once I rend my child,
> the glory of my house –

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

a father's hands are stained,
blood of a young girl streaks the altar.
Pain both ways and what is worse?
Desert the fleets, fail the alliance?
No, but stop the winds with a virgin's blood,
feed their lust, their fury? – feed their fury! –
Law is law! –
Let all go well.

(Aeschylus, *Agamemnon*. Translated by Robert Fagles [Fagles, 1977])

Although Aeschylus does not put it quite this way, Agamemnon is grappling with quasi-cyclical preferences. There is a single option, bringing about the death of Iphigenia, that Agamemnon frames in two different ways – as *Murdering his Daughter*, on the one hand, and as *Following Artemis*'s *Will*, on the other. His alternative is *Failing his Ships and People* (by refusing to make the sacrifice). Agamemnon's dilemma is that he evaluates the death of Iphigenia differently, depending on how it is framed. Both of the following are true.

(A) Agamemnon prefers Following Artemis's Will to Failing his Ships and People.
(B) Agamemnon prefers Failing his Ships and People to Murdering his Daughter.

But he knows, of course, that *Following Artemis*'s *Will* and *Murdering his Daughter* are the same outcome, differently framed.

3.3.2 Macbeth at inverness

Shakespeare's *Macbeth* provides another dramatic example of quasi-cyclical preferences.[6]

In Act 1, Macbeth, Thane of Glamis, is told by the witches that he will become both Thane of Cawdor and King of Scotland. When the first part of the prophecy is fulfilled (by King Duncan's executing the current Thane and granting the title to Macbeth), Macbeth begins to think about making the second part of the prophecy come true by killing Duncan. Providentially Duncan arrives under Macbeth's roof, and Lady Macbeth encourages her husband to assassinate the King. But still Macbeth has his doubts. On the one hand, he recognizes that killing Duncan would make him King – and surely that's worth the risk:

If it were done when 'tis done, then 'twere well
It were done quickly. If the assassination
Could trammel up the consequence, and catch
With his surcease success; that but this blow
Might be the be-all and the end-all here,
But here, upon this bank and shoal of time,
We'd jump the life to come.
(Act I, Sc. 7)

But, as he immediately recognizes, he has two different obligations to Duncan. He is both his host and his pledged kinsman. On both counts his duty is to protect Duncan, not murder him.

He's here in double trust:
First, as I am his kinsman and his subject,
Strong both against the deed; then, as his host,
Who should against his murderer shut the door,
Not bear the knife myself.
(Act I, Sc. 7, lines 1–7)

So, with apologies to Shakespeare, the following two propositions both seem to be true.

(C) Macbeth prefers Fulfilling his Double Duty to Duncan to Murdering the King.
(D) Macbeth prefers Bravely Taking the Throne to Backing Away from his Resolution to Make the Prophecy come True.

He knows, of course, that to fulfill his double duty to Duncan is to back away from his resolution to make the prophecy come true – and likewise that bravely taking the throne is murdering the King. So, he has quasi-cyclical preferences.

Someone might object: Agamemnon and Macbeth certainly have quasi-cyclical preferences. They are each knowingly and consciously subject to a framing effect. But why should we think that they are being rational? Surely, Agamemnon and Macbeth are trapped in a cycle of *irrationality*, from which the only escape is to settle on one framing rather than the other?

The next section replies to this objection.

3.4 The rationality of quasi-cyclical preferences

For classical decision theorists, preferences are rational when, and only when, they are suitably consistent – that is, when, and only when, they are in accordance with the axioms of decision theory, such as transitivity and substitution. This way of thinking about rationality makes most sense if we think of preferences as revealed in choices and as having no content over and above the choices to which they give rise.

Leonard Savage, the founder of modern decision theory, was very clear that such a model of choice and preference is really only applicable to what he called "small worlds" – environments where a decision-maker can be assumed to have exhaustive knowledge of all available actions and outcomes and has preferences completely defined over all those actions and outcomes. Few, if any, real-world decisions fit this model. More typically, decision problems have to be constructed. Decision-makers have to work out for themselves what the available actions are and to what outcomes they might lead.

At the same time, preferences are not basic. They are made for reasons. And the process of constructing a decision problem is simultaneously a process of identifying reason-giving aspects of different possible outcomes and possible actions. Whereas standard decision theory assumes that these processes are either unnecessary, or have somehow been completed, before considerations of rationality come into play, it is plausible that there are normative constraints upon them. In particular, when a decision problem is complex and multi-faceted, a rational decision-maker (someone who is deliberating as they ought to be deliberating) should be sensitive to the full range of potential reasons that there might be for choosing one way rather than another.

This is how frames come into play. Reasons are frame relative. Macbeth is an extreme example. You need to look at the world in a very particular way for regicide to seem a good idea, and it is only when things are framed in that way that Macbeth's ambition can get translated into action. Switch the frame and a different set of reasons come into play. In Macbeth-type decision problems, what counts as a reason from within the perspective of one frame is so recessive as to be almost invisible from the other. But then someone who is trying to do justice to the complexity of the decision situation needs to be sensitive to the possibility of multiple possible framings. With that sensitivity comes the

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

possibility of quasi-cyclical preferences. But those quasi-cyclical preferences have emerged from a decision-maker seeking to satisfy the basic rationality requirement of doing justice to the complexity of the situation. They inherit the rationality of the process that generated them.

The Agamemnon and Macbeth cases bring out in striking relief an abstract structure that reappears in much more everyday situations of direct relevance to cognitive and behavioral scientists. There are plenty of complex and multi-faceted decision problems that do not involve filicide or regicide, but that can be illuminated when understood as involving rational framing effects. The remainder of this paper will discuss three of them.

4. Application 1: Framing and quasi-cyclical preferences in self-control

4.1 Self-control, time-inconsistent preferences, and effortful willpower

Since the influential work of Ainslie, Rachlin, and others, failures of self-control have been conceptualized as preference reversals resulting from a particular way of discounting the future (Ainslie, 1974, 1992, 2001; Rachlin, 2000, 2018).

To discount the future is to assign less utility now to a future good than one expects to derive from it when it is eventually reached. A discounting function describes how the degree to which an agent discounts a future good is related to the delay until the reward is received. There are two broad families of discount function.

> *Exponential discount functions* remain constant over time, with the ratio between how much one discounts a future good at the start and the end of any given temporal interval a function only of interval length. So, the impact of a day's delay will be the same tomorrow as 25 years in the future. For that reason, exponential discounting is described as *time consistent*.
>
> *Hyperbolic discount functions* are *time-inconsistent*, because the ratio of the discount function is not constant. The difference between having $10 today and receiving $11 tomorrow is much greater than the difference between having $10 100 days into the future and having $11 in 101 days.

Hyperbolic discount functions permit preference reversals in which a short-term smaller sooner (SS) reward can, at the moment of temptation and despite the agent's best-laid plans, seem more attractive than the long-term larger later (LL) reward. As the moment of choice approaches, the discount function for SS steepens more rapidly than the discount function for LL (because SS is more imminent than LL), which allows SS's utility to exceed that of LL.

This means that an agent exercising self-control must change the shape of her discount function, which raises two obvious questions:

> (Q1) How does this change in the discount function take place?
> (Q2) Are there techniques that make it easier for subjects to change the shape of their discount functions (and hence to resist temptation)?

Traditionally, Q1 has been answered through the construct of *effortful willpower*, as developed within the theory of *ego depletion* proposed by Baumeister and others (Baumeister, Bratslavsky, Muraven, & Tice, 1998). The basic idea is that self-control requires energy, which is a limited resource and can become depleted. However, the theory of ego depletion does not offer clear guidance in response to Q2. It explains how self-control occurs and why it might fail, but not how it might be improved. Moreover, there are problems with replicating the key effects supporting the ego depletion model.[7]

Framing offers a better approach to both Q1 and Q2. The basic idea is that cases of self-control can have the following structure. An agent committed to a long-term goal (LL) is at risk of succumbing to temptation (SS), because at the moment of choice they prefer SS to LL. They prefer the extra drink to the clear head in the morning, or the extra hour in bed to results of the fitness regime. They can resist the temptation, however, by reframing either or both the temptation or the long-term goal.

Framing approaches to self-control are well-supported experimentally (section 4.2) and also offer practical tools for achieving self-control and resisting temptation (section 4.3).

4.2 Self-control and framing: Evidence from psychology and neuroscience

4.2.1 Framing in the delay of gratification paradigm

The delay of gratification paradigm is an influential paradigm for studying self-control (Mischel & Ayduk, 2004; Mischel & Moore, 1973; Mischel, Shoda, & Rodriguez, 1989). It offers a tool for studying how young children are able to delay immediate gratification (SS) in favor of long-term goals (LL). The children are told that the experimenter needs to go away, but when the experimenter returns, they will receive a delayed reward of, say, two cookies or two marshmallows (i.e., the LL). They can wait for LL or, at any time, while the experimenter is away, they can ring a bell to receive an immediate reward of a single cookie or marshmallow (the SS).[8]

Mischel and collaborators suggested that the experimental behavior can be explained through the interaction between "hot" and "cold" cognitive-affective systems – a version of the dual process theory (Evans, 2008; Kahneman, 2011). Here is how Mischel and Ayduk state the contrast:

> Briefly, the cool system is an emotionally neutral, "know" system: It is cognitive, complex, slow, and contemplative. Attuned to the informational, cognitive, and spatial aspects of stimuli, the cool system consists of a network of informational, *cool* nodes that are elaborately interconnected to each other, and generate rational, reflective, and strategic behavior…. In contrast, the hot system is a "go" system. It enables quick, emotional processing: simple and fast, and thus useful for survival from an evolutionary perspective by allowing rapid fight or flight reactions, as well as necessary appetitive approach responses. The hot system consists of relatively few representations, *hot spots* (e.g., unconditioned stimuli), which elicit virtually reflexive avoidance and approach reactions when activated by trigger stimuli. This hot system develops early in life and is the most dominant in the young infant. (Mischel & Ayduk, 2004, p. 109)

This dual process theory explains why the two discount curves behave as they do. With SS at a safe (temporal) distance, the cool system dominates and so the utility attached to the two outcomes reflects the agent's considered preference for the added benefit of LL. The closer the agent gets to SS, however, the more the hot system kicks in and so the slope of the valuation function steepens, until the SS discount curve eventually intersects the LL discount curve and paves the way for the weak-willed response. Mischel and Ayduk make this very point, describing some of the early delay of gratification studies:

> it became clear that delay of gratification depends not on whether or not attention is focused on the objects of desire, but rather on just how they

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

> are mentally represented. A focus on their hot features may momentarily increase motivation, but unless it is rapidly cooled by a focus on their cool informative features (e.g., as reminders of what will be obtained later if the contingency is fulfilled) it is likely to become excessively arousing and trigger the "go" response. (Mischel & Ayduk, 2004, p. 114)

Here are some of the studies that they identify as pointing to what I will term the frame-dependence of discount curves.

- Mischel and Moore (1973) found that performance on the delay of gratification paradigm varied when children were presented with images of the rewards, as opposed to the rewards themselves. They reasoned that presenting an iconic representation of the reward would present the reward in a "cool" light, highlighting its cognitive and informational features, whereas presenting the reward itself would highlight its motivational features and engage the "hot" system. Children who had the actual reward in front of them performed much worse on the delay of gratification task than children who merely had a picture of the reward in front of them.
- Mischel and Baker (1975) divided children undergoing delay of gratification experiments for marshmallows and pretzels into two groups and cued them to think about the rewards differently. One group, the "cold" group, was primed to think about the marshmallows as "white, puffy clouds" and the pretzels as "little, brown logs." Children in the second, "hot," group were cued to think about obvious motivational features of the marshmallows and pretzels – as "yummy and chewy" and "salty and crunchy respectively." As predicted, children in the cold group were significantly better able to withstand temptation than children in the hot group – a mean of 13 minutes before ringing the bell for the SS reward, as opposed to a mean of 5 minutes.

It seems that the way in which the reward is framed directly affects rate of change of the SS discount curve, and so the point at which the SS discount curve crosses the LL discount curve. The language of "framing" is very natural here, because "white puffy clouds" and "yummy and chewy" are clearly different ways of describing the same reward – and likewise "little brown logs" and "salty and crunchy."

These findings suggest positive strategies for enhancing self-control. Agents can ensure that hot representations of SS are counter-balanced and kept in check by cooler representations that emphasize, for example, the long-term consequences of succumbing to temptation. Likewise, they can represent LL in ways that engage the hot system, thus steepening the LL discount function and preventing the SS discount function from crossing it. (More on this in sect. 4.3.)

4.2.2 Framing in the hidden zeros paradigm

There is evidence that how rewards are valued is modulated by the ventromedial prefrontal cortex, vmPFC, and striatum areas (Hare, Camerer, & Rangel, 2009), while willpower exertion is typically tied to the dorsolateral prefrontal cortex, dlPFC (Figner et al., 2010). These neuroanatomical facts connect up suggestively with the reflections above about conceptualizing self-control as a matter of changing the slopes of the discount curves for SS and LL. To change the slope of the discount curve for a reward is, in essence, to change how that reward is valued. So, if the neural basis for reward valuation is distinct from the neural basis for effortful willpower, then it seems at least in principle possible for self-control to be exercised without engaging willpower. But how?

Magen, Kim, Dweck, Gross, and McClure (2014) tested the hypothesis that self-control can be enhanced by changing how the rewards are framed. They recruited an independently documented framing effect – the *hidden zero effect*. Experiments on discount curves typically present choices such as: "Would you prefer \$5 today or \$10 in a month's time?." This is in hidden zero format, because it does not make explicit that if you opt for \$5 today, you will receive \$0 in a month's time and, correlatively, that if you opt for \$10 in a month's time you will receive \$0 today. To include the relevant non-rewards in the description of the choice is to frame the choice in an explicit zero format – for example, "Would you prefer \$5 today and \$0 in a month's time, or \$0 today and \$10 in a month's time?"

Consistent with earlier results (Magen, Dweck, & Gross, 2008), participants discounted the future at lower rates with outcomes presented in the explicit zero format than in the hidden zero format (even though the outcomes in the two formats can immediately be seen to be equivalent).

Moreover, the neural data confirmed the hypothesis that the reframing is effective in enabling self-control without the exercise of effortful willpower, because variation in activity in the reward areas was sufficient to explain different valuations in the two conditions. Moreover, when subjects were presented with explicit choices, there was significantly less activation in the dlPFC (the area correlated with willpower) when LL was selected in the explicit zero format than when it was chosen in the hidden zero format. In other words, less willpower was required when LL was framed in a cool manner.

4.3 Modeling self-control with quasi-cyclical preferences

To see how to model the connection between framing and self-control in the quasi-cyclical preferences framework, consider Figure 2, which represents the situation as a *sequential choice problem* (McClennen, 1990).

The moment of temptation is marked as t_1. At that moment, the utility of SS outweighs the utility of LL (because of the preference reversal explained in sect. 4.1). According to the basic maxim of decision theory, therefore, the agent should go down at t_1, rather than hold out for LL – because that is the action that will maximize her expected utility at the moment of choice (t_1).

From the perspective of classical decision theory, therefore, the exercise of self-control is problematic. It seems to be counter-preferential and so counter-rational (Bermúdez, 2009, 2018, 2020b). For that reason, decision theorists have developed a number of strategies for explaining how it can be rational to hold out for LL. These include theories of sophisticated choice (Strotz, 1956) and resolute choice (Holton, 2009; McClennen, 1990). None of these has found universal acceptance. Sophisticated choosers essentially avoid the problem through precommitment strategies (e.g., Odysseus tying himself to the mast, to take a much-discussed example). Resolute choice theories are inconsistent with the basic axioms of decision theory, and also have serious difficulties explaining how the counter-preferential choice can actually be made at time t_1 (Bermúdez, 2018).

In contrast, incorporating frames preserves the basic principle of expected utility maximization. Figure 3 shows how quasi-cyclical preferences can come about in self-control cases. The agent frames LL in two different ways. On the one hand, she frames it simply as the long-term reward. Picking up on the

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

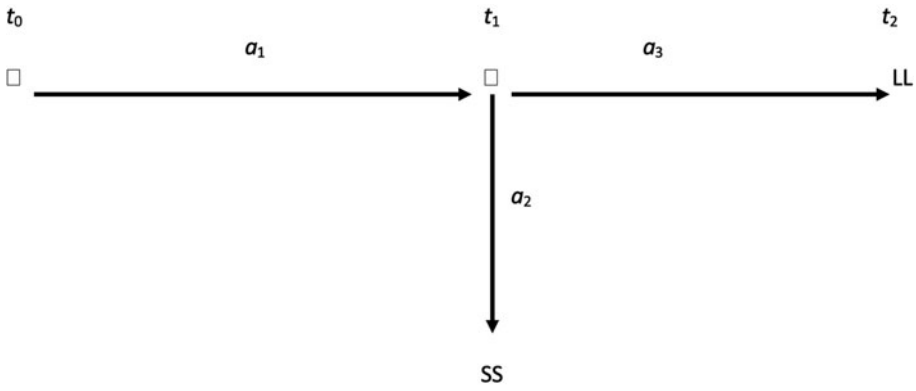


Figure 2. Paradigm case of self-control represented as a sequential choice problem. The moment of planning is at time t_0 with the moment of choice at time t_1, when the agent chooses between the immediate smaller sooner temptation (SS) and the delayed larger later reward (LL).

Mischel discussion from section 5.2.1, this could be termed the cool framing. But at the same time, she also frames LL as *Having Successfully Resisted SS*. This is a hot frame, highlighting the struggle over temptation.

The agent assigns more utility to SS than to LL. This explains why she is in the grip of temptation – if it were not the case, there would not be a problem of self-control at all. At the same time, though, she assigns more utility to *Having Successfully Resisted SS* than she does to SS. The value attached to *Having Successfully Resisted SS* can reflect more than just the LL reward itself. It might, for example, reflect the perceived "virtue" of having overcome temptation. Or it might be taken as a signal of how the agent will react in the future (if I manage to resist temptation here, then I will be more likely to do so in the future).

This explains how she is able to exercise self-control. Moreover, it explains how her self-control is rational, since she is following the option that she most prefers and so is maximizing utility. The preferences are quasi-cyclical because she is perfectly well aware that LL and *Having Successfully Resisted SS* are different ways of framing the same outcome. So, we have a rational framing effect.

This proposal is consistent with the experimental evidence reviewed in section 4.2 as well as with the general model of time-inconsistent preferences outlined in section 4.1. Discount curves are frame-relative and having multiple frames allows for the possibility that there is a framing of one or both SS and LL on which the two discount curves do not cross. More generally, self-control provides a clear illustration of H1 through H3. Self-control situations are typically complex and multi-faceted (H2), sufficiently so that it is rational to have quasi-cyclical preferences (H3). The different framings operate by making one dimension of the decision problem particularly salient.

5. Application 2: Framing in game theory

Game theory is the mathematical theory of strategic choice. In a strategic decision the outcomes for each agent are a function of both what she herself does and what other agents do. Classical decision theory (expected utility theory) is *parametric* (Bermúdez, 2015a, 2015b) – that is, the outcomes are fixed by what the agent does and by the state of the world. Expected utility theory cannot work in strategic decision problems because strategic decision problems have two features:

(1) The actions of the different agents are independent of each other – neither agent is constrained by another to act in a certain way.
(2) Different agents' actions are interdependent with respect to rationality – what it is rational for me to do depends upon what I think it would be rational for other agents to do, but what it would be rational for other agents to do depends upon what they think it is rational for me to do.

The basic solution concept in game theory is *Nash equilibrium*. A Nash equilibrium is a set of strategies such that each player's strategy is a *best response* to the strategies of the others – that is, none of the players can unilaterally improve their position relative to the strategies of the other agents (Shoham & Leyton-Brown, 2008 provide all the details).

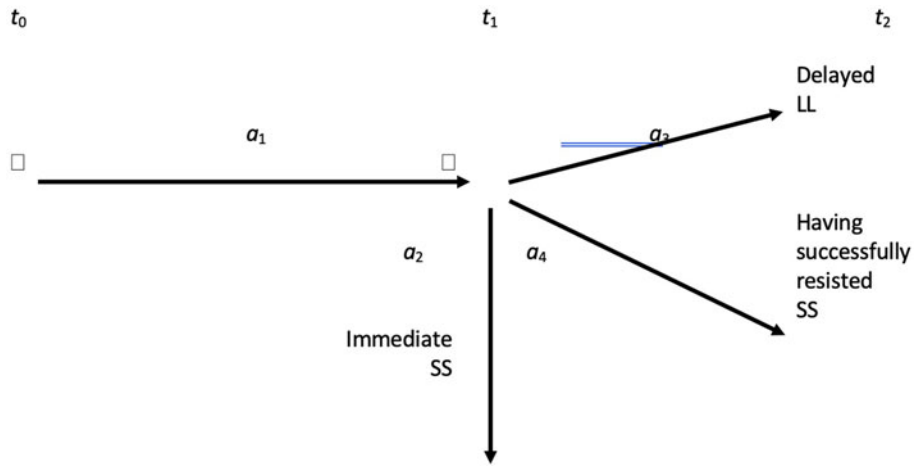


Figure 3. Reframing the decision problem in Figure 2.

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

In two well-known respects, game theory is on a much less firm footing than expected utility theory. Looking at two foundational problems will help us see how frames can be important in game theory.

5.1 Foundational problems in game theory

5.1.1 The equilibrium selection problem

Nash's theorem says that every strategic interaction satisfying some basic conditions has at least one equilibrium solution. But many games have multiple equilibrium solutions. There is no generally accepted method within game theory for identifying one solution as more rational than another, even in situations where many people find it obvious that there is a unique rational solution.

Stag Hunt (SH) provides a good example (Skyrms, 2012). Two players have a choice between hunting hares or hunting a stag. The stag is the better reward, but requires collaboration, while hunting a hare is better if the other player is not doing the same. Here is the payoff table:

		Column	
Row		*Stag*	*Hare*
	Stag	4, 4	0, 2
	Hare	2, 0	1, 1

Each cell represents the payoff to Row, first, and then to Column. There are two pure Nash equilibria (*Stag*, *Stag*) and (*Hare*, *Hare*), as well as a probabilistic mixed strategies equilibrium. Game theory provides tools for identifying the equilibria, but not for choosing between them. This is the equilibrium selection problem.

Many criteria have been proposed for equilibrium selection (Harsanyi & Selten, 1988). Two prominent candidates are:

- *Pareto superiority* (or *Payoff dominance*): Choose the equilibrium such that no player is worse off and at least one player is better off.
- *Risk dominance*: Choose the least risky equilibrium.

In Stag Hunt these two criteria pull in different direction. (*Stag*, *Stag*) is Pareto superior, but (*Hare*, *Hare*) is risk dominant. It is fair to say that neither, nor any other, candidate for equilibrium selection has gained widespread acceptance.

5.1.2 The problem of non-equilibrium solutions

Game theory is a normative theory, and so not straightforwardly opens to empirical counterexample. Nonetheless, it is reasonable to expect normative theories to reflect the realities of practical decision-making, and there is strong experimental and anecdotal evidence that subjects often adopt non-equilibrium solutions in social dilemma games – typically when non-equilibrium solutions reflect considerations of fairness and collaboration to which Nash equilibrium is blind.

The one-shot Prisoner's Dilemma (PD) has been much studied in this context. As is well known, *Cooperate* is a dominated strategy in PD, where mutual *Defection* is the only pure-strategies Nash equilibrium. A well-known meta-analysis (Sally, 1995) looked at 130 experiments on social dilemmas such as PD and found a mean cooperation rate across the studies of 47.4%. (For similar effects see Heuer & Orland, 2019; Janssen, 2008; Pothos, Perry, Corr, Matthew, & Busemeyer, 2011.)

This generates two significant questions.

> The descriptive question:
> Can we give a principled account of why subjects systematically diverge from equilibrium solutions?
> The normative question:
> Can we give a principled account of how it might be rational to diverge from Nash equilibrium?

Framing and quasi-cyclical preferences allow us to answer both questions.

5.2 Framing in games: Bacharach's proposal

Michael Bacharach was a pioneer in this area, focusing primarily on the descriptive question particularly in the posthumously published (Bacharach, 2006). Unusually in game theory, his ideas are richly informed by work on the psychology of groups and cooperation.

Bacharach's basic idea is that strategic interactions can be framed in different ways. He focuses in particular on two frames, which I will call the "I"-frame and the "we"-frame, although this is not his standard terminology.

> **"I"-frame**
> In the "I"-frame, agents look only at their own payoffs, employing the type of best response reasoning that seeks a Nash equilibrium.

> **"We"-frame**
> A team reasoner thinks about the payoff table from the perspective, not of an isolated individual, but instead from the perspective of a team member, or group member.

These two frames are extensionally equivalent. Games are defined by their payoff tables, and the payoff table remains constant across the two frames – the rewards to each player in each possible outcome are the same in the "I"-frame and the "we"-frame. The differences lie in which aspects and properties of the payoff table are salient in each frame. "I"-frame reasoners look only at their portion of the joint outcome, ignoring the outcomes for other players. In contrast, "we"-frame reasoners look at the total outcome profile – the payoffs in each outcome, not just for themselves, but also for the other players.

The "we"-frame makes possible what Bacharach terms *team reasoning* – reasoning based on the total outcome profile, not just the player's individual profile. The simplest form of team reasoning that he considers (*mode-P reasoning*) ranks available strategy combinations according to Pareto superiority – one strategy combination is Pareto-superior to another just if it makes at least one player better off and does not make any player worse off.[9] In the SH game from section 5.1.1, (*Stag*, *Stag*) is Pareto-superior to (*Hare*, *Hare*), and so is the choice of a mode-P reasoner.

Bacharach conjectures that mode-P and other forms of team reasoning are likely to be engaged in interactions with the following characteristics.

> **Common interests**
> There are at least two outcomes such that in one the interests of both players are better served than in the other.

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

Strong interdependence
Each player perceives that they will do well only if the other does something not guaranteed by standard, best-response reasoning (and they perceive that the other player perceives the same thing, etc.).

Typical social dilemmas such as PD, SH, and Chicken have both characteristics, and so prime for the "we"-frame.

Bacharach's account is suggestive. However, Pareto optimality is not the most useful concept in this context. In a zero-sum game (such as the widely studied Ultimatum game where players have to agree on how to divide a good), every strategy-pair is Pareto-optimal. Relatedly, Pareto-optimality is inconsistent with any form of fairness-based redistribution.

Moreover, Bacharach's account only answers the descriptive question and has nothing to say about the normative question of how or why it might be rational to adopt one frame rather than another. Discussions of social dilemmas in game theory often incorporate their own framing effects. Joint action is typically framed as cooperative, collaborative, and desirable. Of course, though, cooperation is not always desirable. The "we"-frame is adopted by genocidal mobs as well as by altruists. Bacharach, however, explicitly steers clear of normative questions.

5.3 Quasi-cyclical preferences in game theory

A more nuanced approach to frames in game theory would open up a space for reasoning across frames. As we will seem, such reasoning can be conceptualized through rational framing effects.

Reasoning across frames might seem impossible, particularly for someone who thinks that the "I"-frame and the "we"-frame are at bottom incommensurable. That belief is promoted by the framing effect just referred to, as reflected by the standard terminology in PD. The outcome of individualistic reasoning labeled as mutual defection and the outcome of team reasoning labeled as mutual cooperation. From an individualistic perspective, the optimal outcome is often described as being a free rider, receiving the benefits without taking the costs. This is all loaded terminology, and it is easy to see why the contrast between the "I"-frame and the "we"-frame might come across as a contrast between selfishness and cooperation.

Against this I suggest that, while frames are symbiotically connected to values, it is possible for different frames to express the same value in a way that provides an *anchoring point* for instrumental reflection. Here is an illustration from the game known as Chicken, showing how the value of fairness can lead a player to reason her way from the "I"-frame to the "we"-frame.

Like many two-person games, Chicken can be framed in multiple ways. Sometimes it is framed as a Hawk-Dove game, of the type made famous in the film *Dr. Strangelove*. It can also be framed as what is sometimes called the Snowdrift game. Two people are stranded by a snowdrift in their car. Each can either *Stay Inside* or leave the warmth of the car and *Dig Snow*, yielding the following payoff table.

		Column	
Row		*Stay inside*	*Dig snow*
	Stay inside	0, 0	4, 1
	Dig snow	1, 4	2, 2

There are two, pure-strategy Nash equilibria (*Stay inside*, *Dig snow*) and (*Dig snow*, *Stay inside*), as well as a mixed-strategies equilibrium in which each player plays *Stay inside* with probability 2/3 and *Dig snow* with probability 1/3.

From the perspective of the "I"-frame, Row ranks the outcomes as follows – in descending order of preference:

(1_I) Stay inside, Dig snow
(2_I) Dig snow, Dig snow
(3_I) Dig snow, Stay inside
(4_I) Stay inside, Stay inside

Column's "I"-frame ranking is the same, but with (1) and (3) reversed.

The "we"-frame ranking is the same for both, assuming that both are motivated by considerations of fairness:

(2_{WE}) Dig snow, Dig snow
(1_{WE}) Stay inside, Dig snow
=
(3_{WE}) Dig snow, Stay inside
(4_{WE}) Stay inside, Stay inside

Game theorists typically take preferences as given, but it is reasonable to ask where they come from. Row ranks (2_I) over (3_I). Why? Presumably partly because (2_I) is fairer – both are sharing in the work of digging out the snowdrift. But, as Row reflects on this, he may well see that his preferred outcome (1_I) is no less unfair. Then, as considerations of fairness start to take hold, Row has a compelling reason to adopt the "we"-frame and to prefer the fair strategy-pair over all others. This leaves him with the quasi-cyclical preferences $(2_{WE}) > (1_I) > (2_I)$. These preferences can be perfectly rational, because they are arrived at through a process of instrumental reasoning anchored in a frame-neutral value.

This is highly schematic, of course. But it is certainly consistent with the extensive experimental literature on Ultimatum games (introduced in Güth, Schmittberger, & Schwarze, 1982). In an Ultimatum game one player proposes a division of some good (typically a sum of money), which the second player can either accept or reject (in which case neither player receives anything). On standard models of economic rationality, a rational player should accept any non-zero offer. It is a very robust result, though, that most players make offers in the 40–50% range, which are typically accepted. As the offers diminish the rejection rate increases dramatically (as reviewed in Camerer, 2003, Ch. 2). The standard explanation is that considerations of fairness drive the effect (Kahneman, Knetsch, & Thaler, 1986).

Again, this provides support for H1 through H3. Games are schematic representations of complex and multi-faceted interactions (H2), in which it is rational to have quasi-cyclical preferences (H3). The different framings operate by making one dimension of the decision problem particularly salient (per H1) – the individual payoffs in the "I"-frame and the joint payoffs in the "we"-frame.

6. Application 3: Framing and quasi-cyclical preferences in interpersonal conflicts

The two previous applications have illustrated how reasoning is possible within and across frames. Frames can be tools for rational problem-solving, not just primes or nudges. The final application offers a more overarching perspective and suggests that frame-based reasoning can be deployed to overcome discursive deadlock,

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

both public and private. As we will see, this creates, and indeed requires, rational framing effects.

6.1 Discursive deadlock as a clash of frames

Discursive deadlock, where ordinary techniques for dispute resolution and collective decision-making fail, is a characteristic of contemporary private and public discourse. Political and social commentators often wring their hands about partisan deadlock and polarization on what are euphemistically called "values issues." These "values issues" are particularly susceptible to framing.

Gun control and gun safety are different ways of framing the same thing (restrictions on gun ownership). The right to life is not just in conflict with the right to choose, but also frames the issue of abortion very differently. Many of those who support taxing inheritances would steer well clear of support for a death tax. That there are many comparable examples, and that every hot-button issue lends itself to multiple framings, is completely unsurprising, in view of H2. These are all complex, multifaceted, and multi-dimensional issues that seem difficult, if not impossible, to subsume under a single attribute/dimension.

In fact, it can be useful to think of discursive deadlock in terms of clashes of frames, rather than clashes of values. This recognizes the possibility that a single value can underlie discursive deadlock. So, for example, some forms of deadlock about taxation can be seen as clashes between two different ways of framing the value of fairness – between fairness as equality and fairness as equity, for example (as distinguished in Deutsch [1975] and, at much greater length, in Rawls [1971, 2001]). Fairness as equality suggests that a fair system of taxation will tax all equally (as in various types of poll tax). Fairness as equity suggests that a fair system of taxation will tax in proportion to income (or wealth more generally). If the conflict is at root a conflict about the nature of fairness, then it would seem in principle more tractable than if it is due to clashes between fundamentally different and opposed values. (Studies show, moreover, that perceptions of the fairness of distributions are themselves subject to experimentally induced framing effects. Gamliel & Peer [2006, 2010] found that non-egalitarian distributions are judged fairer in positive frames [see Diederich, 2020 for an overview]. More generally, frames are more susceptible to scrutiny, debate, and modification than values [Lakoff, 2004]. It seems to be frames all the way down!)

But how might this type of debate and conflict resolution take place? Frame-based reasoning involves a range of skills and techniques that are recognizably similar to skills and techniques that have been studied, generally independently of each other, in different areas of social, clinical, and developmental psychology. What has hitherto been neglected, however, is the role of frames and rational framing effects.

6.2 Frame-based reasoning across discursive deadlock: Basic elements

Imagine that you are locked in a fundamental and seemingly intractable disagreement with someone. At bottom it is a clash of frames. You both agree on the facts but frame them in different and incompatible ways. The facts might concern the biological development of an embryo, the consequences of widespread gun ownership, or the statistics of wealth inequality. The clash of frames might take a familiar form: The right to life versus the denial of the right to choose; gun control versus gun safety; death tax versus inheritance tax.

It is highly unlikely that there will be any general algorithm or set of techniques for resolving this type of dispute in every case. Nonetheless, when there is a resolution, it will involve the participants going through something like the following stages of frame-sensitive reasoning.

6.2.1 Reflexive decentering

Making any progress in resolving clashes of frames requires appreciating that that is what they are. So, the first step is for the participants to turn their attention from the first-level issue on which they are deadlocked and focus instead on how they are each framing that issue. Each needs to step outside their own framing in order to reflect on the frame itself.

6.2.2 Imaginative simulation

Once model frame-sensitive reasoners have used reflexive decentering to appreciate the frame-relativity of their perspective, a next step is to be open to different ways of framing the issue. This second aspect of frame-sensitive reasoning is in effect an exercise in simulation. Frame-sensitive reasoners need to imagine what it would be like to frame things completely differently, and then to simulate actually being in that frame.

6.2.3 Perspectival flexibility

Frame-sensitive reasoning requires being able to hold multiple frames in mind at once, which is how quasi-cyclical preferences can arise. This is because (as explained in sect. 3.4) a rational preference must be held for a reason and different frames bring different reasons into play. Evaluating those reasons and seeing how they interact requires being able to adopt multiple frames simultaneously.

6.2.4 Reason construction and analysis

Reflection across frames involves a decision-maker appreciating how different frames bring different reasons into play. This can happen in multiple ways, for example,

- By foregrounding one reason-giving feature while downplaying another – as emerged dramatically in the Agamemnon and Macbeth examples.
- By highlighting a reason-giving similarity to some other (appropriately framed) action or outcome. For example, GM foods can be viewed (1) under the husbandry frame, priming the similarity to thousands of years of agricultural selective breeding, or (2) under the monopoly capitalism frame, priming the similarity to various types of predatory behavior by large corporations.
- By expressing a particular value in a particular way. The community charge frame for local authority taxation in Great Britain in the 1980s emphasized fairness as equality (if the tax is a charge on services, then it is fair to charge everyone the same for equal access to services). The alternative framing, as a poll tax, highlighted its inequitable dimension.

As we will see, appreciating the force of these competing, frame-relative reasons can directly lead to quasi-cyclical preferences and rational framing effects.

The general framework is depicted in Figure 4.

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

6.3 Rational framing effects in frame-based reasoning

While none of the skills, techniques, and abilities in Figure 4 has been directly studied, they are analogs and developments of phenomena that have been well studied in the cognitive and behavioral sciences (Bermúdez, 2020a, Ch. 11). Looking at them in more detail brings out how they create and depend upon rational framing effects. One cannot engage in the type of frame-sensitive reasoning that will break discursive deadlock without exposing oneself to rational framing effects.

6.3.1 Reflexive decentering and the clinical psychology of decentering

Within clinical psychology decentering is a shift in one's experiential perspective on the world, away from being immersed in one's experience of other people, oneself, and the world toward being able to reflect upon the experience as if from outside it. A fundamental technique in cognitive-behavioral therapy (CBT) is *cognitive distancing*, stepping back from one's own thoughts in order to reflect upon them as psychological events (as opposed to direct guides to the nature of the world and the nature of one's self) (Butler, Chapman, Forman, & Beck, 2006; Kazantzis et al., 2018). *Self-distancing* is a related concept (Kross & Ayduk, 2011).

Bernstein et al. (2015) and Bernstein, Hadash, and Fresco (2019) propose a model on which decentering emerges from three interrelated psychological processes:

> Meta-awareness:
>
> To be meta-aware of an episode of thinking is to be aware of the process of thinking itself (as opposed to its content, what it is about).
>
> Disidentification from internal experience:
>
> This is the experience of internal states as separate from oneself, in contrast to the human tendency to identify with subjective experience and to experience internal states such as thoughts, emotions, and sensations as integral parts of the self.
>
> Reduced reactivity to thought content:
>
> Decentering reduces the affective power of one's thoughts – for example, thinking of oneself as fat without feeling guilt, or thinking that one is being insulted without feeling violent rage.

Within frame-based reasoning, reflexive decentering has analogous components. First, frame-sensitive reasoners need to be able to shift perspective from their involved engagement with the world to the frame that is structuring that engagement. Second, just as the patient undergoing CBT learns to shift from internalizing to externalizing their feelings of their own worthlessness (the shift from "I am worthless" to "there is a feeling of worthlessness"), frame-sensitive reasoners need to put distance between themselves and the evaluative dimensions of the frame. With this comes, third, reduced reactivity. Dispassionate disidentification must go hand in hand with emotional and affective distancing.

6.3.2 Imaginative simulation and perspective-taking

Simulation is standardly discussed by developmental psychologists and cognitive scientists in the context of how young children acquire the complex of skills and representational abilities known as *theory of mind* or *mindreading* (see Bermúdez [2020a] for an overview with references). According to simulation theories (Carruthers & Smith, 1996; Davies & Stone, 1995a, 1995b), we make sense of other people's behavior by simulating them. We run our own decision-making processes off-line, taking as inputs the beliefs and desires that we think another person has. That process tells us what we ourselves would do if we had that person's beliefs and desires. Assuming that they will react similarly gives us a prediction for how they will behave.

Imaginative simulation in the context of frame-based reasoning is somewhat different. The assumption behind simulation theory is that we all have a relatively secure grip on our own psychologies, which we can then use to make sense of other people. This is not helpful for thinking about frame-sensitive reasoning. The whole point of reflexive decentering, as just discussed, is to weaken the grip of one's own framing of the situation in order to make room for alternative framings.

Minimally, a frame-sensitive reasoner must be able to appreciate that a single action or outcome can be apprehended from different perspectival frames. A useful analog from developmental psychology is *visual perspective-taking*, which has been studied as an aspect of how children develop mindreading skills, and in particular of how they come to understand the differences between how things appear and how they really are (Flavell, 1977; Flavell, Everett, Croft, & Flavell, 1981). Flavell distinguishes two levels of visual perspective-taking.

- At the first level, young children have a very partial understanding of visual perspective. They understand the idea of a line of sight and of an object's being occluded or not occluded.
- At the second level, in contrast, children can understand that a single object can be seen differently from different perspectives (e.g., that a picture on a table in front of them will look upside-down to an experimenter sitting opposite them) (Masangkay et al., 1974; Moll & Meltzoff, 2011).

Frame-based reasoning engages skills analogous to second-level perspective-taking – understanding that things can look different to different people even when they have similar information, because they are operating within different frames. For that reason, the transition from frame-blind reasoning to frame-sensitive reasoning is analogous to the transition between first- and second-level perspective-taking.

6.3.3 Perspectival flexibility and the theory of role-taking

Selman's theory of role-taking is a useful starting-point for thinking about how frame-sensitive reasoners need to work simultaneously across two or more frames. Selman studied how children understand and react to social situations presented in short vignettes. After cross-sectional and longitudinal interviews (Gurucharri & Selman, 1982; Selman & Byrne, 1974), Selman and collaborators came up with a hierarchy of four levels of role-taking/perspective-taking. So, for example, at the age of 10–12 they suggest that children become capable of *mutual role-taking*, simultaneously considering their own perspective and that of another, while at the same time understanding that the other person can do the same. The highest level is *social role-taking*, which incorporates the perspectives of different social groupings.

However, even a normal and socially adept decision-maker capable of all these types of perspective-taking will still fall short of the type of perspective-taking that skilled frame-sensitive reasoning requires. Frame-sensitive reasoners must be able to operate simultaneously in multiple frames, not just be aware that issues and decision problems can be multiply framed.

A Selmanian role-taker can treat people with different perspectives and frames as fixed features of the world with which she has to negotiate and, if necessary, compromise, perhaps using

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

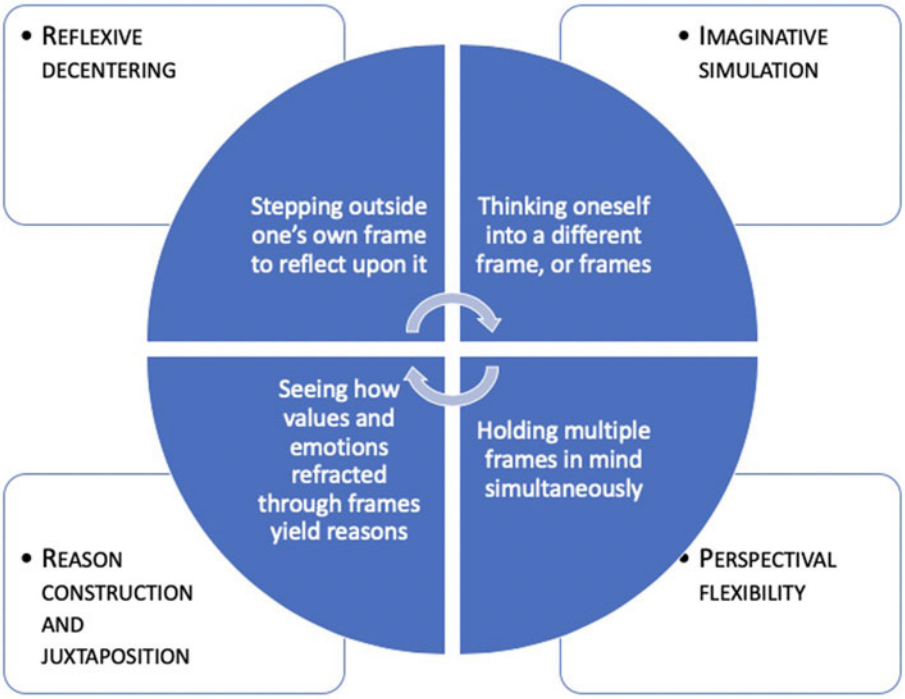


Figure 4. Key framing techniques for frame-sensitive reasoning.

strategies of principled negotiation and non-positional bargaining (e.g., depersonalize the situation; base agreement on objective criteria, etc., as proposed in Fisher & Ury [1981]). But things are very different in an example of discursive deadlock like that between, say, pro-choice and pro-life legislators trying to come to terms on regulation for abortion clinics. The decision problem is too closely bound up with participants' deepest values and sense of their own identity for depersonalizing it to be a realistic instruction. And each participant's sense of what are going to count as objective criteria is determined by their frame.

So, to tackle discursive deadlock it is not enough for a frame-sensitive reasoner simply to understand that a particular action or outcome can be framed in multiple ways. She needs to frame it herself in multiple ways simultaneously, and the mechanisms by which this might take place are ripe for further study. Perhaps the process is not strictly speaking simultaneous but better understood as switching very quickly from one frame to another, because of the bandwidth issues discussed in Chater (2018) and revealed by well-known inattentional blindness phenomena (Mack & Rock, 1998).

It is at this point that rational framing effects can enter the picture. Different ways of framing, say, restrictions on gun ownership, are associated with different preferences. For that reason, someone who internalizes the competing frames that give rise to discursive deadlock will often end up with quasi-cyclical preferences. Even if I know that "gun control" and "gun safety" are different ways of framing legal restrictions on gun ownership, properly internalizing those different frames requires me simultaneously to prefer, say, individual freedom from interference to gun *control*, but gun *safety* to individual freedom from interference.

6.3.4 Reason construction

A rational choice is a choice based on a rational preference, and a rational preference is grounded in a reason. So, how does a rational frame-sensitive reasoner extract reasons from frames? First by understanding the perspectival nature of individual frames – extracting the values they express, and the emotions that drive them. And then by extracting reasons from the frames and making comparisons within and across frames.

Frames are often embedded in narratives, which are themselves constructed in particular ways (see, e.g., Schiller [2019] on narratives in economics). Those narratives can be reason-giving. For example, the construction of a pipeline (the Keystone Pipeline, e.g., or the Trans Mountain Pipeline) can be embedded in multiple different narratives. On one narrative the pipeline might be framed as part of a steady evolution toward energy independence and freedom from dependence on Middle Eastern oil. On another, the pipeline is a further step in raising standards of living by creating jobs and lowering fuel prices. A third narrative might see the pipeline as another step in the lengthy process of dispossessing Native American peoples of their land and heritage, while on a fourth narrative the pipeline is a further increase in environmental damage and environmental risk. Each of these narratives brings with it a different set of reasons.

Here, as before, a reasoner who has succeeded in internalizing multiple narratives, and the corresponding reasons, can easily find themselves with quasi-cyclical preferences. So, for example, I might prefer energy-independence to dependence on Middle Eastern oil, while at the same preferring safeguarding the environment to running the risk of environmental catastrophe. I know, of course, that choosing to safeguard the environment is choosing not to reduce dependence on Middle Eastern oil.

7. Conclusion

The orthodox view in the cognitive and behavioral sciences is that all framing effects are irrational. This paper has argued against the orthodox view, proposing to shift the debate away from the

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

experimentally induced framing effects familiar from the "rationality wars" and toward more complicated situations where agents and decision-makers are well aware that they are framing a single action or outcome in different ways. Such situations can give rise to *quasi-cyclical preferences* (where A is preferred to B under one frame, but B preferred to A when one or both is framed differently). The paper has provided support from across the social, cognitive, and behavioral sciences for three hypotheses:

(H1) Frames and framing factor into decision-making by one dimension/attribute/value of the decision problem highly salient, thereby driving a particular response.

(H2) Framing effects associated with quasi-cyclical preferences are likely to be found in decision problems that are sufficiently complex and multi-faceted that they cannot be subsumed under a single dimension/attribute/value. What happens is that different frames prime different responses. The decision-maker is aware of this without being able to resolve the conflict either by ignoring one frame or by subsuming them both under a higher, overarching frame.

(H3) Framing effects and quasi-cyclical preferences will be rational in circumstances where it is rational to have a complex and multi-faceted response to a complex and multi-faceted situation.

In addition to shedding light on the three focus areas of self-control, game theory, and discursive deadlock, I hope in this paper to have shaken the grip of purely extensional approaches to reasoning and rationality.

Financial support. This work was supported by the American Council for Learned Societies (Fellowship 2018–2019); the National Endowment for the Humanities (Summer Stipend 2018); and the Philosophy and Psychology of Self-Control Project at Florida State University, funded by the Templeton Foundation.

Conflict of interest. None.

Notes

1. For meta-analyses and replications see Kühberger (1998); Piñon and Gambara (2005); and Steiger and Kühberger (2018), all of which confirm the effect. For individual differences and design issues see Mahoney, Buboltz, Levin, Doverspike, and Svyantek (2011). While the original study was a between-subjects design, it has been replicated in a within-subjects design (Diederich, Wyszynski, & Ritov, 2018; Frisch, 1993; Kühberger, 1995). While the effects are robust, not all are convinced that they are genuine *framing* effects. Mandel rejects the assumption that any rational person would recognize the description of 200 people being saved out of 600 as being equivalent to 400 people dying, suggesting instead that linguistic expressions are generally understood imprecisely (Mandel, 2014). It has been claimed that the effect disappears when the frames are presented with the numbers made precise (i.e., *exactly* 200 people will die) – but see Simmons and Nelson (2013) and Chick, Reyna, and Corbin (2016). The experiments reported in Tombu and Mandel (2015), however, may point toward a different interpretation, as they found that framing effects were accompanied by shifts in how the participants evaluated the riskiness of options. In any event, it seems unlikely (and nor does Mandel claim) that all of the framing effects will be explained in this way.
2. Sher and McKenzie have objected that the different options are not *informationally equivalent* (Sher & McKenzie, 2006, 2008, 2011). One frame can bring into play factors not salient in the other. E.g., labeling meat as 75% lean implies that the proportion presented is higher than normal (relative to an implicit *reference point*). It seems unlikely, though, that all valence-consistent framing effects will be explicable in the same way.
3. Well-known statements of the irrationality thesis include Stich (1990) and, more recently, Ariely (2008), with spirited arguments to the contrary in Cohen (1981) (see also the extensive commentary to that BBS target paper) and significant reinterpretations of key data points from the fast and frugal heuristics movement (Gigerenzer & Gaissmaier, 2011; Gigerenzer & Selten, 2001) and from the rational analysis approach (Oaksford & Chater, 2007).
4. Typically, but not always. See Frisch (1993) and Mandel (2014), both of whom report many subjects refusing to accept that the two situations really are equivalent.
5. There are dissenting voices. In philosophy, Larry Temkin has argued against transitivity as a normative requirement (Temkin, 1987, 1996). In economics, Mandler (2005) suggests that rational agents can choose intransitively when their preferences are incomplete.
6. Cf. Schick (1997) for a different analysis in a somewhat similar spirit. Schick emphasizes Macbeth's different "understandings," but does not engage with quasi-cyclicality. Schick was a pioneer in criticizing extensionalist approaches to decision-making (Schick, 1991, 1997, 2003). For differences between his approach and that presented here see Bermúdez (2020a, Ch. 5).
7. See Hagger et al. (2016) for a large-scale multi-lab study that failed to find replication, and Baumeister, Vonasch, and Sjastad (2020) for objections, as well as doubts about the statistical validity of some meta-analyses that seem to support the model (Carter, Kofler, Forster, & McCullough, 2015).
8. The basic results are robust (Watts, Duncan, & Quan, 2018), although questions have been raised about how performance on the marshmallow test predicts long-term outcomes, such as success on standardized tests, but those longitudinal claims are orthogonal to this paper (Michaelson & Munakata, 2020).
9. Bacharach's more complicated notion of *circumspect team reasoning* incorporates a player's estimate of the probability that other players are team reasoners. See Bacharach (2006) and Bermúdez (2020a, Ch. 8) for discussion.

References

Adolphs, R., & Anderson, D. J. (2018). *The neuroscience of emotion: A new synthesis*. Princeton University Press.
Ainslie, G. (2001). *Breakdown of will*. Cambridge University Press.
Ainslie, G. W. (1974). Impulse control in pigeons. *Journal of the Experimental Analysis of Behavior*, *21*(3), 485–489. doi: 10.1901/jeab.1974.21-485
Ainslie, G. W. (1992). *Picoeconomics: The strategic interaction of successive motivational states within the person*. Cambridge University Press.
Ariely, D. (2008). *Predictably irrational: The hidden forces that make our decisions*. Harper Collins.
Bacharach, M. (2006). *Beyond individual choice: Teams and frames in game theory*. Princeton University Press.
Barberis, N., & Huang, M. (2006). The loss aversion/narrow framing approach to the equity premium puzzle. In R. Mehra (Ed.), *Handbook of the equity risk premium* (pp. 199–236). Elsevier.
Barberis, N., & Xiong, W. E. I. (2009). What drives the disposition effect? An analysis of a long-standing preference-based explanation. *The Journal of Finance*, *64*(2), 751–784. doi: 10.1111/j.1540-6261.2009.01448.x
Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology*, *74* (5), 1252–1265. doi: 10.1037//0022-3514.74.5.1252
Baumeister, R. F., Vonasch, A. J., & Sjastad, H. (2020). The long reach of self-control. In A. Mele (Ed.), *Surrounding self-control* (pp. 17–46). Oxford University Press.
Bechara, A., Damasio, H., Damasio, A. R., & Lee, G. P. (1999). Different contributions of the human amygdala and ventromedial prefrontal cortex to decision-making. *Journal of Neuroscience*, *19*(13), 5473–5481.
Bermúdez, J. L. (2009). *Decision theory and rationality*. Oxford University Press.
Bermúdez, J. L. (2015a). Strategic vs. parametric choice in Newcomb's problem and the prisoner's dilemma: Reply to Walker. *Philosophia*, *43*, 787–794. doi: 10.1007/s11406-015-9606-6.
Bermúdez, J. L. (2015b). Prisoner's dilemma cannot be a Newcomb problem. In M. Peterson (Ed.), *The prisoner's dilemma* (pp. 115–132). Cambridge University Press.
Bermúdez, J. L. (2018). *Self-control, decision theory, and rationality: New essays (J. L. Bermúdez Ed.)*. Cambridge University Press.
Bermúdez, J. L. (2020a). *Frame it again: New tools for rational thought*. Cambridge University Press.
Bermúdez, J. L. (2020b). Framing as a mechanism for self-control: Rationality and quasi-cyclical preferences. In A. Mele (Ed.), *Surrounding self-control* (pp. 361–383). Oxford University Press.

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

Bernstein, A., Hadash, Y., & Fresco, D. M. (2019). Metacognitive processes model of decentering: Emerging methods and insights. *Current Opinion in Psychology*, *28*, 245–251. doi: 10.1016/j.copsyc.2019.01.019

Bernstein, A., Hadash, Y., Lichtash, Y., Tanay, G., Shepherd, K., & Fresco, D. M. (2015). Decentering and related constructs: A critical review and metacognitive processes model. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *10*(5), 599–617. doi: 10.1177/1745691615594577

Bernstein, L. M., Chapman, G. B., & Elstein, A. S. (1999). Framing effects in choices between multioutcome life-expectancy lotteries. *Medical Decision Making*, *19*(3), 324–338. doi: 10.1177/0272989x9901900311

Brewer, M. B., & Kramer, R. M. (1986). Choice behavior in social dilemmas: Effects of social identity, group size, and decision framing. *Journal of Personality and Social Psychology*, *50*, 543–549.

Butler, A. C., Chapman, J. E., Forman, E. M., & Beck, A. T. (2006). The empirical status of cognitive-behavioral therapy: A review of meta-analyses. *Clinical Psychology Review*, *26*(1), 17–31. doi: 10.1016/j.cpr.2005.07.003

Camerer, C. F. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

Carruthers, P., & Smith, P. K. (Eds.) (1996). *Theories of theory of mind*. Cambridge University Press.

Carter, E. C., Kofler, L. M., Forster, D. E., & McCullough, M. E. (2015). A series of meta-analytic tests of the depletion effect: Self-control does not seem to rely on a limited resource. *Journal of Experimental Psychology: General*, *144*(4), 796–815. doi: 10.1037/xge0000083

Chater, N. (2018). *Mind is flat*. Penguin Books.

Chick, C. F., Reyna, V. F., & Corbin, J. C. (2016). Framing effects are robust to linguistic disambiguation: A critical test of contemporary theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(2), 238–256. doi: 10.1037/xlm0000158

Cohen, L. J. (1981). Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences*, *4*, 317–370.

Damasio, A. (1994). *Descartes's error: Emotion, reason, and the human brain*. Putnam/Grosset.

Davies, M., & Stone, T. (Eds.). (1995a). *Folk psychology*. Basil Blackwell.

Davies, M., & Stone, T. (Eds.). (1995b). *Mental simulation*. Basil Blackwell.

Deutsch, M. (1975). Equity, equality, and need: What determines which value will be used as the basis of distributive justice? *Journal of Social Issues*, *31*(3), 137–149. https://doi.org/10.1111/j.1540-4560.1975.tb01000.x

Diederich, A. (2020). Identifying needs: The psychological perspective. In S. Traub & B. Kittel (Eds.), *Need-Based distributive justice* (pp. 59–89). Springer.

Diederich, A., Wyszynski, M., & Ritov, I. (2018). Moderators of framing effects in variations of the Asian disease problem: Time constraint, need and disease type. *Judgment and Decision Making*, *13*(6), 529–546.

Evans, J. S. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*, 255–278. doi: 10.1146/annurev.psych.59.103006.093629

Fagles, R. (1977). *The Oresteia (Aeschylus)*. Penguin Classics.

Figner, B., Knoch, D., Johnson, E. J., Krosch, A. R., Lisanby, S. H., Fehr, E., & Weber, E. U. (2010). Lateral prefrontal cortex and self-control in intertemporal choice. *Nature Neuroscience*, *13*(5), 538–539. doi: 10.1038/nn.2516

Fisher, R., & Ury, W. L. (1981). *Getting to yes: Negotiating agreement without giving in*. Penguin.

Flavell, J. H. (1977). The development of knowledge about visual perception. *Nebraska Symposium on Motivation*, *25*, 43–76.

Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young children's knowledge about visual perception: Further evidence for the level 1–level 2 distinction. *Developmental Psychology*, *17*(1), 99–103. doi: 10.1037/0012-1649.17.1.99

Frisch, D. (1993). Reasons for framing effects. *Organizational Behavior and Human Decision Processes*, *54*(3), 399–429. doi: 10.1006/obhd.1993.1017

Gamliel, E., & Peer, E. (2006). Positive versus negative framing affects justice judgments. *Social Justice Research*, *19*(3), 307–322. doi: 10.1007/s11211-006-0009-5

Gamliel, E., & Peer, E. (2010). Attribute framing affects the perceived fairness of health care allocation principles. *Judgment and Decision Making*, *5*(1), 11–20.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*(1), 451–482. doi: 10.1146/annurev-psych-120709-145346

Gigerenzer, G., & Selten, R. (2001). *Bounded rationality: The adaptive toolbox*. MIT Press.

Gurucharri, C., & Selman, R. L. (1982). The development of interpersonal understanding during childhood, preadolescence, and adolescence: A longitudinal follow-up study. *Child Development*, *53*(4), 924–927. doi: 10.2307/1129129

Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, *3*(4), 367–388. https://doi.org/10.1016/0167-2681(82)90011-7

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., … Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, *11*(4), 546–573. doi: 10.1177/1745691616652873

Hare, T. A., Camerer, C. F., & Rangel, A. (2009). Self-control in decision-making involves modulation of the vmPFC valuation system. *Science (New York, N.Y.)*, *324*(5927), 646–648. doi: 10.1126/science.1168450

Harsanyi, J. (1977). Advances in understanding rational behavior. In R. E. Butts & J. Hintikka (Eds.), *Foundational problems in the special sciences* (pp. 315–343). D. Reidel.

Harsanyi, J., & Selten, R. (1988). *A general theory of equilibrium selection in games*. MIT Press.

Heuer, L., & Orland, A. (2019). Cooperation in the prisoner's dilemma: An experimental comparison between pure and mixed strategies. *Royal Society Open Science*, *6*(7), 182142. doi: 10.1098/rsos.182142

Holton, R. (2009). *Willing, wanting, waiting*. Oxford University Press.

Janssen, M. A. (2008). Evolution of cooperation in a one-shot prisoner's dilemma based on recognition of trustworthy and untrustworthy agents. *Journal of Economic Behavior & Organization*, *65*(3), 458–471. https://doi.org/10.1016/j.jebo.2006.02.004

Jeffrey, R. (1983). *The logic of decision* (2nd ed.). University of Chicago Press.

Kahneman, D. (2011). *Thinking fast and thinking slow*. Farrar, Straus, and Giroux.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1986). Fairness and the assumptions of economics. *The Journal of Business*, *59*(4), S285–S300. http://www.jstor.org/stable/2352761

Kazantzis, N., Luong, H. K., Usatoff, A. S., Impala, T., Yew, R. Y., & Hofmann, S. G. (2018). The processes of cognitive behavioral therapy: A review of meta-analyses. *Cognitive Therapy and Research*, *42*(4), 349–357. doi: 10.1007/s10608-018-9920-y

Keeney, R. L., & Raiffa, H. (1976). *Decisions with multiple objectives: Preferences and value trade-offs*. Cambridge University Press.

Keltner, D., & Lerner, J. S. (2010). Emotion. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of social Psychology* (pp. 317-352). Wiley.

Kross, E., & Ayduk, O. (2011). Making meaning out of negative experiences by self-distancing. *Current Directions in Psychological Science*, *20*, 187–191.

Kühberger, A. (1995). The framing of decisions: A new look at old problems. *Organizational Behavior and Human Decision Processes*, *62*(2), 230–240. http://dx.doi.org/10.1006/obhd.1995.1046

Kühberger, A. (1998). The influence of framing on risky decisions: A meta-analysis. *Organizational Behavior and Human Decision Processes*, *75*(1), 23–55. doi: 10.1006/obhd.1998.2781

Lakoff, G. (2004). *Don't think of an elephant: Know your values and frame the debate – The essential guide for progressives*. Chelsea Green.

Lerner, J. S., & Keltner, D. (2000). Beyond valence: Toward a model of emotion-specific influences on judgement and choice. *Cognition and Emotion*, *14*(4), 473–493. doi: 10.1080/026999300402763

Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. *Annual Review of Psychology*, *66*(1), 799–823. doi: 10.1146/annurev-psych-010213-115043

Levin, I. P., & Gaeth, G. J. (1988). How consumers Are affected by the framing of attribute information before and after consuming the product. *Journal of Consumer Research*, *15*(3), 374–378. doi: 10.1086/209174

Levin, I. P., Schneider, S. L., & Gaeth, G. J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. *Organizational Behavior and Human Decision Processes*, *76*(2), 149–188.

Linville, P. W., Fischer, G. W., & Fischoff, B. (1993). AIDS risk perception and decision biases. In J. B. Pryor & G. D. Reeder (Eds.), *The social psychology of HIV infection* (pp. 5–38). Lawrence Erlbaum.

Loewenstein, G., & Lerner, J. S. (2003). The role of affect in decision making. In R. Davidson, H. Goldsmith, & K. Scher (Eds.), *Handbook of affective science* (pp. 619–642). Oxford University Press.

Mack, A., & Rock, I. (1998). *Inattentional blindness*. MIT Press.

Magen, E., Dweck, C. S., & Gross, J. J. (2008). The hidden-zero effect: Representing a single choice as an extended sequence reduces impulsive choice. *Psychological Science*, *19*(7), 648–649. doi: 10.1111/j.1467-9280.2008.02137.x

Magen, E., Kim, B., Dweck, C. S., Gross, J. J., & McClure, S. M. (2014). Behavioral and neural correlates of increased self-control in the absence of increased willpower. *Proceedings of the National Academy of Sciences*, *111*(27), 9786–9791. doi: 10.1073/pnas.1408991111

Mahoney, K. T., Buboltz, W., Levin, I. P., Doverspike, D., & Svyantek, D. J. (2011). Individual differences in a within-subjects risky-choice framing study. *Personality and Individual Differences*, *51*(3), 248–257. https://doi.org/10.1016/j.paid.2010.03.035

Mandel, D. (2014). Do framing effects reveal irrational choices? *Journal of Experimental Psychology: General*, *143*, 1185–1198. doi: 10.1037/a0034207

Mandler, M. (2005). Incomplete preferences and rational intransitivity of choice. *Games and Economic Behavior*, *50*, 255–277.

Masangkay, Z. S., McCluskey, K. A., McIntyre, C. W., Sims-Knight, J., Vaughn, B. E., & Flavell, J. H. (1974). The early development of inferences about the visual percepts of others. *Child Development*, *45*(2), 357–366. doi: 10.2307/1127956

McClennen, E. F. (1990). *Rationality and dynamic choice*. Cambridge University Press.

Mehra, R. (2008). The equity premium puzzle: A review. *Foundations and Trends in Finance*, *2*, 1–81.

Mehra, R., & Prescott, E. (1985). The equity premium: A puzzle. *Journal of Monetary Economics*, *15*, 145–161.

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

Michaelson, L. E., & Munakata, Y. (2020). Same data set, different conclusions: Preschool delay of gratification predicts later behavioral outcomes in a preregistered study. *Psychological Science*, *31*(2), 193–201. doi: 10.1177/0956797619896270
Mischel, W., & Ayduk, O. (2004). Willpower in a cognitive-affective processing system. In R. F. Baumeister & K. D. Vohs (Eds.), *Handbook of self-regulation: Research, theory, and applications* (pp. 99–129). Guilford.
Mischel, W., & Baker, N. (1975). Cognitive appraisals and transformations in delay behavior. *Journal of Personality and Social Psychology*, *31*, 254–261.
Mischel, W., & Moore, B. (1973). Effects of attention to symbolically presented rewards on self-control. *Journal of Personality and Social Psychology*, *28*(2), 172–179.
Mischel, W., Shoda, Y., & Rodriguez, M. I. (1989). Delay of gratification in children. *Science (New York, N.Y.)*, *244*(4907), 933–938.
Moll, H., & Meltzoff, A. N. (2011). How does it look? Level 2 perspective-taking at 36 months of age. *Child Development*, *82*(2), 661–673. doi: 10.1111/j.1467-8624.2010.01571.x
Montier, J. (2010). *How not to be your own worst enemy: The little book of behavioral investing*. Wiley.
Neale, M. A., & Bazerman, M. H. (1985). The effects of framing and negotiator overconfidence on bargaining behaviors and outcomes. *Academy of Management Journal*, *28*, 34–49.
Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.
Paul, L. A. (2014). *Transformative experience*. Oxford University Press.
Piñon, A., & Gambara, H. (2005). A meta-analytic review of framing effect: Risky, attribute and goal framing. *Psicothema*, *17*(2), 325–331.
Pothos, E. M., Perry, G., Corr, P. J., Matthew, M. R., & Busemeyer, J. R. (2011). Understanding cooperation in the prisoner's dilemma game. *Personality and Individual Differences*, *51*(3), 210–215. https://doi.org/10.1016/j.paid.2010.05.002
Rachlin, H. (2000). *The science of self-control*. Harvard University Press.
Rachlin, H. (2018). In what sense are addicts irrational? In J. L. Bermúdez (Ed.), *Self-control, decision theory, and rationality* (pp. 147–166). Cambridge University Press.
Rawls, J. (1971). *A theory of justice*. Harvard University Press.
Rawls, J. (2001). *Justice as fairness: A restatement*. Harvard University Press.
Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*, 1161–1178.
Sally, D. (1995). Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992. *Rationality and Society*, *7*(1), 58–92. doi: 10.1177/1043463195007001004
Schick, F. (1991). *Understanding action*. Cambridge University Press.
Schick, F. (1997). *Making choices*. Cambridge University Press.
Schick, F. (2003). *Ambiguity and logic*. Cambridge University Press.
Schiller, R. (2019). *Narrative economics*. Princeton University Press.
Selman, R. L., & Byrne, D. F. (1974). A structural-developmental analysis of levels of role taking in middle childhood. *Child Development*, *45*(3), 803–806. doi: 10.2307/1127850
Shefrin, H., & Statman, M. (1985). The disposition to sell winners too early and ride losers too long: Theory and evidence. *The Journal of Finance*, *40*(3), 777–790. doi: 10.2307/2327802
Sher, S., & McKenzie, C. R. (2006). Information leakage from logically equivalent frames. *Cognition*, *101*(3), 467–494.
Sher, S., & McKenzie, C. R. (2008). Framing effects and rationality. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for Bayesian cognitive science* (pp. 79–96). Oxford University Press.
Sher, S., & McKenzie, C. R. (2011). Levels of information: A framing hierarchy. In G. Keren (Ed.), *Perspectives on framing* (pp. 35–63). Psychology Press.
Shoham, Y., & Leyton-Brown, K. (2008). *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press.
Simmons, J., & Nelson, L. (2013). "Exactly": The most famous framing effect is robust to precise wording. *Data Colada*. http://datacolada.org/11
Skyrms, B. (2012). *The Stag Hunt and the evolution of social structure*. Cambridge University Press.
Steiger, A., & Kühberger, A. (2018). A meta-analytic re-appraisal of the framing effect. *Zeitschrift für Psychologie*, *226*(1), 45–55. doi: 10.1027/2151-2604/a000321
Stich, S. (1990). *The fragmentation of reason*. MIT Press.
Strotz, R. H. (1956). Myopia and inconsistency in dynamic utility maximization. *The Review of Economic Studies*, *23*, 165–180.
Temkin, L. S. (1987). Intransitivity and the mere addition paradox. *Philosophy and Public Affairs*, *16*(2), 138–187.
Temkin, L. S. (1996). A continuum argument for intransitivity. *Philosophy and Public Affairs*, *25*(3), 175–210.
Thaler, R. H. (1999). Mental accounting matters. *Journal of Behavioral Decision Making*, *12*(3), 183–206. doi: 10.1002/(SICI)1099-0771(199909)12:3<183::AID-BDM318>3.0.CO;2-F
Thaler, R. H., & Johnson, E. J. (1991). Gambling with the house money and trying to break even: The effect of prior outcomes on risky choice. *Management Science*, *36*, 643–660.
Tombu, M., & Mandel, D. R. (2015). When does framing influence preferences, risk perceptions, and risk attitudes? The explicated valence account. *Journal of Behavioral Decision Making*, *28*(5), 464–476. doi: 10.1002/bdm.1863
Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science (New York, N.Y.)*, *211*, 453–458.
Tversky, A., & Kahneman, D. (1986). Rational choice and the framing of decisions. *The Journal of Business*, *59*(4), S251–S278.
Watts, T. W., Duncan, G. J., & Quan, H. (2018). Revisiting the marshmallow test: A conceptual replication investigating links between early delay of gratification and later outcomes. *Psychological Science*, *29*(7), 1159–1177. doi: 10.1177/0956797618761661

Open Peer Commentary

Framing is a motivated process

George Ainslie

Department of Veterans Affairs, Veterans Affairs Medical Center, Coatesville, PA 19320, USA
George.Ainslie@va.gov
www.picoeconomics.org

doi:10.1017/S0140525X22000991, e221

Abstract

Frames group choices into categories, thus modifying the incentives for them. This effect makes framing itself a motivated choice rather than a neutral cognition. In particular, framing an inferior choice with a high short-term payoff as part of a broad category of choices recruits incentive to reject it; but this must be motivated by its being a test case.

Bermudez criticizes decision theorists' "almost unanimous acceptance" (target article, sect. 2.1, para. 1) of the extensionality principle and argues instead that choosing within frames can rationally make different valuations of a given outcome incommensurable. But the "reason-giving aspects of different possible outcomes and possible actions" (target article, sect. 3.4, para. 3) are all based on reward: The prospect-theoretical and behavioral approaches he cites assume that people form cognitions in the service of getting reward, a view that cognitive theorists are also now accepting in the form of "predictive processing" (Gilead, Trope, & Liberman, 2020) – the venerable behaviorist axiom that all learnable processes are actions. People's frames are then tools for organizing their hypotheses to test options by vicarious trial and error (Redish, 2016), the outcomes of which are rewards. Rewards must in principle be commensurable (Ainslie, 1992, pp. 28–32; Montague & Berns, 2002; Shizgal & Conover, 1996), which validates the norm of extensionality.

It is true that different frames can lead to fiercely conflicting conclusions, but Macbeth and Agamemnon manage to reach decisions. One frame finally displaces the other. After extreme failures competing frames may take turns running the whole self, as in alcoholic blackouts where "the alcohol talks," or where real-life Jekylls give way to Hydes. Even such dissociation of identity is a motivated process, as shown by the struggle the normal self often experiences with impending takeover by an alter (Dell, 2009).

An analysis of frames needs to keep the motivation for them in view. In discussing self control (sect. 4) Bermúdez cites me and the late Howard Rachlin, and starts out largely as we did, with hyperbolic delay discount curves, doubt about the effectiveness of effortful willpower, and a model of choice-framing. However, he presents framing as an unmotivated process, as if Mischel's children, for instance, would not have been tempted to think of

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

marshmallows as something yummy instead of puffy clouds. This leads to trouble after he makes the unnecessary assumption that "an agent exercising self-control must change the shape of her discount function," as opposed to recruiting additional larger-later (LL) incentive by framing her choices more broadly. To explain "how counter-preferential choice can actually be made [during temptation]" he has to posit a "hot" motive of "successfully resisting SS," to which the subject "assigns more utility" than to the smaller-sooner (SS) reward itself. The notion that someone can make self-control hot – in effect an occasion for aroused emotion – by assigning it utility would seem to violate the laws of motivational gravity. If you could assign utility just anywhere, why not assign it to the LL alternative in the first place?

Admittedly, the laws of motivational gravity are still in some flux. There is reason to believe that much human reward is endogenous – a fiat currency not backed by external primary rewards (Ainslie, 2013, 2017, in press). Thus people do assign, or even create, utility, rather than just finding it. But self-created utility still has constraints. We seem to work up "hot" thought processes to indulge temptations, not to dampen them (e.g., Mischel, 2014). Bermúdez's other proposal for the value of a current LL choice makes more sense, and would not involve learning to bend what is probably an inborn discount curve (Ainslie, 1992, pp. 123–125; 2001, pp. 36–38): Resisting SS "might be taken as a signal of how the agent will react in the future (if I resist temptation here, then I will be more likely to do so in the future)" (target article, sect. 4.3, para. 5).

This is a solid reward-predicting perception, which would in fact motivate the mechanism Bermúdez did not consider, framing the choice broadly. Here is my *choice bundling* (1975, 2021), or Rachlin's *molar choice* (1995, 2016), or the *broad bracketing* of Read, Lowenstein, and Rabin (1999), or, in effect, the high-level construal of Trope and Liberman (2010). Essentially, if you make a current choice as part of a whole category of later choices, the relatively higher tails of the hyperbolic discount curves from the LL options in those choices will sum together to exceed the summed lower tails of the SS options plus the upward spike of the current SS option (Ainslie, 2001, pp. 78–85). However, a person could just as easily frame the choice at hand as standing alone, unless an SS choice would confront her with evidence that she would keep choosing SS in similar situations, and would thus lose the prospect of the wider category of rewards (Ainslie, 2001, pp. 90–100; see also Fujita, 2011, p. 360). There is always incentive to frame the choice at hand narrowly – "just this once," and this incentive fuels the search for rationalizations – ways of distinguishing the choice at hand from the broader category of choices at stake. What obliges the person to include her present choice in a broad frame is its implication as a test case. And the present LL choice is not just a *signal* – an indicator of intent – but a *sign* – behavioral evidence – of how the agent will react in the future. Given this kind of motivated framing I can again agree with Bermúdez's model.

Incentives for frames are important in Bermúdez's other applications as well. The "I" frame and the "we" frame may be easily interchangeable in experimental games, but not when they lead to major consequences in the real world (sect. 5). In section 6 he proposes cognitive re-framing as a technique for conflict resolution, by analogy to the distancing technique in cognitive behavior therapy. But while it is undoubtedly helpful for negotiators to "appreciate how different frames bring different reasons into play" (target article, sect. 6.2.4, para. 1), he neglects the possibility that framing is itself a motivated choice. The "cognitive distancing" that is the starting place for his re-framing exercises involves withdrawing what may be highly valued emotional investment in a person's stance. Drama more than deduction commits political partisans to their positions – the plights of helpless babies-to-be, for instance, versus the injustice of entrapped pregnant women. The antagonists can probably imagine the others' points of view well enough, but do not want to, and may even feel that not doing so is good impulse control.

Financial support. This material is the result of work supported with resources and the use of facilities at the Department of Veterans Affairs Medical Center, Coatesville, PA, USA. The opinions expressed are not those of the Department of Veterans Affairs or of the US Government.

Conflict of interest. None.

References

Ainslie, G. (1975). Specious reward: A behavioral theory of impulsiveness and impulse control. *Psychological Bulletin*, *82*, 463–496. http://dx.doi.org/10.1037/h0076860

Ainslie, G. (1992). *Picoeconomics: The strategic interaction of successive motivational states within the person*. Cambridge University Press.

Ainslie, G. (2001). *Breakdown of will*. Cambridge University Press. https://doi.org/10.1017/CBO9781139164191

Ainslie, G. (2013). Grasping the impalpable: The role of endogenous reward in choices, including process addictions. *Inquiry*, *56*, 446–469. http://dx.doi.org/10.1080/0020174X.2013.806129

Ainslie, G. (2017). *De gustibus disputare*: Hyperbolic delay discounting integrates five approaches to choice. *Journal of Economic Methodology*, *24*(2), 166–189. http://dx.doi.org/10.1080/1350178X.2017.1309748

Ainslie, G. (2021). Willpower with and without effort. *Behavioral and Brain Sciences*, *44*, e30. https://doi.org/10.1017/S0140525X20000357

Ainslie, G. (in press). The behavioral construction of the future. *Psychology of Addictive Behaviors*.

Dell, P. F. (2009). Understanding dissociation. In P. F. Dell & J. A. O'Neil (Eds.), *Dissociation and the dissociative disorders: DSM-V and beyond* (pp. 709–825). Routledge.

Fujita, K. (2011). On conceptualizing self-control as more than the effortful inhibition of impulses. *Personality and Social Psychology Review*, *15*, 352–366. http://dx.doi.org/10.1177/1088868311411165

Gilead, M., Trope, Y., & Liberman, N. (2020). Above and beyond the concrete: The diverse representational substrates of the predictive brain. *Behavioral and Brain Sciences*, *43*, e121, 1–74. doi: 10.1017/s0140525x19002000

Mischel, W. (2014). *The marshmallow test: Understanding self-control and how to master it*. Bantam.

Montague, P. R., & Berns, G. S. (2002). Neural economics and the biological substrates of valuation. *Neuron*, *36*, 265–284. http://dx.doi.org/10.1016/S0896-6273(02)00974-1

Rachlin, H. (1995). Self-control: Beyond commitment. *Behavioral and Brain Sciences*, *18*, 109–159. http://dx.doi.org/10.1017/S0140525X00037602

Rachlin, H. (2016). Self-control based on soft commitment. *The Behavior Analyst*, *3*, 259–268. http://dx.doi.org/10.1007/s40614-016-0054-9

Read, D., Lowenstein, G., & Rabin, M. (1999). Choice bracketing. *Journal of Risk and Uncertainty*, *19*(1), 171–197. http://dx.doi.org/10.1023/A:1007879411489

Redish, A. D. (2016). Vicarious trial and error. *Nature Reviews Neuroscience*, *17*, 147–159. http://dx.doi.org/10.1038/nrn.2015.30

Shizgal, P., & Conover, K. (1996). On the neural computation of utility. *Current Directions in Psychological Science*, *5*, 37–43. http://dx.doi.org/10.1111/1467-8721.ep10772715

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review*, *117*, 440–463. http://dx.doi.org/10.1037/a0018963

The polyphony principle

Bree Beal

Psychology and Interdisciplinary Sciences (PAIS) Building, Emory University, Atlanta, GA 30322, USA
BreeLBeal@gmail.com
https://www.bree-beal.com/

doi:10.1017/S0140525X2200108X, e222

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

Abstract

Bermúdez's "rational framing effects" are consequences of a counterintuitive phenomenon that I call "normative polyphony": the reality that a single action may, with logical consistency, sustain diverse positive and negative judgments. I show that normative polyphony emerges from "ontological polyphony" – that is, diverse possible framings of relevant details – and illustrate this "polyphony principle" through a reading of Dostoevsky's (1993) *Crime and Punishment*.

Introduction

Bermúdez delivers a compelling argument for the existence of "rational framing effects" and a promising proposal for using frame-sensitive education to reduce polarization. I have no criticism of this brilliant work. Instead, I'll expand upon Bermúdez's insight and show that a rational framing effect is a special consequence of a more widespread phenomenon that I'll call "normative polyphony": the reality that an objectively single action may, with logical consistency, sustain diverse positive and negative judgments. Beyond its potential to explain rational framing effects, the described "polyphony principle" presents a special challenge for theory in moral psychology.

I'll elaborate my thesis by analyzing a passage in Dostoevsky's (1993) *Crime and Punishment* – the story of an impoverished young man named Raskolnikov who develops a moral rationale for committing murder, but who later finds he is unable to live with his guilt. In the novel's climax, Raskolnikov confesses his crime to Sonya – an unlikely ally and close friend of one of the victims. By my count, Raskolnikov offers seven distinct explanations for the crime, responding to pushback from Sonya, who is unable to reconcile who he is with what he has done. In some of these accounts, Raskolnikov justifies his action, while in others he condemns himself. Yet, the remarkable thing about these incompatible confessions is that each is internally coherent. I show below that this juxtaposition of internal coherence against external incompatibility is possible because Raskolnikov's normative evaluations are each supported by a unique narrative that selects relevant details from a host of competing characterizations, motives, contextual factors, counterfactuals, and so on.

The takeaway is this: A subjectively complex situation supporting diverse framings of the relevant details (i.e., ontological polyphony) can logically support a diverse polyphony of normative judgments (i.e., normative polyphony). As the ontological framing shifts, so may the sense of what is normatively best. This translation from ontological polyphony into normative polyphony is precisely what Bermúdez points to in arguing for rational framing effects – in, for instance, his examples from classic literature. However, the polyphony principle carries additional implications, especially for the field of moral psychology. We cannot adequately characterize the complexity of moral cognition as current models do, by appealing to value-pluralism (see Beal, 2020). And we must move beyond the vague claim that people are imperfectly rational. There is an internal rationality to our moral inconsistencies, expressed in the protean logic of the polyphony principle.

Illustration

Bakhtin (1984) developed the concept of "polyphonic" (i.e., "many-voiced") art to characterize a distinctive feature of Dostoevsky's work. Whereas in "dialectical" art, different characters express viewpoints that stand in some implicit relation to a unitary idea of *the* truth, Dostoevsky's characters voice diverse truths – that is, complexes of beliefs and evaluations that each hang together coherently, despite the fact that they are inconsistent with other equally coherent accounts. And readers even find that polyphony exists intrapersonally: Contradictory ideas are sometimes expressed by a single character with such rigor that each idea is internally coherent. Thus, Dostoevsky suggests that humans live in polyphonic worlds and regularly navigate situations with multiple, sometimes contradictory rights and wrongs.

Although Bakhtin did not slice his idea more finely, Dostoevsky actually illustrates a relation between two distinct kinds of polyphony: Ontological (descriptive) polyphony supports normative (prescriptive) polyphony. This observation may raise an objection: How can one derive normative prescription from mere ontological description, inferring "ought" from "is"? The answer is what Shweder (1992) called the reality principle: Sufficient information is often implicitly encoded into mere descriptive statements such that normative prescriptions can "in a straightforward way, be derived" from them – an argument sustained both empirically and anecdotally (e.g., Much & Shweder, 1978). Elsewhere, I've added that we do not even need explicit description, but spontaneously derive a normative sense of what is appropriate from certain perceived details (Beal, 2021; Beal & Gogia, 2021). Without rehashing the finer points of these related arguments, I'll point out that neither involves an illogical conflation of "is" with "ought." Rather, the conditions for the emergence of the normative "ought" are contained within the ontological "is" – whether at the level of description or perception. The ubiquity of this logic is evident in every dispute where opponents ground their conflicting normative positions in divergent beliefs, interpretations, or framings of the details.

Through the confession, Dostoevsky illustrates this constitutive relationship between ontological and normative polyphony in intricate detail. As Raskolnikov's accounts change, Sonya takes on the aspect of a "holy fool," a judge or priest, an enemy or competitor, a friend or lover, a proxy for one of the victims, and a child. In parallel, Raskolnikov's self-portrait morphs from a defender of the innocent to a petty criminal to a desperate victim to a spiteful narcissist to a visionary great man to a moral anarchist to a bug in just 12 pages. And these ontological reframings, both at the levels of description and perception, straightforwardly yield the normative quality of each confession: As a defender of the innocent and a believer in justice, Raskolnikov had been motivated to defend his mother and sister and Sonya (framed as helpless victims) from crippling poverty and sexual extortion; as a petty criminal, Raskolnikov's motive had been simply to steal; as a desperate victim, he had been struggling for survival within a cruelly impersonal social order; as a spiteful narcissist, he had wanted to lash out against the stupidity of those around him; as a visionary, he had committed the crime to realize his own exceptional nature; as a moral anarchist, he had wanted to "dare" to break with all morality; and as a disillusioned "louse," he had committed the crime to punish himself for his own weakness and lack of originality.

Thus, each new account of the crime is marked by incompatible but sincere reframings of Raskolnikov's nature and motives, alongside reframings of his interlocutor, the victims, counterfactual events, and so on. These ontological reframings support, in a straightforward rational way, diverse positive and negative moral interpretations of the deed. Ontological polyphony potentiates normative polyphony. I consider this a theorem.

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

Financial support. The author has no funding to report.

Conflict of interest. None.

References

Bakhtin, M. M. (1984). *Problems of Dostoevsky's poetics*. (Caryl Emerson, Ed. & Trans.). University of Minnesota Press (Original work published 1929).

Beal, B. (2020). What are the irreducible basic elements of morality? A critique of the debate over monism and pluralism in moral psychology. *Perspectives on Psychological Science*, *15*(2), 273–290. https://doi.org/10.1177/1745691619867106

Beal, B. (2021). The nonmoral conditions of moral cognition. *Philosophical Psychology*, *34*(8), 1097–1124. https://doi.org/10.1080/09515089.2021.1942811

Beal, B., & Gogia, G. (2021). Cognition in moral space: A minimal model. *Consciousness and Cognition*, *92*, 103134. https://doi.org/10.1016/j.concog.2021.103134

Dostoevsky, F. (1993). *Crime and punishment*. (Richard Pevear & Larissa Volokhonsky Trans.) Vintage Classics (Original work published 1866).

Much, N. C., & Shweder, R. A. (1978). Speaking of rules: The analysis of culture in breach. *New Directions for Child and Adolescent Development*, *2*, 19–39.

Shweder, R. A. (1992). Ghostbusters in anthropology. In R. G. D'Andrade & C. Strauss (Eds.), *Human motives and cultural models* (pp. 45–58). Cambridge University Press (Reprinted in Kroeber Anthropological Society Paper nos. 69–70, 1989, pp. 100–108, Department of Anthropology, University of California, Berkeley).

Rationality as the end of thought

Nick Chater

Behavioural Science Group, Warwick Business School, University of Warwick, Coventry CV4 7AL, UK
nick.chater@wbs.ac.uk

doi:10.1017/S0140525X22000899, e223

Abstract

Bermúdez convincingly argues that framing effects are ubiquitous and that this is not a sign of human irrationality, but an unavoidable feature of any intelligent system. The commentary adds that framing effects arise even in formal domains, such as chess and mathematics, which appear paradigms of rational thought. Indeed, finding and attempting to resolve clashes between different frames is a major impetus for deliberative cognition.

Bermúdez makes a compelling case that the impact of framing on reasoning and choice is both widespread and entirely reasonable. Yet finding that one is subject to a framing effect does imply that one is, in some sense, in cognitive disequilibrium: Some further thought is required to determine what to think or do. I suggest that theories of rationality, both formal and informal, should rightly be construed as providing conditions for equilibrium (Chater & Oaksford, 2012). So, for example, probability theory, logic, and game theory are all attempts to establishing when our potentially divergent intuitions, prompted by different frames, can simultaneously be embraced. Conversely, these conditions determine when our thoughts are at equilibrium and adjustments are required (these theories do not, crucially, specify which of the many possible adjustments should be made).

Bermúdez highlights how framing effects arise naturally in the political and ethical dilemmas of literature and real life, as well as being a staple of laboratory experimentation. But it is worth stressing that framing effects are likely to arise in any situation in which a boundedly rational agent faces a problem that is too complex to be solved completely.

Let us take the example of chess, where the objective and rules of the game are formally specified, and there are widely agreed standards of what counts as a good move (these days, a good move is operationalized by referring to the "chess engines" that now spectacularly outperform human players). During play, a human (or machine) chess player will continually generate and evaluate conflicting arguments for the virtues of different possible moves – and framing effects will be legion. Suppose, for example, a player is considering an innocuous move (say, advancing a pawn). As different possible consequences of the move are considered (i.e., continually shifting and elaborate its framing), the overall evaluation of its virtues may ebb and flow. If the different frames give wildly different answers, the player may either abandon the move as too risky, or think further to establish which frame should dominate. For example, if the frame is "gain control of the center of the board" the move might seem uninspired but solid; but under the framing "trigger an exchange of pawns and then knights, weakening the defence of the opponent's king," it may seem more attractive. Suppose the player decides to make the move, and then is confronted with a completely unexpected queen sacrifice, which leads to checkmate in three moves. Now the earlier move is seen through a different frame, which was not previously considered – and the dismayed player will realize that this frame is decisive.

Is the player displaying irrationality? It might appear so, from the point of view of an extensional decision theory. After all, the player evaluates "advance pawn" as a good move; and moments later "advance pawn, opening up the possibility of a devastating Queen sacrifice" as a bad move. But these are, of course, the same move, simply described differently. But it would entirely misguided to criticize a person for such a mistake, saying: "don't worry about the description, just make the best move," because we can only evaluate whether or not a move is good or not by considering specific descriptions (including strategic advantages, likely countermoves, etc.). Indeed, a purely extensional approach to playing chess would entirely "abstract away" from the computational challenge of chess – and the reason that chess requires hard thought in the first place.

What is the role of the game theory here (i.e., as providing a rational theory of how strategic interactions should be played)? I suggest that we should view it as providing no more than mild consistency constraints. For example, if moving the pawn leads to certain defeat (after the unexpected queen sacrifice), then reasoning backward, this must have been a bad move, on the assumption that the opponent will choose their move to maximize the chance of winning (and assuming perfect rationality). Similarly, the goodness of the current position should relate directly to the goodness of the position after each player has made the "best" next move; and so on.

But these are minimal constraints that ignore almost everything of interest in the game. Indeed, a pure game-theoretic analysis of chess would simply advise each player to choose a winning strategy from the outset (if White or Black has a winning strategy, which is not known); or otherwise each should play out a certain draw (Schwalbe & Walker, 2001). But the computational complexity of chess is such that no such strategies can be found for either player (Storer, 1983).

The same picture arises across domains. We have, inevitably, a plethora of inconsistent mathematical intuitions (Lakatos, 1976), and framing will be crucial. One moment we might consider a theorem fairly plausible (perhaps by analogy with some similar theorem), but when re-framed as having the consequence that

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

Fermat's Last Theorem is false, the credibility of the theorem reduces sharply. The purpose of mathematical reasoning is surely to help uncover and resolve such cases; and progress in mathematics will involve continually generating and resolving inconsistencies, without obvious limit. Similarly, in less purely formal domains, we can see scientific theories, ethical principles, and indeed the project of philosophy itself, as attempting to find and resolve the endless clashes between our diverging intuitions.

A mind without framing effects would be in perfect equilibrium. Principles of rationality would therefore be satisfied for such a mind. But if rationality constraints were fully satisfied, the need for further thought would have come to an end. In any case, such equilibrium is unattainable. Framing effects, and our continual and partial attempts to resolve them, are not a signature of irrationality; rather they are inevitable consequence of grappling with a world more complex than we can fully understand (cf. Harman, 1986).

Rationality is also the end of thought in a rather different, and more positive sense: The objective of reconciling different frames to be rationally consistent is a driving force behind deliberative cognition. Searching for and evaluating different frames demands intense thought by chess players or mathematicians, and lengthy soliloquizing by Agamemnon and Macbeth. Indeed, it is no exaggeration, perhaps, to see the resolution of conflicts between frames as a major driver of individual and collective cognitive progress.

Financial support. This work was supported by the ESRC Network for Integrated Behavioural Science (grant number ES/K002201/1).

Conflict of interest. None.

References

Chater, N., & Oaksford, M. (2012). Normative systems: Logic, probability, and rational choice. In K. Holyoak & R. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 11–21). Oxford University Press.

Harman, G. (1986). *Change in view: Principles of reasoning*. MIT Press.

Lakatos, I. (1976). *Proofs and refutations: The logic of mathematical discovery*. Cambridge University Press.

Schwalbe, U., & Walker, P. (2001). Zermelo and the early history of game theory. *Games and Economic Behavior*, *34*(1), 123–137.

Storer, J. A. (1983). On the complexity of chess. *Journal of Computer and System Sciences*, *27*(1), 77–100.

Distinguishing self-involving from self-serving choices in framing effects

M. J. Crockett[a] and L. A. Paul[a,b]

[a]Department of Psychology, Yale University, New Haven, CT 06511, USA and [b]Department of Philosophy, Yale University, New Haven, CT 06511, USA
mj.crockett@yale.edu, la.paul@yale.edu
www.crockettlab.org, www.lapaul.org

doi:10.1017/S0140525X22001108, e224

Abstract

We distinguish two types of cases that have potential to generate quasi-cyclical preferences: *self-involving* choices where an agent oscillates between first- and third-person perspectives that conflict regarding their life-changing implications, and *self-serving* choices where frame-based reasoning can be "first-personally rational" yet "third-personally irrational." We argue that the distinction between these types of cases deserves more attention in Bermúdez's account.

We argue that Bermúdez overlooks an important type of framing effect that can lead to quasi-cyclical preferences: the contrasting frames of first-person reasoning and third-person reasoning. While the examples of Macbeth and Agamemnon are compelling, one can argue that in these cases there is a frame-neutral moral rule (murder is wrong) that should dominate a rational agent's reasoning, resolving the incompatibility between frames and undermining Bermúdez's central argument.

We propose that (1) there are better cases of frame-based perspective-taking where, plausibly, no higher, over-arching frame subsumes the conflicting frames. Such cases are *self-involving* choices where an agent oscillates between first- and third-person perspectives that conflict regarding their life changing, or transformative, implications (Paul, 2014, 2018, 2020). However, we also argue that (2) one must carefully distinguish *self-involving* choices from *self-serving* choices, where, given the reasoning of the decision-maker, the frame-based reasoning is "first-personally rational" yet "third-personally irrational."

Consider the following case: Sally, a committed humanitarian who travels to war-torn areas to help people in very great need, does not want to become a parent. Her partner is ambivalent about the choice and wants Sally to make the decision. When Sally reflects on how she feels from within, she finds no desire to have a child. She simply can't see any good reason to give up the valuable, child-free life she is currently leading. She is deeply committed to her successful, demanding career, she finds the small children crying on planes noisy and extremely irritating, and she wants to spend all of her available time pursuing the meaningful work that she finds to be fundamentally satisfying.

In this frame of mind, as she looks within herself, she can't imagine that she would be happier as a parent.

However, all of her friends and family members tell her that, if she were to have a child, she would form a deep and loving attachment to her baby and would enthusiastically endorse her choice. Moreover, she has recently read an argument in favor of relying on science and testimony when making the choice to become a parent (Bloom, 2019). Sally lives in a Scandinavian country with extensive childcare resources and ample support for new parents, could easily change her career focus by shifting to an office-based job that would allow for more time with her child, and knows that the research on people like her suggests that she would maximize her happiness and life satisfaction by becoming a parent. After having dinner at her sister Sera's home and observing Sera's maternal happiness and satisfaction, Sally imagines watching herself as a mother, enacting a similar scene.

In this frame of mind, she finds herself with every reason to become a parent.

As the rosy glow from the evening fades, Sally finds herself switching back and forth between ways to think of the choice. Her reasons, with a first-personal framing, are very persuasively in favor of the choice to remain childfree. She has no reason to accept "I should have a baby." Yet her reasons, with a third personal framing, are very persuasively in favor of the choice to become a parent, as she has many reasons to accept, from a third personal perspective, "Sera's sister should have a baby."

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

We argue that this is a better example of the type of case that Bermúdez wishes to use to defend the argument that framing effects can lead to quasi-cyclical preferences that are not resolvable in a frame-neutral way.

In particular, such a choice is "self-involving," in the sense that the choice depends, and should depend, on Sally's reasons. Moreover, either choice would be morally, legally, and practically permissible for the agent. Yet the frames are inequivalent, and fundamentally so, leading to quasi-cyclical preferences. ("I should have a baby" and "Sera's sister should have a baby" do not mean the same thing, because the first-person mode of reasoning does not translate into the third-person mode of reasoning, and vice versa.) We think that this type of case provides a strong argument in favor of the considerations that Bermúdez raises in his argument for the existence of framing effects that lead to quasi-cyclical preferences.

However, not all cases of conflicting first- and third-person reasoning support Bermúdez's argument, and they also deserve more attention. Consider cases that, given the reasoning of the decision-maker, can be described as "first-personally rational" yet "third-personally irrational." Such cases are self-involving, but importantly, they are also self-serving.

Return to the objection we raised at the start: The self-serving nature of particular decision frames can be opaque to the decision-maker while at the same time painfully obvious to third-party observers. Macbeth might be able to convince himself he is "bravely taking the throne" while observers see straight through his murderous power grab; Agamemnon assures himself he's "following Artemis' will" while the audience looks on in horror as he kills his child. These examples occupy the pantheon of high drama because the audience can clearly see that the protagonist is fooling himself (meaning his decision is third-personally irrational) but can also empathize with the dilemma of the protagonist (because his decisions are first-personally rational).

Supporting this idea, psychology research shows how "ethical blind spots" make it more difficult for people to detect, acknowledge, and remember their own moral transgressions than those of others (Carlson, Maréchal, Oud, Fehr, & Crockett, 2020; Kouchaki & Gino, 2016; Sezer, Gino, & Bazerman, 2015), and that people judge themselves less harshly than others for the same actions (Valdesolo & DeSteno, 2007), perhaps because they deploy self-serving narratives that enable them to justify their actions (Bénabou, Falk, & Tirole, 2018). These behaviors can have disastrous social consequences because everyone hates hypocrites (Jordan, Sommers, Bloom, & Rand, 2017). Thus, the normativity of framing effects cannot be defined merely by the reasoning of the decision-maker, who might fail to recognize how self-serving frames that are alluring from the first-person perspective can have disastrous reputational consequences from the third-person perspective. If the consequences of our choices depend not just on ourselves but also the wider social world, we ought to be suspicious of self-serving frames because of their ability to exploit blind spots in anticipating how we'll be seen by others.

Financial support. MJC was supported by a grant from the John Templeton Foundation (No. 61495).

Conflict of interest. None.

References

Bénabou, R., Falk, A., & Tirole, J. (2018). Narratives, imperatives, and moral reasoning (No. w24798). National Bureau of Economic Research.

Bloom, P. (2019). Arguing with the vampire. Symposium on Transformative Experience. In S. Dellantonio & A. Varzi (Eds.), *Rivista internazionale di filosofia e psicologia* (vol. X, n. 3) (pp. 320–329).

Carlson, R. W., Maréchal, M. A., Oud, B., Fehr, E., & Crockett, M. J. (2020). Motivated misremembering of selfish decisions. *Nature Communications*, *11*(1), 1–11.

Jordan, J. J., Sommers, R., Bloom, P., & Rand, D. G. (2017). Why do we hate hypocrites? Evidence for a theory of false signaling. *Psychological Science*, *28*(3), 356–368.

Kouchaki, M., & Gino, F. (2016). Memories of unethical actions become obfuscated over time. *Proceedings of the National Academy of Sciences*, *113*(22), 6166–6171.

Paul, L. A. (2014). *Transformative experience*. Oxford University Press.

Paul, L. A. (2018). "Transformative treatments" (with Kieran Healy). *Noûs*, *52*, 320–335.

Paul, L. A. (2020). Who will I become?. In J. Schwenkler & E. Lambert (Eds.), *Becoming someone new: Essays on transformative experience, choice, and change* (pp. 16–36). Oxford University Press.

Sezer, O., Gino, F., & Bazerman, M. H. (2015). Ethical blind spots: Explaining unintentional unethical behavior. *Current Opinion in Psychology*, *6*, 77–81.

Valdesolo, P., & DeSteno, D. (2007). Moral hypocrisy: Social groups and the flexibility of virtue. *Psychological Science*, *18*(8), 689–690.

Four frames and a funeral: Commentary on Bermúdez (2022)

Carsten K. W. De Dreu[a,b]

[a]Leiden University, Leiden, Netherlands and [b]University of Amsterdam, Amsterdam, Netherlands
c.k.w.de.dreu@fsw.leidenuniv.nl

doi:10.1017/S0140525X22000917, e225

Abstract

There is much to like in Bermúdez's analysis, yet it is incomplete and at times problematic for social decision making and, by extension, interpersonal conflict. Here I explain how four frames – gains, losses, me, we – operate in conjunction and how humans gravitate toward a "me–loss" frame that, without intervention, leads to a breakdown of cooperation and an arguably tragic funeral of the commons.

Situations in which two or more individuals make choices that affect each other's future can be captured in games of strategy, the simplest ones having two players each with a choice among two options. Well-known examples include the prisoner's dilemma, stag-hunt, and hawk-dove games, and contain a social dilemma – whereas choosing C(ooperate) maximizes social welfare and is "collectively rational," choosing D(efect) maximizes personal welfare and is "individually rational" (Kollock, 1998; Van Dijk & De Dreu, 2021). Whether Cooperate or Defect is labeled as rational or as irrational depends on the perspective taken – by the outsider or the players themselves. Whether the perspective is taken deliberately or intuitively is irrelevant.

Collective and individual rationality can be about minimizing collective or personal loss, or about maximizing collective or personal gain. In social dilemmas, a psychological focus on either gains or losses can be due to reference-dependent framing (Kahneman & Tversky, 1979), framing public good provision as give-some or take-some (Gaechter, Kolle, & Quercia, 2017; Van Lange, Joireman, Parks, & Van Dijk, 2013), or by mindfully taking one rather than the other perspective (Bermúdez, target article). Gain–loss framing does not alter the rank-ordering of preferences and leaves intact the logic and equilibrium properties of the game.

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

Gain–loss framing can, however, change the strength of preferences. Humans pursue individual rationality when in a "me" frame, and because of loss aversion more rigorously so in loss-frames. Experiments corroborate that individuals more likely Defect in loss-framed than gain-framed social dilemmas (Brewer & Kramer, 1986; Sun et al., 2022). Individuals pursue collectively rationality when in a "we" frame – in the behavioral sciences captured and extensively studied under headers such as pro-social motivation (De Dreu, Weingart, & Kwon, 2000; Van Lange et al., 2013) or cooperative orientation (Deutsch, 1960). And indeed, experiments robustly revealed that individuals in a "we" frame more likely Cooperate in loss- rather than gain-framed social dilemmas (Carnevale, 2008; De Dreu & McCusker, 1997; Ispano & Schwardmann, 2017). From a normative perspective, as in the target article by Bermúdez, frames can be pondered in isolation. From an analytic and empirical point of view, however, frames need to be considered in combination – in social decision making, the impact of gain–loss frames depends on the "we–me" frame.

In any game of strategy, individual outcomes are a function of the choices made by oneself and at least one other player. Bermúdez remains silent on the possibility that, therefore, individuals are influenced by their own frames, and by those of the other player(s). Already in two-player games of strategy, there are no less than 2 (players) × 2 (gain or loss frame) × 2 (we or me frame) = eight possible scenarios. This complicates the analysis in particular for games with repeated play where players learn about and adapt to their partner's behavioral strategies and, possibly, their partner's frame(s). Someone in a gain frame may interact with someone in a loss frame, and someone in a "we" frame may interact with someone in a "me" frame. Put differently, someone trying to minimize personal loss as rationally as possible may interact with someone trying to maximize collective gain as rationally as possible.

Although Bermúdez offers little to understand how such interactions unfold, extant work in the behavioral sciences does. For example, knowing that the other player is under a loss rather than gain frame raises empathy and motivates cooperation (De Dreu, Emans, & Van de Vliert, 1992; Fiedler & Hillenbrand, 2020; Van Beest, Van Dijk, De Dreu, & Wilke, 2005; Weber, Kopelman, & Messick, 2004). This incentivizes individually rational players, such as those in a "me frame," to adopt and communicate a loss frame themselves, as it helps to extract cooperation from the counterpart. Experiments with two-person bargaining games have revealed such asymmetric convergence on loss rather than gain frames (De Dreu, Carnevale, Emans, & Van de Vliert, 1994). With players in a "me" frame, loss framing dominates gain framing.

When Cooperate is the dominated strategy, cooperators and those in a "we" frame will be exploited by counterparts in a "me" frame. In mixed populations, "we" players only survive if they adopt their counterpart's "me" frame. Conversely, non-cooperators paired to cooperators have little incentive to switch from their "me" into a "we" frame. Accordingly, we should see asymmetric convergence on "me" frames. This prediction resonates with Pruitt and Kimmel's (1977) goal-expectation hypothesis, with experiments on conditional cooperation in *N*-person social dilemmas (Fehr & Gächter, 2000; Van Dijk & De Dreu, 2021), and with evolutionary agent-based modeling of cooperation in mixed populations (Axelrod & Hamilton, 1981; Gross & De Dreu, 2019a, 2019b; Nowak, 2006). Except when all players are in a "we" frame, social decisions in repeated play will be dominated by a "me" frame.

It is of note that the domination of loss over gain, and "me" over "we" frames cannot be solved with rational (re)framing as suggested by Bermúdez. When Cooperate is the dominated strategy, as in the prisoner's dilemma, considering Cooperate as creating social welfare or as safeguarding one's personal outcomes does not solve the dilemma as long as it is not (1) common knowledge that (2) the other player(s) also take a social welfare perspective. Rational framing is also of no help when neither Cooperate nor Defect is the dominated strategy, as in games with their equilibrium in mixed strategies such as the Stag-Hunt game discussed by Bermúdez (also see De Dreu & Gross, 2019). Consider a simple Hide-and-Seek game in which player H can hide in location 1 or 2 and player S can seek in location 1 or 2. H wants not to be found by S, and S wants to find H. If S is expected to choose 1, H should choose 2, upon which S is incentivized to choose 2, leading H to opt for location 1, and so on *ad infinitum*. It is unclear how choosing a particular frame, or considering decision-options from different perspectives, can help players to achieve whatever goal they have – what to choose remains conditional on what others choose (Arad & Rubinstein, 2012; Camerer, Ho, & Chong, 2004). Quasi-cyclical preferences may be deliberated, rational, and defendable. They do not, however, solve the dilemma.

Experiments in the psychological and economic sciences converge on the possibility that some frames discussed by Bermúdez can exist in theory, yet rarely survive in practice – human psychology gravitates toward minimizing *my loss*, and this explains the break-down of cooperation. In the end, loss framing explains how much effort humans invest, and the "me–we" frame on what the effort is invested in. From a social welfare perspective, and to prevent the tragedy – or funeral – of the commons (Gross & De Dreu, 2019a, 2019b; Hardin, 1968), it is collectively rational to collectively adopt and rationally stick to a "minimize *our* loss" frame.

Financial support. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (AdG agreement no. 785635).

Conflict of interest. None.

References

Arad, A., & Rubinstein, A. (2012). Multi-dimensional iterative reasoning in action: The case of the Colonel Blotto game. *Journal of Economic Behavior and Organization*, *84*, 571–585.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, *211*, 1390–1396.

Brewer, M. B., & Kramer, R. M. (1986). Choice behavior in social dilemmas: Effects of social identity, group size, and decision framing. *Journal of Personality and Social Psychology*, *50*, 543–549.

Camerer, C. F., Ho, T. H., & Chong, J. K. (2004). A cognitive hierarchy model of games. *Quarterly Journal of Economics*, *119*, 861–898.

Carnevale, P. J. (2008). Positive affect and decision frame in negotiation. *Group Decision and Negotiation*, *17*, 51–63.

De Dreu, C. K. W., Emans, B. J. M., & Van de Vliert, E. (1992). Frames of reference and cooperative social decision making. *European Journal of Social Psychology*, *22*, 297–302.

De Dreu, C. K. W., Carnevale, P. J. D., Emans, B. J. M., & Van de Vliert, E. (1994). Effects of gain–loss frames in negotiation: Loss aversion, mismatching, and frame adoption. *Organizational Behavior and Human Decision Processes*, *60*, 90–107.

De Dreu, C. K. W., & Gross, J. (2019). Revisiting the form and function of conflict: Neurobiological, psychological and cultural mechanisms for attack and defense within and between groups. *Behavioral and Brain Sciences*, *42*, 1–44.

De Dreu, C. K. W., & McCusker, C. (1997). Gain–loss frames on cooperation in two-person social dilemmas: A transformational analysis. *Journal of Personality and Social Psychology*, *72*, 1093–1106.

De Dreu, C. K. W., Weingart, L. R., & Kwon, S. (2000). Influence of social motives on integrative negotiation: A meta-analytical review and test of two theories. *Journal of Personality and Social Psychology*, *78*, 889–905.

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

Deutsch, M. (1960). The effect of motivational orientation upon trust and suspicion. *Human Relations*, *13*, 123–139.
Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *The American Economic Review*, *90*, 980–994.
Fiedler, S., & Hillenbrand, A. (2020). Gain–loss framing in interdependent choice. *Games and Economic Behavior*, *121*, 232–251.
Gaechter, S., Kolle, F., & Quercia, S. (2017). Reciprocity and the tragedies of maintaining and providing the commons. *Nature Human Behavior*, *1*, 650–656.
Gross, J., & De Dreu, C. K. W. (2019a). Individual solutions to shared problems create a modern tragedy of the commons. *Science Advances*, *5*, eaau7296.
Gross, J., & De Dreu, C. K. W. (2019b). The rise and fall of cooperation through reputation and group polarization. *Nature Communications*, *10*, e776.
Hardin, G. (1968). Tragedy of commons. *Science*, *162*, 1243.
Ispano, A., & Schwardmann, P. (2017). Cooperating over losses and competing over gains: A social dilemma experiment. *Games and Economic Behavior*, *105*, 329–348.
Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–291.
Kollock, P. (1998). Social dilemmas: The anatomy of cooperation. *Annual Review of Sociology*, *24*, 183–214.
Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*, *314*, 1560–1563.
Pruitt, D. G., & Kimmel, M. J. (1977). Twenty years of experimental gaming: Critique, synthesis, and suggestions for the future. *Annual Review of Psychology*, *28*, 363–392.
Sun, Q. Z., Guo, H. Z., Wang, J. R., Zhang, J., Jiang, C. M., & Liu, Y. F. (2022). Differences in cooperation between social dilemmas of gain and loss. *Judgment and Decision Making*, *16*, 1506–1524.
Van Beest, I., Van Dijk, E., De Dreu, C. K. W., & Wilke, H. A. M. (2005). Do-no-harm in coalition formation: Why losses inhibit exclusion and promote fairness cognitions. *Journal of Experimental Social Psychology*, *41*, 609–617.
Van Dijk, E., & De Dreu, C. K. W. (2021). Experimental games and social decision-making. *Annual Review of Psychology*, *72*, 415–438.
Van Lange, P. A. M., Joireman, J., Parks, C. D., & Van Dijk, E. (2013). The psychology of social dilemmas: A review. *Organizational Behavior and Human Decision Processes*, *120*, 125–141.
Weber, J. M., Kopelman, S., & Messick, D. M. (2004). A conceptual review of decision making in social dilemmas: Applying a logic of appropriateness. *Personality and Social Psychology Review*, *8*, 281–307.

A reputational perspective on rational framing effects

Charles Adam Dorison

Kellogg School of Management, Northwestern University, Evanston, IL 60208, USA
charles.dorison@kellogg.northwestern.edu
charlesdorison.com

doi:10.1017/S0140525X22001054, e226

Abstract

To assess whether behaviors like framing effects are rational, researchers need to consider decision makers' goals. I argue that researchers should broaden the scope of analysis to include impression management goals. Under predictable conditions, behaviors traditionally considered irrational (e.g., loss–gain framing effects on risk preferences) can be reputationally rewarding, casting doubt on strict claims of irrationality.

Framing effects refer to how making one dimension of a decision problem salient influences subsequent choices. For example, one classic study revealed that negotiators who framed negotiations in terms of gains used more concessionary processes (and had more successful outcomes) than negotiators who framed negotiations in terms of losses (Neale & Bazerman, 1985). Over the past half century, such framing effects are among the most widely documented and influential findings in the social and behavioral sciences (Ruggeri et al., 2020; Tversky & Kahneman, 1981). Traditionally, although with important exceptions (e.g., Gigerenzer & Gaissmaier, 2011; Sher & McKenzie, 2006), framing effects are considered irrational biases because they violate the statistical axiom of invariance. In his target article, Bermúdez extends prior research by arguing that the role of frames in reasoning can give rise to rational framing effects. Here, I argue for the existence of rational framing effects for a complementary reason: Frames can shift reputational incentives for decision makers.

Both traditional research and Bermúdez's extension focus primarily on individual cognition. Yet, to assess when framing effects are rational, researchers need to consider the full set of goals individuals hold. Most relevant to the present article, individuals hold strong impression management goals. A large and interdisciplinary body of research studies reputation as a *cause* of behavior, arguing that attention to reputation shapes judgment and decision making (Lerner & Tetlock, 1999; Tetlock, 2002). For example, prior research has made clear that accountability (e.g., making decisions in public vs. private) systematically shifts behavior (Lerner & Tetlock, 1999). To fully assess the rationality of framing effects, researchers must therefore broaden the scope of analysis to include reputational incentives. Specifically, researchers must examine the reputational *consequences* of framing effects.

To illustrate, consider perhaps the most famous example of framing effects: loss–gain framing effects on risk preferences (Kahneman & Tversky, 1979; Tversky & Kahneman, 1981). The canonical finding is well-documented: Decision makers are more likely to make risk-seeking choices when the potential outcomes of alternative choices are framed as losses. In contrast, they are more likely to make risk-averse choices when outcomes are framed as gains, even when choice sets are otherwise equivalent. Could it be the case that loss versus gain frames also shape which choice is perceived more positively by observers?

Recent advances support this theorizing. In a series of three online experiments, Dorison and Heller (2022) found that third-party observers penalize decision makers whose risk preferences are unaffected by loss–gain framing. Specifically, observers reputationally punished leaders who made risk-averse (vs. risk-seeking) decisions when outcomes were framed as losses. In contrast, this result reversed when outcomes were framed as gains: risk-seeking leaders were punished. This reversal was robust across a variety of social dimensions (e.g., warmth vs. competence), choice contexts (e.g., public health decisions vs. monetary gambles), and even with financial stakes on the line. These effects occurred because observers themselves fell victim to loss–gain framing effects, and in turn reputationally derogated decision makers who did not. Of note, such patterns are not limited to framing effects: Across contexts, violating traditional prescriptions for rationality can yield reputational benefits, casting doubt on whether such violations should always be considered irrational in the first place (e.g., Cao, Kleiman-Weiner, & Banaji, 2019; Dorison, Umphres, & Lerner, 2022; Jordan, Hoffman, Nowak, & Rand, 2016; Tenney, Meikle, Hunsaker, Moore, & Anderson, 2019).

When should framing effects thus be considered rational? It may depend on the balance between reputational incentives (e.g., appearing trustworthy) and non-reputational incentives (e.g., maximizing profits for an organization). Rational framing effects are more likely to occur in contexts where differences in reputational incentives are large and differences in non-reputational incentives are small. Consider a public leader who is equivocal about the effects of a COVID-19 policy, but knows

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

that presenting a risky, negatively framed choice to the public would bolster her approval ratings before an upcoming election. The further away the election, the less a framing effect can be considered rational. Conversely, rational framing effects are less likely to occur when differences in reputational incentives are small, but differences in non-reputational incentives are large. Consider a low-profile CEO of a start-up, whose unpopular risk-seeking decisions avoid the public gaze but are critical for the success of her budding company. The more public the decisions, the more a framing effect can be considered rational. Future research is needed to tease apart how such incentives can be balanced and integrated into formal decision models.

Finally, understanding the reputational consequences of framing effects sheds light on new paths for behavior change. Prior work has identified strategies focused on training individual leaders (Morewedge et al., 2015). To the extent that decision makers are attuned to reputational incentives, a complement to such a cognitive approach could involve examining – and shifting – reputational incentives. Rational framing effects may be more versus less likely to occur depending on the broader social or organizational context in which they take place. For example, a leader could set an organizational culture that values identifying, in the moment, how a choice could be framed differently (for related work, see Daniels & Zlatev, 2019). Future research is needed on this possibility.

Bermúdez provided an important advance in identifying how the role of frames in reasoning can yield rational framing effects. Here, I suggest that broadening the scope of analysis to consider reputational incentives provides a complementary reason for rational framing effects. Looking ahead, critical and exciting questions remain regarding boundary conditions for when framing effects should be considered rational and how irrational framing effects can be reduced. Answering these questions is essential to understanding not only rational framing effects, but also rationality in judgment and decision making generally. Applied applications abound for leadership, organizations, and public policy.

Acknowledgments. I appreciate constructive feedback from Elizabeth Huppert, Ariella Kristal, Christopher To, and Dylan Wiwad.

Financial support. Funding was provided by the Kellogg School of Management, Northwestern University.

Conflict of interest. None.

References

Cao, J., Kleiman-Weiner, M., & Banaji, M. R. (2019). People make the same Bayesian judgment they criticize in others. *Psychological Science*, *30*(1), 20–31.

Daniels, D. P., & Zlatev, J. J. (2019). Choice architects reveal a bias toward positivity and certainty. *Organizational Behavior and Human Decision Processes*, *151*, 132–149.

Dorison, C. A., & Heller, B. H. (2022). Observers penalize decision makers whose risk preferences are unaffected by loss–gain framing. *Journal of Experimental Psychology: General*. https://psycnet.apa.org/record/2022-29895-001

Dorison, C. A., Umphres, C. K., & Lerner, J. S. (2022). Staying the course: Decision makers who escalate commitment are trusted and trustworthy. *Journal of Experimental Psychology: General*, *151*(4), 960–965.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*(1), 451–482. https://doi.org/10.1146/annurev-psych-120709-145346

Jordan, J. J., Hoffman, M., Nowak, M. A., & Rand, D. G. (2016). Uncalculating cooperation is used to signal trustworthiness. *Proceedings of the National Academy of Sciences*, *113*(31), 8658–8663.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291. https://doi.org/10 .2307/1914185

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, *125*(2), 255.

Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H., & Kassam, K. S. (2015). Debiasing decisions: Improved decision making with a single training intervention. *Policy Insights from the Behavioral and Brain Sciences*, *2*(1), 129–140.

Neale, M. A., & Bazerman, M. H. (1985). The effects of framing and negotiator overconfidence on bargaining behaviors and outcomes. *Academy of Management Journal*, *28*, 34–49.

Ruggeri, K., Alí, S., Berge, M. L., Bertoldo, G., Bjørndal, L. D., Cortijos-Bernabeu, A., … Folke, T. (2020). Replicating patterns of prospect theory for decision under risk. *Nature Human Behaviour*, *4*(6), 622–633.

Sher, S., & McKenzie, C. R. (2006). Information leakage from logically equivalent frames. *Cognition*, *101*(3), 467–494.

Tenney, E. R., Meikle, N. L., Hunsaker, D., Moore, D. A., & Anderson, C. (2019). Is overconfidence a social liability? The effect of verbal versus nonverbal expressions of confidence. *Journal of Personality and Social Psychology*, *116*(3), 396.

Tetlock, P. E. (2002). Social functionalist frameworks for judgment and choice: Intuitive politicians, theologians, and prosecutors. *Psychological Review*, *109*(3), 451.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science (New York, N.Y.)*, *211*, 453–458.

Defining preferences over framed outcomes does not secure agents' rationality

Sarah A. Fisher

Department of Philosophy, University of Vienna, 1010 Vienna, Austria
sarah.fisher@univie.ac.at
https://sites.google.com/view/sarahafisher/

doi:10.1017/S0140525X22001029, e227

Abstract

Bermúdez claims that agents think about framed outcomes, not outcomes themselves; and that seemingly incoherent preferences can be rational, once defined over framed outcomes. However, the agents in his examples know that alternative frames describe the same outcome, neutrally understood. This undermines the restriction of their preferences to framed outcomes and, in turn, the argument for rational framing effects.

Bermúdez claims that agents think about framed outcomes, not outcomes themselves. As he puts it: "We cannot help but see the objects of choice as framed, or described, or conceptualized in certain ways" (target article, sect. 3.1, para. 5). Accordingly, he argues that our preferences range over framed outcomes, not outcomes themselves.

We are supposed to apply this lesson to Bermúdez's literary case studies of Aeschylus's Agamemnon and Shakespeare's Macbeth. Agamemnon is dealing with two distinct framed outcomes, *Following Artemis's Will* and *Murdering his Daughter*, which simply happen to share an extension in his particular context. Whether Agamemnon prefers *Following Artemis*'s *Will* to the alternative of *Failing his Ships and People* is one question. An independent question is whether Agamemnon prefers the framed outcome *Murdering his Daughter* to the alternative of *Failing his Ships and People*. Agamemnon prefers *Following Artemis's Will* to *Failing his Ships and People* while also preferring *Failing his Ships and People* to *Murdering his Daughter*. This is despite the fact that following Artemis's will involves murdering his daughter. It is because Agamemnon's preferences range over framed outcomes, rather than outcomes themselves, that he is supposed to escape the charge of having inconsistent preferences.

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

Similarly, Macbeth is dealing with two distinct framed outcomes, *Murdering the King* and *Bravely Taking the Throne*, which happen to share an extension in his context. In this case, there are also two distinct framed alternatives, *Fulfilling his Double Duty to Duncan* and *Backing Away from his Resolution to Make the Prophecy come True*. These too are co-extensive. Macbeth's preference between *Murdering the King* and *Fulfilling his Double Duty to Duncan* is supposed to be independent of his preference between *Bravely Taking the Throne* and *Backing Away from his Resolution to Make the Prophecy come True*. Macbeth prefers *Fulfilling his Double Duty to Duncan* to *Murdering the King*, while also preferring *Bravely Taking the Throne* to *Backing Away from his Resolution to Make the Prophecy come True*. Again, since Macbeth's preferences range over framed outcomes, rather than outcomes themselves, it is argued that his preferences are consistent.

Whatever the other merits of defining preferences over framed outcomes, I deny that doing so can justify patterns of preferences like Agamemnon's or Macbeth's. This is because both agents are stipulated to *know* that they are dealing with pairs of frames which describe the same outcome. In each case, then, that common outcome, must be an object of the agent's intentional state of knowledge. Whether it is an extensional phenomenon or just another framed outcome, it must at least be neutral between the two target frames; otherwise it would be impossible to know that it is shared by each. So, knowing that two frames have the same outcome involves having an intentional state, the object of which is a relevantly frame-neutral outcome. And, once we have accepted that agents are able to think about relevantly frame-neutral outcomes, there is no reason to suppose that their thoughts and preferences should be restricted only to the two framed outcomes.

We can unpack this a little by re-examining Agamemnon's dilemma. As Bermúdez explicitly asserts, Agamemnon knows full well that *Following Artemis's Will* and *Murdering his Daughter* are the same outcome, differently framed. Yet if Agamemnon knows this, then an object of his intentional state of knowing is an outcome that is necessarily independent of – or neutral between – these two alternative ways of framing it. In other words, Agamemnon can and does think about the outcome independently of the two frames in question. Given this, there is no reason to suppose that his preferences should range only over the framed outcomes *Following Artemis's Will* and *Murdering his Daughter*. Instead, Agamemnon's preferences can concern the frame-neutral outcome – say, *Killing Iphigenia*.

In fact, it seems entirely right that Agamemnon's preferences do concern the frame-neutral outcome. He fully appreciates that there are strong reasons for and against killing Iphigenia. These competing reasons do not remain frame-relative, even if they are initially made more salient by one or other frame. Instead, Agamemnon recognises that the reasons pertain simultaneously to the single shared outcome. The great difficulty he faces is in how to weigh them up and decide which should take precedence. Thus, Agamemnon's dilemma is substantive, not merely linguistic. Indeed, it may be precisely Agamemnon's ability to reason beyond the two frames – and not remain bound by them – that makes his dilemma so acute.

Note that, if this analysis is correct, Agamemnon's preferences are straightforwardly cyclical. He oscillates between two diametrically opposed preference orderings, sometimes preferring the frame-neutral outcome *Killing Iphigenia* to the alternative of *Failing His Ships and People*, and sometimes the reverse. By Bermúdez's own lights, such cyclical preferences cannot be rationally maintained.

A parallel analysis can be run for Macbeth, and for each of Bermúdez's examples concerning self-control, strategic coordination, and discursive deadlock. This is no accident, as the criticism generalises across the set of framing effects Bermúdez is interested in. After all, he explicitly focuses on situations where agents are well aware that they are framing a single outcome in different ways. As I have argued, such awareness requires agents to be conceptualising the outcome in a relevantly frame-neutral way. It is then no longer clear why they cannot or should not have preferences about the frame-neutral outcome. On the contrary, it seems absolutely right that they can and should.

Bermúdez's argument, then, does not allow us to conclude that there are rational quasi-cyclical preferences. Instead, I believe Bermúdez must acknowledge that agents can and do conceptualise outcomes in relevantly frame-neutral ways; and that their preferences can and do range over these frame-neutral outcomes.

There might still be some other route to the conclusion that preferences like Agamemnon's and Macbeth's are ultimately rational. However, I believe this would require an entirely different line of argument. For a sketch of how it could look, see Fisher (2022). In the meantime, the jury remains out on such cases.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

Reference

Fisher, S. A. (2022). A Critical Notice of: Frame It Again: New Tools for Rational Decision-Making. *International Journal of Philosophical Studies*. doi: 10.1080/09672559.2022.2057685

Even simple framing effects are rational

Stephen J. Flusberg[a], Paul H. Thibodeau[b] and Kevin J. Holmes[c]

[a]Department of Psychology, State University of New York Purchase College, Purchase, NY 10577, USA; [b]Department of Psychology, Oberlin College, Oberlin, OH 44074, USA and [c]Department of Psychology, Reed College, Portland, OR 97202, USA
stephen.flusberg@purchase.edu
https://www.purchase.edu/live/profiles/662-stephen-flusberg
paul.thibodeau@oberlin.edu
https://www.oberlin.edu/paul-thibodeau
kjholmes@reed.edu
https://www.reed.edu/psychology/faculty/holmes.html

doi:10.1017/S0140525X22000942, e228

Abstract

Bermúdez persuasively argues that framing effects are not as irrational as commonly supposed. In focusing on the reasoning of individual decision-makers in complex situations, however, he neglects the crucial role of the social-communicative context for eliciting certain framing effects. We contend that many framing effects are best explained in terms of basic, rational principles of discourse processing and pragmatic reasoning.

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

Bermúdez presents a persuasive case that framing effects are not as "irrational" as commonly supposed. It is interesting to consider intra-individual variation in preferences over time and across contexts as a kind of framing effect, where complex decision-making is cast as an iterative process of reasoning from different perspectives. Yet by focusing on the reasoning of individual decision-makers, Bermúdez's account neglects the crucial role of the social-communicative context in explaining why (at least some) framing effects arise. Language is the central medium for communicating our beliefs and attitudes and persuading others to adopt them. We argue, as a result, that many framing effects are best explained in terms of basic principles of discourse processing and pragmatic reasoning. This framework highlights a key mechanism by which framing operates: Subtle linguistic cues communicate the speaker's knowledge and perspective on a target problem, and decision-makers rely on those cues to draw reasonable inferences about the problem. Therefore, even seemingly "simple" framing effects are rational.

To differentiate his account from the existing literature, Bermúdez describes certain "classic" framing effects as the consequence of a basic "priming" mechanism, where exposure to a frame "activates" a dimension/attribute of the target problem, driving reasoning. This may be a textbook account of framing – and a useful way to frame the target article – but it paints an oversimplified picture of how people process language. It also fails to capture certain findings in the framing literature. For example, much research has shown that framing social issues using metaphors can shape attitudes in a metaphor-congruent fashion (e.g., Thibodeau & Boroditsky, 2011; Thibodeau, Crow, & Flusberg, 2017). When people read a news story that frames crime as a *beast* (vs. a *virus*) ravaging a city, they are more likely to propose enforcement-related solutions to the crime problem that are consistent with how people would address a literal beast problem (Thibodeau & Boroditsky, 2011). In these same studies, however, simply priming participants with the metaphorical source domain (beast or virus) has no effect on their responses. Rather, the metaphor must be used *in context* to describe the social issue in order to impact reasoning. These findings situate common framing effects under the rubric of basic discourse processing (Graesser, Millis, & Zwaan, 1997; Thibodeau & Flusberg, in press; Zwaan & Radvansky, 1998). Language comprehension involves dynamically integrating linguistic input with prior knowledge to generate a mental representation of the topic of discussion. When the topic is unfamiliar, abstract, or complicated – like crime – metaphors serve as useful scaffolding, structuring the listener's representation of the target domain. While exposure to different metaphors may result in different representations, this is a rational response to (subtle) variation in message content – analogous to the quasi-cycles of iterated reasoning Bermúdez describes for individual decision-makers.

Effective language processing also requires that listeners make certain assumptions about the communicative intentions of the speaker. For example, listeners infer that specific words and phrases were chosen because they are relevant and informative (Goodman & Frank, 2016; Grice, 1975; Sperber & Wilson, 1986). Recent evidence suggests that this ability to "read between the lines" and grasp the pragmatic implications of a linguistic frame is critical for many framing effects to obtain (e.g., Flusberg et al., 2022; Holmes, Doherty, & Flusberg, 2021; Leong, McKenzie, Sher, & Müller-Trede, 2017). In one set of studies, we examined the impact of "victim framing" on attitudes toward sexual assault. Participants read a news report that described an alleged sexual assault, often in vivid detail. The report also included a quote from a friend, reflected in the headline, that framed either the accuser as the victim (of assault) or the alleged perpetrator as the victim (of false accusations). Relative to a baseline condition, participants expressed more support for the victim-framed character and less support for the other character. However, this was only the case for those who explicitly cited the framing language as influencing their evaluations – suggesting they surmised that the writer chose to cast one individual as a victim *for good reason* (i.e., to signal who deserves support; Flusberg et al., 2022).

In another set of studies, we assessed people's ability to pick up on the pragmatic implications of subject–complement statements of equality. Sentences like "girls are just as good as boys at math" appear to express an equivalence between two social groups, yet people tend to infer that the group in the complement position – in this case, "boys" – is superior (Chestnut & Markman, 2018). As a result, these sentences can perpetuate, counteract, and even generate new stereotypes in framing studies that manipulate which groups occupy the subject versus complement positions (Chestnut & Markman, 2018; Chestnut, Zhang, & Markman, 2021; Holmes et al., 2021). In a recent study, we measured participants' ability to discern the pragmatics of this syntax by asking them, for example, to infer the beliefs of a journalist who uses a particular subject–complement statement of equality (e.g., "Balurians are just as good as Arigans at cooking" implies that the journalist believes Arigans are the superior chefs). Only those who could successfully recognize these subtle pragmatics showed significant framing effects in an experiment that used similar statements to frame the math abilities of various social groups (e.g., "children from Wyoming do just as well as children from Montana at math"; Holmes et al., in prep; Wu, Elpers, Doherty, Flusberg, & Holmes, 2021). This is consistent with other work showing that even logically equivalent frames (e.g., a basketball player who "makes 40%" vs. "misses 60%" of his shots) communicate subtly different speaker appraisals, which sensitive listeners readily incorporate into their decision-making (e.g., Leong et al., 2017; McKenzie & Nelson, 2003; Sher & McKenzie, 2006).

Taken together, such findings suggest a rational basis for seemingly simple framing effects: Decision-makers infer that specific labels or syntactic constructions communicate relevant information about the target issue and – quite sensibly – use this information in the course of their decision-making. Iterative, quasi-cycles of reasoning about complex situations, while fascinating, are not necessary to reveal the rationality of framing.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Chestnut, E. K., & Markman, E. M. (2018). "Girls are as good as boys at math" implies that boys are probably better: A study of expressions of gender equality. *Cognitive Science*, *42*(7), 2229–2249. https://doi.org/10.1111/cogs.12637

Chestnut, E. K., Zhang, M. Y., & Markman, E. M. (2021). "Just as good": Learning gender stereotypes from attempts to counteract them. *Developmental Psychology*, *57*(1), 114–125. https://psycnet.apa.org/doi/10.1037/dev0001143

Flusberg, S. J., van der Vord, J., Husney, S. Q., & Holmes, K. J. (2022). Who's the "real" victim? How victim framing shapes attitudes towards sexual assault. *Psychological Science*, *33*(4), 524–537. https://doi.org/10.1177/09567976211045935.

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829. https://doi.org/10.1016/j.tics.2016.08.005

Graesser, A. C., Millis, K. K., & Zwaan, R. A. (1997). Discourse comprehension. *Annual Review of Psychology*, *48*(1), 163–189. https://doi.org/10.1146/annurev.psych.48.1.163

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics* (vol. 3, pp. 41–58). Academic Press.

Holmes, K. J., Doherty, E. M., & Flusberg, S. J. (2021). How and when does syntax perpetuate stereotypes? Probing the framing effects of subject–complement statements of equality. *Thinking & Reasoning*. https://doi.org/10.1080/13546783.2021.1963841

Holmes, K. J., Wu, S. H., Elpers, N., Doherty, E. M., & Flusberg, S. J. (in prep). How syntax shapes beliefs about social groups: The role of stereotypes and pragmatic reasoning.

Leong, L. M., McKenzie, C. R., Sher, S., & Müller-Trede, J. (2017). The role of inference in attribute framing effects. *Journal of Behavioral Decision Making*, *30*(5), 1147–1156. https://doi.org/10.1002/bdm.2030

McKenzie, C. R. M., & Nelson, J. (2003). What a speaker's choice of frame reveals: Reference points, frame selection, and framing effects. *Psychonomic Bulletin & Review*, *10*(3), 596–602. https://doi.org/10.3758/BF03196520

Sher, S., & McKenzie, C. (2006). Information leakage from logically equivalent frames. *Cognition*, *101*(3), 467–494. https://doi.org/10.1016/j.cognition.2005.11.001

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition* (vol. 142). Harvard University Press.

Thibodeau, P. H., & Boroditsky, L. (2011). Metaphors we think with: The role of metaphor in reasoning. *PLoS ONE*, *6*(2), e16782. https://doi.org/10.1371/journal.pone.0016782

Thibodeau, P. H., Crow, L., & Flusberg, S. J. (2017). The metaphor police: A case study of the role of metaphor in explanation. *Psychonomic Bulletin & Review*, *24*(5), 1375–1386. https://doi.org/10.3758/s13423-016-1192-5

Thibodeau, P. H., & Flusberg, S. J. (in press). Metaphor and elaboration in context. In H. L. Colston, T. Matlock, & G. J. Steen (Eds.), *Dynamism in metaphor and beyond* (pp. 223–240). John Benjamins Publishing.

Wu, S. H., Elpers, N., Doherty, E. M., Flusberg, S. J., & Holmes, K. J. (2021). *Pragmatic reasoning ability predicts syntactic framing effects on social judgments*. Poster presented at the 43rd Annual Meeting of the Cognitive Science Society [online].

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, *123*(2), 162–185. https://doi.org/10.1037/0033-2909.123.2.162

Consistent preferences, conflicting reasons, and rational evaluations

Francesco Guala

Department of Philosophy, University of Milan, 20122 Milano, Italy
francesco.guala@unimi.it
http://users.unimi.it/guala/

doi:10.1017/S0140525X22000978, e229

Abstract

Bermúdez's arguments in favour of the rationality of quasi-cyclical preferences conflate reasons, desires, emotions, and responses with genuine preferences. Rational preference formation requires that the decision-makers not only identify reasons, but also weigh them in a coherent way.

In what sense is standard decision theory a theory of *rational* decision? Although it is neutral about the content of people's preferences, the theory imposes a few constraints on their shape: It requires that preferences are "well-ordered" or "consistent," as specified by axioms such as completeness, transitivity, and independence. This axiomatic conception of rationality, however, conceals a deeper sense in which the agents of decision theory make rational decisions. Significant decisions typically involve conflicting reasons – there are reasons to do X, but also reasons to do Y, when X and Y are available. An abundant snowfall followed by a spell of sunshine may give me a good reason to go skiing, for example, but the fact that I haven't visited my parents for a while may give me a good reason not to do so. As Bermúdez correctly points out, we often become aware of such conflicts by looking at the same situation from different perspectives – the perspective of the ski-lover versus the perspective of the good son, in the example above. Endorsing (partially, and preliminarily) different frames is indeed an effective way to make sure that all reasons – the reasons that can possibly matter to us – are taken into consideration in the process of decision-making.

Why "partially" and "preliminarily"? Rational decision theory requires that preferences are consistent. Consistency, in turn, requires that each option is assigned a stable value, and that the value of each option reflects the relative value of different *aspects* of the option. The value of each aspect and each option must then be weighed against the values of other options (and their aspects). This weighing process is the truly difficult part of decision-making, as we all know from personal experience. It is a common complaint that the standard theory does not offer much help in making up one's own mind and weighing different options. But it does give *some* help, if only as a warning: When the preferences that are produced by the weighing process turn out to be inconsistent, then we know that something wrong must have happened. Some aspect of an option, for example, must have been evaluated differently in contexts that are effectively identical – as in the framing effects described by Bermúdez in his paper.

The point of framing, to put it differently, is not simply to see things from a different perspective, however intellectually pleasing this may be. The point is to make up one's own mind, to decide what the relative value of different options (and aspects of the options) *really* is. A rational agent thus cannot simply endorse one frame and then another. The rational decision-maker must *compare* the (partial, frame-dependent) reasons or valuations that each frame has elicited, and come up with an all-things-considered evaluation of the alternative options.

According to a prominent proposal, a preference just *is* an "all-things-considered evaluation" (Hausman, 2011). Whether this conception of preference is adequate in the descriptive psychological sense is controversial (e.g., Angner, 2018), but there is little doubt that it fits the standard account of *rational* decision-making. Bermúdez unfortunately constantly conflates preferences with cognate entities, such as emotions, desires, or "responses." One of his working hypotheses goes, for example:

> (H3) Framing effects and quasi-cyclical preferences can be rational in circumstances where it is rational to have a complex and multi-faceted response to a complex and multi-faceted situation.

But a "response" is not a preference. It may be a gut reaction, an emotion, or a *pro tanto* evaluation, in which case it may constitute part of the input for preference-formation. It may even be a choice made impulsively before the process of due diligence has been properly completed. But, alas, in such a case it would be an irrational choice, not a preference in the proper, all-things-considered sense.

Examples can be found easily in Bermúdez's paper: Agamemnon may *want* to follow Artemis's will (under the grip of frame A), and may *want* to fail his ships and people (under the grip of B), but he cannot *prefer* both. Macbeth may have a *desire* or a *reason* to fulfil his double duty to Duncan, and another desire or reason to take the throne, but he cannot *prefer* both, in the sense of rational preference. Another way to put it is that a

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

rational decision-maker must not only "be sensitive to the full range of potential reasons that there might be for choosing one way rather than another" (sect. 3.4, para. 3); she must also *make up her mind* about the relative importance of such reasons. A rational decision-maker reasons across frames, and not just within each frame. She looks at the various reasons (pros and cons) from a bird's-eye point of view, and decides which ones are more important for her, all-things-considered. Although "reasons are frame relative" (sect. 3.4, para. 4), in other words, the comparative evaluation of reasons cannot be made within a single frame.

This is, in a nutshell, why quasi-cyclical preferences cannot be rational preferences. As Bermúdez correctly points out, a rational preference must inherit the rationality of the process that generated them, and quasi-cyclical preferences are the output of a process that fails to accomplish what rationality requires.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Angner, E. (2018). What preferences really are. *Philosophy of Science*, *85*(4), 660–681.
Hausman, D. M. (2011). *Preference, value, choice, and welfare*. Cambridge University Press.

Self-control modulates information salience

Polaris Koi

Philosophy, University of Turku, Turku 20014, Finland
polaris.koi@utu.fi
www.polariskoi.com

doi:10.1017/S0140525X22001066, e230

Abstract

Bermúdez suggests that agents use framing to succeed in self-control. This commentary suggests that frames are effective in steering behavior because they modulate information salience. This analysis extends to self-control strategies beyond framing, raising the question whether there remains an explanatory role for dual process theories for self-control.

> When you find yourself, in a morning, averse to rise, have this thought at hand: I arise to the proper business of a man: And shall I be averse to set about that work for which I was born, and for which I was brought into the universe? Have I this constitution and furniture of soul granted me by nature, that I may lye among bed-cloaths and keep my self warm? (Aurelius, 2008, p. 58)

In the above quote, Marcus Aurelius, emperor of Rome, relies on a familiar idea: The way in which we construe of a given situation will influence how we conduct ourselves. Like Aurelius, much of Hellenistic philosophy was engaged with an approach to philosophy where philosophy does not only help us understand the world, but also enables us to lead better lives (Hadot, 1995). Extending Bermúdez's analysis to Aurelius, by framing the Smaller Sooner (SS) reward of lying in bed as contrary to his very purpose and thereby as aversive, Aurelius effects changes on his behavior.

The literature on framing effects has focused on the use of framing to steer other people, rather than in using frames to steer oneself. But as the case of Marcus Aurelius illustrates, while the discussion of frames in decision theory and psychology has not emphasized the use of frames in regulating our own behavior, the idea and phenomenon are far from new.

While some, including the present author, extend the scope of the concept of self-control beyond delayed gratification, many instances of self-control include a delay-discounting component. Ingeniously, Bermúdez suggests that construing of an instance of delaying gratification as an instance of self-control is in itself a frame that conveys a "struggle over temptation": even if the agent prefers SS over the Larger Later (LL) reward, she may prefer overcoming this struggle to LL (Bermúdez, 2022). This analysis suggests that not only do frames direct our attention to various aspects of our environment and the objects of choice over others, they also direct our attention to certain aspects of our own agency over others.

Frames are not uniquely effective in shaping our actions. Rather, frames impact what information we attend to. A similar impact on behavior can be expected from other factors that modulate information salience. For example, the notion of "choice architecture" (Thaler & Sunstein, 2008) underscores how factors that obscure or highlight certain features of the environment impact behavior, an effect that is best understood in terms of modulating which information is most salient to agents.

The action-modulating force of the environment is also highlighted by pluralists about self-control who argue that self-control can be accomplished by a variety of means, including both intrapsychic means (construal/framing, inhibitory control, self-distraction) and situation selection and situational modification (Duckworth, Gendler, & Gross, 2016; Koi, 2021a). Situational strategies, including situation selection and situation modification, can be used to decrease the salience of SS cues, and to add to the salience of cues promoting LL-conforming behaviors. Strategies like self-distraction and the avoidance of tempting situations both help to decrease the salience of the SS reward, whereas construal and setting up cues and reminders can also operate by increasing the salience of certain features of the LL and/or SS reward. While situational strategies are often analyzed in terms of modulating the effortfulness and hence the cost of the competing courses of action, situational strategies also modulate information salience; some, such as setting reminders, only operate on salience rather than significantly modulating effort. As a result, it appears that the modulation of information salience is a general feature of most self-control strategies, and framing is one of the many ways to accomplish this.

Watzl (forthcoming) argues that self-control is not special in this regard, and that attentional processes are essential for all action. As Watzl points out, agents rely on framing and other attentional processes also when no self-control is required. Considering the richness of our lived circumstances, where there is more actionable information than agents can feasibly process (Wu, 2011), the modulating effect of attentional processes on action should be unsurprising. Understanding this relationship also helps explain why disorders of attention, such

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

as attention deficit/hyperactivity disorder, often result in difficulties in supposedly volitional processes, including self-control (Koi, 2021b).

The broader question that arises is, then, not whether attentional systems play any explanatory role in action and choice, but rather to what extent attention and salience leave an explanatory gap for other systems to fill in. In Bermúdez's analysis, frames impact our decision-making because they highlight features of the choice at hand that align with the "hot" and "cold" cognitive–affective systems (Bermúdez, 2022). Is there, in fact, an explanatory role for the dual process theory here? As agents in complex worlds cannot attend to all relevant information, the salience of certain information can be expected to be predictive of our decision-making, regardless of how it maps onto a dual process conception of cognition. Moreover, decisions are sometimes difficult because the "cold" systems recommend more than one course of action. The framing effect's explanatory power over self-control would scarcely be diminished by considering it without recourse to dual process theories.

The dual process theory is not the only theory whose explanatory force is called into question by the role of information salience in self-control. Much of the literature on self-control has posited a dedicated "willpower," "ego depletion," or "muscle" (Baumeister & Exline, 1999; Baumeister, Tice, & Vohs, 2018) or else reduced self-control to inhibitory control (Cohen, Berkman, & Lieberman, 2013) or a sequence of cognitive control (Sripada, 2021). The ego depletion hypothesis has been questioned (Inzlicht & Friese, 2019), whereas there are concerns that inhibitory and cognitive control processes are too general to help account for the decisional aspect of self-control. If it is plausible that self-control operates, at least in part, by modulating information salience, future research on self-control should seek to devise means to control for the role of attentional processes in self-control, and to see to what extent there remains an explanatory gap for other proposed mechanisms to fill.

Financial support. This contribution was written in the context of a research project funded by the Strategic Research Council at the Academy of Finland (grant number 335186).

Conflict of interest. None.

References

Aurelius, M. (2008). *The meditations of the Emperor Marcus Aurelius Antoninus*. (F. Hutcheson & J. Moor, Trans.) Liberty Fund.

Baumeister, R. F., & Exline, J. J. (1999). Virtue, personality and social relations: Self-control as the moral muscle. *Journal of Personality*, *67*(6), 1165–1194. doi: 10.1111/1467-6494.00086

Baumeister, R. F., Tice, D. M., & Vohs, K. D. (2018). The strength model of self-regulation: Conclusions from the second decade of willpower research. *Perspectives on Psychological Science*, *13*(2), 141–145. doi: 10.1177/1745691617716946

Cohen, J. R., Berkman, E. T., & Lieberman, M. D. (2013). Intentional and incidental self-control in ventrolateral PFC. In D. T. Stuss & R. T. Knight (Eds.), *Principles of frontal lobe function: Second Edition* (pp. 417–440). Oxford University Press.

Duckworth, A., Gendler, T. S., & Gross, J. J. (2016). Situational strategies of self-control. *Perspectives on Psychological Science*, *11*(1), 35–55. doi: 10.1177/1745691615623247

Hadot, P. (1995). *Philosophy as a way of life*. (M. Chase, Trans.) Blackwell.

Inzlicht, M., & Friese, M. (2019). The past, present, and future of ego depletion. *Social Psychology*, *50*(5–6), 370–378. doi: 10.1027/1864-9335/a000398

Koi, P. (2021a). Accessing self-control. *Erkenntnis*. doi: 10.1007/s10670-021-00500-y

Koi, P. (2021b). Born which way? ADHD, situational self-control, and responsibility. *Neuroethics*, *14*, 205–218. doi: 10.1007/s12152-020-09439-3

Sripada, C. (2021). The atoms of self-control. *Noûs*, *55*, 800–824. doi: 10.1111/nous.12332

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth and happiness*. Yale University Press.

Watzl, S. (forthcoming). Self-control, attention, and how to live without special motivational powers. In M. Brent & L. M. Titus (Eds.), *Mental action and the conscious mind*. Routledge.

Wu, W. (2011). Confronting many–many problems: Attention and agentive control. *Noûs*, *45*, 50–76. doi: 10.1111/j.1468-0068.2010.00804.x

Why framing effects can be rational

Anton Kühberger

Department of Psychology, University of Salzburg, 5020 Salzburg, Austria
anton.kuehberger@plus.ac.at
https://ccns.sbg.ac.at/people/kuehberger/

doi:10.1017/S0140525X22001133, e231

Abstract

When communication is not disinterested, seemingly inconsistent preferences are predictable from language pragmatics and information non-equivalence. In addition, the classic risky choice framing effect found in the Asian disease task – risk-aversion with gains and risk-seeking with losses – applies to gambles, but tends to be overgeneralized to non-gambling situations.

Bermúdez argues (see especially sect. 2.4) that the consensus view on classic framing effects a la Asian disease is that they are irrational, and that the irrationality view would be consensual among psychologists of reasoning, and in the cognitive sciences. I disagree: Rather, there is evidence that classic risky choice framing effects (i.e., risk-seeking with losses and risk-aversion with gains) can be construed as being rational. Notably, recent research has shown that risky choice framing effects can result from various cognitive processes, all being entirely intelligible and rational. Central is the idea that, rather than passively taking in information, people actively select and process information, taking also background knowledge into consideration. For instance, the Asian disease task includes a background of fighting diseases, rendering an option described as "200 people are saved" unlikely to be interpreted as "exactly 200 people are saved." If, however, the bilateral "exactly" reading does not apply, but a lower-bound "at least"-reading does, extensionality of differently described option breaks down (see Mandel, 2014).

In addition, the options in most tasks modelled after the Asian disease task are incompletely described: It is only stated that "200 people are saved," but nothing is said about the remaining people. Research shows that symmetric description (i.e., explicitly specifying all outcomes) makes the framing effect disappear (Kühberger, 1995), indicating that there is more to the different descriptions than simply the framing. Finally, information equivalence of framed options implies that the choice of frame is arbitrary. This generally is not true, since people tend to choose descriptions, and frames, non-randomly. Specifically, they tend to describe objects by attributes that exceed, rather than fall short, of some reference point. In addition, frame selection may leak information about a speaker's attitude (Sher & McKenzie, 2006).

All this goes to show that what initially was considered a most impressive demonstration of irrationality in many risky choice framing tasks follows from the naive, and often unsubstantiated, intuition that arithmetics (i.e., 200 of 600 saved = 400 of 600 dying) is all that counts in framing tasks. Pertinent research has

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

shown this to be shortsighted: At semantic and pragmatic levels number expressions can have importantly different meanings. Thus, rather than showing irrational choices, the findings show that the (untested) assumption of equivalence stands on shaky grounds. If extensionality cannot be assured, the extensionality principle can hardly be violated.

Another curiosity in framing research should also be pointed out. In the Asian disease task, and all tasks modelled accordingly, one option (200 people saved; or 400 people die) is framed as "the sure option," and the second option (600 people saved with $p = 1/3$, or no one saved with $p = 2/3$) is said to be "the risky option." Bermúdez argues that "frames *prime* responses," such that the gain frame primes risk-aversion (i.e., the choice of the sure option) and the loss frame primes risk-seeking (i.e., the choice of the risky option). Why is the option "200 people are saved" (or "400 people will die") a sure option? Consider the respective situation: We have a population of 600 people contracting the disease. Two outcomes are possible: save or die; repeated 600 times. This is true for the risky option, but also for the sure option: any individual can be saved or not. Each option thus consists of 600 risky events. In other words, both options are risky, and they are identical in risk. They are only framed as sure or risky. The framing is done by using the word "probability" for the so-called risky option, while avoiding it (or using the notion "for sure") for describing the so-called sure option. The impression of a sure option only follows from hiding the risk part. Taken together, in tasks following the Asian disease structure, a distinction of options in terms of risk does not make sense. If the task is modelled as a gamble, things are different. Imagine you have to choose among (A) winning €200 for sure, or (B) winning €600 with $p = 1/3$ or nothing with $p = 2/3$. Here, option A has only one possible outcome (€200), while option B has two (€600, or €0). Thus, A is sure, and B is risky.

The gambling situation is different from the disease situation in many respects. Most notably, the semantic and pragmatic aspects are weaker, or even nonexistent, with gambles. Gambles are critical for testing the irrationality argument, since with gambles extensionality can be best preserved. Using between-subjects designs, robust evidence exists for risk-aversion with gains and risk-seeking with losses also for gambles (e.g., Kühberger, Schulte-Mecklenbeck, & Perner, 2002). However, little is known whether preferences also reflect in within-subjects designs. Note, however, that gambles are a very specific domain (unless you are a decision researcher). Rational cognition may be more adapted to general, rather than to specific situations. Finding irrationality in gambles may be too weak an argument for the verdict that human choices are irrational in general.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Kühberger, A. (1995). The framing of decisions: A new look at old problems. *Organizational Behavior and Human Decision Processes*, *62*, 230–240. https://doi.org/10.1006/obhd.1995.1046

Kühberger, A., Schulte-Mecklenbeck, M., & Perner, J. (2002). Framing decisions: Real and hypothetical. *Organizational Behavior and Human Decision Processes*, *89*, 1162–1175. https://doi.org/10.1016/S0749-5978(02)00021-3

Mandel, D. R. (2014). Do framing effects reveal irrational choice? *Journal of Experimental Psychology: General*, *143*, 1185–1198. https://doi.org/10.1037/a0034207

Sher, S., & McKenzie, C. R. M. (2006). Information leakage from logically equivalent frames. *Cognition*, *101*, 467–494. https://doi.org/10.1016/j.cognition.2005.11.001

Ceteris paribus preferences, rational farming effects, and the extensionality principle

Joe Y. F. Lau

Department of Philosophy, The University of Hong Kong, Pok Fu Lam, Hong Kong
jyflau@hku.hk
http://philosophy.hku.hk/joelau

doi:10.1017/S0140525X22001017, e232

Abstract

Bermúdez argues for rational framing effects in the form of quasi-cyclical preferences. This is supposed to refute the extensionality principle in standard decision theory. In response, I argue that it is better to analyze seemingly quasi-cyclical preferences as *ceteris paribus* preferences. Furthermore, if frames are included as objects of choice, we can acknowledge rational framing effects without rejecting extensionality.

Many cognitive biases are due to framing. Bermúdez agrees, but he thinks (1) there are rational framing effects in the form of quasi-cyclical preferences. Furthermore, (2) these effects refute the extensionality principle – the widely accepted normative assumption that rational preferences should not depend on how outcomes are framed. I shall argue against both conclusions. Seemingly quasi-cyclical preferences are better construed as *ceteris paribus* preferences. Moreover, if frames are included as objects of choice, we can accept rational framing effects without rejecting extensionality.

Consider Bermúdez's example of Agamemnon. Bermúdez claims that the following sentences are true:

(A) Agamemnon prefers *Following Artemis's Will* to *Failing his Ships and People*.
(B) Agamemnon prefers *Failing his Ships and People* to *Murdering his Daughter*.

Furthermore, Bermúdez thinks that Agamemnon is not irrational, even though Agamemnon knows that *Following Artemis's Will* and *Murdering his Daughter* are two different ways of framing the same outcome.

I think Bermúdez is mistaken to treat this as demonstrating quasi-cyclical preferences. A better analysis is that (A) and (B) express *ceteris paribus* general preferences, rather than strict preferences about specific outcomes (Hansson, 1996; Van Benthem, Girard, & Roy, 2009). If I say I prefer coffee over tea, we normally take this to involve an implicit qualification – all else being equal, I prefer coffee to tea. This preference is defeasible and not absolute. I am not being inconsistent if I happen to choose tea over an overpriced, watery coffee. If (A) and (B) express *ceteris paribus* preferences, we can see why they can both be true, even when Agamemnon decides to sacrifice his daughter. All else being equal, Agamemnon prefers *Failing his Ships and People* to *Murdering his Daughter*. But in this unfortunate instance, taking everything into account, it is indeed rational for him to murder his daughter. It does not matter whether the outcome is framed

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

in terms of *Murdering his Daughter* or *Following Artemis's Will*. Extensionality is consistent with (A) and (B) being true. The same analysis applies to Bermúdez's Macbeth example, and his examples of gun-violence and energy-independence, as they all share the same structure.

Bermúdez discusses rational framing effects in three types of everyday situations. The first relates to self-control. For example, an agent might be tempted to choose a smaller immediate reward over a larger future reward. Bermúdez proposes that the agent can resist temptation through self-control by reframing the future reward as a case of having successfully resisted the immediate reward. This is because the agent might consider resisting temptation as demonstrating highly valuable traits such as virtue and commitment. I suspect this misrepresents typical cases of temptations, where considerations about virtues often lack motivational power ("Lord make me chaste but not yet," as Saint Augustine might say).

In any case, even if Bermúdez is correct that reframing allows the agent to overcome temptation, this only goes to show that the agent underestimated the full value of the future reward earlier on. The value of overcoming a temptation should also have been included, as they belong to the same outcome, regardless of how it is framed. Failure to compute utility correctly does not constitute evidence for rational quasi-cyclical preferences.

The same problem arises for Bermúdez's second type of cases concerning strategic decision in game theory. In his snowdrift example, cooperation is not the preferred outcome under an "I"-frame ranking, but it eventually becomes the preferred outcome when the subject changes to a "we"-frame ranking. It is unclear why this supports rational quasi-cyclical preferences. Bermúdez says the subject changes frame "as considerations of fairness start to take hold" (target article, sect. 5.3, para. 9). This suggests that the subject failed to fully assess the fairness of the situation when cooperation was rejected under "I"-frame thinking. But this does not show that the rationality of cooperation is frame-relative. As Rawls (1971) has pointed out, reflective equilibrium requires working back and forth among our considered judgments. Reflecting on the principles of justice can lead us to revise our earlier beliefs as to whether something is fair, but this is consistent with extensionality.

Bermúdez's third type of example concerns interpersonal conflicts. He argues that discursive deadlocks in the public domain often involve clashes of frames, and their resolution requires techniques such as reflexive decentering and imaginative simulation. I am sympathetic to Bermúdez's framework, as I think it complements related proposals, such as the use of emotional regulation in resolving intractable conflicts (Halperin, Cohen-Chen, & Goldenberg, 2014). These proposals can help us promote rationality and objectivity in public reasoning. However, we can acknowledge these insights without accepting rational quasi-cyclical preferences.

I agree with Bermúdez that there can be rational framing effects, but this is compatible with extensionality. Consider how mindsets affect learning and performance. A mindset at its core is a set of frames, incorporating a system of concepts, principles, and values for interpreting the world. There is considerable evidence that a "growth" mindset that views intelligence as a malleable rather than fixed trait is more likely to lead to success (Dweck, 2006; Yeager & Dweck, 2012). Similarly, having a "stress-is-enhancing" mindset – perceiving stress as functional and adaptive – might improve coping behavior and performance (Jamieson, Nock, & Mendes, 2012). If these findings are correct, there will be times when a project will be successful only if we embrace it with an appropriate mindset. In these situations, whether it is rational to pursue the project seems to be frame-relative.

This does not refute extensionality, for the simple reason that the proper object of choice in these situations is not solely the project but the project coupled with a commitment to a particular mindset. Bermúdez says reasons are always frame-relative, that it is "frames all the way down" (target article, sect. 6.1, para. 3). But then whether a decision is rational relative to a given frame is presumably also relative. Does it mean there is no objective fact as to whether a decision is rational? However, if mindset can be an object of rational choice, we can then preserve extensionality and the objectivity of reason, and still recognize the existence of rational framing effects.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Dweck, C. S. (2006). *Mindset: The new psychology of success*. Random House.

Halperin, E., Cohen-Chen, S., & Goldenberg, A. (2014). Indirect emotion regulation in intractable conflicts: A new approach to conflict resolution. *European Review of Social Psychology*, *25*(1), 1–31.

Hansson, S. O. (1996). What is ceteris paribus preference?. *Journal of Philosophical Logic*, *25*(3), 307–332.

Jamieson, J. P., Nock, M. K., & Mendes, W. B. (2012). Mind over matter: Reappraising arousal improves cardiovascular and cognitive responses to stress. *Journal of Experimental Psychology: General*, *141*(3), 417.

Rawls, J. (1971). *A theory of justice*. Harvard University Press.

Van Benthem, J., Girard, P., & Roy, O. (2009). Everything else being equal: A modal logic for ceteris paribus preferences. *Journal of Philosophical Logic*, *38*(1), 83–125.

Yeager, D. S., & Dweck, C. S. (2012). Mindsets that promote resilience: When students believe that personal characteristics can be developed. *Educational Psychologist*, *47*(4), 302–314.

Framing provides reasons

Neil Levy[a,b]

[a]Department of Philosophy, Macquarie University, Sydney 2109, Australia and [b]Uehiro Centre for Practical Ethics, University of Oxford, Oxford OX1 1PT, UK
neil.levy@philosophy.ox.ac.uk
https://scholar.google.com/citations?user=hqLeZWcAAAAJ&hl=en

doi:10.1017/S0140525X22000875, e233

Abstract

Framing effects are held to be irrational because preferences should remain stable across different descriptions of the same state of affairs. Bermúdez offers one reason why this may be false. I argue for another: If framing provides implicit testimony, then rational agents will alter their preferences accordingly. I show there is evidence that framing should be understood as testimonial.

Framing effects are usually held to be irrational because they violate the principle Bermúdez calls *extensionality*: Preferences should be stable across different descriptions of the same

outcomes. Bermúdez argues that extensionality is false: Preferences can be *quasi-cyclical*. An agent has quasi-cyclical preferences when she prefers A to B under one description while preferring B to A under another description, while knowing that the descriptions are different ways of framing the same outcome but believing that one way of framing the options is preferable to the other. Quasi-cyclical preferences are stable and therefore avoid many of the problems that might arise from cyclical preferences (such as failures of transitivity). More importantly, rational agents may reason their way to stable preferences by coming to see that one frame is normatively preferable to another. Some preferences adopted via framing are therefore rationally justifiable and we may rationally employ framing in decision-making, deliberation, and discussion.

Bermúdez's aim is to show that framing effects *can* be rational, especially in the "complicated" conditions that arise "outside the laboratory." He does not attempt to vindicate the rationality of framing *inside* the laboratory, and in fact his strategy is unlikely to show that participants in classic experiments on framing choose rationally. Nevertheless, I suggest, there's a strong case for thinking that they *do* choose rationally. Frames provide agents with genuine evidence, and participants respond rationally to this evidence.

The extensionality principle is false when descriptions are coupled with recommendations. In most circumstances, agents with no prior knowledge of or preference between options who are told that A is better than B acquire a genuine reason to prefer A to B. Such an agent might rationally prefer A to B even though they also know they would have preferred B to A had their informant given them a different recommendation. If framing of options is testimonial, its rationality is vindicated.

There is evidence, both experimental (McKenzie, Liersch, & Finkelstein, 2006) and from modelling (Carlin, Gervais, & Manso, 2013) that options selected as the default are understood as authoritatively recommended. This is a special case of a broader phenomenon: To make an option salient – by making it the default, by placing it prominently, and so on – is in many contexts to recommend it or to implicate its relevance or importance. Framing is one more way of making options salient to individuals. In fact, there is evidence that framing is understood as conveying recommendations and that ordinary people use framing to this end. For example, someone might say that a researcher has had 40% of her journal submissions *accepted* (rather than conveying the same information by mentioning the proportion *rejected*) in order to recommend her to a recipient (Sher & McKenzie, 2006). Further, the testimony will be understood as a recommendation, at least functionally – the recipient would form a more favourable impression of the person when her acceptance rate is mentioned, rather than her rejection rate (Fisher, 2020, 2022; McKenzie et al., 2006).

A cognitive process might be assessed as rational against an *ecological* or a more direct standard. Ecological rationality is assessed by how well designed a process is to achieve agents' goals (Gigerenzer & Selten, 2002; Todd & Gigerenzer, 2012). A process is directly rational, as I define it here, if it transforms inputs in a way that reflects their actual informational content, such that it causes Bayesian belief update. Gigerenzer's influential defence of ecological rationality turns on the fact that the two kinds of rationality can dissociate: Human beings might systematically depart from Bayesian rationality yet do so in a way that allows us to achieve our goals, given the kinds of challenges we typically face.

Defences of human rationality that emphasise its ecological appropriateness face the problem that the environment in which we make decisions may depart significantly from the environment for which our decision-making processes are adapted, either because we deliberate and act in environments that depart significantly from the environment of evolutionary adaptiveness, or because hostile agents take advantage of our vulnerabilities to engineer cognitive traps for us (Evans & Stanovich, 2013; Stanovich, 2018). If our responses to framing are rational in the manner suggested, however, they are directly and not merely ecologically rational. Framing offers participants genuine *evidence* – technically, higher-order evidence; evidence about evidence – and it is rational for agents to respond to it accordingly (Fisher, 2020, 2022; Levy, 2019, 2021).

It is somewhat controversial under what conditions we ought to trust testimony; in particular, whether we need evidence of trustworthiness before we are justified in taking someone's word for something (Lackey & Sosa, 2006). There is no doubt, however, that no matter how demanding the conditions for justified trust might be, they are satisfied in the classic experiments on framing effects. The stakes for participants are low, because nothing turns for them on their response. They have nothing else to base their decision on. And the testimony comes from experimenters, who participants have some reason to regard as more knowledgeable on the topic in question than they are (other things equal, the very act of offering testimony implicates knowledge). If framing functions as testimony, we ought to see participants being sensitive to the same features they are sensitive to when it comes to explicit testimony. For example, they should prefer frames associated with people who are competent or benevolent informants, or those associated with numerically more rather than fewer informants (Harris, 2012; Sperber et al., 2010). In other work, I've suggested that inference to the best explanation supports understanding framing effects as functioning via the provision of testimony. Manipulation of properties like the apparent trustworthiness or competence of those seen to frame the information might provide a way to experimentally test the hypothesis.

Financial support. Neil Levy was supported by grants from the European Research Council (Grant number 819757) and the Australian Research Council (DP180102384).

Conflict of interest. None.

References

Carlin, B. I., Gervais, S., & Manso, G. (2013). Libertarian paternalism, information production, and financial decision making. *The Review of Financial Studies*, *26*(9), 2204–2228.

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223–241.

Fisher, S. A. (2020). Rationalising framing effects: At least one task for empirically informed philosophy. *Crítica: Revista Hispanoamericana de Filosofía*, *52*(156), 5–30.

Fisher, S. A. (2022). Meaning and framing: The semantic implications of psychological framing effects. *Inquiry*, *65*(8), 967–990. https://doi.org/10.1080/0020174X.2020.1810115

Gigerenzer, G., & Selten, R. (2002). *Bounded rationality: The adaptive toolbox*. MIT Press.

Harris, P. (2012). *Trusting what you're told*. Harvard University Press.

Lackey, J., & Sosa, E. (2006). *The epistemology of testimony*. Oxford University Press.

Levy, N. (2019). Nudge, nudge, wink, wink: Nudging is giving reasons. *Ergo: An Open Access Journal of Philosophy*, *6*, 281–302. https://doi.org/10.3998/ergo.12405314.0006.010

Levy, N. (2021). *Bad beliefs: Why they happen to good people*. Oxford University Press.

McKenzie, C. R. M., Liersch, M. J., & Finkelstein, S. R. (2006). Recommendations implicit in policy defaults. *Psychological Science*, *17*(5), 414–420.

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

Sher, S., & McKenzie, C. R. M. (2006). Information leakage from logically equivalent frames. *Cognition*, *101*(3), 467–494.
Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind & Language*, *25*(4), 359–393.
Stanovich, K. E. (2018). Miserliness in human cognition: The interaction of detection, override and mindware. *Thinking & Reasoning*, *24*(4), 423–444.
Todd, P. M., & Gigerenzer, G. (2012). *Ecological rationality: Intelligence in the world*. Oxford University Press.

Framing, equivalence, and rational inference

David R. Mandel

Intelligence, Influence, and Collaboration Section, Defence Research and Development Canada, Toronto, Ontario M3K 2C9, Canada
drmandel66@gmail.com
https://sites.google.com/site/themandelian/home

doi:10.1017/S0140525X22000954, e234

Abstract

Bermúdez's case for rational framing effects, while original, is unconvincing and gives only parenthetical treatment to the problematic assumptions of extensional and semantic equivalence of alternative frames in framing experiments. If the assumptions are false, which they sometimes are, no valid inferences about "framing effects" follow and, then, neither do inferences about human rationality. This commentary recaps the central problem.

What distinguishes alternative frames from mere re-descriptions of a referent is the requirement that frames represent elements of a set that are semantically, if not pragmatically, identical. Thus, when Tversky and Kahneman (1981) introduced the "Asian disease" problem, they claimed, "it is easy to see that the two problems [i.e., alternative frames] are effectively identical" (p. 453). Some years later, they again stated that "it's easy to verify that options C and D in Problem 2 are undistinguishable in real terms from options A and B in Problem 1" (Kahneman & Tversky, 1984, p. 434).

This intuitive argument is an appeal to a watered-down extensional equivalence claim: To avoid contradictions, alternative frames constitute semantically equivalent inputs and should therefore yield identical behavioral outputs. Of course, extensional equivalence does not assume intensional equivalence, as is well known in computational theory (Bruni, Giacobazzi, Gori, Garcia-Contreras, & Pavlovic, 2020). Prospect theory (Kahneman & Tversky, 1979) exploits examples such as the Asian disease problem to make the case that intensional non-equivalences (e.g., the psychophysics of valuation) from "extensionally equivalent" inputs causes extensionally nonequivalent outputs. The scare quotes are meant to call out the sleight-of-hand trick in which objectively different inputs are nevertheless said to be the same. In the application of extensional equivalence in the theory of computation, *same inputs* means identical inputs and not merely semantically synonymous: for example, IV is not the same input as 4, even if these symbols represent the same quantity.

In order to show a rationality-violating contradiction, it is, therefore, vital (a) that semantic equivalence be the only equivalence that matters and (b) that at least semantic equivalence is guaranteed. Neither of these conditions is generally met. McKenzie and colleagues (e.g., McKenzie & Nelson, 2003; Sher & McKenzie, 2006) have been clear in articulating that alternative frames such as in the Asian disease problem are informationally non-equivalent, even if they happen to be semantically equivalent because, pragmatically, the alternatives are indisputably non-equivalent. For instance, a positive frame will convey more optimism than a negative frame even if the two share the same quantity semantics (i.e., 200 saved out of 600 implies 400 not saved out of the same number).

However, even if we put the informational nonequivalence of frames aside, it is, in fact, not "easy to verify" that the alternative frames are semantically equivalent. Tversky and Kahneman certainly did not offer any such verification, only the claim itself (which, sadly, seems to have been enough to have convinced most psychologists and behavioral economists for over four decades). It did not take long for some to question this claim. Notably, Macdonald (1986) hypothesized that people tend to interpret quantifiers as lower bounds, but remarkably, the idea remained untested for a quarter century. However, in experiment 3 of Mandel (2014), I found that Macdonald's hypothesis had support. Participants responded to an Asian disease problem variant (substituting war for disease as a cause of deaths) and after making their choice, they indicated whether they thought the quantifier they had encountered meant *at least*, *exactly*, or *at most* that value. Most participants (64%) in the Asian disease problem variant indicated that they interpreted the quantifiers "200" or "400" in options A and C, respectively, as meaning *at least* that amount (30% said *exactly* and 6% said *at most*). Since saving at least 200 out of 600 (option A in the gain frame) is objectively better than letting die at least 400 out of 600 (option C in the loss frame), the preference for A over B (the gain-frame gamble) and D (the loss-frame gamble) over C is hardly a preference reversal. Indeed, it cannot be a preference reversal because lower-bounding of the quantifiers destroys semantic equivalence across "frames."

By now, the reason for the scare quotes should be obvious: options A and C are not frames at all. The Asian disease problem is not a framing problem. Therefore, it can yield no framing effect, no matter how replicable the behavioral effect is (sorry, inductivists). The effect does not demonstrate a violation of description invariance in risky choice because the descriptions are not semantically invariant and cannot merely be assumed to be so through hand-waving exercises. The said effect, therefore, does not demonstrate a rationality-violating contradiction.

Now to Bermúdez's argument. Bermúdez does not question whether the behavioral evidence from the central base of framing research implies irrationality. Rather, his approach is to carve away the so-called small-world of toy problems entirely and focus on what he describes as the larger complex world in which multi-attribute decisions are the result of conflicting perspectives. Here, we are told, quasi-cyclical preferences induced by the consideration of alternative frames are rational. To be sure, such preferences exist (as his Agamemnon and Macbeth examples illustrate), but the case for why they should be viewed as rational is uncompelling because the contrast to the small-world case is never pinned down tightly. In the complex case, we are told that "preferences are not basic. They

© The Crown in Right of Canada (Defence R&D Canada), 2022

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

are made for reasons" (sect. 3.4, para. 3), and "reasons are frame relative" (sect. 3.4, para. 4). But, surely, the same applies in the small world. The individual who prefers to save 200 lives for sure rather than gamble on a 1/3 chance of saving all 600 (with a corresponding 2/3 chance that all will die) will have a reason for this preference (e.g., "people who could surely be saved *should* be saved") even though the same individual will have a different reason for preferring the gamble had the outcomes been framed as losses (e.g., "you *shouldn't* accept any deaths that could be surely prevented"). This individual may have differing reasons even if the options were semantically equivalent (which they are not).

In short, Bermúdez's carve-out is unsustained and we are left with a contradiction: Bermúdez accepts the irrationality of framing effects in small-world cases but rejects it in complex-world cases for reasons that often apply equally well to the small world. Meanwhile, the conceptual quagmire sketched earlier (also see Fisher & Mandel, 2021; Mandel, 2021) deserves full attention, but is largely neglected.

Financial support. Funding for this research was provided by the Canadian Safety and Security Program project CSSP-2018-TI-2394.

Conflict of interest. None.

References

Bruni, R., Giacobazzi, R., Gori, R., Garcia-Contreras, I., & Pavlovic, D. (2020). Abstract extensionality: On the properties of incomplete abstract interpretations. *Proceedings of the ACM on Programming Languages*, *4*(POPL), Article 28, 1–28. https://doi.org/10.1145/3371096

Fisher, S., & Mandel, D. R. (2021). Risky-choice framing and rational decision-making. *Philosophy Compass*, *16*(8), e12763. https://doi.org/10.1111/phc3.12763

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*, 263–291. http://dx.doi.org/10.2307/1914185

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, *39*, 341–350. http://dx.doi.org/10.1037/0003-066X.39.4.341

Macdonald, R. R. (1986). Credible conceptions and implausible probabilities. *British Journal of Mathematical and Statistical Psychology*, *39*, 15–27. http://dx.doi.org/10.1111/j.2044-8317.1986.tb00842.x

Mandel, D. R. (2014). Do framing effects reveal irrational choice? *Journal of Experimental Psychology: General*, *143*, 1185–1198. http://dx.doi.org/10.1037/a0034207

Mandel, D. R. (2021). Theories, queries, "frames" and linguistic games: commentary on Wall, Crookes, Johnson & Weber (2020) (and the literature on risky-choice framing). March 19. https://doi.org/10.31234/osf.io/c5bf4

McKenzie, C. R. M., & Nelson, J. D. (2003). What a speaker's choice of frame reveals: Reference points, frame selection, and framing effects. *Psychonomic Bulletin & Review*, *10*, 596–602. https://doi.org/10.3758/BF03196520

Sher, S., & McKenzie, C. R. M. (2006). Information leakage from logically equivalent frames. *Cognition*, *101*, 467–494. http://dx.doi.org/10.1016/j.cognition.2005.11.001

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*, 453–458. http://dx.doi.org/10.1126/science.7455683

Probably, approximately useful frames of mind: A quasi-algorithmic approach

Mihnea C. Moldoveanu

Rotman School of Management & Desautels Centre for Integrative Thinking, University of Toronto, Toronto, ON M5S 3E6, Canada
mihnea.moldoveanu@rotman.utoronto.ca

doi:10.1017/S0140525X22001078, e235

Abstract

Frames for interpreting situations are necessary in the face of time constraints for action and indeterminacy of the "right or optimal thing to do" given multiple objectives but not all frames are equally useful. We need a way of modeling representational frames according to the informational gain of using them and the computational cost of synthesizing a decisive reason for acting from them.

Bermúdez (2020; target article) makes persuasive arguments against taking extensionality ("irrelevance of an agent's representation of options to choices among them") to be an inviolable part of rationality – and highlights the "rational" use of different representations of "the same" option space to describe and prescribe (human) action. Specific frames are useful for selecting among equilibria in competitive games. The *existence* of an equilibrium does not entail knowledge of how one gets there, let alone how one gets there quickly. Can we proceed beyond an acceptance of reasonable violations of extensionality to study the *degree to which* and *probability that* any particular frame is useful in a context?

Think of a frame F as a representation of a state of affairs that generates a decisive reason for undertaking a particular action in a specific context – one that functions as a cause for taking an action, in the counterfactual sense: If R were not true or valid, then action A would not be undertaken. An intuitive way to measure its usefulness is to ask: How *quickly* does it produce such a reason? – or – How much thinking is required to generate a reason R for acting from a frame F that structures the representation of the facts relevant to a situation? The computational complexity (Cook, 1971; Papadimitriou, 1994) of calculating R from F can operationalize the "cost of rationalizable action" for a frame by measuring the dependence of the number of operations required to get from F to R on the number of variables picked out by F as relevant. Also, how much closer to R do we get with every operation that proceeds from F as a starting point – how much *information* per unit of calculation do we get as we think our way from F to R?

Informational gain and computational cost. Take frame (F) to be: "competitive game with perfect information, N players and M strategies each": 'being a strategy in the (right) Nash equilibrium NE of strategies most likely to be selected by all $N-1$ others' is the decisive reason (R) the frame can supply – upon due calculation of NE. How difficult is it to get from F to R? In the worst case, the number of operations required to compute a Nash equilibrium grows exponentially in the number of players and strategies (Daskalakis, 2008). The cost may be acceptable for the (infrequently encountered) 2 player +2-strategy games used to teach and talk about theoretical games in which all agents maximize their own utilities, know the strategies and payoffs of all other players and know that all others have the same knowledge of everyone's utility–strategy profiles and are capable of the calculations required to get to (the right) equilibrium, and are rational. But, in situations involving many players and many strategies (more frequently encountered in practice) that are not susceptible to a shortcut, the cost of getting from F to R may be prohibitive or crippling.

The benefit of "thinking one step further" down the path from F to R may not be monotonically increasing in the number of calculative steps performed (Moldoveanu, 2009). It depends on the

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

frames used by *other* interactants and on the logical depth to which *they* reason. When the computational capacity or computational cost–benefit profiles of other players are such that they do not think their way(s) to *any* equilibrium, the use of frames that address the computational prowess of other players via "cognitive hierarchies" ("'savvy strategists," "neophytes") can help create more accurate representations (Ho, Camerer, & Chong, 2004).

"I–we" re-framing of games featuring interactions among joint and individual payoffs (Bacharach, 2006) is a way of simplifying the process of generating a reason for acting in a certain way from a frame for representing a social situation by helping one select from a number of equilibria via *priming* a particular response. To act consistently, one would have to assume the prime works in the same way for the others, so, at least *some* interactive reasoning is required. But "game models" of interactive decisions may be replaced by "social frames" (Fiske, 1992): In an "authority ranking" frame in which A sees her and B's actions as confirming or disconfirming A's (respectively, B's) power, rank, or prestige in the eyes of *N* observers watching the interaction, a(n easily justifiable) option – for which A has a decisive reason predicated on the use of the frame – is to take the action most likely to "put *B* in his place." This may be to act so as to upend the rules of the game or to walk away from it.

Efficiency can trump representational accuracy in *individual* deliberations. Take the case of an individual deciding among outcomes that must fulfill *many* objectives (Keeney & Raiffa, 1993). If emotions truly influence decisions in ways that varies with at least six different attributes of an option (uncertainty, pleasantness, attentional activity, anticipated effort, controllability of self *of*, and responsibility of others *for* outcome) (Lerner, Li, Valdesolo, & Kassam, 2015) – that supplies an example of a six-dimensional objective function. Given the computational complexity of multi-objective optimization, the use of context-adaptive frames that prime a small subset of the attributes to induce the quick generation of a reason makes sense, given time constraints: One tries to avoid being "caught in mid-thought" by the hazardous flow of life (Moldoveanu, 2011).

Making the computational complexity- and informational gain-aware selection of frames to match contexts that require decisive, quick action part of a "rationality toolbox" challenges more of what we thought we knew about rational choice than just the extensionality of its representations. Given it is *not* reasonable to require a rational agent to know all of the logical consequences of what she knows and also not reasonable for one to not think about the first-order consequences of a representation of a situation ("The exam is on Tuesday" & "Today is Tuesday"→"The exam is today") – where, along the "depth of reasoning" dimension, do we demarcate between "rational" and "irrational"?

Financial support. This research was funded by the Desautels Centre for Integrative Thinking, Rotman School of Management, University of Toronto.

Conflict of interest. None.

References

Bacharach, M. (2006). *Beyond individual choice: Teams and frames in game theory*. Princeton University Press.

Bermúdez, J. L. (2020). *Frame it again: New tools for rational thought*. Cambridge University Press.

Cook, S. (1971). *The complexity of theorem proving procedures*. Proceedings of the Third annual ACM Symposium on Theory of Computing, pp. 151–158.

Daskalakis, K. (2008). *The complexity of Nash equilibria*. Doctoral dissertation, University of California, Berkeley.

Fiske, A. P. (1992). The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychological Review*, *99*, 689–723.

Ho, T.-H., Camerer, C. F., & Chong, J.-K. (2004). A cognitive hierarchy model of games. *Quarterly Journal of Economics*, *119*, 861–898.

Keeney, R. L., & Raiffa, H. (1993). *Decisions with multiple objectives*. Cambridge University Press.

Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. *Annual Review of Psychology*, *66*, 799–823.

Moldoveanu, M. C. (2009). Thinking strategically about thinking strategically: The computational structure and dynamics of managerial problem selection and formulation. *Strategic Management Journal*, *30*, 737–763.

Moldoveanu, M. C. (2011). *Inside man: The discipline of modeling human ways of being*. Stanford University Press.

Papadimitriou, C. (1994). *Computational complexity*. Addison Wesley.

Competing reasons, incomplete preferences, and framing effects

Richard Pettigrew

Department of Philosophy, University of Bristol, Bristol BS6 6JL, UK
Richard.Pettigrew@bristol.ac.uk
richardpettigrew.com

doi:10.1017/S0140525X22000887, e236

Abstract

The quasi-cyclical preferences that Bermúdez ascribes to Agamemnon and others in analogous situations do not best represent them. I offer two alternative accounts. One works best if the preference ordering is taken to be the agent's personal betterness ordering of acts; the other works best if it is taken to provide a summary of the agent's dispositions to act.

According to Bermúdez, it is rational for Agamemnon to prefer the act of killing Iphigenia (act o) to the act of not killing her (act o') when the former is framed as following Artemis's will (frame $F_2(o)$) and the latter is framed as failing his ships and people (frame $F(o')$), while at the same time preferring not to kill her (o') to killing her (o) when the former is framed as failing his ships and people (frame $F(o')$) and the latter is framed as murdering his daughter (frame $F_1(o)$). And, Bermúdez says, such preferences are rational even when Agamemnon is fully aware that $F_1(o)$ and $F_2(o)$ are frames for the same option. That is, he claims that Agamemnon's preferences are these, and they are rational:

$$F_1(o) \prec F(o') \quad F(o') \prec F_2(o) \quad F_1(o) \prec F_2(o) \qquad (1)$$

I agree that there are rationally permissible attitudes in the vicinity of (1), but the ones Bermúdez ascribes to Agamemnon aren't them. In this note, I'll describe them and say why I think they better represent Agamemnon's situation.

Bermúdez argues that Agamemnon's quasi-cyclical preferences are rational because two competing reasons are in play: the value of following Artemis's will, on the one hand, and the value of his daughter's life, on the other. Agamemnon cannot settle how he wishes to weigh these two reasons against each other. If he were to give more weight to the former, he'd kill his daughter; if the latter, he would not. But he can't decide which he favours.

Here's another way we might describe Agamemnon's situation. Unable to decide how to weigh the two competing reasons against

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

one another, his preferences are simply incomplete. Knowing that following Artemis's will is the same act as murdering his daughter, he should be indifferent between those two framings. But he neither prefers the act o that these two framings frame to the act o' of not killing Iphigenia, framed as failing his ships and his people, nor disprefers it, nor is indifferent between the two.

That is, his preferences are as follows:

$$F_1(o) \sim F_2(o), \text{ and no further comparisons.} \quad (2)$$

Now, how would Agamemnon choose if these were his preferences? When you face a choice between two acts and you prefer neither to the other, you are rationally permitted to choice either. But notably in the cases Bermúdez describes, the agents don't simply choose at random. If the value of following Artemis's will is made salient and the value of his daughter's life is not, Agamemnon will kill Iphigenia; if the value of his daughter's life is made salient and the value of following Artemis's will is not, he will not. So, Agamemnon is different from someone with incomplete preferences between two acts who simply chooses at random when faced with a choice between them. You might think it is an advantage of Bermúdez's account that it captures that.

Whether this is an advantage or not depends on what we take the agent's preference ordering to be. Here are two alternatives: (a) it records the agent's judgments of betterness (cf. Broome, 1991); (b) it provides a summary of the agent's actual behaviour or their behavioural dispositions (cf. Samuelson, 1938, 1948). If $\prec$ records Agamemnon's judgments of betterness, then (2) gives the correct account. After all, Agamemnon knows that it is acts that are better or worse than one another, not framings, so while he might have to define his betterness ordering over framings of acts because he is not able to represent an act to himself without framing it in some way, when he knows that two frames are presentations of the same act, he should be indifferent between them because he knows they are equally good. On the other hand, if $\prec$ records his dispositions to behave in certain ways, Bermúdez's proposal (1) is more plausible.

Nonetheless, I think there is a third account that is more plausible than (1) in the cases Bermúdez considers and when the preference ordering is taken to record behaviour. On this account, we say that Agamemnon's preferences are in fact complete at any time, but they change from one time to another depending on which of the two competing reasons is salient to him. So, when the value of following Artemis's will is salient, his preferences are:

$$F(o') \prec F_1(o) \quad F(o') \prec F_2(o) \quad F_1(o) \sim F_2(o) \quad (3.1)$$

But when the value of his daughter's life is salient, his preferences are:

$$F_1(o) \prec F(o') \quad F_2(o) \prec F(o') \quad F_1(o) \sim F_2(o) \quad (3.2)$$

On this account, there is no quasi-cyclicity and yet Agamemnon's pattern of behaviour is recorded, unlike in (2). But it also has the advantage that it highlights something troubling about Agamemnon's preferences from the point of view of rationality. If your preferences change, they are exploitable. Suppose you prefer act o to act o' at one time and act o' to act o at a later time. Then I can offer you a choice between o and o' at the first time, when you'll choose o; then, at the later time, there will be some price such that, if I offer you the option to switch to o' for that price, you'll take it, because you then prefer o' to o. So you'll end up with o' minus the price, when you could have chosen o' costlessly at the outset and then not switched later. Of course, in order for someone to reliably exploit you like this, your change of preferences must be predictable or manipulable. But that is exactly how things are for Agamemnon. One need only make one or other of the competing reasons salient to him to manipulate his decision.

So I think that (2) is a better account of Agamemnon's preferences than (1) if those preferences give a betterness ordering, while (3) is better than (1) if they give a summary of behaviour.

Financial support. The research was funded by a British Academy Mid-Career Fellowship.

Conflict of interest. None.

References

Broome, J. (1991). *Weighing goods*. Oxford University Press.

Samuelson, P. A. (1938). A note on the pure theory of consumers' behaviour. *Economica*, *5*(17), 61–71.

Samuelson, P. A. (1948). Consumption theory in terms of revealed preference. *Economica*, *15*(60), 243–253.

Explaining bias with bias

Krzysztof Przybyszewski[a], Dorota Rutkowska[b] and Michał Białek[c]

[a]Department of Economic Psychology, Kozminski University, 03-301 Warsaw, Poland; [b]Faculty of Psychology, Warsaw University, 00-183 Warsaw, Poland and [c]Faculty of Historical and Pedagogical Sciences, Institute of Psychology, University of Wroclaw, 50-527 Wrocław, Poland
crispy@kozminski.edu.pl
dorota.rutkowska@psych.uw.edu.pl
michal.bialek3@uwr.edu.pl

doi:10.1017/S0140525X22001091, e237

Abstract

Bermúdez argues that a framing effect is rational, which will be true if one accepts that the biased editing phase is rational. This type of rationality was called procedural by Simon. Despite being procedurally rational in the evaluation phase framing effect stems from biased way we set a reference point against which outcomes are compared.

The very concept of rational framing effect deserves a thorough analysis. The global rationality of economic man as Simon (1955) puts it (or absolute, Olympian, economic rationality) is definitely not the one contained in the "rational framing effect" phrase used by Bermúdez. The type of rationality in Bermúdez's paper is what in Simon's work is dubbed (1972, 1976) a procedural rationality. In this view, when a decision is procedurally rational it means that it is the outcome of an appropriate deliberation, or as dual-system proponents would say reasoned and effortful process (e.g., Stanovich & West, 2000, see

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

also Stanovich, West, & Toplak, 2011). The stress is put on the choice of the proper reasoning rules regardless of the quality of data the decision was taken upon. Hence, the economically irrational decision, which was effortfully reasoned, should be considered procedurally rational. What follows, increase in the cognitive effort should enhance the procedural rationality and in turn should reduce the framing effect (e.g., McElroy & Seta, 2003; Simon, Fagley, & Halleran, 2004). However, when the Asian disease scenario was labelled as "medical" as opposed to "statistical," the framing effect was increased although the decision took longer and was more effortful (Igou & Bless, 2007). This example shows that procedurally rational process may produce the biased outcome when the very core of the problem is wrongly understood because of its biased formulation.

The framing effect has been described in the prospect theory (Kahneman & Tversky, 1979). In this framework, a decision is a sequential process in which the methods of achieving a goal are elicited, then edited and finally evaluated. Each of the stages may be judged against some rationality criteria and found to be faulty in many ways. Prospect theory separates the editing phase from the evaluation phase. In the editing phase the representation of a problem is created: The prospects are coded against a reference point (i.e., the outcome is framed). The prospects become gains or losses and then as such are subject to evaluation. Prospect theory states that the value (or utility) carriers are the changes of utility, rather than the value of utility itself. This implies that the prospects should be treated as deviations from a reference point rather than as absolute values, and this opens the possibilities for many biases stemming from the fact that reference points may be the status quo but also an aspiration level (e.g., an employee who gets the bonus of 1,000 USD would consider it a gain, while they would consider it a loss if they expected the bonus of 2,000 USD).

The list of potential culprits – biases related to the editing – is quite long. The least complex of them is the anchoring effect (Tversky & Kahneman, 1974). It consists in imposing the reference point, which in turn produces the biased value of a prospect. Among others and more complex one can find the translation effect (Abelson & Levi, 1985), the endowment effect (Kahneman, Knetsch, & Thaler, 1991), the disposition effect (Dacey & Zielonka, 2008), and the certainty effect (Kahneman & Tversky, 1979).

Most of the arguments of Bermúdez about the rationality of the framing effect refer to the subsequent evaluation phase. Here, people tend to violate normative models, for example, by differently treating gains and losses (losses appear larger than gains) and distort probabilities (sure outcomes are weighted as more important than the uncertain ones). Bermúdez claims these distortions are rational. There is a possibility that the decision maker can be involved in "rational" evaluation of the prospects that are inherently faulty because of being the product of a biased editing.

A procedurally rational agent acts to maximize chances to achieve a goal, regardless of whether the goal was set rationally in the editing phase. As we argue, the goal setting itself is biased by a frame of a problem, and so is the entire decision. Bermúdez argues, and we agree, that loss-avoidance and gain-approach seem to be universally correct and beneficial. The problem with the framing effect, however, is that it is irrational because the same prospects are not only treated differentially, but also the same prospects are malleable – they change when framed differentially on the basis of individual's aspiration levels or imposed reference points.

We argue that the way we set a reference point against which outcomes are compared is temporarily unstable, subject to external manipulation (not only by gain/loss frames, see also anchoring effect, see Tversky & Kahneman, 1974), and thus unreliable to the extent that the decisions are unpredictable. Hence, displaying framing effect is merely procedurally rational consequence of biased editing phase. We conclude that Bermúdez is presenting the framing effect as rational, simply because the evaluation phase is correct. As we argue, there is another bias involved at an earlier stage, and the whole process of decision making may still be irrational. Bermúdez simply moved the source of the bias elsewhere.

Financial support. This study was funded by ALK grant: BST ALK 908.2.10 and NCN grant: 2020/38/E/HS6/00282.

Conflict of interest. None.

References

Abelson, R. P., & Levi, A. (1985). Decision making and decision theory. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology* (Vol. 1, 3rd ed., pp. 231–309). Random House.

Dacey, R., & Zielonka, P. (2008). A detailed prospect theory explanation of the disposition effect. *Journal of Behavioral Finance*, *9*, 43–50. https://doi.org/10.1080/15427560801897758

Igou, E. R., & Bless, H. (2007). On undesirable consequences of thinking: Framing effects as a function of substantive processing. *Journal of Behavioral Decision Making*, *20*(2), 125–142. https://doi.org/10.1002/bdm.543

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic Perspectives*, *5*(1), 193–206. doi: 10.1257/jep.5.1.193

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–292. https://doi.org/10.2307/1914185

McElroy, T., & Seta, J. J. (2003). Framing effects: An analytic–holistic perspective. *Journal of Experimental Social Psychology*, *39*(6), 610–617. https://doi.org/10.1016/S0022-1031(03)00036-2

Simon, A. F., Fagley, N. S., & Halleran, J. G. (2004). Decision framing: moderating effects of individual differences and cognitive processing. *Journal of Behavioral Decision Making*, *17*(2), 77–93. http://dx.doi.org/10.1002/(ISSN)1099-0771

Simon, H. (1955). A behavioural model of rational choice. *The Quarterly Journal of Economics*, *69*(1), 99–118. https://doi.org/10.2307/1884852

Simon, H. (1972). Theories of bounded rationality. In C. B. McGuire & R. Radner (Eds.), *Decision and organization* (pp. 161–176). North-Holland Publishing Company.

Simon, H. (1976). From substantive to procedural rationality. In S. J. Latsis (Ed.), *Method and appraisal in economics* (pp. 129–148). Cambridge University Press. https://doi.org/10.1017/CBO9780511572203

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate?. *Behavioral and Brain Sciences*, *23*(5), 645–665. http://dx.doi.org/10.1017/S0140525X00003435

Stanovich, K. E., West, R. F., & Toplak, M. E. (2011). The complexity of developmental predictions from dual process models. *Developmental Review*, *31*(2-3), 103–118. http://dx.doi.org/10.1016/j.dr.2011.07.003

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. http://dx.doi.org/10.1126/science.185.4157.1124

Quasi-cyclical preferences in the ethics of Plato, Aristotle, and Kant

Adam J. Roberts

Department of Philosophy, King's College London, London WC2R 2LS, UK
adam.roberts@oxon.org

doi:10.1017/S0140525X22001042, e238

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

Abstract

Bermúdez describes the extensionality principle as being "almost unquestioned." This claim might come as a surprise to philosophers who work on agency and ethics. In Kantian deontological ethics and in Platonic or Aristotelian virtue ethics, our preferences for outcomes can be rationally affected by how those outcomes are framed in terms of maxims and character traits.

Bermúdez describes the extensionality principle – that "preferences, values, and decisions should be unaffected by how outcomes are framed" – as receiving "almost unanimous acceptance" (target article, sect. 2.1), and even as being "almost unquestioned" (target article, sect. 1, para. 2). These claims might come as a surprise to philosophers who work on agency and ethics. In the language of moral philosophy, it sounds as if almost all decision theorists are consequentialists.

Moral theories are conventionally categorised as being either consequentialist, deontological, or virtue ethical. According to consequentialists, what fundamentally make an action morally good or bad are its consequences. According to deontologists, what fundamentally matter are instead the rules or principles which are embodied in an action. According to virtue ethicists, what fundamentally matter are instead the traits of character embodied in the action.

It might seem natural for a consequentialist to accept the extensionality principle. The facts of how an agent frames the outcomes of an action are facts about that agent before they are facts about those outcomes. A perfectly rational consequentialist agent would only be guided by facts about outcomes. In that sense, there is no place for frames in the rational ideal of a consequentialist.

There is a place for frames, however, in the rational ideals of deontologists and virtue ethicists. The principles and the character traits which matter on their views seem able to be thought of as framing outcomes. The same act can be good if it is thought of under the guise of one principle or trait, and bad if it is thought of under another, despite the outcome or the consequences being the same.

The most famous deontological moral philosopher is Kant. Kant believed that we must only act on principles or maxims which we could consent to anyone anywhere following. When we find ourselves unable to give that consent, one of the options open to us is to try to come up with a more nuanced version of that unacceptable principle. For example, I may propose to myself that I pursue a career in forestry just because I want to. The principle I need to be able to consent to is that of anyone pursuing a career in forestry should they want to. This principle is too simple, however: I would not consent to it in a situation in which everyone wanted to pursue a career in forestry. However, I might be able to consent to a principle of pursuing a career in forestry in times when, simultaneously, foresters are needed, I have the skills and means to become one, and I want to become one. In being guided by this more complex principle, I act just as I would have on the simpler one. The outcome is the same: I pursue a career in forestry. It is only an acceptable, preferable outcome, however, when it is chosen as framed by the more complex principle. In this way, Kantian moral theory can violate the extensionality principle.

Bermúdez's (target article, sect. 3.3.1) example of Agamemnon's dilemma might be reimagined in Kantian terms. If Agamemnon can consent to the principle of following Artemis's will in even this case, he can consent it universally. At the same time, he finds himself unable to consent to the principle of murdering his daughter, even though his act would be the same were he following Artemis's will or murdering his daughter.

This dilemma might be phrased in virtue ethical terms, too. Agamemnon is caught between the frames of faithfulness to his gods and of faithfulness to his family. In the *Republic*, Plato has Socrates imagine a situation in which you have borrowed weapons from a friend, but that friend has since gone mad (Plato, 2012, p. 8/331c). Honouring your debts requires you to return those weapons when your friend demands them. Handing weapons to a madman, though, would be viciously reckless. Honouring your debts is preferable to keeping the weapons from him, but keeping the weapons from him is preferable to recklessly giving them back.

Korsgaard (1996, p. 108, 2009, p. 15) uses this example as part of an argument that what agents must assess and choose between are what she terms "actions" rather than "acts." An action incorporates the principle or end for the sake of which the act is done. Returning the weapons to your friend is an act. Returning the weapons to your friend because they are the property of your friend, and they have asked you to return them, is an action.

Korsgaard reads both Kant and Aristotle, the philosopher most famous as a virtue ethicist, as subscribing to a theory of agency on which actions are our objects of choice and assessment. Bermúdez also appears to subscribe to something like this picture: He suggests that "there is no such thing as making choices over a purely extensional opportunity set, independent of any way of describing or framing the things in it" (target article, sect. 3.1, para. 5). In this respect and in respect of his belief in the coherence of quasi-cyclical preferences, Bermúdez would seem able to claim powerful allies beyond the disciplinary pale of decision theory.

I do not want to overstate the common ground: Bermúdez is working with a different conception of rationality to these philosophers, if one which is different by being less substantive and demanding. There might be empirical reasons to treat moral quasi-cyclical preferences as being special cases, or even not to think of moral reasoning in terms of frames and preferences. In addition, some of these philosophers might be unwilling to admit the possibility of our being unable to resolve conflicts between frames by subsuming them under higher frames. As Aristotle (2012, p. 36/1226b) puts it in the *Eudemian Ethics*, "those who have no aim before them are not liable to deliberate." If we are to choose between the urgings of two frames, we must possess some ground on which that choice can be made. The question can only be whether to think of that ground in terms of a higher frame.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Aristotle (2012). *Eudemian ethics*. (B. Inwood, & R. Woolf, Trans.) Cambridge University Press.

Korsgaard, C. M. (1996). *The sources of normativity*. Cambridge University Press.

Korsgaard, C. M. (2009). *Self-constitution: Agency, identity, and integrity*. Oxford University Press.

Plato (2012). *Republic*. (C. Rowe, Trans.) Penguin.

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

The framing of decisions "leaks" into the experiencing of decisions

Barry Schwartz[a] and Nathan N. Cheek[b]

[a]Haas School of Business, University of California, Berkeley, Berkeley, CA 94720, USA and [b]Peretsman Scully Hall, Princeton University, Princeton, NJ 08544, USA
bschwar1@swarthmore.edu
nncheek@princeton.edu
https://haas.berkeley.edu/faculty/schwartz-barry/
https://www.natecheek.com

doi:10.1017/S0140525X22000905, e239

Abstract

We connect Bermúdez's arguments to previous theorizing about "leaky" rationality, emphasizing that the decision process (including decision frames) "leaks" into the experience of decision outcomes. We suggest that the implications of Bermúdez's analysis are broadly applicable to the study of virtually all real-world decision making, and that the field needs a substantive and not just a formal theory of rationality.

> "[F]raming has an effect on decisions because it has an effect on experience."
> — Frisch (1993, p. 402)

Bermúdez argues that the objects of preference are framed outcomes, so that the frame is an ineliminable part of the objects under consideration. Further, this feature of preference is often a very good thing – frames highlight some dimensions and downplay others, and without them, many decisions might seem impossibly difficult to make. We think this is a wise and important insight, and it echoes points made by Keys and Schwartz (2007), who further emphasized that what people *experience* after making decisions are also *framed outcomes*. In Keys' and Schwartz's terminology, the decision process "leaks" into the subsequent experience of the decision's results (see also Frisch, 1993; Kahneman & Tversky, 1984). This "leakage" can be complicated, as in many of the "complex and multi-faceted" cases on which Bermúdez focuses, but it can also be simple, as when a "95% fat free" yogurt tastes less rich than a yogurt labeled "only 5% fat" (Sanford, Fay, Stewart, & Moxey, 2002). Either way, "leakage" means that frames shape not just decisions but also experiences in substantive ways.

When framing "leaks" into experience, apparently inconsistent decisions made under different frames may actually be consistent with the experience of the decisions themselves. Most approaches to rationality seem to assume that heuristics and biases exert their effects while a decision is being contemplated, but that once the decision is made, experiencing the results of the decision will be "path independent." The object of a decision will be experienced on its own, carrying no trace of how the decision was arrived at (but see Kahneman, 1994). Gilbert and Ebert (2002) compared this "illusion of intrinsic satisfaction" to the perceptual illusion of direct access to the world, describing the tendency to behave as if "hedonic experiences were due entirely to the enduring intrinsic properties of [decisions'] outcomes" (p. 511). But this variant of "naïve realism" is mistaken, as Bermudez and Keys and Schwartz (2007) argue.

The ideas explored across many examples by Bermúdez have broad and deep implications for thinking about the rationality of decisions beyond cases focused on how options are described or otherwise framed. Research on choice overload, for example, has shown that people may feel worse about their choice if it was made from a larger assortment of options, even if they would have made the same choice from a smaller assortment (Iyengar & Lepper, 2000; Schwartz, 2016). And people in poverty may be more likely to pursue government assistance after reflecting on aspects of their identity they are proud of, even though the amount of their benefits is no different (Hall, Zhao, & Shafir, 2014). In the first case, regrets about the losses from tradeoffs among a larger number of alternatives "leak" into the prospective consumption of the chosen option. In the second, affirming the self prevents the potential stigmatization of accepting government benefits from "leaking" into the experience of receiving them. In both cases, factors that might seem outside the scope of rational deliberation obviously matter in a substantive, though perhaps non-normative, way.

What Bermúdez calls "quasi-cyclical preferences" and what we call "leakage" imply that there are important limits on what formal principles of rationality can tell us, since there are surprisingly few cases where the formal principles apply in their strictest form. Even seemingly inconsequential changes to the situation may "leak" into experience, affecting one's subjective outcomes and hence the reasonableness of different choices. Formal rules of rationality may allow researchers to draw important normative conclusions based on minimal, widely accepted structural claims about rationality. However, once the importance of subjective experiences and the prevalence of "leakage" are taken into account, it becomes clear that much more needs to be known before anything approaching a satisfactory theory of rationality is in hand. What is needed is a *substantive theory of rationality* – a theory that considers the content and not just the structure of decisions, and that evaluates that content in light of the decision-maker's goals, experiences, and life as a whole.

A substantive theory of rationality must consider the consequences of decisions, very broadly construed. It must consider short- and long-term consequences, consequences to the self and to others, consequences that are central to the decision at hand and those that may be peripheral. It also takes seriously the notion, raised by Bermúdez as well as Schwartz (1994), that there is no such thing (in the real world) as a decision in a vacuum. Frames, prior experiences, incidental environmental cues, and more will always be present in some form, and their effects may often "leak" into experiences. It seems irrational (and silly) to pay more for 7UP in a yellow-green bottle than for the same drink in a green bottle if the standard of rationality assumes the 7UP is experientially the same regardless of its container (Gladwell, 2005). But a more psychological and practical view of rationality might point out that paying more for a drink that tastes different – in this case, more like a mix of lemon and lime – may be reasonable if it better satisfies a person's tastes.

What we suggest is that the entire field of judgment and decision making has to a large degree answered questions about how well our decisions conform to formal principles of rationality instead of questions about how well our decisions serve substantive rationality (Keys & Schwartz, 2007). The field is not necessarily confused about what questions are being asked, though perhaps it is somewhat confused about the questions that fundamentally matter for a full understanding of decision-making as it

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

is actually experienced in everyday life – the questions that the rest of us want answered. The main reason for this, we suspect, is that the "real" questions are much harder. But we hope that Bermúdez's many insights will fuel a broader shift in how we as a field think about not only framing, but also rationality more broadly, spurring us address the "hard" questions as best we can.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Frisch, D. (1993). Reasons for framing effects. *Organizational Behavior and Human Decision Processes*, *54*, 399–429.

Gilbert, D. T., & Ebert, J. E. (2002). Decisions and revisions: The affective forecasting of changeable outcomes. *Journal of Personality and Social Psychology*, *82*, 503–514.

Gladwell, M. (2005). *Blink: The power of thinking without thinking*. Little Brown.

Hall, C. C., Zhao, J., & Shafir, E. (2014). Self-affirmation among the poor: Cognitive and behavioral implications. *Psychological Science*, *25*, 619–625.

Iyengar, S. S., & Lepper, M. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology*, *79*, 995–1006.

Kahneman, D. (1994). New challenges to the rationality assumption. *Journal of Institutional and Theoretical Economics*, *150*, 18–36.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, *39*, 341–350.

Keys, D. J., & Schwartz, B. (2007). "Leaky" rationality: How research on behavioral decision making challenges normative standards of rationality. *Perspectives on Psychological Science*, *2*, 162–180.

Sanford, A. J., Fay, N., Stewart, A., & Moxey, L. (2002). Perspective in statements of quantity, with implications for consumer psychology. *Psychological Science*, *13*, 130–134.

Schwartz, B. (1994). *The costs of living: How market freedom erodes the best things in life*. Norton.

Schwartz, B. (2016). *The paradox of choice: Why more is less* (2nd ed.). Ecco Press.

Incomplete preferences and rational framing effects

Shlomi Sher[a] and Craig R. M. McKenzie[b]

[a]Department of Psychological Science, Pomona College, Claremont, CA 91711, USA and [b]Department of Psychology and Rady School of Management, University of California, San Diego, La Jolla, CA 92093, USA
Shlomi.Sher@pomona.edu, https://www.pomona.edu/directory/people/shlomi-sher
cmckenzie@ucsd.edu, https://pages.ucsd.edu/~mckenzie/

doi:10.1017/S0140525X2200111X, e240

Abstract

The normative principle of description invariance presupposes that rational preferences must be complete. The completeness axiom is normatively dubious, however, and its rejection opens the door to rational framing effects. In this commentary, we suggest that Bermúdez's insightful challenge to the standard normative view of framing can be clarified and extended by situating it within a broader critique of completeness.

Bermúdez raises an important challenge to the traditional view that rational choice must be invariant to framing. We are sympathetic to his primary conclusions – that framing effects need not be irrational, and in some cases rationality may require sensitivity to different frames. However, we believe that these conclusions can be put on a firmer foundation by deriving them from a more general critique of *completeness* (defined below), a core axiom in traditional models of rational choice. This reformulation clarifies the scope of the analysis (i.e., regarding the conditions under which rational actors may exhibit framing effects) and sharpens its applications (e.g., regarding game-theoretic equilibria).

First, a terminological clarification: In relating normative principles to empirical findings, it is useful to distinguish between *descriptive frames* and *conceptual frames* (cf. Druckman, 2001). The former refers to the overt description of a choice problem that is given to an agent (e.g., beef described as "80% lean"). The latter refers to the agent's internal representation of the problem – that is, their conception of relevant options, values, and reasons. In framing experiments, the descriptive frame is manipulated. The conceptual frame is a theoretical construct, which is sometimes invoked by theorists in explaining the descriptive frame's observed effects.

The normative principle of *description invariance* concerns descriptive frames. It says that equivalent descriptions should not lead to different decisions. However, the normative validity of description invariance depends on two critical implicit assumptions: (1) descriptive frames must not "leak" distinct, choice-relevant information; and (2) rational preferences must be complete (Sher & McKenzie, 2011).

In prior work (Sher & McKenzie, 2006), we have argued that some widely studied descriptive frames are not information equivalent, violating (1). But for the purposes of this commentary, we assume that information is constant across descriptive frames, and examine the second implicit assumption.

The completeness axiom states that the normative ranking of alternatives is everywhere well-defined: For any options, a, b, either a is definitely better than b vis-à-vis the agent's values ($a \succ b$), b is better than a ($b \succ a$), or the two options are precisely equivalent ($a \sim b$). Completeness has unreasonable normative implications – for example, monetary indifference points for all goods must be defined to infinite precision. Accordingly, some economists and philosophers have argued that, despite its mathematical convenience, the completeness axiom is not a plausible requirement of rationality (e.g., Aumann, 1962; Raz, 1985). In recent years, economists have developed increasingly sophisticated models of rational choice that allow for incomplete preferences (e.g., Mandler, 2005).

Rejecting the completeness requirement immediately opens the door to rationally permissible framing effects. In a finite choice menu, there may be distinct alternatives, a, b, unranked relative to one another, neither of which is outranked by any other option in the menu. If a is chosen under one descriptive frame and b under another, choices are frame-dependent but never suboptimal.

Why are rational preferences sometimes incomplete, and when may rational framing effects occur? Incomplete normative rankings may trace back to two distinct sources – value imprecision and value conflict. (Owing to space limitations, we omit Schick's [1997] value "ambiguity," which may be regarded as a third source of incompleteness.) In cases of *imprecision*, the agent's underlying values are coarse-grained (e.g., a mug worth between $5 and $10, with no well-defined indifference point). Framing effects in one-shot choice are then normatively neutral, neither good nor bad, provided that definitely outranked options

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

are never chosen. (In *repeated* choice, however, subtler normative issues arise; Sher, Müller-Trede, & McKenzie, 2022.)

In cases of *conflict*, the agent accepts two "schemes of valuation," V_1 and V_2, when all else is equal, yet is not committed to a superordinate principle that reconciles them when they come into conflict. For example, in Sartre's (1946/2007) famous example of a young man torn between supporting his mother and taking up arms against the Nazis, V_1 may rank acts according to a son's duties, V_2 according to a citizen's.

V_1 and V_2 may be regarded as distinct *conceptual frames*, which, in isolation, generate distinct preference orders, $\succcurlyeq_1$ and $\succcurlyeq_2$. When $a \succ_1 b$ but $b \succ_2 a$, the normative ranking of alternatives may not be well-defined. As Bermúdez suggests, rationality then requires a kind of *joint frame-sensitivity*: To understand what is at stake in the problem, and what is required to solve it, the agent must be able to enter into both evaluative frames, identifying points of both contact and divergence ("perspectival flexibility"). Insofar as different descriptive frames make different conceptual frames salient, some behavioral framing effects may perhaps be regarded as manifestations of the requisite joint sensitivity.

Recast in these terms, some of Bermúdez's applications come into clearer focus. For example, in game theory, the payoff matrix represents the agents' subjective utilities, not objective material outcomes. When preferences are incomplete, a given objective outcome need not have a uniquely defined utility; hence a game need not have a unique payoff matrix. Different schemes of valuation (e.g., Bermúdez's "I-frame" vs. "we-frame") will be represented by different matrices; some may have pathological properties (e.g., an undesirable equilibrium), others not. In some cases, agents may then resolve their internal value conflict (i.e., complete their incomplete preferences) in a way that makes the resulting game non-pathological.

Of course, a rational analysis of value conflict – and of the framing effects it may generate – must ultimately go further, and formulate normative principles of value integration. For Sartre's pupil, which methods of reconciling conflicting values are rational, which are not, and why? Leading formal models in the decision sciences are silent on this question, because they assume that preferences are complete, and hence that choice-relevant values have already, *somehow*, been integrated. The problem of value integration thus remains in the misty pre-theoretical realm of "decision structuring," where our understanding of rationality is more art than science (von Winterfeldt, 1980). In accord with the target article, we suspect that the most complex and important framing effects reside in this normatively uncharted territory – where incomplete preferences, arising from value conflict, must be completed in the act of choice.

Financial support. S.S. is supported by a Scholar Award from the James S. McDonnell Foundation. C.R.M.M. is supported by a National Science Foundation grant (SES-2049935).

Conflict of interest. None.

References

Aumann, R. (1962). Utility theory without the completeness axiom. *Econometrica*, *30*, 445–462.

Druckman, J. N. (2001). The implications of framing effects for citizen competence. *Political Behavior*, *23*, 225–256.

Mandler, M. (2005). Incomplete preferences and rational intransitivity of choice. *Games and Economic Behavior*, *50*, 255–277.

Raz, J. (1985). Value incommensurability: Some preliminaries. *Proceedings of the Aristotelian Society*, *86*, 117–134.

Sartre, J.-P. (1946/2007). *Existentialism is a humanism*. Yale University Press.

Schick, F. (1997). *Making choices*. Cambridge University Press.

Sher, S., & McKenzie, C. R. M. (2006). Information leakage from logically equivalent frames. *Cognition*, *101*, 467–494.

Sher, S., & McKenzie, C. R. M. (2011). Levels of information: A framing hierarchy. In G. Keren (Ed.), *Perspectives on framing* (pp. 35–63). Psychology Press.

Sher, S., Müller-Trede, J., & McKenzie, C. R. M. (2022). *Choices without preferences: Principles of rational arbitrariness*. Manuscript in preparation.

von Winterfeldt, D. (1980). Structuring decision problems for decision analysis. *Acta Psychologica*, *451*, 71–93.

The ecological benefits of being irrationally moral

Elisabetta Sirgiovanni

Department of Molecular Medicine, Museum of the History of Medicine, Sapienza University of Rome, 00185 Rome, Italy
elisabetta.sirgiovanni@uniroma1.it

doi:10.1017/S0140525X2200098X, e241

Abstract

Trolley-like dilemmas are other cases of what Bermúdez refers to as (conscious) quasi-cyclical preferences. In these dilemmas, identical outcomes are obtained through morally non-identical actions. I will argue that morality is the context where descriptive invariance and ecological relevance may be crucially distinguished. Logically irrational moral choices in the short term may promote greater social benefits in the longer term.

Framing effects offer a classic example of irrationality. Cases like Aeschylus's *Agamemnon* and Shakespeare's *Macbeth*, brilliantly discussed by José Luis Bermúdez, may subvert this consensus view. In my impression, the reason why this is evidently the case is that these are moral dilemmas as the agents face conflicts of moral nature. I will argue that morality is specifically the context where descriptive invariance and ecological relevance may be crucially distinguished.

The famous trolley-like dilemmas are other cases of what Bermúdez refers to as (conscious) quasi-cyclical preferences. Notoriously, switching the lever to deviate the trolley and sacrificing one person (F1o) is preferred to letting the five people on the tracks die (A), which is itself preferred to pushing a man out from the footbridge to stop the trolley (F2o). Even if frame-insensitive responses (F1o > A < F2o) can be primed through order presentation (Schwitzgebel & Cushman, 2012), responses to "complex" frames are perceived as *highly conflictual* (Koenigs et al., 2007). Complex frames are the crying baby (i.e., would you smother your crying newborn to save yourself and four townspeople hiding from a Nazi incursion?) (Cushman & Greene, 2012), the transplant (i.e., would you harvest the organs from a pizza boy who entered the ER to save five patients in need of a transplant?) (Thomson, 1976), the cabin boy (i.e., would you kill and eat the cabin boy to save yourself and other starving shipwrecked sailors?) (Harris, 2020), triage in a pandemic (Kneer & Hannikainen, 2022), and many others (Edmonds, 2014; Koenigs

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

et al., 2007). These situations are logically indistinguishable from the bystander case (i.e., switching the lever). The irrationality of intuitive non-utilitarian responses in these complex frames (i.e., a "no" response) is the reason why these responses, associated with greater activation of emotional brain areas, are discarded by some as ethically unreliable (Greene, 2016; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001, 2004). Despite general appeals attributing a crucial role to emotions in morality (Greene & Haidt, 2002) and even if within a utilitarian (rather than a deontological) perspective (Greene, 2008), this so-called dual-system approach ends up in a traditional merger between morality and rationality.

Emotionally different frames of the *same* outcome (one vs. five) in moral scenarios, I instead believe, may differ informationally by translating into dissociable moral attitudes. Framing it into Bermúdez's discussion, we may call this the *non-extensionality* of moral preferences, values, and actions. This discourse falls within the concept of "ecological rationality" proposed and defended over the years by Gigerenzer and colleagues (Gigerenzer, 2015; Todd & Gigerenzer, 2012). Accordingly, we may distinguish *logical* rationality or "axiomatic definitions of rationality that economic models draw upon" (Todd & Gigerenzer, 2012, p. 425) from *ecological* rationality. The latter consists of contextual evaluations of decision processes that go beyond the identification of logical consistency, but rather refer "to how well they match the environments in which they are used" (ibid.). More precisely, ecological rationality expresses in encoding covert scenarios or surplus information between the lines (Sher & McKenzie, 2006), a significant ability for moral cognition. In trolley-like dilemmas, identical outcomes are obtained through morally non-identical actions (Cushman, 2013). No less importantly, informational differentiation in these dilemmas is mostly evoked in a native language, where the emotionality of content sounds clearer (Costa et al., 2014; Geipel, Hadjichristidis, & Surian, 2016).

Sacrificial moral dilemmas are extraordinary cases where the logical accuracy of a frame does not necessarily imply its moral applicability. Interestingly enough, some utilitarian scholars (Kahane, 2012; Kahane, Everett, Earp, Farias, & Savulescu, 2015; Kahane et al., 2017) have defended this position. They reject the idea that cool harm-endorsing responses to complex frames are expressions of genuine utilitarianism, that is, "impartial concern for the greater good" (Kahane et al., 2015, p. 194), as they less arguably are in the bystander case. The scale they designed, the Oxford Utilitarian Scale (OUS), shows that responses to the two models through which utilitarianism is framed (impartial beneficence vs. instrumental harm) measure different individual traits and moral attitudes. Consistent with previous studies (Bartels & Pizarro, 2011; Glenn, Koleva, Iyer, Graham, & Ditto, 2010; Koenigs, Kruepke, Zeier, & Newman, 2012), the scale associates non-clinical psychopathic tendencies with instrumentally harmful responses but, more importantly, *not* with an endorsement of impartial beneficence. Other morally questionable attitudes, such as rational egoism or lenient moral attitude, were also linked to "purely extensional approaches." Besides, utility calculation – as intended by its originators – must consider also distal consequences of an action, referring to the class of actions to which the action belongs, and not only proximal consequences as if the action were single and insulated (Warnock, 2003).

Sensitivity to emotionally salient frames in the moral domain, for example, to *personal force* (Greene et al., 2009) and instrumental harm (Everett, Faber, Savulescu, & Crockett, 2018), is an efficient heuristic to prevent harm to others in real-life situations that also predicts expectations toward other people's moral behavior. Evolution may have selected for reactive harm-aversion as this favors cooperation, reciprocation, and trust (Cosmides, Guzmán, & Tooby, 2018). Harm-aversion is linked to a social cognition brain area (i.e., right temporoparietal junction [RTPJ]) (Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010), and it indeed increases the chances of being chosen as romantic or business partners (Everett, Pizarro, & Crockett, 2016).

We may also cast doubt on the association – mentioned by Bermúdez – between the recruitment of a brain area (i.e., dorsolateral prefrontal cortex [DLPFC]) and cognitive control processing, especially as this association justifies the appropriateness of utilitarian choices in complex frames (Greene, Nystrom, Engell, Darley, & Cohen, 2004, 2008). Brain stimulation studies suggest the opposite of what this model predicts, that is, disrupting the right-DLPFC actually increases utilitarian responses (Tassy, Oullier, Mancini, & Wicker, 2013), while enhancing the left-DLPFC increases non-utilitarian responses (Kuehne, Heimrath, Heinze, & Zaehle, 2015). Agreeing with Bermúdez, the literature is now converging on the idea that self-control is closer to emotional regulation than cognitive effort (Fujita, 2011; Inzlicht & Friese, 2021; Magen, Kim, Dweck, Gross, & McClure, 2014). Accordingly, cognitive load delays reaction time, but does not determine any change in footbridge responses (Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008). We discern two aspects of self-control – inhibiting intuitive response and delaying gratification (Duckworth & Kern, 2011). Failing to inhibit automatic non-utilitarian choices may be seen as gratification delay if looked at through ecological lenses. These choices may be logically irrational in the short term, as they imply a greater number of deaths, but promote greater social benefits in the longer term, as they limit pernicious attitudes like justifying the sacrifice of innocents.

Financial support. The author received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Bartels, D. M., & Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition*, *121*, 154–161. doi: 10.1016/j.cognition.2011.05.010

Cosmides, L., Guzmán, R. A., & Tooby, J. (2018). The evolution of moral cognition. In A. Zimmerman, K. Jones, & M. Timmons (Eds.), *The Routledge handbook of moral epistemology* (pp. 174–228). Routledge Publishing.

Costa, A., Foucart, A., Hayakawa, S., Aparici, M., Apesteguia, J., Heafner, J., & Keysar, B. (2014). Your morals depend on language. *PLoS ONE*, *9*(4), e94842. doi: 10.1371/journal.pone.0094842.

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review*, *17*(3), 273–292. doi: 10.1177/1088868313495594

Cushman, F., & Greene, J. D. (2012). Finding faults: How moral dilemmas illuminate cognitive structure. *Social Neuroscience*, *7*(3), 269–279. doi: 10.1080/17470919.2011.614000.

Duckworth, A. L., & Kern, M. L. (2011). A meta-analysis of the convergent validity of self-control measures. *Journal of Research in Personality*, *45*(3), 259–268. doi: 10.1016/j.jrp.2011.02.004

Edmonds, D. (2014). *Would you kill the fat man? The trolley problem and what your answer tells us about right and wrong*. Princeton University Press.

Everett, J. A. C., Faber, N. S., Savulescu, J., & Crockett, M. J. (2018). The costs of being consequentialist: Social inference from instrumental harm and impartial beneficence. *Journal of Experimental Social Psychology*, *79*, 200–216. doi: 10.1016/j.jesp.2018.07.004

Everett, J. A. C., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology*, *145*(6), 772–787. doi: 10.1037%2Fxge0000165

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

Fujita, K. (2011). On conceptualizing self-control as more than the effortful inhibition of impulses. *Personality and Social Psychology Review*, *15*(4), 352–366. doi: 10.1177/1088868311411165

Geipel, J., Hadjichristidis, C., & Surian, L. (2016). Foreign language affects the contribution of intentions and outcomes to moral judgment. *Cognition*, *154*, 34–39. doi: 10.1016/j.cognition.2016.05.010

Gigerenzer, G. (2015). On the supposed evidence for libertarian paternalism. *Review of Philosophy and Psychology*, *6*, 361–383. doi: 10.1007/s13164-015-0248-1

Glenn, A. L., Koleva, S., Iyer, R., Graham, J., & Ditto, P. H. (2010). Moral identity in psychopathy. *Judgment and Decision Making*, *5*, 497–505.

Greene, J. D. (2008). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral psychology, Vol. 3. The neuroscience of morality: Emotion, brain disorders, and development* (pp. 35–80). MIT Press.

Greene, J. D. (2016). Beyond point-and-shoot morality: Why cognitive (neuro)science matters for ethics. In S. M. Liao (Ed.), *Moral brains: The neuroscience of morality* (pp. 119–149). Oxford University Press.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*(3), 364–371.

Greene, J. D., & Haidt, J. D. (2002). How does moral judgment work? *Trends in Cognitive Sciences*, *6*(12), 517–523. doi: 10.1016/S1364-6613(02)02011-9

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, *107*(3), 1144–1154. doi: 10.1016/j.cognition.2007.11.004

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*, 389–400. doi: 10.1016/j.neuron.2004.09.027

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105–2108. doi: 10.1126/science.1062872

Harris, J. (2020). Why kill the cabin boy? *Cambridge Quarterly of Healthcare Ethics*, *30*(1), 4–9. doi: 10.1017/S0963180120000420

Inzlicht, M., & Friese, M. (2021). Willpower is overrated. *Behavioral and Brain Sciences*, *44*, e42. doi: 10.1017/S0140525X20000795

Kahane, G. (2012). On the wrong track: Process and content in moral psychology. *Mind and Language*, *25*(5), 519–545.

Kahane, G., Everett, J., Earp, B., Caviola, L., Faber, N., Crockett, M., & Savulescu, J. (2017). Beyond sacrificial harm: A two dimensional model of utilitarian psychology. *Psychological Review*, *125*(2), 131–164. doi: 10.1037/rev0000093

Kahane, G., Everett, J., Earp, B., Farias, M., & Savulescu, J. (2015). Utilitarian judgment in sacrificial dilemmas does not reflect impartial concern for the greater good. *Cognition*, *134*, 193–209. doi: 10.1016/j.cognition.2014.10.005

Kneer, M., & Hannikainen, I. R. (2022). Trolleys, triage and Covid-19: The role of psychological realism in sacrificial dilemmas. *Cognition & Emotion*, *36*(1), 137–153. doi: 10.1080/02699931.2021.1964940

Koenigs, M., Kruepke, M., Zeier, J., & Newman, J. P. (2012). Utilitarian moral judgment in psychopathy. *Social Cognitive and Affective Neuroscience*, *7*(6), 708–714. doi: 10.1093/scan/nsr048

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, *446*(7138), 908–911. doi: 10.1038/nature05631

Kuehne, M., Heimrath, K., Heinze, H.-J., & Zaehle, T. (2015). Transcranial direct current stimulation of the left dorsolateral prefrontal cortex shifts preference of moral judgments. *PLoS ONE*, *10*(5), e0127061. doi: 10.1371/journal.pone.0127061

Magen, E., Kim, B., Dweck, C. S., Gross, J. J., & McClure, S. M. (2014). Behavioral and neural correlates of increased self-control in the absence of increased willpower. *Proceedings of the National Academy of Sciences*, *111*(27), 9786–9791.

Schwitzgebel, E., & Cushman, F. (2012). Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. *Mind & Language*, *27*, 135–153. doi: 10.1111/j.1468-0017.2012.01438.x

Sher, S., & McKenzie, C. R. (2006). Information leakage from logically equivalent frames. *Cognition*, *101*(3), 467–494.

Tassy, S., Oullier, O., Mancini, J., & Wicker, B. (2013). Discrepancies between judgment and choice of action in moral dilemmas. *Frontiers in Psychology*, *4*, 1–8. doi: 10.3389/fpsyg.2013.00250.

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, *59*(2), 204–217. doi: 10.5840/monist197659224

Todd, P. M., & Gigerenzer, G. (2012). *Ecological rationality: Intelligence in the world*. Oxford University Press.

Warnock, M. (2003). Introduction. In M. Warnock (Ed.), *Utilitarianism and on liberty* (pp. 1–16). Blackwell.

Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 6753–6758. doi: 10.1073/pnas.0914826107

Reframing rationality: Exogenous constraints on controlled information search

Yi Yang Teoh[a], Ian D. Roberts[b] and Cendri A. Hutcherson[b,c]

[a]Department of Psychology, University of Toronto, Toronto, ON M5S 1A1, Canada; [b]Department of Psychology, University of Toronto Scarborough, Scarborough, ON M1C 1A4, Canada and [c]Department of Marketing, Rotman School of Management, University of Toronto, Toronto, ON M5S 3E6, Canada
yang.teoh@mail.utoronto.ca
iandavidroberts@gmail.com
c.hutcherson@utoronto.ca

doi:10.1017/S0140525X22001030, e242

Abstract

Bermúdez argues that framing effects are rational because particular frames provide goal-consistent reasons for choice and that people exert some control over the framing of a decision-problem. We propose instead that these observations raise the question of whether frame selection itself is a rational process and highlight how constraints in the choice environment severely limit the rational selection of frames.

Classical theories of rationality often assume complete knowledge of decision-relevant factors. However, this assumption contradicts apparent constraints on decision-making in the real world, where people need to simultaneously search for information and determine its relevance for choices at hand. Because the quantity of information in many decisions far outstrips an individual's information processing capacity, selective attention is required to maintain representations of information one piece at a time, essentially highlighting different frames at different times during choice (Kiyonaga & Egner, 2013; Moore & Zirnsak, 2017; Myers, Stokes, & Nobre, 2017; Smith & Krajbich, 2019). While this can theoretically result in a process of sequential frame selection using rational goal-driven attention, attention is also frequently exogenously constrained by the environment: What is attended is as often as not stimulus-driven as opposed to goal-directed (Corbetta & Shulman, 2002; Vanunu, Hotaling, Le Pelley, & Newell, 2021). Importantly, these attentional processes may interact in dynamic ways over time: the decision context primes particular frames of evaluation (Diederich & Trueblood, 2018; Maier, Raja Beharelle, Polanía, Ruff, & Hare, 2020), prior frames differentially enhance and constrain the accessibility of subsequent framings (Johnson, Häubl, & Keinan, 2007; Nook, Satpute, & Ochsner, 2021), and executed decisions frame and bias post-choice evaluation (Chaxel, Russo, & Kerimi, 2013; Navajas, Bahrami, & Latham, 2016). We argue here that determining when, and if, framing effects are rational requires a thorough consideration of these components of frame selection.

First, we argue that the external environment may disproportionately impact initial frames compared to internally represented goals, because attention tends to be drawn first toward salient information in the environment. Indeed, this is the mechanism for most framing studies, which induce frames by highlighting specific information in the decision problem itself (Kühberger,

1998; Levin, Schneider, & Gaeth, 1998; McDonald, Graves, Yin, Weese, & Sinnott-Armstrong, 2021). This is true not only in classical framing studies, but even in the studies that Bermúdez cites as evidence for the potential rationality of framing effects. For example, studies showing that framing marshmallows as "puffy clouds" facilitates rational choice and self-regulation work by explicitly encouraging participants to adopt these frames. It is not clear that people would typically select such "cool" frames in real-world contexts, particularly as the initial frame. Instead, research suggests that foods' appetitive qualities (e.g., sweet and tasty) are the most immediately salient dimension of evaluation (i.e., "hot frames"; see Maier et al., 2020; Sullivan, Hutcherson, Harris, & Rangel, 2015), and that these appetitive frames may be rapidly represented regardless of people's efforts to refocus on healthy frames (HajiHosseini & Hutcherson, 2021). Effortful attentional control is thus usually required to refocus attention away from initially appetitive frames in order to regulate one's choice (Papies, Stroebe, & Aarts, 2008; Rangel, 2013). Studies like the ones Bermúdez cites circumvent the need for regulatory control by presenting "cool" frames in advance, effectively off-loading the work of controlled attention onto the environmental context.

Second, although we fully agree that self-control may facilitate the decision-maker's ability to reframe decision problems in alignment with their goals, we note that exogenously determined initial frames can also constrain the accessibility of subsequent frames. For example, query theory accounts of the endowment effect show that framing a mug first as being owned by the decision-maker led people to consider its value-enhancing aspects more than when the mug was first framed as one of two possible choice options (Johnson et al., 2007). Moreover, certain frames of evaluation may be even more strongly constrained by sequence due to their emergent nature. For example, Bermúdez discusses two frames of evaluation in a strategic interpersonal interaction where people can cooperate for maximal joint outcomes or selfishly choose to minimize potential losses for themselves. Here, the "I-frame" provides strategic reasons for selfish behavior by emphasizing self-relevant outcomes of options while the "We-frame" provides reasons for cooperative behavior by emphasizing joint outcomes of options. Yet Bermúdez's discussion does not consider that the sequence of frames in this decision problem is directionally constrained: People have to first separately acquire information about their own outcomes ("I-frame") and their partner's outcomes ("You-frame") in order to evaluate joint outcomes ("We-frame"). This is highly consequential for decisions because recent work suggests that the information search necessary to adopt a frame incurs a cost (i.e., time and effort; see Callaway, Rangel, & Griffiths, 2021; Jang, Sharma, & Drugowitsch, 2021). Construction of the complex "We-frame" incurs a greater cost by requiring two separate information samples. People may thus be less inclined to spontaneously adopt this frame, especially under time constraints that limit the number of possible frames and increase the cost of each frame (Roberts, Teoh, & Hutcherson, 2022; Teoh & Hutcherson, in press; Teoh, Yao, Cunningham, & Hutcherson, 2020).

Finally, we highlight that constraints on frame-selection extend beyond the immediate choice and into processes of post-choice evaluation. Evidence suggests that people may continue to acquire more information about forgone options and sometimes reframe the chosen option in light of new information (Shani & Zeelenberg, 2007; Teodorescu, Sang, & Todd, 2018). This process could promote more rational choice, by providing decision-makers with the opportunity to pre-emptively frame future decisions in service of rational goals (see Braver, 2012; Brick, MacIntyre, & Campbell, 2016 for discussions of proactive and reactive control). However, just as prior frames constrain the accessibility of subsequent frames during choice, frames adopted during the decision process may also continue to bias post-choice framing of the decision problem. For example, research finds that people tend to seek out confirmatory information to justify the decisions they made (Brehm & Wicklund, 1970; Qin et al., 2011; Scherer, Windschitl, & Smith, 2013), diminishing the potential for goal-directed attention to explore alternative frames that would promote more rational framings in future decisions.

Therefore, while different frames may provide different reasons that lead to distinct choices and people may strategically reframe choices in alignment with their goals, we emphasize here that exogenous environmental and contextual factors strongly constrain these strategies. We have highlighted thus far how these constraints gate the accessibility of particular frames during and after choice. Understanding these constraints on frame-selection will prove critical to any theoretical account of how framing effects can be strategically used to promote human rationality.

Financial support. Y.Y.T. is supported by funding from the Social Sciences and Humanities Research Council of Canada. I.D.R. is supported by funding from a Silvio O. Conte Center Grant to C.A.H. from the National Institutes of Mental Health. C.A.H. is supported by funding from the Social Sciences and Humanities Research Council and Natural Science and Engineering Research Council of Canada, as well as funding from the Canada Research Chairs program. All views expressed in this article represent the views of the authors, and not of SSHRC or NSERC.

Conflict of interest. None.

References

Braver, T. S. (2012). The variable nature of cognitive control: A dual-mechanisms framework. *Trends in Cognitive Sciences*, *16*(2), 106–113. https://doi.org/10.1016/j.tics.2011.12.010

Brehm, J. W., & Wicklund, R. A. (1970). Regret and dissonance reduction as a function of postdecision salience of dissonant information. *Journal of Personality and Social Psychology*, *14*(1), 1–7. https://doi.org/10.1037/h0028616

Brick, N. E., MacIntyre, T. E., & Campbell, M. J. (2016). Thinking and action: A cognitive perspective on self-regulation during endurance performance. *Frontiers in Physiology*, *7*, 159. https://doi.org/10.3389/fphys.2016.00159

Callaway, F., Rangel, A., & Griffiths, T. L. (2021). Fixation patterns in simple choice reflect optimal information sampling. *PLoS Computational Biology*, *17*(3), e1008863. https://doi.org/10.1371/journal.pcbi.1008863

Chaxel, A.-S., Russo, J. E., & Kerimi, N. (2013). Preference-driven biases in decision makers' information search and evaluation. *Judgment and Decision Making*, *8*(5), 561–576.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, *3*(3), 201.

Diederich, A., & Trueblood, J. S. (2018). A dynamic dual process model of risky decision making. *Psychological Review*, *125*(2), 270.

HajiHosseini, A., & Hutcherson, C. A. (2021). Alpha oscillations and event related potentials reflect distinct dynamics of attribute construction and evidence accumulation in dietary decision making. *eLife*, *10*, e60874.

Jang, A. I., Sharma, R., & Drugowitsch, J. (2021). Optimal policy for attention-modulated decisions explains human fixation behavior. *eLife*, *10*, 1–31. https://doi.org/10.7554/eLife.63436

Johnson, E. J., Häubl, G., & Keinan, A. (2007). Aspects of endowment: A query theory of value construction. *Journal of Experimental Psychology: Learning Memory and Cognition*, *33*(3), 461–474. https://doi.org/10.1037/0278-7393.33.3.461

Kiyonaga, A., & Egner, T. (2013). Working memory as internal attention: Toward an integrative account of internal and external selection processes. *Psychonomic Bulletin & Review*, *20*(2), 228–242.

Kühberger, A. (1998). The influence of framing on risky decisions: A meta-analysis. *Organizational Behavior and Human Decision Processes*, *75*(1), 23–55. https://doi.org/10.1006/obhd.1998.2781

Levin, I. P., Schneider, S. L., & Gaeth, G. J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. *Organizational Behavior and Human Decision Processes*, *76*(2), 149–188. https://doi.org/10.1006/obhd.1998.2804

Maier, S. U., Raja Beharelle, A., Polanía, R., Ruff, C. C., & Hare, T. A. (2020). Dissociable mechanisms govern when and how strongly reward attributes affect

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

decisions. *Nature Human Behaviour*, *4*(9), 949–963. https://doi.org/10.1038/s41562-020-0893-y
McDonald, K., Graves, R., Yin, S., Weese, T., & Sinnott-Armstrong, W. (2021). Valence framing effects on moral judgments: A meta-analysis. *Cognition*, *212*(March), 104703. https://doi.org/10.1016/j.cognition.2021.104703
Moore, T., & Zirnsak, M. (2017). Neural mechanisms of selective visual attention. *Annual Review of Psychology*, *68*, 47–72. https://doi.org/10.1146/annurev-psych-122414-033400
Myers, N. E., Stokes, M. G., & Nobre, A. C. (2017). Prioritizing information during working memory: Beyond sustained internal attention. *Trends in Cognitive Sciences*, *21*(6), 449–461. https://doi.org/10.1016/j.tics.2017.03.010
Navajas, J., Bahrami, B., & Latham, P. E. (2016). Post-decisional accounts of biases in confidence. *Current Opinion in Behavioral Sciences*, *11*, 55–60. https://doi.org/10.1016/j.cobeha.2016.05.005
Nook, E. C., Satpute, A. B., & Ochsner, K. N. (2021). Emotion naming impedes both cognitive reappraisal and mindful acceptance strategies of emotion regulation. *Affective Science*, *2*(2), 187–198. https://doi.org/10.1007/s42761-021-00036-y
Papies, E. K., Stroebe, W., & Aarts, H. (2008). The allure of forbidden food: On the role of attention in self-regulation. *Journal of Experimental Social Psychology*, *44*(5), 1283–1292. https://doi.org/10.1016/j.jesp.2008.04.008
Qin, J., Kimel, S., Kitayama, S., Wang, X., Yang, X., & Han, S. (2011). How choice modifies preference: Neural correlates of choice justification. *NeuroImage*, *55*(1), 240–246. https://doi.org/10.1016/j.neuroimage.2010.11.076
Rangel, A. (2013). Regulation of dietary choice by the decision-making circuitry. *Nature Neuroscience*, *16*(12), 1717–1724. https://doi.org/10.1038/nn.3561
Roberts, I. D., Teoh, Y. Y., & Hutcherson, C. A. (2022). Time to pay attention?: Information search explains amplified framing effects under time pressure. *Psychological Science*, *33*(1), 90–124. https://doi.org/10.1177/09567976211026983
Scherer, A. M., Windschitl, P. D., & Smith, A. R. (2013). Hope to be right: Biased information seeking following arbitrary and informed predictions. *Journal of Experimental Social Psychology*, *49*(1), 106–112. https://doi.org/10.1016/j.jesp.2012.07.012
Shani, Y., & Zeelenberg, M. (2007). When and why do we want to know? How experienced regret promotes post-decision information search. *Journal of Behavioral Decision Making*, *20*(3), 207–222. https://doi.org/10.1002/bdm.550
Smith, S. M., & Krajbich, I. (2019). Gaze amplifies value in decision making. *Psychological Science*, *30*(1), 116–128. https://doi.org/10.1177/0956797618810521
Sullivan, N., Hutcherson, C., Harris, A., & Rangel, A. (2015). Dietary self-control is related to the speed with which attributes of healthfulness and tastiness are processed. *Psychological Science*, *26*(2), 122–134. https://doi.org/10.1177/0956797614559543
Teodorescu, K., Sang, K., & Todd, P. M. (2018). Post-decision search in repeated and variable environments. *Judgment and Decision Making*, *13*(5), 17.
Teoh, Y. Y., & Hutcherson, C. A. (in press). The games we play: Prosocial choices under time pressure reflect context-sensitive information priorities. *Psychological Science*.
Teoh, Y. Y., Yao, Z., Cunningham, W. A., & Hutcherson, C. A. (2020). Attentional priorities drive effects of time pressure on altruistic choice. *Nature Communications*, *11*, 3534. https://doi.org/10.1038/s41467-020-17326-x
Vanunu, Y., Hotaling, J. M., Le Pelley, M. E., & Newell, B. R. (2021). How top-down and bottom-up attention modulate risky choice. *Proceedings of the National Academy of Sciences*, *118*(39), e2025646118. https://doi.org/10.1073/pnas.2025646118

The study of rational framing effects needs developmental psychology

Jared Vasil

Department of Psychology and Neuroscience, Duke University, Durham, NC 27707, USA
jared.vasil@duke.edu

doi:10.1017/S0140525X22000930, e243

Abstract

Experimental research is reviewed which suggests that rational framing effects influence young children's social activities according to a logic of interdependence. However, young children are unlikely to possess some of the elaborate cognitive skills argued in the Target Article to be prerequisite for rational framing effects. Understanding rational framing effects requires understanding their ontogenetic origins.

Understanding rational framing effects requires understanding their ontogenetic origins. Reflecting the centrality of interdependence in ontogeny (Tomasello, 2019), the tasks used in framing research in developmental psychology often rely on a logic of interdependence. Interdependent partners' fates are intertwined such that rewards/punishments for one are rewards/punishments for the other (Roberts, 2005; Tomasello, Melis, Tennie, Wyman, & Herrmann, 2012). Children's behavior following interdependent framing illuminates the fundamentally cooperative basis of human rationality. Specifically, interdependent frames increase the sense of commitment toward partners in children as young as 3 years of age.

However, despite their sensitivity to framing, children are unlikely to possess some of the elaborate cognitive skills that the Target Article argues are prerequisite for rational framing effects, such as sophisticated reason-giving skills and a framing analog of reflexive decentering. A developmental approach to rational framing effects is needed to resolve these issues.

In Koomen, Grueneisen, and Herrmann (2020), 5- and 6-year-old peer dyads met before going into separate rooms and being separately instructed about the task. Instructions were delivered using either interdependent, solo, or dependent language. Interdependent participants were instructed that both would receive another cookie if both waited the full time without eating (10 minutes) but, if either ate, then neither would get another. Solo or dependent participants' fates were fully or partly decoupled, respectively. Interdependent participants were more likely to wait the full time without eating compared to those in the solo condition (dependence condition intermediate, nonsignificantly different). This pattern suggests that the interdependent condition motivated participants to inhibit their proximal desires in favor of distal rewards. From the classical perspective, participants' behavior in the interdependent condition was "irrational" in that – on a purely extensional reading – uncertainty in the partner's decision rendered the expected payoff of waiting the full time less than that of the solo condition (in which partner's behavior was irrelevant with respect to the probability of obtaining distal rewards). Consequently, the classical perspective predicts that children should less often wait the full time in the interdependent than dependent condition. As the opposite pattern was found, a more useful explanatory perspective is one in which the rationality of waiting inherits from the rationality of upholding commitments in interdependent contexts.

In a study with younger children, Butler and Walton (2013) investigated the effect of framing on children's motivation to persist on difficult tasks. In a psychologically together condition, 4- and 5-year-olds were given a puzzle and told that a (fictive) child was actively working on the same puzzle in another room "right now" and that the two children were working "together." In a psychologically separate condition, participants were told that the other child worked on the puzzle "a few weeks ago" and that it was the participant's turn to work on it. Following framing, children in the psychologically together condition worked on the puzzle longer than did those in the psychologically separate condition. Moreover, children's self-reported liking of the task was greater in the psychologically together than separate condition (see also Cimpian, Arce, Markman, & Dweck, 2007). These findings suggest that framing difficult tasks as social endeavors increases children's motivation to persist in them. Though Butler and Walton (2013) do not use this language, their findings support the idea that children felt more committed to their partner in the psychologically together than separate condition. Arguably, however, the results of Butler and Walton (2013) only weakly support this interpretation,

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

as participants (1) never met the fictive child (though, they saw a prerecorded video of the fictive child working on the puzzle beforehand) and (2) were not interdependent (i.e., success or failure on the task was independent of that of the fictive child's).

Vasil and Tomasello (2022) provided stronger evidence for the commitment interpretation. These authors investigated 3- and 4-year-olds' commitment, sharing, and helping toward partners during a dyadic activity framed as either a collaborative ("we"-framing) or individualistic endeavor ("you"-framing). Participants and a puppet colored alongside one another on their own sheets of paper at a table, ostensibly to "help decorate for a party later." In the "we"-framing condition, the puppet told children "**We** will color **our** papers with **our** markers" ("you"-framing: replace bolded forms with **you** and **your**). Thus, children (1) actively co-participated alongside their partner and (2) were interdependent only in the "we"-framing condition (i.e., because "success" in that condition required that "we" complete the activity). While partners colored, another person began to play a fun game in the same room; children were free to abandon their partner to play with them. When they abandoned their partner, 3-year-olds more often took leave following "we"-framing compared to "you"-framing (i.e., by nonverbally or verbally excusing themselves before leaving). Moreover, only following "we"-framing did 4-year-olds abandon their partner less often than 3-year-olds. Thus, following "we"-framing, 3-year-olds were more polite when they abandoned partners, whereas 4-year-olds simply did not leave. Framing did not influence children's sharing or helping behavior. Overall, these results converge with those above by suggesting that interdependent framing increases children's sense of commitment. The rationality of commitment inherits from interdependent thinking, in which "we" succeed only if "we" work together.

Despite their sensitivity to rational framing effects, children in the studies above were unlikely to possess elaborate cognitive skills of the type suggested in the Target Article to be prerequisite for rational framing effects. Two such skills are reflective reason construction and analysis and reflexive decentering. While 3-year-olds sometimes produce appropriate reasons for their actions in simple situations (Köymen, Rosenbaum, & Tomasello, 2014), only 5-year-olds – 2 years older than the youngest children who were sensitive to framing, above – competently reason in more complicated situations, for example, involving reflective meta-talk about the reasoning process (Köymen & Tomasello, 2018). Moreover, it is unlikely that young children possess the ability to engage in reflexive decentering, although this requires empirical examination. In fact, a fruitful approach may be to jointly study the emergence of framing effects alongside the emergence of potentially prerequisite psychological skills to see which of these or other skills are in fact prerequisite for rational framing effects. The study of rational framing effects needs developmental psychology.

Acknowledgment. The author expresses gratitude to Michael Tomasello for insightful feedback on early versions of the manuscript.

Financial support. This research reviewed no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Butler, L. P., & Walton, G. M. (2013). The opportunity to collaborate increases preschoolers' motivation for challenging tasks. *Journal of Experimental Child Psychology*, *116*(4), 953–961. https://doi.org/10.1016/j.jecp.2013.06.007

Cimpian, A., Arce, H.-M. C., Markman, E. M., & Dweck, C. S. (2007). Subtle linguistic cues affect children's motivation. *Psychological Science*, *18*(4), 314–316. https://doi.org/10.1111/j.1467-9280.2007.01896.x

Koomen, R., Grueneisen, S., & Herrmann, E. (2020). Children delay gratification for cooperative ends. *Psychological Science*, *31*(2), 139–148. https://doi.org/10.1177/0956797619894205

Köymen, B., Rosenbaum, L., & Tomasello, M. (2014). Reasoning during joint decision-making by preschool peers. *Cognitive Development*, *32*, 74–85. https://doi.org/10.1016/j.cogdev.2014.09.001

Köymen, B., & Tomasello, M. (2018). Children's meta-talk in their collaborative decision making with peers. *Journal of Experimental Child Psychology*, *166*, 549–566.

Roberts, G. (2005). Cooperation through interdependence. *Animal Behaviour*, *70*(4), 901–908. https://doi.org/10.1016/j.anbehav.2005.02.006

Tomasello, M. (2019). *Becoming human: A theory of ontogeny*. Harvard University Press.

Tomasello, M., Melis, A. P., Tennie, C., Wyman, E., & Herrmann, E. (2012). Two key steps in the evolution of human cooperation: The interdependence hypothesis. *Current Anthropology*, *53*(6), 673–692.

Vasil, J., & Tomasello, M. (2022). Effects of "we"-framing on young children's commitment, sharing, and helping. *Journal of Experimental Child Psychology*, *214*, 105278. https://doi.org/10.1016/j.jecp.2021.105278

The received view of framing

Paul Weirich

Philosophy Department, University of Missouri, Columbia, MO 65211
weirichp@missouri.edu
https://philosophy.missouri.edu/people/weirich

doi:10.1017/S0140525X22000929, e244

Abstract

The received view of framing has multiple interpretations. I flesh out an interpretation that is more open-minded about framing effects than the extensionality principle that Bermúdez formulates. My interpretation attends to the difference between preferences held all things considered and preferences held putting aside some considerations. It also makes room for decision principles that handle cases without a complete all-things-considered preference-ranking of options.

Framing may highlight some considerations and because of selective highlighting may generate "quasi-cyclical preferences." If in a decision problem *A*, *B*, and *C* are options described certain ways, an agent may prefer *A* to *B* and prefer *B* to *C* despite knowing that *A* and *C* are the same option under different descriptions. Framing's generation of such quasi-cyclical preferences is rational, according to Bermúdez.

The received view in decision theory claims, roughly, that framing should not influence preferences among the options in a decision problem. A framing effect is irrational because options may be framed in multiple ways, and selection of a frame is arbitrary. Arbitrary matters should not influence an agent's preference-ranking of options.

To reconcile Bermúdez's view with the received view, I present a more precise version of the received view. Because the received view is a vague aggregation of theorists' views, it has multiple interpretations. I present an interpretation that limits its opposition to framing's effects.

The core principles of the received view evaluate choices. Suppose that an agent adopts an option at the top of the agent's preference-ranking of options and so meets a common standard

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

of rational choice. Suppose also that the choice would have been different had the agent framed options differently. The core principles do not deem the choice irrational because different framing would have changed it. Their evaluation of a choice ignores the effect of framing on the choice.

The core principles evaluate not only single choices, but also sets of choices. A standard decision principle states that a pair of choices is irrational if the choices are inconsistent. It takes two choices as inconsistent if they arise from the same decision problem framed two ways. However, this principle prohibits inconsistency, not framing effects.

Evaluative principles outside the received view's core review the causes of a choice and prohibit some framing effects. The version of the received view I present includes such a principle but formulates it guardedly.

The principle relies on a familiar distinction concerning preferences. A traveler may want all-things-considered to fly economy-class even though putting aside price the traveler wants to fly first-class. Some desires and preferences are held all-things-considered, whereas others are held putting aside some considerations.

Bermúdez rejects an extensionality principle asserting that preferences, values, and decisions should be unaffected by how outcomes are framed. However, the received view's guarded principle prohibits the influence of framing on an agent's all-things-considered preferences among options, and through them an influence on the agent's choice. The principle does not address framing's effect on preferences that are not circumspect.

Another refinement adds more precision. Decision theory uses idealizing assumptions to form models of rational choice. It progresses by removing idealizations and generalizing decision principles. The received view, fully formulated, makes explicit the background assumptions of its principles. A standard decision principle states that a rational option in a decision problem is at the top of the agent's all-things-considered preference-ranking of options. The principle assumes an ideal agent in ideal circumstances facing a standard decision problem and possessing rational all-things-considered preferences among options. Without these assumptions, an agent's choice may be rational despite failing to comply with the principle, or it may be irrational despite complying with the principle. In a case of quasi-cyclical preferences, an agent treats an outcome differently under different descriptions despite awareness that the descriptions designate the same outcome; a lack of awareness would supply an easy excuse for the difference in treatment.

Standard decision principles address decision problems in which the decider has a complete all-things-considered preference-ranking of options. Bermúdez entertains decision problems in which the ranking is incomplete. As he notes, incommensurable options create such decision problems. A decider may be aware that different frames favor different options but, even with reflection, may fail to resolve the conflict and so fail to form an all-things-considered preference between two options.

In such a decision problem, suppose that an agent chooses according to some frame rather than other frames in play. Imagine that the guiding frame affects the choice without affecting all-things-considered preferences among options. The principle of extensionality Bermúdez formulates condemns the framing effect, but the received view is open-minded.

The received view, as I interpret it, welcomes a decision principle for decision problems with an incomplete all-things-considered preference-ranking of options. Although no such principle enjoys a consensus, at least one candidate permits a role for framing.

To set the stage for the principle, first consider a random process's effect on a choice in a decision problem. A prohibition of this effect is too strong because flipping a coin is a reasonable way of breaking a tie among options at the top of an all-things-considered preference-ranking of options. For example, it is reasonable to flip a coin to decide between chocolate and vanilla if these flavors are at the top. Although tie-breaking by flipping a coin affects a choice, it does not affect the agent's all-things-considered preferences among options, and so has a permitted effect on the choice. Framing may also break ties, but it has a more substantial role in decision problems without a complete all-things-considered preference-ranking of options.

A common principle for such decision problems classifies an option as rational if it is a top-preference according to a possible completion of the agent's all-things-considered preference-ranking of options (see, e.g., Weirich, 2021). Framing may suggest a possible completion of a preference-ranking of options and an option at the top of this possible completion. The decision principle allows choosing according to the frame's suggestion. It permits this framing effect despite the arbitrariness of framing.

The received view leaves a role for framing's effect on a choice. It prohibits framing's influence on all-things-considered preferences among options but does not prohibit other effects of framing on choices. The received view, as I interpret it, reaches a reconciliation with Bermúdez's view.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

Reference

Weirich, P. (2021). *Rational choice using imprecise probabilities and utilities*. Cambridge University Press.

Frames, trade-offs, and perspectives

Ori Weisel[a] and Ro'i Zultan[b]

[a]Coller School of Management, Tel Aviv University, Tel Aviv 6997801, Israel and [b]Department of Economics, Ben-Gurion University of the Negev, Beer Sheva 8410501, Israel
oriw@tauex.tau.ac.il; https://english.tau.ac.il/profile/oriw
zultan@bgu.ac.il; https://www.bgu.ac.il/~zultan/

doi:10.1017/S0140525X22001005, e245

Abstract

Bermúdez argues for rational framing effects based on normatively appropriate quasi-cyclical preferences. We suggest that this argument conflates preferences over specific outcomes with preferences over outcome aspects. Instead of implying quasi-cyclical preferences, framing affects decisions through standard economic trade-offs. Nonetheless, we demonstrate that framing can affect behavior through altering perceptions of particular outcome aspects when framing effects are not decomposable.

In his target article, Bermúdez makes a case for "rational framing effects," according to which quasi-cyclical preferences – that

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

violate transitivity – can be normatively correct and appropriate. Here, we (1) argue that a reasonable interpretation of the type of preferences that Bermúdez considers does not imply quasi-cyclical preferences at all, but in fact reflect ordinary economic trade-offs, and (2) refer to previous work to demonstrate framing effects that are not consistent with such trade-offs in the context of inter-group conflict.

Bermúdez treats preferences over actions, outcomes, and outcome categories interchangeably. While a statement such as "Macbeth would like to bravely take the throne" reasonably carries an implicit ceteris paribus ("all other things being equal"), Bermúdez treats it as a coarse partition of the outcome space, in which Macbeth prefers any outcome that involves being the king to any outcome that does not. Quasi-cyclical preferences arise when considering a choice between actions for which different such partitions contrast. In the action space that Macbeth faces, the outcome that involves becoming the king necessarily involves breaching his double duty to Duncan. The cycle disappears when viewing Macbeth's preferences as ceteris paribus rather than categorical preferences.

To illustrate, consider the following statements, which are structurally equivalent to statements (C) and (D) that appear in the target article regarding Macbeth:

(C*) Jose prefers having more money to having less money.
(D*) Jose prefers eating over remaining hungry.

Should Jose buy lunch? Since "having more money," in this context, means "remaining hungry," and "having less money" is the same as "eating," this situation reflects, according to Bermúdez, a choice between a "money frame" and a "food frame." But what Jose is facing is, in fact, a standard economic decision. Statements C* and D* can be held concurrently as ceteris paribus statements without implying quasi-cyclical preferences. The decision (to buy lunch or not to buy?) does not reflect a choice between frames but results from resolving a trade-off. Such trade-offs can be decomposed in a counterfactual action space. Macbeth would have been happy to become the king without murdering Duncan, just as Jose would be happy with a free lunch. Similarly, it is possible to decompose the timing and reward in self-control problems or the effects on the self and others in the social dilemma examples analyzed in the target article. Indeed, "it is rational to have a complex and multi-faceted response to a complex and multi-faceted situation" (H3; target article, sect. 3.2, para. 6). However, the natural conclusion is not that framing effects are rational but that a complex rational response considers all aspects of the outcome to reach a consistent decision in each given choice task.

Framing can, nonetheless, influence how single-choice tasks are perceived. What does fulfilling Macbeth's double duty to Duncan mean? When viewed from different perspectives, fulfilling Macbeth's double duty can at times mean diverse and even opposing things; the frame can change the perspective. We found evidence for such an effect in the context of inter-group conflict and group identity.

We used the inter-group prisoner's dilemma (Bornstein, 2003) to examine how framing affects the willingness of individual group members to contribute to their group in the social dilemma that arises in inter-group conflict. The inter-group prisoner's dilemma models conflict between two groups. Each group member chooses between keeping resources for themselves and contributing to their group at a personal cost. Contributions benefit the in-group and simultaneously harm the out-group. In other words, cooperation in one group poses a threat to the other.

We used framing to manipulate the target of the threat, which was either the group as a whole or the individual group members (the two frames are identical in terms of the objective underlying strategic situations, i.e., the mapping from actions to payoffs). These frames lead to opposing reactions, increasing cooperation in the group frame but decreasing cooperation in the individual frame. The unifying principle, in this case, is "help those under threat" (Weisel & Zultan, 2016, 2021a). In the group frame the perception is that the group is under threat, so group members help the group by increasing their contributions. In the individual frame group members perceive themselves – as individuals – to be under threat, so they help themselves by keeping resources for private use. Thus, the frame does not increase the salience of a particular outcome aspect associated with the action. Instead, framing leads to a reinterpretation of the same aspect.

One may argue that this is also a case of a simple trade-off that can be eliminated by decomposing the outcome, similar to the "I" versus "we" frames analyzed by Bacharach (2006). Weisel and Zultan (2021b) tested such decomposition by allowing players to choose whether cooperation harms the out-group or not, keeping the effects of cooperation on the self and the in-group constant across the two options. This decomposition should eliminate the framing effect if frames affect the different narratives' relative salience, such as "cooperation" or "conflict." However, the framing effect remained strong in the new game: Group members were more likely to harm the out-group under a group frame than under an individual frame.

Conflict of interest. We declare no conflict of interest.

References

Bacharach, M. (2006). *Beyond individual choice: Teams and frames in game theory*. Princeton University Press.
Bornstein, G. (2003). Intergroup conflict: Individual, group, and collective interests. *Personality and Social Psychology Review*, *7*(2), 129–145.
Weisel, O., & Zultan, R. (2016). Social motives in intergroup conflict: Group identity and perceived target of threat. *European Economic Review*, *90*, 122–133.
Weisel, O., & Zultan, R. (2021a). Perceived level of threat and cooperation. *Frontiers in Psychology*, *12*, 704338.
Weisel, O., & Zultan, R. (2021b). Perceptions of conflict: Parochial cooperation and out-group spite revisited. *Organizational Behavior and Human Decision Processes*, *167*, 57–71.

Biases and suboptimal choice by animals suggest that framing effects may be ubiquitous

Thomas R. Zentall

Department of Psychology, University of Kentucky, Lexington, KY 40506, USA
zentall@uky.edu
Uky.edu/~zentall

doi:10.1017/S0140525X22000966, e246

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

Abstract

Framing effects attributed to "quasi-cyclical" irrational complex human preferences are ubiquitous biases resulting from simpler mechanisms that can be found in other animals. Examples of such framing effects vary from simple learning contexts, to an analog of human gambling behavior, to the value added to a reinforcer by the effort that went into obtaining it.

Bermúdez says "Framing effects associated with quasi-cyclical preferences are likely to be found in decision problems that are sufficiently complex and multi-faceted that they cannot be subsumed under a single dimension/attribute/value" (target article, sect. 3.2, para. 8 (H2)).

The fact that animals show framing effects very similar to those of humans suggests that what appears to be irrational or suboptimal behavior may not require human complexity and that many framing effects may have biological/evolutionary bases that offer simpler accounts of the behavior.

In the simplest case one can find, framing can be found in the learning of simple associations. What is learned depends on one's frame, the discrepancy between what is expected and what is obtained (e.g., Rescorla & Wagner, 1972). There is much learning early in training, much less learning later in training.

In a less obvious example, one can offer an animal a single pellet of food and give it one pellet, or one can offer the animal two pellets but give it only one. If now the animal is given a choice between the two offers, either of which results in one pellet, the animal will show a strong preference for the single pellet offered (loss aversion; Sturgill et al., 2021). On the other hand, animals have a strong preference for receiving more than expected – presenting them with one pellet but giving them two pellets, over presenting them with two pellets and giving them two pellets (gain attraction; Clayton, Brantley, & Zentall, in press).

The delay-discounting framing effect mentioned by the author occurs when subjects prefer the smaller-sooner (SS) outcome over the larger-later (LL) outcome. Because we humans view self-control as an asset, we value choice of the LL. What we fail to recognize is that in the natural world in which humans and other animals evolved, generally, delayed rewards can rarely be guaranteed, so historically, impulsivity may be a trait that has been selected for. There is evidence, however, that one can change the frame of such a discrimination by requiring humans or other animals to make a "prior commitment" – an earlier response that makes the SS and LL appear to be closer to each other in time (Rachlin & Green, 1972). The reversal of preference from SS to LL can be attributed to the difference in the way the discrimination is framed, or as a natural consequence of the fact that delay-discounting functions are naturally hyperbolic, and when extended in time, the hyperbolic delay-discounting functions naturally cross over (Mazur, 1987).

Impulsive choice is thought to be the frame that makes choice of the SS over the LL suboptimal. It also appears to make the reverse contingency procedure (in which one must choose the smaller amount in order to receive the larger amount) difficult, even for chimpanzees (Boysen & Berntson, 1995). If, however, one uses symbols to represent the quantities, impulsivity is reduced (converting what Bermúdez refers to as a hot representation into a cool representation), altering the frame, and resulting in optimal choice (Sturgill et al., 2021).

It has been found that humans can learn to modify the framing of a problem to reduce suboptimal choice. We might refer to this as learning to exert self-control by distracting oneself (Mischel, Ebbesen, & Raskoff-Zeiss, 1972). But we have found that animals can do the same. In the choice procedure involving an SS versus an LL in which the SS is preferred, similar to humans, providing pigeons with an irrelevant response-independent alternative can result in pigeons that are less impulsive and that choose less suboptimally (Mueller et al., submitted). One might say that the irrelevant, response-independent stimulus "distracts" the pigeon from the SS reinforcer long enough to receive the LL reinforcer. Is that a framing effect or a learning effect?

Unskilled gambling, thought to be uniquely human, is another "irrational" behavior that appears to result from framing, but it too can be found in pigeons (Zentall & Stagner, 2011). For example, pigeons prefer a signal for a 20% chance of getting 10 pellets of food (a jackpot) over the superior signal for a 100% chance of getting three pellets (no uncertainty). But perhaps the pigeons prefer the uncertainty of the 10 pellets. However, they also prefer the 20% occurrence of a signal for reinforcement over a signal for 50% reinforcement (now the optimal alternative is even more uncertain).

Another framing effect shown by humans is their (presumed cultural) bias to prefer *outcomes* that require more effort, over comparable outcomes that require less effort (e.g., for students an A grade in a difficult subject is often judged to have more value than an A grade in an easy subject). It might be proposed that this bias comes from the cultural adage that "work is its own reward." Alternatively, humans may value hard work because it is reinforced socially, and those social reinforcers may even become internalized (become self-rewarding). On the other hand, there is considerable evidence that pigeons show a very similar bias (Clement, Feltus, Kaiser, & Zentall, 2000; Zentall, 2010). If pigeons have to work hard for alternative A but not so hard for alternative B and they both result in the same reward, when given a choice between A and B, pigeons often prefer A (the one that they had to work harder for) over B.

All of this research suggests that one can see similar framing effects in animals that likely do not have the capacity for perspective taking or a theory of mind. However, in spite of these examples of suboptimal behavior, with relatively simple modifications of the animals' environment, one can alter the animal's frame to manipulate the optimality of its behavior. Furthermore, if one can easily manipulate the environment to result in changes in the animals' frame, it suggests that one should consider the possibility that similar simpler mechanisms may be involved in the modification of many human frames.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Boysen, S. T., & Berntson, G. G. (1995). Responses to quantity: Perceptual vs. cognitive mechanisms in chimpanzees (*Pan troglodytes*). *Journal of Experimental Psychology: Animal Behavior Processes*, *21*, 83–86.

Clayton, W. D., Brantley, S. M., & Zentall, T. R. (in press). Decision making under risk: Framing effects in pigeon risk preferences. *Animal Cognition*.

Clement, T. S., Feltus, J., Kaiser, D. H., & Zentall, T. R. (2000). "Work ethic" in pigeons: Reward value is directly related to the effort or time required to obtain the reward. *Psychonomic Bulletin & Review*, *7*, 100–106.

Mazur, J. E. (1987). An adjusting procedure for studying delayed reinforcement. In M. L. Commons, J. E. Mazur, J. A., Nevin, & H. Rachlin (Eds.), *Quantitative analyses of behavior: The effect of delay and of intervening events on reinforcement value* (Vol. 5, pp. 55–73). Erlbaum.

Mischel, W., Ebbesen, E. B., & Raskoff-Zeiss, A. (1972). Cognitive and attentional mechanisms in delay of gratification. *Journal of Personality and Social Psychology*, *21*(2), 204–218.

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

Mueller, P. M., Peng, D., & Zentall, T. R. (submitted). Pavlovian processes in "distractor" effects in a self-control task by pigeons.
Rachlin, H., & Green, L. (1972). Commitment, choice, and self-control. *Journal of the Experimental Analysis of Behavior*, *17*(1), 15–22.
Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II* (pp. 64–99). Appleton-Century-Crofts.
Sturgill, J., Bergeron, C., Ransdell, T., Colvin, T., Joshi, G., & Zentall, T. R. (2021). "What you see may not be what you get": Reverse contingency and perceived loss aversion in pigeons. *Psychonomic Bulletin & Review*, *28*, 1015–1020.
Zentall, T. R. (2010). Justification of effort by humans and pigeons: Cognitive dissonance or contrast? *Current Directions in Psychological Science*, *19*, 219–300.
Zentall, T. R., & Stagner, J. P. (2011). Maladaptive choice behavior by pigeons: An animal analog of gambling (sub-optimal human decision making behavior). *Proceedings of the Royal Society B: Biological Sciences*, *278*, 1203–1208.

Rational framing effects and morally valid reasons

Tomasz Żuradzki

Institute of Philosophy & Interdisciplinary Centre for Ethics, Jagiellonian University, ul. Grodzka 52, 31-044 Kraków, Poland
t.zuradzki@uj.edu.pl
https://incet.uj.edu.pl/tomasz-zuradzki

doi:10.1017/S0140525X22001121, e247

Abstract

I argue that the scope of *rational framing effects* may be broader than Bermúdez assumes. Even in many "canonical experiments," the explanation of the judgment reversals or shifts may refer to reasons, including moral ones. Referring to the Asian disease paradigm (ADP), I describe how non-consequentialist reasons related to fairness and the distinction between doing and allowing may help explain and justify the typical pattern of choices in the cases like ADP.

Bermúdez contrasts simple cases of framing effects (e.g., the Asian disease paradigm – ADP) where "frames prime responses" with more complicated situations where "frames can function reflectively, by making salient particular reason-giving aspects of a thing, outcome, or action" (target article, abstract and sect. 1, para. 3). He concludes that the focus on these simple cases "has blinded us to the existence of *rational framing effects*" (target article, sect. 2.4, para. 2). I agree with Bermúdez that, surprisingly, most research in moral psychology, behavioral economics, and experimental ethics on framing effect concentrate on the very existence and the scale of this effect while neglecting its explanatory and justificatory dimension. In particular, many studies neglect to investigate valid reasons people may have to reverse or change their judgments regarding morally salient choices.

For example, Tversky and Kahneman (1981) seemed to be fully aware of the ethical implications of their findings ("When framing influences the experience of consequences, the adoption of a decision frame is an ethically significant act"). Therefore, it is surprising that most research on ADP does not even mention an ethical dimension of this choice situation, although it consists of making a life-saving decision. This observation also applies to Bermúdez's works (2021, target article), which seem to treat the standard pattern of choices in "the canonical experiments" (including ADP) as irrational. Although Bermúdez discusses some highly stylized moral reasons in the examples of Agamemnon and Macbeth, he mentions more morally relevant examples (i.e., abortion) in merely one sentence.

I assume that two types of moral non-consequentialist reasons may be particularly helpful in *explaining* and *justifying* the standard pattern of choices in the cases like ADP: fairness (egalitarian reasons) and the distinction between doing and allowing. First, programs B and D ("1/3 probability that *600 people will be saved* [*nobody will die*], and 2/3 probability that *no people will be saved* [*600 people will die*]") are fully egalitarian (cf. Segall, 2016). Not only has every individual *ex ante* equal chance for survival (1/3), but there is also full equality *ex post*: Everyone is alive or dead. In turn, programs A and C (*200 people will be saved* [*400 people will die*]) are only partially egalitarian. Even if one understands these two options as cases of random distribution (so again, every individual has *ex ante* equal chance for survival, i.e., 1/3), there is inevitable *ex post* inequality (some will be alive and some dead).

Second, the next non-consequentialist dimension of this choice concerns the distinction between doing and allowing (Woollard, 2015). Programs A and B in the ADP are formulated in terms of "people will be saved," so the reference point is 600 people dead (or precisely speaking, "almost dead"). Therefore, many may assume that the choice concerns the distribution of benefits since the harm has already been done, and choosing program A is actively saving ("doing something for") 200 people, that is, causing them to survive while only allowing the other 400 people to die. Thus, many respondents may want to avoid program B because they do not want to gamble with the life of these 200 people that can be actively saving by choice of program A.

In turn, the choice between the programs C and D is formulated in terms of "people will die," so the reference point is 600 people alive. Therefore, many may assume the choice concerns the distribution of harms because no harm as yet has been done. Consequently, many respondents may avoid program C because they do not want to harm ("doing something for") 400 people actively and may prefer to allow a chance to decide their fate. Moreover, some authors even speculated that the descriptions of the programs C and D might suggest that people will die because of "*something lethal about the intervention itself*" (Dreisbach & Guevara, 2017). Even though there is no reference to any vaccine in the original version of ADP, it is surprising that many authors speculated that programs in ADP refer to (or suggest) some vaccine whose side effects will bring about the death of most (A, C) or whose effectiveness is very uncertain (B, D).

In our experimental studies (total $n = 1{,}106$), we asked participants who had previously seen all four programs simultaneously to evaluate them one by one (Żuradzki, Szwed, & Maj, 2022). Specifically, we asked to respond to the following claims (on a scale of 1–7; totally agree to disagree): "This program is fair for each of 600 people," "This program will harm many people," "This program will kill many people," and so on. Fairness seems to be not only an important reason, but also risky options (B and D), which are fully egalitarian in our understanding, were indeed marked by the participants as much fairer (Fig. 1).

In contrast, we observed no significant differences between programs when participants evaluated whether "This program will harm many people" or "This program will kill many people." In our other study, we asked participants who had previously been presented with the standard version of ADP to provide reasons why they decided on a specific program. They were presented with multiple reasons, and they could mark up to three. The reason "Because this program seemed to me less harmful" was chosen

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

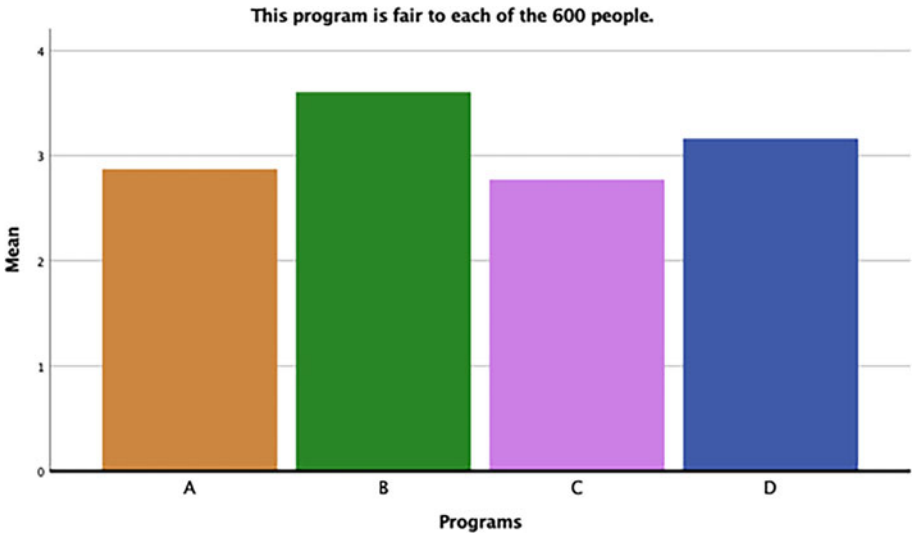


Figure 1. (Żuradzki) In this study, participants ($n = 164$) were presented with a hypothetical scenario of a rare Asian disease approaching their country. They were told that it would kill 600 people, but four alternative programs for disease control were invented. The descriptions of the programs were presented (all at once) as in the original Asian disease paradigm. Next, participants saw descriptions of each program separately, and they were asked to evaluate them. Specifically, they were asked to respond to the following claim (on a scale of 1–7; totally agree to disagree): "This program is fair to each of the 600 people." Program B was evaluated as the fairest to all 600 of people, in comparison with program, A ($p < 0.001$), program C ($p < 0.001$), and program D ($p = 0.007$).

relatively often (about 20–30% of those who chose subsequent programs marked this as one of their reasons). However, we have not observed significant differences in marking this reason between the followers of different programs. Our studies show how various moral reasons may interact and counterbalance each other even in "canonical experiments" on framing effects. To sum up, our theoretical hypotheses, partially verified in the experimental studies, show that the scope of *rational framing effects* may be even broader than Bermúdez assumes in his brilliant paper.

Financial support. This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 805498).

Conflict of interest. None.

References

Bermúdez, J. J. (2021). *Frame it again: New tools for rational decision-making*. Cambridge University Press.

Dreisbach, S., & Guevara, D. (2017). The Asian disease problem and the ethical implications of prospect theory. *Noûs*, *53*(3), 613–638.

Segall, S. (2016). *Why inequality matters*. Cambridge University Press.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458.

Woollard, F. (2015). *Doing and allowing harm*. Oxford University Press.

Żuradzki, T., Szwed, P., & Maj, M. (2022). A risky choice framing effect, doing/allowing, and fairness in uncertain prospects (manuscript). https://osf.io/mj3yp

Author's Response

Frames and rationality: Response to commentators

José Luis Bermúdez

Department of Philosophy, Texas A&M University, College Station, TX 77843, USA
jbermudez@tamu.edu

doi:10.1017/S0140525X22001418, e248

Abstract

The thoughtful and rewarding peer commentaries on my target article come from a broad range of disciplinary perspectives. I engage with the commentaries in three groups. First, I discuss the commentaries that apply my basic approach to new cases not considered in the target article. Second, I explore those that helpfully extend and refine my arguments. Finally, I offer replies to those that object either to the overall framework or to specific arguments.

R1. Overview

The peer commentaries on my target article come from a broad range of disciplinary perspectives. Studying the individual commentaries has been thought-provoking and rewarding.

The target article explores *rational framing effects* emerging from what I term *quasi-cyclical preferences*. Quasi-cyclical preferences occur when a thinker or agent prefers A to B and B to C, despite knowing fully well that A and C are different ways of framing the same outcome or thing (because the agent knows that $A = F_1(D)$ and $B = F_2(D)$, for a single outcome D and two distinct frames F_1 and F_2). The target article develops the following three basic ideas.

> (H1) Frames and framing factor into decision-making by making one dimension/attribute/value of the decision problem highly salient, which influences how the subject engages emotionally and affectively.
>
> (H2) Quasi-cyclical preferences are likely to be found in decision problems that are sufficiently complex and multi-faceted that they cannot be subsumed under a single dimension/attribute/value. Different frames engage different affective and emotional responses, which the decision-maker cannot resolve either by ignoring all the frames except one or by subsuming them into a larger frame.
>
> (H3) Framing effects and quasi-cyclical preferences can be rational in circumstances where it is rational to have a complex and multi-faceted response to a complex and multi-faceted situation.

In the paper I motivate H1 through H3 with two literary examples – Aeschylus' Agamemnon and Shakespeare's Macbeth. Several of the

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

R2	R3	R4
Extending the conclusions	Extending and refining the arguments	Objecting to arguments and framework
Chater	Ainslie	Crockett and Paul
Dorison	Beal	Fisher
Flusberg, Thibodeau, and Holmes	De Dreu	Guala
Kühberger	Koi	Lau
Levy	Moldoveanu	Mandel
Roberts	Schwartz and Cheek	Pettigrew
Sirgiovanni	Sher and McKenzie	Przybyszewski, Rutkowska, and Białek
Vasil	Weirich	Teoh, Roberts, and Hutchison
Zentall		Weisel and Zultan
Żuradzki		

commentaries directly engage with my literary examples – for example, **Crockett and Paul**, **Fisher**, and **Lau**. Others contribute new ones. **Koi** discusses some wonderful passages from Marcus Aurelius and **Beal** brings Dostoevsky's Raskolnikov into the discussion.

The target article discusses rational framing effects in a number of non-literary domains, including self-control, game theory, and interpersonal conflict and coordination more generally. **Ainslie**, **Lau**, **Teoh**, **Roberts**, **and Hutcherson (Teoh et al.)**, and **Zentall** engage with self-control, while **Sher & McKenzie**, **Moldoveanu**, **Weisel & Zultan**, and **De Dreu** address game theory and the interpersonal dimension. Other authors open up new areas for discussion. **Flusberg, Thibodeau, and Holmes (Flusberg et al.)** look at framing effects in discourse processing and pragmatic reasoning. Zentall looks at animal learning. **Roberts** and **Sirgiovanni** relate the discussion to themes in moral philosophy, while **Vasil** brings in developmental psychology and **Beal** literary theory. **Chater** finds framing effects even in formal domains such as chess and mathematics.

Within the behavioral sciences, the most discussed framing effects have been experimentally induced, such as the Asian disease paradigm. Several authors address the classic experiments, including **Dorison**, **Levy**, **Flusberg et al.**, **Mandel**, and **Żuradzki**, all of whom argue, in different ways, that traditional interpretations of those experiments need to be revisited. **Przybyszewski, Rutkowska, and Białek (Przybyszewski et al.)** take issue with the framework of prospect theory that was originally proposed by Kahnemann and Tversky as a descriptive framework for making sense of the experimental observed behavior.

Within decision theory, discussion of experimental framing effects has focused on the principle of extensionality/invariance and its normative validity. **Schwartz and Cheek** share my skepticism about extensionality, while **Guala**, **Weisel and Zultan**, and **Lau** all push back against the idea that quasi-cyclical preferences can be normatively permissible. All three commentaries take issue with how I am understanding rational preferences. **Weirich** looks for a middle ground, offering a version of extensionality that he claims does not rule out the types of rational framing effect that I discuss.

For the purposes of this response, I will divide the commentaries into three groupings, as in the table below. The first group (to be discussed in sect. R2) offers ways of extending the basic idea of rational framing effects to new areas. Commentaries in the second group (in sect. R3) are largely sympathetic to the basic position developed in the target article and offer me ways of refining and developing my arguments. Section R4 discusses and attempts to reply to those commentaries that directly object to aspects of my position.

R2. Extending the conclusions

R2.1. Rational framing effects in the canonical experiments

One of the key claims in the target article is a sharp divide between the "easy" experimental framing effects and the "hard" cases where, I claim, rational framing effects can be found. Several commentators argue that framing effects can be rational even in the canonical experiments.

- **Levy** argues that frames can provide genuine evidence, so that framing effects are the result of subjects responding rationally to evidence. Certain frames might be understood, for example, as conveying a recommendation – describing someone's publication record in terms of her *acceptance* rate rather than her *rejection* rate is more likely to generate a positive impression. Being sensitive to such forms of implicit recommendation can be ecologically rational, according to Levy. If a subject has no initial preference between two options A and B and option A is implicitly recommended by somebody who might reasonably be taken to be an authority, the subject thereby acquires a reason to choose A over B. (See also **Flusberg et al.,** who emphasize more broadly the role of communicative effect, and pragmatic language processing more generally in creating framing effects.)
- **Kühberger** points to a number of ways in which "risky choice framing effects can result from various cognitive processes, all being entirely intelligible and rational. Central is the idea that, rather than passively taking in information, people actively select and process information, taking also background knowledge into consideration." There are multiple semantic and pragmatic factors that can be appealed to show that the different options in, for example, the Asian disease paradigm are not really informationally equivalent (see also **Sher & McKenzie**, **Mandel**, and **Schwartz & Cheek**).
- **Dorison** reports experiments showing that third-party observers systematically penalize decision-makers who are *not* sensitive to gain/loss framing effects. It follows, then, that there can be reputational rewards for being susceptible to such framing effects. When those reputational rewards outweigh the non-reputational costs, then being susceptible to gain/loss framing effects can be rational.

All these observations are good ones. They add to the roster of plausible explanations of what is actually going on in the canonical framing effects. However, I don't really think that either succeeds in breaking down the contrast between rational framing effects and irrational ones. As formulated in the target article, the contrast is between (a) cases where frames *prime* responses,

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

as in many of the classic framing experiments, and (b) more complicated situations where frames function reflectively by *making salient* particular reason-giving aspects of a thing or outcome.

The **Levy** analysis seems to be a form of priming (the implicit recommendation primes a positive response). There is a sense in which the subject *has* a reason, but nothing at all comparable to the complex reflection involved in the (b)-type cases. Similarly for **Dorison's** example – he shows that it can make sense reputationally to be susceptible to gain–loss framing effects, but that is not the same as showing them to be rational, in the sense of correct and normatively appropriate. For one thing, the subject does not make the preference-reversing choices *because of* the reputational consequences.

Flusberg et al. argue that the pragmatic phenomena they point to are not instances of priming. They cite studies showing that when crime is framed as a beast (rather than a virus) ravaging a city, subjects are more likely to propose law-enforcement solutions – an effect that disappears if subjects are simply primed with the concept *beast*. This is not convincing, however. A priming effect that only works in a particular context is still a priming effect. Metaphors and pragmatic processing are important tools for framing, but rational quasi-cyclical preferences require more reflective decision-making.

Żuradzki appreciates this and explicitly tries to show how even in the Asian disease case the different frames actually bring different moral considerations into play. The preference for a certain outcome in the positive frame (200 will be *saved*, as opposed to a 1/3 probability that all will be saved) reflects a preference for doing rather than allowing – the outcome is actively brought about, rather than being the luck of the draw. In contrast, he suggests, the negative frame focuses the attention on fairness. People choose the lottery because it is more egalitarian – either all will live or all will die, and the probability is the same for each. This is much closer to quasi-cyclical preferences, rather than just priming.

R2.2. New domains for rational framing effects?

The commentaries by **Chater**, **Vasil**, and **Zentall** offer suggestive proposals for extending the scope of rational framing effects to formal domains such as chess and mathematics (Chater), developmental psychology (Vasil), and animal learning (Zentall). Chater is broadly supportive, while the other two authors claim that the rational framing effects they identify are potentially problematic for my position.

According to H2 above, framing effects will occur in decision problems that are too complicated and multi-faceted to be subsumed under a single evaluative dimension. In my target article, I focused primarily on emotional and evaluative complexity, with different frames bringing different and incompatible responses into play. **Chater** points out convincingly that the same structure can emerge in the (relatively) emotion-free realms of mathematics and chess. Computational complexity makes framing a necessity. No individual move in chess can be evaluated on its own terms – it must be framed as part of a strategy, and quasi-cyclical preferences quickly emerge when the same move is (knowingly) considered under multiple frames/strategies.

This is surely correct, and offers another way of thinking about the rational power of frames – as heuristic tools for managing computational complexity, with rational framing effects an inevitable consequence.

Zentall's references to the animal learning literature are also very much to the point, since the preference reversal from long-term rewards (LL) to short-term ones (SS) typically used to frame discussions of human self-control has been well studied in the animal case, as have some of the models used to explain it (e.g., hyperbolic delay discounting). Zentall points to numerous relevant phenomena. For example, when pigeons are offered a choice between (i) an offer of one pellet of food resulting in a reward of one pellet, and (ii) an offer of two pellets of food resulting in a reward of one pellet, they will reliably choose (i), displaying some analog of loss aversion, because they disprefer the outcome where what they receive is less than what they chose (Sturgill et al., 2021). This is plausibly a framing effect. The same outcome (delivery of one pellet of food) is framed in two different ways – as involving a loss and as not involving a loss.

Also very interesting is the finding (Sturgill et al., 2021) that pigeons find it significantly easier to deal with reverse contingencies (when the animal has to choose a smaller amount in order to receive a larger amount) if they are presented with symbols rather than with the actual quantities. As **Zentall** points out, this is in some respects analogous to the reframing strategy I explore for self-control – either (as in this case) a "cool" reframing of the impulsive choice, or (as in the target article) a "hot" reframing of the long-term goal. Certainly, this shows that animals can be sensitive to the mechanisms that can be deployed in the rational framing effects that I discuss. However, it seems a stretch to describe these as rational framing effects. They are primed responses, rather than thought-strategies the animal engages in.

Vasil's discussion of developmental psychology is somewhat closer to the topics of the target article. Reporting joint work with Michael Tomasello, he finds interesting framing effects in young children. Framing tasks as social endeavors (e.g., by telling them that another child was working on the same puzzle in an adjacent room) seems to increase 5–6 year olds' motivation to persist in the tasks (Butler & Walton, 2013). Higher degrees of commitment to their partners were observed in 3–4 year old's during activities that were framed collaboratively (as an activity that *we* engage in together), rather than as individualistically (Vasil & Tomasello, 2022). Vasil suggests that these are rational framing effects, because, as he puts it, "we" succeed only if "we" work together.

These are interesting results (relevant also to the "I"-frame vs. "we"-frame contrast in game theory – see further below). **Vasil** is correct that children of this age lack the sophisticated cognitive skills discussed in the last section of my target article (reflexive decentering, perspectival flexibility, etc.). I am not sure that this is a problem for me, however. I did not say, and nor do I believe, that those cognitive abilities are necessarily implicated in rational framing effects. I was arguing in the opposite direction – namely, that those skills and abilities are going to be needed in any successful attempts to resolve frame-based interpersonal conflict (what I called *discursive deadlock*) and that exercising those skills and abilities will bring with it rational framing effects.

Still, **Vasil**'s commentary raises an important question. As discussed in section R2.1, I rely heavily on the distinction between framing effects produced by "mere" priming, and the rational framing effects that derive from competing reasons and require a degree of reflective understanding. I certainly need to spell out in more detail what exactly reflective understanding consists of. Looking at the ontogeny of framing effects will be an important tool here in the attempt to disentangle the different components – so too will aspects of the animal learning literature, which might point to a bottom-line set of frame-sensitive capacities without which rational framing effects could not possibly occur.

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

R2.3. Rational framing effects and moral reasoning

Roberts and **Sirgiovanni** both focus on the arguments of the target article in the area of moral philosophy. This brings refreshing new perspectives, as does **Beal**'s exploration of what Bakhtin has called the polyphony principle in the context of Dostoyevsky's *Crime and Punishment*.

Sirgiovanni observes that there is an approach to the much-discussed trolley problem that seems to involve quasi-cyclical preferences. So, for example, I might prefer switching the lever as the trolley comes thundering down the track, which will save the five people on the tracks at the cost of sacrificing one person. At the same time, though, I might prefer letting the trolley kill the five people on the tracks to pushing one person off a bridge to stop the trolley. These preferences (which I and many others find quite appealing) would count as quasi-cyclical if we equate both choices as between the loss of one life and the loss of five lives.

This is not a very plausible way of looking at the matter, however. It would be hard to defend a view of outcomes on which the death of one person counts as the same outcome as the death of a completely different person. The number of lives lost is the same in the two cases, but surely the objects of choice are different. Choosing an action that leads to someone being pushed off a bridge is not the same as choosing an action that leads to a different person being run over by a trolley – so there is no way in which we can be dealing with different ways of framing the same outcome, as would be required for quasi-cyclical preferences. As we will see at greater length in section R4, there are important questions about how exactly to understand outcomes and what it is to prefer one outcome to another. At a minimum, though, for A to be the same outcome as B there must be at least a core of common consequences.

This brings us back to the aim of the target article, namely, the principle of extensionality, considered as a binding principle of rationality. Within the social and behavioral sciences, the validity of extensionality is almost unquestioned. **Roberts** observes, though, that extensionality is only held to be binding within moral philosophy by consequentialists (and decision theory, of course, typically reflects a strong form of consequentialism). From the perspective of deontology and virtue ethics, the idea that outcomes or actions can only be assessed and chosen when framed is not news.

As **Roberts** points out, Christine Korsgaard, the well-known contemporary Kantian, has claimed that agents must choose between actions, not between acts, where an action incorporates aspects of the reasons for which the act is performed. In my terminology, an action is an act under a particular reason-giving frame. Virtue ethicists, who value character dispositions, also typically think of actions in a way that incorporates the dispositions and character traits that they reflect. The same reasoning can be extended to outcomes. There are no bare outcomes – an outcome reflects the action that gave rise to it.

While this is an important insight, it is important not to exaggerate the parallels. Kantians such as Korsgaard and virtue ethicists are not known for their liberality when it comes to the admissibility of multiple frames. This means that neither is likely to accept the admissibility of quasi-cyclical preferences, for which being able to view a single action or outcome under multiple frames is a prerequisite. It is true that a Kantian can and should accept that an action can reflect multiple maxims or principles. But only one can be correct. I think that Kant would find deeply alien the view that it would be rationally permissible to prefer A to B and B to C where A and C are the same act or outcome framed according to different maxims or principles. The task of moral reasoning is to identify the current maxim or principle for the given situation, and once that has been found there is no room for further framing. In contrast, frame-sensitive reasoners cannot be tied to a single frame. Frame-bound moral reasoners are, in fact, one of the causes of the discursive deadlock discussed in the last part of the target article. It is no accident that the paradigm cases of discursive deadlock are all in the so-called values issues.

As **Beal** points out, the type of reasoning characteristic of frame-sensitive reasoners is much closer to the concept of *polyphony*, developed by the literary critic and philosopher Mikhail Bakhtin in connection with Dostoyevsky's *Crime and Punishment*. This is the idea that "an objectively single action may, with logical consistency, sustain diverse positive and negative judgments." Raskolnikov offers a series of incompatible but internally coherent explanations of why he killed the pawnbroker and her sister. Each explanation is supported by a unique narrative that selects relevant details from a host of competing characterizations, motives, contextual factors, counterfactuals, and so on. Reality, according to Bakhtin and Beal, is too complex to be distilled into a single narrative. It is, one might say, frames all the way down. (For another artistic development of this basic idea, consider Kurosawa's well-known film *Rashomon*, which presents multiple tellings of a rape and a murder.)

R3. Refining the arguments

The second group of commentaries offers what I take to be friendly amendments and refinements of my general line of argument. **Sher and McKenzie** and **Schwartz and Cheek** embed my case for rational framing effects in the context of a more general critique of classical theories of choice and decision. **Ainslie** and **Koi** offer alternative theoretical frameworks for thinking about self-control, while **De Dreu** proposes tools for refining my arguments in the realm of game theory.

R3.1. Extending the critique of classical theories

Sher and McKenzie have long argued against the traditional view that choice and decision should be frame-invariant (e.g., Sher & McKenzie, 2006, 2011, both of which maintained that the frames implicated in some of the classical framing effects are not informationally equivalent – on which also see **Schwartz & Cheek**). Here they turn their attention to how the case against rational framing effects rests upon the assumption that preferences are complete, in the following sense: That, given any two available options A and B, the agent either prefers A to B, B to A, or is indifferent between them. Failures of completeness can derive both from *imprecision* and from *conflict*.

Sher and McKenzie suggest that lifting the completeness requirement opens the door to rational framing effects: "In a finite choice menu, there may be distinct alternatives, A, B, unranked relative to one another, neither of which is outranked by any other option in the menu. If A is chosen under one descriptive frame and B under another, choices are frame-dependent but never suboptimal." This point is important, but I should stress that the pattern of choices they describe are not rational framing effects in the sense I describe them. In my cases of quasi-cyclical preferences, there is only one outcome and the agent knows that there is only one outcome. In other words, the agent knows that $A = F_1(D)$ and $B = F_2(D)$, for a

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

single outcome D and two distinct frames F_1 and F_2. So, Sher & McKenzie's description does not apply. Either what they characterize as A and B are framed outcomes, in which case it is falsie that they are unranked relative to each other. Or they are unframed outcomes, in which case A = B and so they cannot possibly be unranked relative to each other. (Nonetheless, completeness is very important in this area, and is discussed also by **Pettigrew** and **Weirich**.)

Schwartz and Cheek offer a very distinct perspective on the idea that distinct frames fail to be informationally equivalent. Various authors, including **Sher & McKenzie**, have suggested that frames are informationally "leaky" – that is, they "leak" choice-relevant information (see also **Mandel** and **Kühberger**). Schwartz & Cheek take this a step further, arguing that frames leak into experience. In other words, it is not just that what people choose between are framed outcomes and framed actions, they also experience framed outcomes, particularly following a frame-sensitive choice.

This, they suggest, has direct implications for rationality, even in the simplest framing effects. Even if we take a genuine framing effect, where there is complete informational equivalence between the two options, there can be experiential non-equivalence further downstream, when subjects experience the result of their decision. In other words, difference-making reasons can emerge downstream of actual decision-making. **Schwartz and Cheek** suggest, surely correctly, that a substantive theory of rationality must take into account not just the anticipated consequences of actions but also their actual consequences, where those *actual* consequences can experientially reflect the frame-driven choices that gave rise to them.

R3.2. New analyses of rational framing effects

Ainslie, **Koi**, and **De Dreu** offer helpful suggestions for deepening the analyses I offer of specific rational framing effects, tackling self-control (**Ainslie** and **Koi**) and game theory (**De Dreu**).

Ainslie suggests that I present framing as an unmotivated process in the context of self-control. The specific effect I discuss is when an agent comes to prefer LL (the long-term reward) to SS (the short-term reward) relative to the *having successfully resisted temptation* frame, although she prefers SS to LL framed more neutrally (e.g., as an abstract benefit in the future). He objects: "The notion that someone can make self-control hot – in effect, an occasion for aroused emotion – by assigning it utility would seem to violate the laws of motivational gravity. if you could assign utility just anywhere, why not assign it to the LL alternative in the first place."

This criticism is a little unfair, however. My point is that different frames engage in different ways with values and emotions. The *resisting temptation* frame brings into play a new set of reasons for the agent. It is *that* process that results in different assignments of utilities (in accordance with what **Ainslie** engagingly describes as the laws of motivational gravity).

Having said that, I appreciate the suggestions that **Ainslie** makes for additional motivational mechanisms. He points out, surely correctly, that a choice looks very differently when it is framed as one of a series of choices, rather than as a one-off choice. Framed thus, the choice of LL can provide behavioral evidence (be a *sign*) of how the agent will act in the future. If the value of the summed LLs outweighs the value of the summed SSs, with both suitably discounted, then that will provide a powerful motivation that overrides the immediate attractiveness of SS.

Also on the topic of self-control, **Koi** sees framing as one tool among many. Framing works by modulating information salience (as in H1). But there are other ways of modulating information salience that can be deployed to help with self-control. Framing is an internal strategy, but there are also external ones, such as situation selection and situational modification. I can choose not to expose myself to SS cues, for example – or set up reminders and other mechanisms to increase the salience of LL.

Extending still further, self-control, **Koi** observes is just one area where informational salience can be modulated and manipulated. The overarching phenomenon is attention (although in a more high-level sense than standardly discussed in cognitive science and cognitive psychology). The role of attentional processes in modulating perception has been much more studied than its role in modulating action. But, as Koi observes, attention may be the most fundamental psychological mechanism. I argue in the target article that framing may be a better way of thinking about self-control than postulating mechanisms of willpower or ego depletion. The interesting possibility Koi raises is that what we are really talking about with dual process (e.g., hot/cold) theories are the results of attentional modulation of information salience.

Turning to game theory, **De Dreu** broadens my discussion of the "I" and "we" frames in social interactions to consider how those two frames interact with "gain" and "loss" frames. In the target article, I focused on the four-dimensional problem of two players, each of whom has two possible frames. As he observes, though, the dialectic between "I"-frame and "We"-frame becomes more complex and more interesting if we take more possible framings into account. Players can be influenced by the frames of others. I considered only how one player's movement between "I"-frame and "We"-frame is a function of the other player's movement between "I"-frame and "We"-frame, but "gain" and "loss" frames can also be relevant.

For example, **De Dreu** cites multiple studies showing that (in my terms) the move to the "We"-frame is primed when a player realizes that the other player is operating under a loss frame (e.g., De Dreu & Gross, 2019). Empathy is a powerful force, and one that can be manipulated, as rational players in the "I"-frame adopt a loss framing to elicit cooperation from others, which they can then exploit. These are valuable insights. I am unconvinced, however, by his more general statement that the kind of rational reframing I discuss in the target article is unlikely to work, on the grounds that "human psychology gravitates towards minimizing *my* loss." That may be so, but the very formulation presupposes an "I"–frame. **De Dreu** is too quick to equate the "I"-frame with the Defect strategy in, for example, the prisoner's dilemma (PD), and the "we"-frame with Cooperate. The way I distinguish them is through which columns in the pay-off table are taken into account. The "We"-framer takes the pay-offs to all players into account, which is why, despite what **De Dreu** says, it permits a solution in Stag Hunt. (See Chs. 8 and 9 of Bermúdez [2020] for more details.)

R4. Objecting to arguments and frameworks

Unsurprisingly, a sizeable group of commentaries directly criticize one or more aspect of the argument and framework. **Mandel** and **Fisher** suggest that my basic approach is self-defeating. Several commentators take issue with how I am understanding preference (**Guala**, **Lau**, and **Weisel & Zultan**), while **Crockett & Paul**, and **Przybyszewski et al.** object to how I am thinking about rationality.

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

R4.1. Is the project misconceived?

Fisher offers a thought-provoking argument that my basic project is misconceived. The problem comes, she claims, with my initial characterization of *quasi-cyclical preferences*. Recall that quasi-cyclical preferences occur when the agent prefers A to B and B to C despite knowing that A and C are the same outcome framed in different ways (e.g., as A = *Following Artemis*'s *Will* and B = *Murdering His Daughter*). Fisher's objection is, in effect, that I can't eat my cake and have it. An agent can only have quasi-cyclical preferences if he or she knows that A and C are different ways of framing the same outcome. At the same time, Fisher argues, agents can only know that A and C are different ways of framing the same outcome if they have frame-neutral ways of thinking about that outcome. But if agents can think about outcomes in a frame-neutral manner then the framed outcomes that I claim to be the objects of preference and value simply fall out of the picture – a rational agent can ignore the framings and simply focus on the frame-neutral outcome.

This argument is ingenious, but unfortunately not compelling. Consider a time-honored analogy. I am aware that there is one person whom I can think about in one of two ways – as Clark Kent or as Superman. But it certainly does not follow from my knowing the identity of Clark Kent and Superman that I have some independent way of thinking about that person that does not involve either thinking of him as Clark Kent or thinking of him as Superman. By analogy, knowing that two different framed outcomes correspond to a single frame-neutral outcome does not *require* that I be able to think about that outcome in a frame-neutral way.

And in fact, there are good reasons for thinking that there could *not* be such a frame-neutral way of thinking about that outcome. That is the whole point of H2. Quasi-cyclical preferences arise in situations that are sufficiently complex and multi-faceted that they need to be viewed and considered from multiple perspectives, none of which is dominant and none of which subsumes the others. For this sort of situation, unlike the highly simplified and stylized examples standardly considered in formal theories of decision and choice, the notion of thinking about them in a frame-neutral manner makes neither descriptive nor normative sense.

This aspect of my overall approach is well understood by **Mandel**, who writes that my approach "is to carve away the so-called small-world of toy problems entirely and focus on what he describes as the larger complex world in which multi-attribute decisions are the result of conflicting perspectives." He does not deny that there are such things as quasi-cyclical preferences. However, he takes exception to my claim that quasi-cyclical preferences can be rational. His reason for saying so is that "the contrast to the small world case is never pinned down tightly."

With respect simply to the structure of my argument, this objection seems misplaced. Nothing that I say about the complex cases rests upon the (alleged) irrationality of the small-world cases. It is a (sociological) fact, I believe, that most people in the area take the laboratory experiments to manifest irrationality, which is why I framed my own position in contrast to what I take to be the standard view. I am perfectly happy, though, to accept that the standard view may be mistaken, which is why I do not take the arguments of, for example, **Dorison**, **Żuradzki**, **Flusberg et al.**, and **Levy** (all discussed above as challenging the alleged irrationality of the classic experimental behavior) to be objections to my view – and nor were they put forward as objections.

Still, stepping back a little, there is a good point to be extracted from **Mandel**'s paper. His critiques of the classic experiments are all variations on the (important) theme that the different scenarios are not semantically equivalent, because of the type of "leakage" discussed by **Schwartz and Cheek**, **Kühberger**, and **Sher and McKenzie** (not to mention presented in his own work – e.g., Mandel, 2014). I think that his real concern, although it is not explicitly presented as such, is that it is simply impossible for rational framing effects to occur, because of the following lines of reasoning:

(1) In order to qualify as a framing effect, the different options must be semantically/informationally equivalent.
(2) In order to qualify as rational, a choice or preference must be made for (good) reasons.
(3) There cannot be (good) reasons for choosing between two different options that are semantically/informationally equivalent.

These lines of reasoning may be applicable to the Asian disease paradigm, and similar experiments, because subjects are typically asked to choose between descriptions. Crucially, however, it does not apply to the cases I consider.

It is absurd to ask whether the two options of *Following Artemis*'s *Will* and *Murdering My Daughter* are semantically or informationally equivalent. Of course, they are not! The point is that they are, and are known by Agamemnon to be, different ways of framing the same outcome. It is the known identity of the outcome (the death of Iphigenia) that does the work done in the classic experiments by the semantic/informational equivalence of the described scenarios. In fact, this is precisely the contrast that **Mandel** is looking for between the small-world cases and the real-world cases.

R4.2. The nature of preference

The notion of preference is key to my argument. **Guala**, **Lau**, **Weisel and Zultan**, and **Pettigrew** propose (somewhat overlapping) ways of thinking about preference that, they claim, will undercut my claims about the rationality of quasi-cyclical preferences. **Weirich** is more conciliatory. He analyzes preference with a view to building a bridge between my account and standard decision theoretic models of rationality.

Guala suggests that I am confusing preferences with other psychological states, such as reasons, desires, emotions, and so on, when I state that the rationality of framing effects is a function of the rationality of a complex and multi-faced response to a complex and multi-faceted situation. Rational preferences have to be "all-things-considered." The complex and multi-faceted responses that I discuss are not themselves preferences, but rather the inputs to a process of reflection that yields an "all-things-considered" preference. He writes: "Agamemnon may want to follow Artemis's will (under the grip of Frame A), and may want to fail his ships and people (under the grip of B), but he cannot prefer both. Macbeth may have a desire or a reason to fulfil his double duty to Duncan, and another desire or reason to take the throne, but he cannot prefer both, in the sense of rational preference."

There is no doubt that **Guala** is correctly characterizing *an* orthodox account of rational preference (we concur in rejecting the theory of revealed preference). However, it does seem a little question-begging simply to quote back at me the theory I am criticizing. My claim is that preferences cannot be frame-neutral in

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

the way that the Guala and the orthodox view hold. We both accept that there are frame-relative emotional responses, reasons, desires, and so forth. I argue, though, that it will not always be possible to turn those frame-relative reasons into frame-neutral all-things-considered preferences. This will happen when the force and appeal of the reasons are tied to the frame in which they emerge, so that stepping back from the frame weakens their hold.

An idea worth exploring, although not developed in the target article, or in Bermúdez (2020), is that the requirement that rational preferences be all-things-considered can in fact be applied within my own context – that is, by requiring that rational, frame-relative preferences be all-things-considered (relative to that frame). In fact, my position is compatible with an even stronger claim: Which is that rational preferences must be *maximally all-things-considered*. That is to say, a rational preference ordering must be complete over all comparable outcomes and reasons (i.e., over all the things that can be considered together). I am not motivated to revise my fundamental claim, though, that a rational agent can have more than one *maximally all-things-considered preference ordering*, in cases where there are reasons, emotional responses, desires, and so on, that cannot be considered together. (On this see the discussion of **Pettigrew** below.)

I would emphasize similar points in response to **Lau** and **Weisel and Zultan**, both of whom propose that my *quasi-cyclical preferences* are best viewed as *ceteris paribus* preferences (i.e., preferences, all other things being equal). **Lau** suggests that *ceteris paribus* preferences are defeasible and general, whereas rational preferences are absolute and general. As he points out, it can be perfectly rational to have conflicting *ceteris paribus* preferences: "If I say I prefer coffee over tea, we normally take this to involve an implicit qualification – all else being equal, I prefer coffee to tea. This preference is defeasible and not absolute. I am not being inconsistent if I happen to choose tea over an overpriced, watery coffee." Weisel and Zultan offer a similar set of considerations: "While a statement such as 'Macbeth would like to bravely take the throne' reasonably carries an implicit ceteris paribus ('all other things being equal'), Bermúdez treats it as a coarse partition of the outcome space, in which Macbeth prefers any outcome that involves being the king to any outcome that does not. Quasi-cyclical preferences arise when considering a choice between actions for which different such partitions contrast. In the action space that Macbeth faces, the outcome that involves becoming the king necessarily involves breaching his double duty to Duncan. The cycle disappears when viewing Macbeth's preferences as ceteris-paribus rather than categorical preferences."

The distinction both commentaries make is perfectly valid, but it is not really applicable here. With respect to **Lau**, I actually take quasi-cyclical preferences to be highly specific. Agamemnon does not, generally speaking, think that following the will of Artemis is to be preferred to failing one's ships and allies. Rather, in this highly specific context (being becalmed at Aulis), he prefers the outcome that he frames as following Artemis's will in this highly specific way (by sacrificing Iphigenia). The specific versus general contrast is somewhat of a red herring, therefore, and the real weight of his argument is taken by the defeasible versus absolute contrast – which takes us back to the discussion of **Guala** above, because **Lau**'s absolute preferences are very similar to **Guala**'s all-things-considered preferences. Likewise for **Weisel & Zultan**. My claim about Macbeth is highly specific. I don't think it's true that he prefers any outcome in which he is king to any outcome that does not. My whole point is that it all depends upon how the outcome is framed – there are frames (e.g., the loyalty frame) where what he prefers is not to be king, where that is framed as doing his double duty to Duncan.

Pettigrew offers an alternative account of Agamemnon's preferences, on which they do not come out as quasi-cyclical. As he notes, a natural response to my description of the case would be to say that Agamemnon fails to have complete preferences, because he is unable to weigh two competing reasons against each other. This would really be another variation on the theme pursued by **Guala**, **Lau**, and **Weisel and Zultan**, because of the close connections between being *ceteris paribus*, being defeasible, being incomplete, and failing to be all-things-considered. As **Pettigrew** notes, though, rationality seems to prescribe indifference between options that cannot be compared and this does not really seem to capture what is going on with Agamemnon, who does end up killing his daughter, but not because he tossed a coin. So, **Pettigrew** proposes an alternative. Perhaps, he suggests, Agamemnon has different complete preference orderings at different times – at any given moment his preferences are complete, but he jumps from one ordering to another depending on which set of reasons and corresponding affective responses are most salient at that time.

I am not sure that this is really a drastic alternative to my own description of the situation (particularly given the suggestion earlier that each frame-relative ordering is maximally all-things-considered complete). The difference between simultaneously having different frame-relative preference orderings and cycling through them in quick succession may not really amount to much. I do, however, want to take issue with the objection that he makes, which is, in effect, that having multiple preference orderings makes Agamemnon susceptible to a money pump/Dutch book (i.e., a series of bets that is guaranteed to lose him money). Quite apart from the fact that losing money is probably the least of Agamemnon's worries, there is a fundamental problem with money pump arguments. A money pump argument would only show that a particular preference structure was irrational if rationality *mandates* accepting the problematic series of bets. But of course it does not, because a rational agent can (and should) simply refuse to play the game.

Weirich also pursues the theme of all-things-considered preferences, but he does so to try to show that a version of the principle of extensionality is compatible with my arguments. The basic point he makes is that the principle of extensionality, considered as a basic principle of rationality, is really only applicable in the ideal case where an agent has an all-things-considered preference ordering (i.e., a unique one). As he puts it, "The principle assumes an ideal agent in ideal circumstances facing a standard decision problem and possessing rational all-things-considered preferences among options." He continues: "Without these assumptions, an agent's choice may be rational despite failing to comply with the principle, or it may be irrational despite complying with the principle."

To explore the first option, **Weirich** suggests the standard theory ought to be happy to entertain decision rules for agents without a unique all-things-considered preference ordering. He proposes the following rule. It is rational to choose an option if it is at the top of some completion of an incomplete preference-ordering. Imagine an agent with an incomplete preference-ordering. He might, for example, prefer beetroot to cauliflower to daikon radish, but not be entirely sure how to decide between eggplant and avocado, although he knows that he prefers each of them to the first three. One completion

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

might have avocado at the top, followed by eggplant. Another might have eggplant at the top, followed by avocado. Both, by Weirich's lights could be rational, and, as he points out, frames can serve as a tool for suggesting one completion rather than another. In this way, then, it could be rational to have preferences that are frame-dependent.

It is churlish to look a gift horse in the mouth, and I am happy to accept **Weirich**'s proposal. However, I am not sure that orthodox decision theorists would be happy with his maneuver, because applying it to the cases I discuss entails that a rational agent can simultaneously have multiple different (and incompatible) completions of an incomplete preference ordering. This is rather different from the claim, which an orthodox decision theorist would surely accept, that there can be multiple rational completions of an incomplete preference ordering.

R4.3. The nature of rationality

The commentaries discussed in section R4.2 challenge my claim that quasi-cyclical preferences can be rational by criticizing how I understand preference. **Crockett and Paul**, **Przybyszewski et al.**, and **Teoh et al.** take issue in different ways with how I understand rationality.

Crockett and Paul accept that framing effects can give rise to rational quasi-cyclical preferences, but they argue that my account glosses over an important distinction. There are, they claim, examples of quasi-cyclical preferences that seem to meet all my requirements for rationality (e.g., by satisfying H1 through H3) but nonetheless seem to be irrational. They reach this conclusion by distinguishing between two types of quasi-cyclical preference, emerging from two different types of choice:

> Self-involving choices: "where an agent oscillates between first and third person perspectives that conflict regarding their life changing, or transformative, implications."
> (Paul, 2014)

> Self-serving choices: where the structure of preferences is rational in the first-person sense, but irrational in a third-person sense.

To illustrate, someone considering whether to have a child might end up with quasi-cyclical preferences, because the considerations may look very different relative to a first-personal framing (relative to her current, rewarding, child-free life) as opposed to a third-person framing (which would incorporate externalities, the testimony and advice of others, etc.). This would be self-involving, in the sense that it is a choice about the type of self one wants to be. But within that general category, some self-involving choices are also self-serving.

For example, "Macbeth might be able to convince himself he is 'bravely taking the throne' while observers see straight through his murderous power grab; Agamemnon assures himself he's 'following Artemis's will" while the audience looks on in horror as he kills his child. These examples occupy the pantheon of high drama because the audience can clearly see that the protagonist is fooling himself (meaning his decision is third-personally irrational) but can also empathize with the dilemma of the protagonist (because his decisions are first-personally rational). In other words, my central examples of quasi-cyclical preferences fail to be rational, not because there is anything wrong with quasi-cyclicality, but because of self-deception and ethical blind spots.

Stepping back from the details of Macbeth and Agamemnon (to whom I suspect I am much more sympathetic than **Crockett and Paul**), this raises interesting questions about the scope of a theory of rationality. In Bermúdez (2020) (but not, admittedly, in the target article) I discuss substantive constraints upon an account of rationality, in particular a version of the "no false belief" requirement: A model, frame-sensitive reasoner should not believe any false *factual* propositions, where a factual proposition is one for which there is a standardly accepted method for determining its truth value. It is a factual proposition, for example, that a 7-day-old fetus has a heart (false), but the proposition that a 7-day-old fetus is a person is non-factual. It would be an interesting result if it turned out that self-deception (and the other ethical/moral failings that Crockett & Paul discuss) could be shown to involve some breach of the "no false belief" requirement.

Przybyszewski et al. are also interested in the process by which rational preferences are reached and feel that I leave important evaluative questions unconsidered. They suggest, with some justice, that I am committed to a procedural notion of rationality – for example, that the outcome of a process of reasoning and reflection inherits the rationality of the process that yields it. The objection they raise is that I do not push the requirements of procedural rationality back far enough. In particular, I do not offer tools for evaluating the rationality of frames. They make the point through the lens of prospect theory and associated experimental work, which formalizes the construction of frames in terms of an editing process by which, for example, a reference point is set relative to which losses and gains are calculated. As they observe, the editing process is susceptible to multiple biases, such as the anchoring effect, the disposition effect, the endowment effect, and the certainty effect. All of these have the potential to introduce irrationality into the process.

This line of objection runs parallel to **Crockett and Paul**, because **Przybyszewski et al.** are in effect identifying a parallel set of ways in which frames can be generated irrationally, in a way that can contaminate the putative rationality of frame-sensitive reasoning. I think that this is a very valid point. I would be inclined, as above, to raise the possibility that constraints such as the "no false belief" requirement might screen out some or all of these biases. I would also be inclined to add that some of these biases are themselves the result of framing. Different frames will generate different anchors, for example. Similarly for loss aversion, which is a close relative of the endowment effect. What counts as a loss depends upon the horizon of evaluation. Do I calculate investment success relative to the last 12 months (yielding a modest loss) or relative to the 12 years since I made the investment (yielding a significant gain)? That depends upon how I frame my overall investment strategy. (For more on this, see Ch. 3 of Bermúdez, 2020).

Teoh et al. also emphasize how framing can be a non-rational process, but from a very different perspective. They emphasize computational complexity and the importance of attention (on which compare **Koi**). Real-world decision-making involves complex trade-offs with respect to the benefits of information versus the costs of gathering it. As they point out, the quantity of information available in the environment can only be managed through selective attention, and attention is often modulated in exogenous and stimulus-driven ways. They give numerous examples. The external environment is very influential, because attention is typically drawn to salient information in the environment. Appetites influence what is taken to be salient. And the way in which information is initially framed/attended to constrains and places limits on subsequent framings. They give an interesting illustration from the game theory discussion. In order to construct

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press

the "we"-frame a player needs not just to acquire information about their own pay-offs and the pay-offs of players (which would be required for best response reasoning in the "I"-frame), but also to combine them, using Pareto-optimality or some other criteria. This act of combination poses information costs that may prevent an agent from ever arriving at the "we"-frame.

These are important points. As indicated above and in the target article, however, I by no means want to say that all framing effects are rational, or even goal-driven. There is a spectrum of sophisticated and complexity, with primed responses at one end and the most complicated and reflective forms of what **Crockett and Paul** call self-involving choices at the other. Much of what **Teoh et al.** point to is most applicable at the primed response end of the spectrum. But the point is well taken that framing is not an abstract activity undertaken by computationally unbounded agents. It is a real-world activity subject to a multitude of constraints, both endogenous and exogenous.

Moldoveanu extends this line of reasoning, approaching the matter from the perspective of game theory. He agrees with me that frames can be useful tools for selecting equilibria in competitive games. For example, from the perspective of "I"-frame best-response reasoning, there is no way of choosing between the two equilibrium solutions in Stag Hunt – both players hunting hare and both players hunting stag, whereas the latter is clearly preferable from the "we"-frame. But, as **Moldoveanu** points out, this is the starting-point for investigations that I do not discuss in the target article. In particular (and complementing the points raised by **Teoh et al.**), we can and should raise questions of informational gain and computational cost with respect to frames themselves. The key question is: How much thinking is required to generate a reason R for acting from a frame F that structures the representation of the facts relevant to a situation? The computational complexity of the process of frame selection brings into play different notions of rationality (ecological/adaptive) that intersect with the computational benefits of using frames as efficient tools to shorten deliberation.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Conflict of interest. None.

References

Bermúdez, J. L. (2020). *Frame it again: New tools for rational thought*. Cambridge University Press.

Butler, L. P., & Walton, G. M. (2013). The opportunity to collaborate increases preschoolers' motivation for challenging tasks. *Journal of Experimental Child Psychology*, *116*(4), 953–961.

De Dreu, C. K. W., & Gross, J. (2019). Revisiting the form and function of conflict: Neurobiological, psychological, and cultural mechanisms for attack and defense within and between groups. *Behavioral and Brain Sciences*, *42*, e116.

Mandel, D. R. (2014). Do framing effects reveal irrational choice? *Journal of Experimental Psychology General*, *143*(3), 1185–1198.

Paul, L. A. (2014). *Transformative experience*. Oxford University Press.

Sher, S., & McKenzie, C. R. (2011). Levels of information: A framing hierarchy. In G. Keren (Ed.), *Perspectives on framing* (pp. 35–63). Psychology Press.

Sher, S., & McKenzie, C. R. M. (2006). Information leakage from logically equivalent frames. *Cognition*, *101*(3), 467–494.

Sturgill, J., Bergeron, J., Ransdell, T., Colvin, T., Joshi, G., & Zentall, T. R. (2021). "What you see may not be what you get": Reverse contingency and perceived loss aversion in pigeons. *Psychonomics Bulletin Review*, *28*(3), 1015–1020.

Vasil, J., & Tomasello, M. (2022). Effects of "we"-framing on young children's commitment, sharing, and helping. *Journal of Experimental Child Psychology*, *214*, 105278.

https://doi.org/10.1017/S0140525X2200005X Published online by Cambridge University Press