Theory and Decision (2021) 91:337–356
https://doi.org/10.1007/s11238-021-09801-8

# Rationality, preference satisfaction and anomalous intentions: why rational choice theory is not self-defeating

**Roberto Fumagalli**[1,2,3]

Accepted: 18 January 2021 / Published online: 31 March 2021


**Abstract**
The critics of rational choice theory (henceforth, RCT) frequently claim that RCT is self-defeating in the sense that agents who abide by RCT's prescriptions are less successful in satisfying their preferences than they would be if they abided by some normative theory of choice other than RCT. In this paper, I combine insights from philosophy of action, philosophy of mind and the normative foundations of RCT to rebut this often-made criticism. I then explicate the implications of my thesis for the wider philosophical debate concerning the normativity of RCT for both ideal agents who can form and revise their intentions instantly without cognitive costs and real-life agents who have limited control over the formation and the dynamics of their own intentions.

**Keywords** Rationality · Preference satisfaction · Intentions · Normativity · Decision-making

## 1 Introduction

The critics of rational choice theory (henceforth, RCT) frequently claim that RCT is self-defeating in the sense that agents who abide by RCT's prescriptions are less successful in satisfying their preferences than they would be if they abided by some normative theory of choice other than RCT (e.g. Bratman 1999, 2000; Gauthier 1984, 1997; Kavka 1978, 1983; McClennen 1990, 1997). The idea is that abiding by RCT's prescriptions hampers (rather than enhances) agents' ability to satisfy their

✉ Roberto Fumagalli
roberto.fumagalli@kcl.ac.uk; R.Fumagalli@lse.ac.uk
https://www.kcl.ac.uk/people/roberto-fumagalli; http://personal.lse.ac.uk/fumagall/

[1] Lecturer and Director of the Philosophy, Politics and Economics Programme, King's College London, London, UK

[2] London School of Economics, London, UK

[3] University of Pennsylvania, Philadelphia, USA

preferences, and that an agent who abides by RCT's prescriptions will often "end up satisfying his preferences less well than he would have done, had he [abided by] some other [theory]" (Sugden 1991, 752; also Bradley 2007 and 2017; Broome 2007a and 2007b; Dietrich et al. 2013 and 2019; Rabinowicz 1995 and 2019; Spohn 2009 and 2012, for recent discussions). In this paper, I combine insights from philosophy of action, philosophy of mind and the normative foundations of RCT to rebut this often-made criticism. I then explicate the implications of my thesis for the wider philosophical debate concerning the normativity of RCT for both ideal agents who can form and revise their intentions instantly without cognitive costs and real-life agents who have limited control over the formation and the dynamics of their own intentions.[1]

The paper is organized as follows. In Sect. 2, I examine one issue that figures centrally in the debate as to whether RCT is self-defeating, namely whether an agent who abides by RCT's prescriptions can rationally form what I call *anomalous intentions*, i.e. intentions to perform actions that maximize the total stream of payoffs the agent can get over the entire course of a decision problem, but fail to maximize the payoffs the agent can get from some subsequent choice nodes onwards (e.g. Bratman 1999, 2000; Gauthier 1984, 1997; Kavka 1978, 1983; McClennen 1990, 1997). I then explicate my thesis that, despite prominent criticisms of RCT, agents can rationally form anomalous intentions, and therefore prominent attempts to demonstrate that RCT is self-defeating do not withstand scrutiny. In Sects. 3–6, I defend my thesis that agents can rationally form anomalous intentions against four major objections put forward in the specialized literature, namely: the objection from *temporal situatedness* (e.g. Bratman 1998; Mintoff 1997); the objection from *bootstrapping* (e.g. Bratman 2009; Broome 2013); the objection from *psycho-physical inability* (e.g. Farrell 1989; Shah 2009); and the *overdemandingness* objection (e.g. Mongin 2000; Steele 2006).[2]

My thesis that agents can rationally form anomalous intentions has at least three implications of general interest for the wider philosophical debate concerning the normativity of RCT (e.g. Bradley 2007, 2017; Broome 2007a, b; Dietrich et al. 2013, 2019; Rabinowicz 1995, 2019; Spohn 2009, 2012). First, anomalous intentions figure in a vast range of decision problems where the payoffs agents

[1] My defence of RCT focuses on RCT as a normative (rather than descriptive) theory of choice. Among normative theories of choice, some theories specify what consistency conditions (e.g. transitivity) preferences ought to satisfy, other theories specify how one ought to act (e.g. what intentions one ought to form) given her preferences. Below I primarily focus on theories that specify how one ought to act given her preferences since most debates concerning the putative self-defeating character of RCT target such theories (Sects. 2–6). For a discussion of theories that specify what consistency conditions preferences ought to satisfy, e.g. Fumagalli 2013, 2019; Hausman 2000, 2012. For a discussion of the interrelations between theories that specify what consistency conditions preferences ought to satisfy and theories that specify how one ought to act given her preferences, e.g. Cantwell 2003, 2001; Rabinowicz 1997; Steele 2010.

[2] I focus on individual decision problems (as opposed to strategic decision problems) since applications of RCT to strategic decision problems raise several concerns tangential to my evaluation (e.g. McClennen 1985, on reputation). Also, I target sequential decision problems (as opposed to non-sequential decision problems) since most debates concerning the putative self-defeating character of RCT target sequential decision problems (Sects. 2–6). For a discussion targeting non-sequential decision problems, e.g. Joyce 2007; Spohn 2012, on Newcomb's problems.

can get if they form the intention to perform an action are at least partly independent of the payoffs agents can get if they actually perform such action (e.g. Andreou 2006; Clarke 2007; van Hees and Roy 2009). Hence, showing that agents can rationally form anomalous intentions would have significant bearing on the normativity of RCT across several decision problems. Second, the claim that agents cannot rationally form anomalous intentions is commonly premised on the assumption that instrumental rationality requires agents to maximize the payoffs they can get from the choice nodes they face onwards irrespective of whether maximizing these payoffs maximizes the total stream of payoffs agents can get over the entire course of a decision problem (e.g. Bratman 1998, 62–66, Mintoff 1997, 624–5, Williams 1981, 35). Below I critically examine this widely endorsed conception of instrumental rationality and argue that in presence of anomalous intentions such conception has implications that contrast with independently plausible requirements of payoff maximization. And third, several authors build on the claim that RCT is self-defeating to argue that this theory must be revised or even rejected (e.g. Bratman 1999, 2000; Gauthier 1984, 1997; Kavka 1978, 1983; McClennen 1990, 1997). If my thesis that agents can rationally form anomalous intentions is correct, prominent attempts to demonstrate that RCT is self-defeating do not withstand scrutiny. This result does not per se vindicate RCT as our best available normative theory of choice. Still, it challenges RCT's critics to put forward more convincing reasons and evidence to support their claim that RCT is self-defeating.[3]

Before proceeding, one preliminary remark is in order. Various characterizations of preferences and intentions have been advocated in the economic and philosophical literatures (e.g. Cozic and Hill 2015; Dietrich and List 2016; Guala 2019; Hausman 2011; Jeffrey 1965; Savage 1954; Thoma 2017, on preferences; Anscombe 1963; Bratman 1987; Davidson 1978; Holton 2009; Roy 2009; Searle 1983; Tenenbaum 2018, on intentions). I shall expand on these characterizations wherever my evaluation directly rests on those characterizations (e.g. footnote no.7 on the so-called belief constraint on intending; also footnote no.17 for a comparison between intentions and beliefs). For now, it suffices to note that although intentions do not figure in all applications of RCT, many applications of RCT model decision-making in terms of the formation and the dynamics of intentions (e.g. Audi 1991; Bales 2020; Cullity 2008; Mele 2000; Pink 1991), and the debate concerning the putative self-defeating character of RCT often targets such applications (Sects. 2–6).

[3] Parfit famously distinguishes between 'self-defeating' theories, which "give us certain aims, but also tell us to act in ways that frustrate these aims", and 'self-effacing' theories, which "[tell] everyone to cause himself to believe some other theory" if one is able to do so (1984, abstract and 24). The notion of self-defeating theories I target closely resembles Parfit's notion of self-defeating theories. Some authors doubt that the alleged fact that RCT is self-defeating in the sense I target would per se require significant revisions of RCT (e.g. Hedden 2015; Levi 1987; Rabinowicz 1995). Below I grant, for the sake of argument, several critics' assumption that if RCT was shown to be self-defeating in such sense, this would require significant revisions of RCT.

## 2 RCT and anomalous intentions

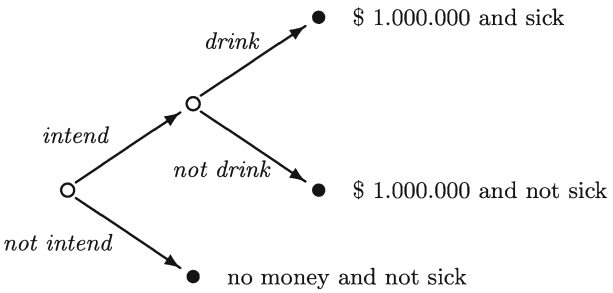
In this section, I examine the issue whether an agent who abides by RCT's prescriptions can rationally form what I call *anomalous intentions*, i.e. intentions to perform actions that maximize the total stream of payoffs the agent can get over the entire course of a decision problem, but fail to maximize the payoffs the agent can get from some subsequent choice nodes onwards. To clarify this issue, I focus on a putative paradox that is commonly taken to indicate that RCT is self-defeating, namely Kavka's (1983) 'toxin puzzle'. I then explicate my thesis that, despite prominent criticisms of RCT, agents can rationally form anomalous intentions even in this putative paradox, and therefore prominent attempts to demonstrate that RCT is self-defeating do not withstand scrutiny.

Kavka explicates his toxin puzzle as follows:

> "An eccentric billionaire [offers] you the following deal. He places before you a vial of toxin that, if you drink it, will make you painfully ill for a day, but will not threaten your life or have any lasting effects. […] The billionaire will pay you one million dollars tomorrow morning if, at midnight tonight, you intend to drink the toxin tomorrow afternoon. […] You need not drink the toxin to receive the money; in fact, the money will already be in your bank account hours before the time for drinking it arrives, if you succeed. […] All you have to do is sign the agreement and then intend at midnight tonight to drink the stuff tomorrow afternoon. You are perfectly free to change your mind after receiving the money and not drink the toxin. [However, arranging] external incentives is ruled out, as are such alternative gimmicks as hiring a hypnotist to implant the intention, forgetting the main relevant facts of the situation, and so forth. […] The presence or absence of the intention is to be determined by the latest 'mind-reading' brain scanner [which] will correctly detect the presence or absence of the relevant intention. [The question is whether today you can rationally form] the intention to drink the toxin [tomorrow]" (1983, 33–34).

Kavka's puzzle is often represented by the following decision tree with two choice nodes (e.g. Van Hees and Roy 2009, Fig. 4), one before midnight where the agent decides whether or not to form the intention to drink the toxin (assuming the agent is psycho-physically able to form such intention) and the other tomorrow afternoon where the agent decides whether or not to drink the toxin (assuming the agent has formed the intention to drink the toxin)[4]:

[4] Standard decision trees assign utilities at each terminal node of decision problems and presuppose that the modelled agents' preferences satisfy consistency conditions sufficient to allow for a utility representation (e.g. Cubitt and Sugden 2001). Below I follow most authors discussing Kavka's decision problem in assigning consequences (rather than utilities) to each terminal node of such decision problem, without presupposing that the modelled agent's preferences satisfy consistency conditions sufficient to allow for a utility representation.




Before assessing the proposed solutions of this puzzle, let us distinguish between what I call *global payoffs* (henceforth, GP), i.e. the total stream of (expected or actual) payoffs an agent can get over the entire course of a decision problem, and *subsequent payoffs* (henceforth, SP), i.e. the stream of (expected or actual) payoffs an agent can get from the choice nodes she faces onwards.[5] At the initial choice node of a decision problem, the set of actions that maximize one's GP and the set of actions that maximize one's SP coincide. At later choice nodes, the two sets of actions may significantly differ, and maximizing one's GP may require one to "perform an action other than the one that at the time of performance would [maximize her SP]" (Gauthier 1994, 697).[6]

To illustrate this, consider again the decision problem envisaged by Kavka. In this decision problem, maximizing GP requires the agent to form the intention to drink the toxin despite anticipating that, when it comes to drinking the toxin, drinking it will fail to maximize her SP. In particular, forming the intention to drink the toxin and drinking the toxin is the course of action that maximizes the agent's GP. To be sure, the agent could conceivably get even higher GP if she succeeded in forming the intention to drink the toxin today and later revised such intention before drinking the toxin. However, under Kavka's (1983, 34) assumptions, the agent anticipates since the first choice node whether she will revise the intention to drink the toxin before drinking the toxin. Moreover, if the agent anticipates that she will revise the intention to drink the toxin, she will not be able to form such intention and she will fail to get the million (ibid., 34). Hence, of the courses of action that the agent can successfully plan to perform, forming the intention to drink the toxin and drinking it are part of the course of action that maximizes the agent's GP.[7]

[5] In Kavka's toxin puzzle, the agent's expected payoffs coincide with her actual payoffs. I consider in Sects. 3–6 variants of the toxin puzzle where the agent's expected payoffs and actual payoffs diverge. For the purpose of my evaluation, I leave it open whether agents ought to maximize expected (as opposed to actual) payoffs, and I do not take a position concerning the rationality of risk attitudes (e.g. Okasha 2007; Schulz 2008, for a debate).

[6] Kavka's toxin puzzle differs from so-called temptation cases, where agents' preferences change in the course of the decision problem, since no such preference change figures in the toxin puzzle. I do not expand on temptation cases in my evaluation. For a discussion of temptation cases and the puzzles they pose to various conceptions of instrumental rationality, e.g. Andreou 2006; Holton 2004; Thoma 2017.

[7] Several authors posit a belief constraint on intending according to which one cannot intend to perform an action while believing that she will not perform such action (e.g. Mele 1995; Sobel 1994). The idea is that intending to perform an action requires one to believe at least that she will probably perform such action (e.g. Audi 1973; Bratman 1987, ch.3). Some doubt the plausibility of this constraint (e.g. Holton

Is it *rational* for an agent who abides by RCT's prescriptions to form the intention to drink the toxin in the decision problem envisaged by Kavka? Two sets of solutions are prominent in the literature. On the one hand, the *globalists* hold that an agent can *always* rationally form the intention to drink the toxin because, of the courses of action that the agent can successfully plan to perform, forming the intention to drink the toxin and drinking it are part of the course of action that maximizes the agent's GP (e.g. Gauthier 1994, 721, McClennen 1990, 230–1). The globalists' reasoning proceeds as follows. One can rationally form an intention if one has reason to expect that forming and acting on this intention are part of the course of action that maximizes her GP (e.g. Gauthier 1998, 50). Of the courses of action that the agent can successfully plan to perform in the toxin puzzle, forming the intention to drink the toxin and drinking it are "part of the best course of action that [the agent] could embrace as a whole" (Gauthier 1998, 48). For these actions are the only way for the agent to get the million, which by assumption makes her better off even at the cost of drinking the toxin (Gauthier 1994, 702–9; also McClennen 1990, 230–1). Therefore, an agent can always rationally form the intention to drink the toxin.[8]

On the other hand, the *localists* hold that an agent *cannot* rationally form the intention to drink the toxin in the decision problem envisaged by Kavka because the agent knows that, when it comes to drinking the toxin, drinking it will fail to maximize her SP, and one cannot rationally form the intention to perform an action that she knows will fail to maximize her SP (e.g. Kavka 1983, 34–5; also Bratman 1998, 62 and 72–73; Quinn 1985, 371). The localists' reasoning proceeds as follows. One can rationally form an intention if one has reason to expect that forming and acting on this intention are part of the course of action that maximizes her SP. In the toxin puzzle, the million gives the agent reason to form the intention to drink the toxin because, by assumption, the agent will be better off if she gets the million and drinks the toxin, than if she does not form the intention to drink the toxin.[9] Yet, when it comes to drinking the toxin, the agent will know whether or not she has got the million. And at that point, it will be irrational for the agent to drink

Footnote 7 continued
2008; McCann 1991). Below I assume, for the sake of argument, that intending to perform an action requires one to believe that there is some positive probability that she will perform this action (e.g. Harman 1986; Velleman 1989).

[8] Globalists often advocate the so-called resolute approach to sequential decision problems, according to which agents who choose between courses of action ought to consider which course of action they prefer at the initial node of a decision problem and then stick to such course of action (e.g. McClennen 1990). However, one may endorse the globalist solution without advocating the resolute approach to sequential decision problems (e.g. Gauthier 1997; Machina 1989; Meacham 2010). The solution I advocate below differs from both the resolute approach and the globalist solution.

[9] I speak of reasons for attitudes broadly to indicate pro tanto considerations that count in favour of forming and acting on such attitudes (e.g. Audi 1986; Broome 2004). In doing so, I mention categorizations of distinct kinds of reasons for attitudes (e.g. Sect. 4 on Parfit's 2001, contrast between object-given and putative state-given reasons for attitudes) whenever these categorizations directly bear on my evaluation. For a debate on the relationship between rationality and reasons for attitudes and on how reasons for attitudes purportedly depend on one's internal attitudes and on external facts independent of one's attitudes, e.g. Alvarez 2018; Arpaly 2000; Broome 2007a; b; Kolodny 2005, 2007; Littlejohn 2016, 2018; Parfit 1997.

the toxin. For drinking the toxin will bring the agent a day of illness without yielding any apparent benefit. Moreover, the agent is assumed to know all these facts today. Hence, today the agent cannot rationally form the intention to drink the toxin tomorrow, and she fails to get the million.[10]

The globalist and the localist solutions are grounded on different conceptions of instrumental rationality, which take actions (e.g. forming specific intentions and acting on such intentions) to be instrumentally rational to the extent that these actions maximize agents' GP and SP, respectively. These two solutions concur in cases where the set of actions that maximize SP and the set of actions that maximize GP coincide, but diverge in cases where these sets of actions differ. In the remainder of this paper, I critically examine those two solutions and argue that neither solution is sufficiently sensitive to the extent RCT's prescriptions vary depending on agents' ability to control the formation and the dynamics of their own intentions. More specifically, I shall argue that: (1) contrary to the localist solution, "an action may be rational even though at the time of performance it [does] not, and is not believed to [maximize SP]" (Gauthier 1994, 701), and so it *can* be rational for agents to form anomalous intentions such as the intention to drink the toxin in the decision problem envisaged by Kavka; and (2) contrary to the globalist solution, whether agents can rationally form anomalous intentions crucially depends on the extent to which agents are able to control the formation and the dynamics of their own intentions, and so it is *not* the case that agents can *always* rationally form anomalous intentions such as the intention to drink the toxin in the decision problem envisaged by Kavka.

My claims (1) and (2) agree with the globalists that the rationality of forming intentions is to be assessed in terms of whether forming and acting on such intentions are part of the course of action that maximizes agents' GP. However, they reject the globalist solution that agents can always rationally form anomalous intentions. In particular, they hold that whether it is rational for agents to form such intentions crucially depends on the extent to which these agents are able to control the formation and the dynamics of their own intentions. In Sects. 3–6, I defend my claims (1) and (2) against four major objections put forward in the specialized literature, namely: the objection from *temporal situatedness* (e.g. Bratman 1998; Mintoff 1997); the objection from *bootstrapping* (e.g. Bratman 2009; Broome 2013); the objection from *psycho-physical inability* (e.g. Farrell 1989; Shah 2009); and the *overdemandingness* objection (e.g. Mongin 2000; Steele 2006). In doing so, I explicate the implications of my claims for the wider philosophical debate concerning the normativity of RCT for both ideal agents who can form and revise their intentions instantly without cognitive costs and real-life agents who have limited control over the formation and the dynamics of their own intentions.[11]

[10] Localists often advocate the so-called sophisticated approach to sequential decision problems, according to which agents who choose between courses of action ought to consider what actions they will choose at the final choice nodes and then proceed backwards through the decision tree to identify what actions they will choose at earlier choice nodes (e.g. Hammond 1988). However, one may endorse the localist solution without advocating the sophisticated approach to sequential decision problems (e.g. Cubitt 1996; Levi 1991; Thoma 2020). The solution I advocate below differs from both the sophisticated approach and the localist solution.

[11] I expand on the defence of my claims (1) and (2) in Sects. 3–6 (rather than here) to make it clear in what respects exactly the position I advocate differs from the positions advocated by prominent authors in the literature. My claims (1) and (2) are not the only solution which differs from both the globalist and the

## 3 Objection from temporal situatedness

The objection from *temporal situatedness* holds that agents cannot rationally form anomalous intentions because instrumental rationality requires agents to maximize their SP irrespective of whether maximizing SP enables agents to maximize their GP (e.g. Bratman 1998, 62–66, Mintoff 1997, 624–5). The objection proceeds as follows. Forming an intention is instrumentally rational for an agent only if forming and acting on this intention enables the agent to maximize her SP (e.g. Williams 1981, 35). However, anomalous intentions are defined as intentions to perform actions that—while maximizing one's GP—fail to maximize one's SP. Therefore, agents cannot rationally form anomalous intentions. Paraphrasing Bratman's remarks about instrumentally rational choice in sequential decision problems, "the agent may well rank her alternatives with respect to past [intentions, but] what is now under her control are her alternatives from now on [and one ought to base her decisions on] one's ranking of options that are at that time in one's control" (1998, 66).

This objection points to a widely endorsed localist conception of instrumental rationality, which takes actions (e.g. forming specific intentions and acting on such intentions) to be instrumentally rational to the extent that these actions maximize agents' SP. This localist conception of instrumental rationality has plausible implications in decision problems where the set of actions that maximize one's SP and the set of actions that maximize one's GP coincide. In presence of anomalous intentions, however, these two sets of actions differ. In these situations, it would be implausible to hold that instrumental rationality *invariably* requires one to maximize SP (rather than GP). For in presence of anomalous intentions, agents who aim to maximize their SP often obtain lower (expected or actual) payoffs than the (expected or actual) payoffs they would obtain if they aimed to maximize their GP (e.g. DeHelian and McClennen 1993; Machina 1989). To be sure, the fact that agents who aim to maximize their SP often obtain lower (expected or actual) payoffs than the (expected or actual) payoffs they would obtain if they aimed to maximize their GP does not imply that agents can *always* rationally form the intention to perform actions that maximize their GP (e.g. Sect. 5 on cases where agents know that they are psycho-physically unable to form such intentions). Yet, it indicates that one *can* rationally form anomalous intentions.

To illustrate this, consider the following *transfer variants* of Kavka's toxin puzzle, where the agent gets the million tomorrow if today she forms the intention to give up (e.g. give back to the billionaire) a specified part of the million upon

Footnote 11 continued
localist solutions. Another such solution is Rabinowicz's (2014) 'unified choice', according to which an agent ought to act, at each time, on her current evaluation of the entire decision tree she faces. This position counts as neither localist (e.g. it recommends agents to evaluate entire decision trees rather than just subsequent parts of such trees) nor globalist (e.g. it recommends agents to act on their current evaluations of decision trees rather than their past evaluations of such trees). Still, my solution differs from unified choice in several decision problems that involve anomalous intentions, including the toxin puzzle. For unified choice denies that agents can rationally form anomalous intentions in decision problems such as the toxin puzzle (e.g. Rabinowicz 2019; analogous remarks apply to Rabinowicz's 1995, 'wise choice').

receiving the million the day after. These variants retain the assumptions of Kavka's original puzzle (e.g. if the agent forms the relevant intention, the million will be in her bank account before the time of acting on such intention; the agent is free to change her mind after receiving the million and before acting on the relevant intention; irreversible arrangements that bind the agent to act on the relevant intention are ruled out by assumption), but eliminate reference to the toxin and its illness-related effects. In this way, they bypass various tangential issues affecting Kavka's original puzzle (e.g. commensurability of money prizes and states of illness) and make it easier to assess the merits of the proposed solutions of such puzzle.[12]

Is it *rational* for an agent who abides by RCT's prescriptions to form the intention to give up a specified part of the million upon receiving the million the day after? According to the localist solution, an agent cannot rationally form the intention to give up any positive part of the million upon receiving the million the day after, since giving up any positive part of the million would fail to maximize the agent's SP. This solution might seem plausible in transfer variants of the puzzle where the agent is required to form the intention to give up the whole (or most of the) million, but is far less plausible in variants where the agent is required to form the intention to give up a small part of the million. In fact, it would be rather implausible to claim that an agent cannot rationally form the intention to give up *any* positive part of the million upon receiving the million the day after.

To see this, consider a *minimum* transfer variant of the puzzle, where the agent gets the million tomorrow if today she forms the intention to give up 1 dollar upon receiving the million. In this context, the mere fact that giving up 1 dollar will fail to maximize the agent's SP falls short of implying that the agent cannot rationally form the intention to give up such dollar. For such dollar will be of extremely low value to the agent upon receiving the million, and the agent can gain 999,999$ by forming the intention to give up such dollar compared to a situation where she does not form such intention. This point holds not merely for a conveniently selected subset of transfer variants of the toxin puzzle, but generalizes across a wide subset of such variants. That is to say, given an arbitrarily small amount of money $\varepsilon > 0$ that (while being lower than 1 million) is *large enough* to make a difference to the agent's payoff valuations, one can construct a transfer variant of the puzzle where the agent gets the million if today she forms the intention to give up ε. To claim that the agent cannot rationally form such intention *irrespective* of how small ε is amounts to a *reductio ad absurdum* of the localist solution. To put it differently, one can rationally form the intention to perform actions that maximize her GP but fail to

[12] In building these variants of the toxin puzzle, I retain Kavka's assumption that agents' preferences target only monetary payoffs (illness-related payoffs are assumed away in such variants). The objects of real-life agents' preferences often include several factors besides monetary payoffs such as desire to stick to previous decisions, aversion to regret, and so on (e.g. Fumagalli 2020a, b). Still, transfer variants of the puzzle can be constructed where agents' preferences target also these additional factors. That is to say, if agents ascribe value to factors such as sticking to previous decisions and avoiding feelings of regret, then one can construct transfer variants of the puzzle where the payoffs reflect such value. And as I argue below, in many such variants of the puzzle the localist conception of instrumental rationality has implications that contrast with independently plausible requirements of payoff maximization.

maximize her SP, and the claim that instrumental rationality invariably requires one to maximize her SP (rather than GP) does not withstand scrutiny.[13]

## 4 Objection from bootstrapping

The objection from *bootstrapping* holds that agents cannot rationally form anomalous intentions because intending to perform an action cannot per se "bootstrap a new reason into existence" that makes it rational to perform such action (Broome 2007b, 354; also Bratman 1987, 24–27, 2009, 415–6). The objection proceeds as follows. Forming the intention to perform an action can *indirectly* create reasons to perform this action (e.g. Sobel 1994, on cases where forming an intention makes one's choice situation change in ways that make it rational to act on such intention). However, forming the intention to perform an action cannot per se create any new reason to perform this action, i.e. any reason to perform such action that one did not have before forming the intention. For "if it did, we could give ourselves a reason to [perform an action] just by intending to [perform] it; and that cannot be right" (Holton 2004, 513; also Broome 2013, ch.5). Now, the objection goes, the fact that forming the intention to perform an action cannot per se create any new reasons to perform this action, together with the fact that agents have no reason to perform actions that fail to maximize their SP, implies that agents have no reason to form the intention to perform such actions. Hence, agents cannot rationally form anomalous intentions.[14]

This objection correctly notes that in many decision problems, the mere fact that one forms the intention to perform an action does not per se make it rational for her to perform such action. However, this does not exclude that "[one's] reasons for performing an action can derive from her reasons for forming the preceding intention" (Gauthier 1994, 709; also Harman 1998, 84). Moreover, realizing that forming and acting on an anomalous intention are part of the course of action that maximizes one's GP provides one with reason to form and act on such intention (e.g. Gauthier 1994, 721, Smith 2016, 2263–4). This reason, despite the globalists, is not always strong enough to make it rational for one to form anomalous intentions

[13] One may object that for many real-life agents, giving up 1 dollar is such a small loss that it makes no difference to agents' payoff valuations, and so is plausibly regarded as no loss at all. This objection can be addressed by noting that transfer variants of the toxin puzzle target amounts of money that, while being arbitrarily small, 'are *large enough* to make a difference to agents' payoff valuations'. That is to say, if 1 dollar makes no difference to agents' payoff valuations, then one can construct transfer variants of the puzzle involving amounts of money that (while being lower than 1 million) are significantly higher than 1 dollar (e.g. 1000 dollars, 10,000 dollars, etc.). And in many such variants of the puzzle, the localist conception of instrumental rationality has implications that contrast with independently plausible requirements of payoff maximization.

[14] The objection from bootstrapping is so called because it builds on various authors' claims about bootstrapping. This should not be taken to indicate that all these authors endorse all the tenets of such objection (e.g. Broome, personal correspondence, agrees that 'forming the intention to perform an action cannot per se create any new reason to perform this action', but does not endorse the claim that 'agents have no reason to perform actions that fail to maximize their SP').

(e.g. Holton 2004).[15] Yet, despite the localists, in cases where aiming to maximize GP enables one to obtain higher (expected or actual) payoffs than the (expected or actual) payoffs she would obtain if she aimed to maximize SP, such reason can be strong enough to make it rational for an agent to form anomalous intentions. In fact, one may envision several decision problems where very large discrepancies between the (expected or actual) payoffs one would obtain by aiming to maximize her GP and the (expected or actual) payoffs one would obtain by aiming to maximize her SP cast doubt on the localists' claim that instrumental rationality invariably requires agents to maximize their SP rather than GP (e.g. Sect. 5 below on large prize variants of the toxin puzzle).[16]

A proponent of the objection from bootstrapping may concede that forming anomalous intentions may yield significant benefits to an agent (e.g. Bratman 2000, on coordination with one's future selves). However, she may object that these benefits fall short of licensing the claim that agents can rationally form anomalous intentions. The objection proceeds as follows. Whether an agent can rationally form specific intentions is "a function of the […] rationality of performing the *actions* that they are intentions to perform" rather than the *benefits* derivable from forming such intentions (Farrell 1989, 293; also Parfit 2001, 21–22, on the contrast between object-given reasons for attitudes, which derive from facts about the objects of agents' propositional attitudes, and putative state-given reasons for attitudes, which derive from facts about agents' having such propositional attitudes). Hence, to settle the question whether it is rational for her to intend to drink the toxin, an agent must settle the question whether it is rational for her to drink the toxin (e.g. Goetz 1998; Shah 2009). Unfortunately, the agent anticipates that, when it comes to drinking the toxin, it will be irrational for her to drink the toxin because drinking the toxin fails to maximize her SP (e.g. Mintoff 1997; Quinn 1985). Therefore, the agent cannot rationally form the intention to drink the toxin.

This objection invites two rejoinders. First, the mere fact that an action such as drinking the toxin fails to maximize an agent's SP does not per se indicate that such action is irrational unless one already presupposes a localist conception of instrumental rationality. Hence, reiterating that drinking the toxin fails to maximize

[15] A globalist may object that anomalous intentions, once formed, often have some inertial force in virtue of which they resist reconsideration (e.g. Den Hartogh 2004). This objection is plausible, but falls short of licensing the claim that agents can always rationally form anomalous intentions. For it is a "causal" (rather than normative) matter that "once you have an intention, you usually retain it until you carry it out" (Broome 2001, 113). And it is an open empirical question whether such inertia is so strong that forming anomalous intentions enables agents to maximize their GP (e.g. Ferrero 2010).

[16] One may object that if an instrumentally rational agent intentionally drinks the toxin, then this agent must associate higher SP to drinking the toxin (and suffering one day of illness) than to not drinking it. Hence, the intention to drink the toxin is not plausibly regarded as anomalous. This objection invites two rejoinders. First, the objection presupposes a localist conception of instrumental rationality, and so does not per se provide any independent reasons to endorse such conception. And second, an instrumentally rational agent may intentionally drink the toxin even if she associates lower SP to drinking the toxin (and suffering one day of illness) than to not drinking it. To see this, suppose an agent associates higher SP to not drinking the toxin than to drinking it (and suffering one day of illness). Assume further that the agent has reason to believe that her intending to drink the toxin is a sufficient reason to drink it. This agent may intentionally drink the toxin even if she associates lower SP to drinking the toxin (and suffering one day of illness) than to not drinking it.

an agent's SP does not per se provide any independent reasons to think that the agent cannot rationally form the intention to drink the toxin. And second, in presence of anomalous intentions, an agent may have reasons to form an intention (e.g. maximizing one's GP) which make it rational to form and act on this intention even if an agent's reasons to form an intention do not generally coincide with the agent's reasons to act on such intention (e.g. Clarke 2008; Pink 1998). More specifically, realizing that forming and acting on the intention to drink the toxin are part of the course of action that maximizes the agent's GP provides the agent with reason to form and act on such intention. This reason, despite the globalists, is not always strong enough to make it rational for an agent to form the intention to drink the toxin (e.g. Sobel 1994). Yet, despite the localists, such reason can make it rational for the agent to form the intention to drink the toxin even if the agent knows that drinking the toxin fails to maximize her SP.

To illustrate this, consider the following *psychological variant* of the toxin puzzle, where the billionaire offers the toxin deal to a *confident agent* whose psychology is so constituted that she believes that she will drink the toxin tomorrow irrespective of whether today she forms the intention to drink the toxin (e.g. Mele 1992). Today, the confident agent can rationally form the intention to drink the toxin tomorrow even if she knows that tomorrow drinking the toxin will fail to maximize her SP. For today this agent has strong reasons both to form the intention to drink the toxin tomorrow and to believe that she will drink the toxin. In this respect, a localist may well object that the confident agent is irrational on the alleged ground that a rational agent, knowing that drinking the toxin will fail to maximize her SP, would regard drinking the toxin as irrational and would believe that she will not drink the toxin. However, as noted in the previous paragraph, the mere fact that an action such as drinking the toxin fails to maximize an agent's SP does not per se indicate that such action is irrational unless one already presupposes a localist conception of instrumental rationality. Hence, reiterating that drinking the toxin fails to maximize an agent's SP does not per se provide any independent reasons to think that a rational agent would regard drinking the toxin as irrational and would believe that she will not drink the toxin. This point holds not merely for the aforementioned confident agent, but generalizes for several agents whose degree of belief that they will drink the toxin is lower than 1. That is to say, given an arbitrarily small probability $p > 0$ that an agent thinks she will drink the toxin, one can construct a variant of the toxin puzzle where the (expected or actual) payoffs the agent can get by forming and acting on the intention to drink the toxin are *large enough* to license the claim that forming such intention maximizes the agent's GP. To claim that the agent cannot rationally form such intention *irrespective* of how large these payoffs are amounts to a *reductio ad absurdum* of the localist solution (e.g. Sect. 5 below for similar remarks concerning large prize variants of the toxin puzzle).[17]

[17] A localist may further object that "our intentions are constrained by our reasons for action [as] our beliefs are constrained by our evidence" (Kavka 1983, 36; also Velleman 2007). I do not aim here to assess whether my claims about the rationality of forming anomalous intentions hold also for anomalous beliefs, i.e. beliefs that one will perform actions that she knows maximize her GP, but fail to maximize her SP. Still, to see how such assessment may proceed, consider the following *hard doxastic variant* of

## 5 Objection from psycho-physical inability

The objection from *psycho-physical inability* holds that real-life agents cannot rationally form anomalous intentions because they lack sufficient control over the formation and the dynamics of their own intentions to be able to reliably form anomalous intentions (e.g. Farrell 1989; also Shah 2009, on cases where one would benefit greatly from forming an intention, yet is unable to form it because she doubts that she will be able to act on it). The objection proceeds as follows. Consider an anomalous intention such as the intention to drink the toxin in the decision problem envisaged by Kavka. Suppose some real-life agent realizes that forming and acting on this anomalous intention is part of the course of action that maximizes her GP. By itself, this realization does not enable the agent to form such anomalous intention. For the agent knows that, when it comes to performing the relevant action (i.e. drinking the toxin), performing this action will fail to maximize her SP. And this knowledge, in turn, undermines the agent's ability to form the intention to perform such action. For real-life agents are psycho-physically unable to form intentions to perform actions that they know will fail to maximize their SP. Paraphrasing Farrell, an agent's intention to perform an action that she knows will fail to maximize her SP "would necessarily be unstable [since] reflection on the fact that [this intention] is directed towards an action which [fails to maximize SP] will inevitably undermine it" (1989, 288).

This objection correctly notes that the realization that performing a particular action fails to maximize one's SP may reduce (or even undermine) one's ability to form the intention to perform such action. Still, it is an open empirical question how often (and to what extent) this realization reduces real-life agents' ability to form anomalous intentions (e.g. Holton 2004). In fact, variations in the payoffs involved in specific variants of the toxin puzzle may significantly affect real-life agents' ability to form anomalous intentions in such variants. For instance, in transfer variants of the toxin puzzle, relatively few agents would presumably be able to form today the intention to give up the specified part of the million if this part amounted to 999,999$. Yet, comparatively more agents would be able to form this intention for lower amounts of money, and even more agents would be able to form such intention if they were required to give up just 1 dollar upon receiving the million. For as noted in Sects. 2–3, such dollar will be of extremely low value to agents upon

Footnote 17 continued
the toxin puzzle, where an agent gets the million if she forms the belief that she will perform an action that she knows she cannot possibly perform. The benefits derivable from forming this belief give the agent a reason to form the belief, but do not enable the agent to form the belief because the agent knows that such belief is false. Hence, it is dubious that the agent can rationally form such belief (e.g. Farrell 1989, 286–8, Gauthier 1998, 51). Conversely, consider the following *weak doxastic variant* of the toxin puzzle, where an agent gets the million if she forms the belief that she will perform an action whose likelihood is uncertain or indeterminate. One can envision many situations where the agent can rationally form such belief (e.g. Bovens 1995, on an agent who forms the belief that she can perform desirable actions because forming this belief significantly increases the chance that she will be able to perform such actions). For a discussion of the extent to which real-life agents are able to control the formation and the dynamics of their own beliefs, e.g. Paul 2015. For some experimental evidence suggesting that real-life agents can and sometimes do voluntarily believe, e.g. Turri et al. 2018.

receiving the million, and agents can gain 999,999$ by forming the intention to give up such dollar compared to a situation where they do not form such intention.

More generally, the point remains that real-life agents' purported inability to form specific anomalous intentions constitutes an *empirical* limitation in agents' ability to control their own intentions, but does not directly bear against the *normative* claim that real-life agents can rationally form anomalous intentions. To be sure, whether real-life agents can rationally form specific anomalous intentions may significantly depend on their ability to form such intentions (e.g. footnote no.20 on putative cases where agents cannot rationally form specific anomalous intentions because they know that it is psycho-physically impossible for them to form such intentions). Moreover, real-life agents are often unable to form the intention to perform an action for the sole reason that forming and acting on this intention maximizes their GP (e.g. Shah 2009). Still, these limitations concern the empirical issue whether real-life agents are psycho-physically able to form specific anomalous intentions, and do not directly bear against the normative claim that real-life agents can rationally form anomalous intentions.[18]

A proponent of the objection from psycho-physical inability may object that real-life agents' control over their own intentions is so limited that it is *rather unlikely* that these agents are able to form anomalous intentions for *the sole* reason that forming and acting on these intentions maximizes their GP (e.g. Farrell 1989). Suppose, for the sake of argument, that this objection is correct. Assume further that real-life agents are aware of this limitation on their ability to control their own intentions. Even this does not bear against the claim that real-life agents can rationally form anomalous intentions. To see this, consider situations where an agent knows that it is rather unlikely that she will be able to form an anomalous intention (e.g. the intention to drink the toxin). In some of these situations, the agent cannot rationally try to form this anomalous intention because the probability that she will be able to form such intention is overly low to license the claim that trying to form such intention maximizes her GP.[19] In other such situations, instead, the agent can rationally try to form this anomalous intention because the (expected or actual) payoffs she can get by forming and acting on the intention are large enough to license the claim that trying to form such intention maximizes her GP (e.g. McCann 1986). That is to say, given an arbitrarily small probability $p > 0$ that the agent is able to form a given anomalous intention (e.g. the intention to drink the toxin), one can construct a *large prize variant* of the toxin puzzle where the

[18] Real-life agents' ability to form anomalous intentions may depend not only on the extent to which such agents can control their own intentions, but also on the epistemic access they have to their own future intentions and actions (e.g. Joyce 2007; Levi 2007; Rabinowicz 2002; Spohn 2012, on the debate as to whether real-life agents can simultaneously deliberate about what actions they ought to perform and predict what actions they will perform). I gloss over this complication for the purpose of my evaluation.

[19] Here I speak of the rationality of 'trying to' form anomalous intentions (as opposed to the rationality of 'forming' these intentions) because the objection I address in this paragraph questions whether real-life agents are psycho-physically able to form such intentions. Also, I assume that trying to form anomalous intentions is cognitively (or otherwise) costly for real-life agents. If trying to form anomalous intentions involved no cost whatsoever for real-life agents, then the probability that these agents succeed in forming specific anomalous intentions would not directly bear on the issue whether those agents can rationally try to form anomalous intentions.

(expected or actual) payoffs the agent can get by forming and acting on the intention are *large enough* to license the claim that trying to form such intention maximizes the agent's GP. To claim that the agent cannot rationally try to form such intention *irrespective* of how large these payoffs are amounts to a *reductio ad absurdum* of the localist solution.[20]

## 6 Overdemandingness objection

The *overdemandingness* objection holds that real-life agents cannot rationally form anomalous intentions because calculating whether forming specific anomalous intentions maximizes their GP is cognitively too costly for them or simply beyond the epistemic access they have to earlier choice nodes of sequential decision problems (e.g. Mongin 2000; Steele 2006). The objection proceeds as follows. The mere fact that anomalous intentions are intentions to perform actions that maximize agents' GP does not per se imply that real-life agents can calculate whether forming specific anomalous intentions maximizes their GP in the sequential decision problems they face. In particular, calculating whether forming specific anomalous intentions enables one to maximize her GP in the sequential decision problems she faces would require one to "keep track of earlier [choice nodes of such problems] that are no longer possible" (Steele 2006, 9). The cognitive costs of keeping track of these choice nodes might be assumed away in idealized decision problems such as the toxin puzzle (Sect. 2). However, for real-life agents it is often cognitively too costly to keep track of earlier choice nodes of the sequential decision problems they face (e.g. Gilboa 2009, ch.11–12). In fact, real-life agents frequently lack the epistemic access to earlier choice nodes required to calculate whether forming specific anomalous intentions enables them to maximize their GP in the sequential decision problems they face (e.g. Mongin 2000). These limitations, in turn, make it dubious that real-life agents can rationally form anomalous intentions.[21]

This objection correctly notes that calculating whether forming specific anomalous intentions enables real-life agents to maximize their GP in the sequential decision problems they face often involves high cognitive costs for such agents. Still, it is doubtful that these cognitive costs are *generally* so high that they prevent

[20] One may object that agents cannot rationally form anomalous intentions in cases where they know that it is psycho-physically impossible for them to form such intentions (e.g. think of *miracle variants* of the toxin puzzle, where an agent gets the million if she forms the intention to perform some action that she knows she is nomologically unable to perform like travelling faster than light or turning water instantly into wine). I do not expand on this objection because my claim that agents can rationally form anomalous intentions in several decision problems (e.g. transfer variants and large prize variants of the toxin puzzle) is compatible with the claim that agents cannot rationally form anomalous intentions in cases where they know that it is psycho-physically impossible for them to form such intentions.

[21] Both GP maximization and SP maximization may be more or less cognitively demanding depending on what sequential decision problems are faced by real-life agents. The overdemandingness objection targets both highly extended sequential decision problems spanning the whole life of the involved agents (e.g. Savage 1954, 83, on idealized decision problems where the involved agent "has only one decision to make in his whole life […] namely, decide how to live") and less extended sequential decision problems having a clearly defined beginning and end. I take my reply to the overdemandingness objection to hold for both highly extended and less extended sequential decision problems.

real-life agents who perform such calculations from maximizing their GP. Moreover, one may consistently hold that in many sequential decision problems keeping track of earlier choice nodes is 'burdensome' (Steele 2006, 19), yet enables real-life agents to obtain higher (expected or actual) payoffs than the (expected or actual) payoffs they would obtain if they did not keep track of such choice nodes. In fact, one can envision several sequential decision problems where the benefits derivable from keeping track of earlier choice nodes outweigh the cognitive costs involved in keeping track of such choice nodes (e.g. Sect. 5 above on large prize variants of the toxin puzzle). In this respect, it would be of little import to object that in *several* sequential decision problems, it is cognitively too costly for real-life agents to calculate whether keeping track of earlier choice nodes enables them to maximize their GP. For substantiating the claim that real-life agents can rationally keep track of earlier choice nodes only requires one to show that there are *some* sequential decision problems where the benefits derivable from keeping track of these choice nodes outweigh the cognitive costs involved in keeping track of such choice nodes. That is to say, both GP maximization and SP maximization may be more or less cognitively demanding depending on what sequential decision problems are faced by real-life agents (e.g. footnote no.21). This variability makes it dubious that real-life agents are always able to maximize their (expected or actual) payoffs in sequential decision problems (e.g. think of highly extended sequential decision problems with more stages and nodes than the involved real-life agents are able to keep track of). However, it does not provide any reason to think that, in general, real-life agents cannot rationally form anomalous intentions (e.g. think of less extended sequential decision problems like Kavka's puzzle).

A proponent of the overdemandingness objection may concede that keeping track of earlier choice nodes of sequential decision problems frequently enables real-life agents to obtain higher (expected or actual) payoffs than the (expected or actual) payoffs they would obtain if they did not keep track of such choice nodes. Still, she may object that assessing the rationality of forming anomalous intentions requires agents to calculate the cognitive costs and benefits involved in forming such intentions, and that the rationality of performing these calculations itself depends on further calculations, leading to a *regress* (e.g. Mongin 2000, 95). Suppose, for the sake of argument, that assessing the rationality of forming anomalous intentions requires agents to calculate the cognitive costs and benefits involved in forming such intentions. This does not per se imply that the rationality of performing these calculations itself depends on further calculations. Moreover, even if the rationality of performing these calculations itself depended on further calculations, this would not per se entail a regress. For there are empirical limitations on how many orders of calculations real-life agents can perform (e.g. van Hees and Roy 2009, on agents' cognitive and computational limitations). These empirical limitations make it dubious that real-life agents are always able to maximize their (expected or actual) payoffs in sequential decision problems (e.g. Mongin 2000, 102). However, they do not provide any reason to think that, in general, real-life agents cannot rationally form anomalous intentions. To put it differently, regress-based considerations cast doubt on the claim that real-life agents can *always* rationally form anomalous

intentions, but do not bear against the thesis that real-life agents *can* rationally form anomalous intentions.

## 7 Conclusion

Over the last few decades, several prominent authors have built on the claim that agents cannot rationally form anomalous intentions to criticize RCT for being self-defeating. In this paper, I argued that despite these prominent criticisms of RCT, both ideal agents who can form and revise their intentions instantly without cognitive costs and real-life agents who have limited control over the formation and the dynamics of their own intentions can rationally form anomalous intentions. If my thesis is correct, prominent attempts to demonstrate that RCT is self-defeating do not withstand scrutiny. This result does not per se vindicate RCT as our best available normative theory of choice. Still, it challenges RCT's critics to put forward more convincing reasons and evidence to support their claim that RCT is self-defeating.

**Acknowledgements** I thank two anonymous referees, Luc Bovens, John Broome, Susanne Burri, Mikaël Cozic, Franz Dietrich, Ryan Doody, Wlodek Rabinowicz, Wolfgang Spohn and Johanna Thoma for their comments on previous versions of this paper. I also received helpful feedback from audiences at the University of Oxford, Tufts University, the University of Hamburg, the 3rd Meeting of the PPE Society (New Orleans), King's College London and the London School of Economics.



## References

Alvarez, M. (2018). Reasons for action, acting for reasons, and rationality. *Synthese, 195,* 3293–3310.
Andreou, C. (2006). Temptation and deliberation. *Philosophical Studies, 131,* 583–606.
Anscombe, G. (1963). *Intention* (2nd ed.). Cambridge, MA: Harvard University Press.
Arpaly, N. (2000). On acting rationally against one's better judgement. *Ethics, 110,* 488–513.
Audi, R. (1973). Intending. *Journal of Philosophy, 70,* 387–403.
Audi, R. (1986). Acting for reasons. *Philosophical Review, 95,* 511–546.
Audi, R. (1991). Intention, cognitive commitment, and planning. *Synthese, 86,* 361–378.
Bales, A. (2020). Intentions and instability: a defence of causal decision theory. *Philosophical Studies, 177,* 793–804.

Bovens, L. (1995). The intentional acquisition of mental states. *Philosophy and Phenomenological Research, 55,* 821–840.
Bradley, R. (2007). A unified Bayesian decision theory. *Theory and Decision, 63,* 233–263.
Bradley, R. (2017). *Decision theory with a human face*. Cambridge: Cambridge University Press.
Bratman, M. (1987). *Intention, plans and practical reason*. Cambridge, MA: Harvard University Press.
Bratman, M. (1998). Toxin, temptation, and the stability of intention. In J. Coleman, C. Morris, & G. Kavka (Eds.) *Rational commitment and social justice: essays for Gregory Kavka* (pp. 59–83). Cambridge University Press.
Bratman, M. (1999). *Faces of intention: selected essays on intention and agency*. Cambridge: Cambridge University Press.
Bratman, M. (2000). Reflection, planning, and temporally extended agency. *Philosophical Review, 109,* 35–69.
Bratman, M. (2009). Intention, practical rationality, and self-governance. *Ethics, 119,* 411–443.
Broome, J. (2001). Normative practical reasoning. *Proceedings of the Aristotelian Society, 85,* 175–193.
Broome, J. (2004). Reasons. In J. Wallace, P. Pettit, S. Scheffler, & M. Smith (Eds.), *Reason and value: themes from the moral philosophy of Joseph Raz* (pp. 28–55). New York: Oxford University Press.
Broome, J. (2007a). Wide or narrow scope? *Mind, 116,* 359–370.
Broome, J. (2007b). Does rationality consist in responding correctly to reasons? *Journal of Moral Philosophy, 4,* 349–374.
Broome, J. (2013). *Rationality through reasoning*. Wiley-Blackwell.
Cantwell, J. (2003). On the foundations of pragmatic arguments. *Journal of Philosophy*, *100*, 383–402.
Clarke, R. (2007). Commanding intentions and prize-winning decisions. *Philosophical Studies, 133,* 391–409.
Clarke, R. (2008). Autonomous reasons for intending. *Australasian Journal of Philosophy, 86,* 191–212.
Cozic, M., & Hill, B. (2015). Representation theorems and the semantics of decision-theoretic concepts. *Journal of Economic Methodology, 22,* 292–311.
Cubitt, R. (1996). Rational dynamic choice and expected utility theory. *Oxford Economic Papers, 48,* 1–19.
Cubitt, R., & Sugden, R. (2001). On money pumps. *Games and Economic Behavior, 37,* 121–160.
Cullity, G. (2008). Decisions, reasons and rationality. *Ethics, 119,* 57–95.
Davidson, D. (1978). Intending. *Philosophy of History and Action, 11,* 41–60.
DeHelian, L., & McClennen, E. (1993). Planning and the stability of intention. *Minds and Machines, 3,* 319–333.
Den Hartogh, G. (2004). The authority of intention. *Ethics, 115,* 6–34.
Dietrich, F., & List, C. (2013). A reason-based theory of rational choice. *Nous, 47,* 104–134.
Dietrich, F., & List, C. (2016). Mentalism versus behaviourism in economics: a philosophy-of-science perspective. *Economics and Philosophy, 32,* 249–281.
Dietrich, F., Staras, A., & Sugden, R. (2019). A Broomean model of rationality and reasoning. *Journal of Philosophy, 116,* 585–614.
Farrell, D. (1989). Intention, reason, and action. *American Philosophical Quarterly, 26,* 283–295.
Ferrero, L. (2010). Decisions, diachronic autonomy, and the division of deliberative labor. *Philosophers' Imprint, 10,* 1–23.
Fumagalli, R. (2013). The futile search for true utility. *Economics and Philosophy, 29,* 325–347.
Fumagalli, R. (2019). (F)utility exposed. *Philosophy of Science, 86,* 955–966.
Fumagalli, R. (2020a). On the individuation of choice options. *Philosophy of the Social Sciences*, *50*, 338–365.
Fumagalli, R. (2020b). How thin rational choice theory explains choices. *Studies in History and Philosophy of Science*, *83*, 63–74.
Gauthier, D. (1984). Deterrence, maximization and rationality. *Ethics, 94,* 474–495.
Gauthier, D. (1994). Assure and threaten. *Ethics, 104,* 690–721.
Gauthier, D. (1997). Resolute choice and rational deliberation: a critique and a defense. *Nous, 31,* 1–25.
Gauthier, D. (1998). Rethinking the toxin puzzle. In J. Coleman, C. Morris, & G. Kavka (Eds.), *Rational commitment and social justice: essays for Gregory Kavka* (pp. 47–58). Cambridge University Press.
Gilboa, I. (2009). *Theory of decision under uncertainty*. Cambridge: Cambridge University Press.
Goetz, S. (1998). Reasons for forming an intention: a reply to Pink. *Mind, 107,* 205–213.
Guala, F. (2019). Preferences: neither behavioural nor mental. *Economics and Philosophy, 35,* 383–401.
Hammond, P. (1988). Consequentalist foundations for expected utility. *Theory and Decision, 25,* 25–78.

Harman, G. (1986). Willing and intending. In R. Grandy & R. Warner (Eds.), *Philosophical grounds of rationality* (pp. 363–380). Oxford: Oxford University Press.
Harman, G. (1998). The toxin puzzle. In J. Coleman, C. Morris, & G. Kavka (Eds.), *Rational commitment and social justice: essays for Gregory Kavka* (pp. 84–89). Cambridge University Press.
Hausman, D. (2000). Revealed preference, belief, and game theory. *Economics and Philosophy, 16,* 99–115.
Hausman, D. (2011). Mistakes about preferences in the social sciences. *Philosophy of the Social Sciences, 41,* 3–25.
Hausman, D. (2012). *Preference, value, choice, and welfare*. Cambridge: Cambridge University Press.
Hedden, B. (2015). Options and diachronic tragedy. *Philosophy and Phenomenological Research, 40,* 423–451.
Holton, R. (2004). Rational resolve. *Philosophical Review, 113,* 507–535.
Holton, R. (2008). Partial belief, partial intention. *Mind, 117,* 27–58.
Holton, R. (2009). *Willing, wanting, waiting*. Oxford: Oxford University Press.
Jeffrey, R. (1965). *The logic of decision*. Chicago: University of Chicago Press.
Joyce, J. (2007). Are Newcomb problems really decisions? *Synthese, 156,* 537–562.
Kavka, G. (1978). Some paradoxes of deterrence. *Journal of Philosophy, 75,* 285–302.
Kavka, G. (1983). The toxin puzzle. *Analysis, 43,* 33–36.
Kolodny, N. (2005). Why be rational? *Mind, 114,* 509–563.
Kolodny, N. (2007). State or process requirements? *Mind, 116,* 371–385.
Levi, I. (1987). The demons of decision. *The Monist, 70,* 193–211.
Levi, I. (1991). Consequentialism and sequential choice. In S. Hurley & M. Bacharach (Eds.), *Foundations of decision theory: issues and advances*. Oxford: Blackwell.
Levi, I. (2007). Deliberation does crowd out prediction. In: T. Rønnow-Rasmussen, B. Petersson, J. Josefsson & D. Egonssson (Eds.), *Homage à Wlodek. Philosophical Papers Dedicated to Wlodek Rabinowicz*. Lund University.
Littlejohn, C. (2016). Do reasons and evidence share the same residence? *Philosophy and Phenomenological Research, 93,* 720–727.
Littlejohn, C. (2018). Stop making sense? On a puzzle about epistemic rationality. *Philosophy and Phenomenological Research, 96,* 257–272.
Machina, M. (1989). Dynamic consistency and non-expected utility models of choice under uncertainty. *Journal of Economic Literature, 27,* 1622–1668.
McCann, H. (1986). Rationality and the range of intention. *Midwest Studies in Philosophy, 10,* 191–211.
McCann, H. (1991). Settled objectives and rational constraints. *American Philosophical Quarterly, 28,* 25–36.
McClennen, E. (1985). Prisoner's dilemma and resolute choice. In R. Campbell, & L. Sowden (Eds.), *Paradoxes of rationality and cooperation*. Vancouver: University of British Columbia Press.
McClennen, E. (1990). *Rationality and dynamic choice*. Cambridge: Cambridge University Press.
McClennen, E. (1997). Pragmatic rationality and rules. *Philosophy & Public Affairs, 26,* 210–258.
Meacham, C. (2010). Binding and its consequences. *Philosophical Studies*, *149*, 49–71.
Mele, A. (1992). Intentions, reasons, and beliefs: morals of the toxin puzzle. *Philosophical Studies, 68,* 171–194.
Mele, A. (1995). Effective deliberation about what to intend: or striking it rich in a toxin-free environment. *Philosophical Studies, 79,* 85–93.
Mele, A. (2000). Deciding to act. *Philosophical Studies, 100,* 81–108.
Mintoff, J. (1997). Rational cooperation, intention and reconsideration. *Ethics, 107,* 612–643.
Mongin, P. (2000). Does optimization imply rationality? *Synthese, 124,* 73–111.
Okasha, S. (2007). Rational choice, risk aversion, and evolution. *Journal of Philosophy, 104,* 217–235.
Parfit, D. (1984). *Reasons and persons*. Oxford: Oxford University Press.
Parfit, D. (1997). Reasons and motivation. *Proceedings of the Aristotelian Society, Supplementary Volume, 71,* 99–129.
Parfit, D. (2001). Rationality and reasons. In D. Egonsson, J. Josefsson, B. Petersson, & T. Rønnow-Rasmussen (Eds.), *Exploring practical philosophy: from action to values* (pp. 19–39). Ashgate.
Paul, S. (2015). Doxastic self-control. *American Philosophical Quarterly, 52,* 145–158.
Pink, T. (1991). Purposive intending. *Mind, 100,* 343–359.
Pink, T. (1998). Reply to Goetz. *Mind, 107,* 215–218.
Quinn, W. (1985). The right to threaten and the right to punish. *Philosophy & Public Affairs, 14,* 327–373.

Rabinowicz, W. (1995). To have one's cake and eat it, too: sequential choice and expected utility violations. *Journal of Philosophy, 92,* 586–620.
Rabinowicz, W. (1997). On Seidenfeld's criticism of sophisticated violations of the independence axiom. *Theory and Decision*, *43*, 279–292.
Rabinowicz, W. (2002). Does practical deliberation crowd out self-prediction? *Erkenntnis, 57,* 91–122.
Rabinowicz, W. (2014). Safeguards of a disunified mind. *Inquiry, 57,* 356–383.
Rabinowicz, W. (2019). Between sophistication and resolution - wise choice. In R. Chang, K. Sylvan (Eds.), *The Routledge Handbook of practical reason*. London: Routledge.
Roy, O. (2009). Intentions and interactive transformations of decision problems. *Synthese, 169,* 335–349.
Savage, L. (1954). *The foundations of statistics*. New York: Dover Publications Inc.
Schulz, A. (2008). Risky business: evolutionary theory and human attitudes toward risk - a reply to Okasha. *Journal of Philosophy, 105,* 156–165.
Searle, J. (1983). *Intentionality*. Cambridge: Cambridge University Press.
Shah, N. (2009). How action governs intention. *Philosophers' Imprint, 8,* 1–19.
Smith, M. (2016). One dogma of philosophy of action. *Philosophical Studies, 173,* 2249–2266.
Sobel, H. (1994). Useful Intentions. In H. Sobel (Ed.), *Taking chances* (pp. 237–254). New York: Cambridge University Press.
Spohn, W. (2009). Why the received models of considering preference change must fail. In T. Grüne-Yanoff & S. Hansson (Eds.), *Preference change: approaches from philosophy, economics and psychology*. Dordrecht: Springer.
Spohn, W. (2012). Reversing 30 years of discussion: why causal decision theorists should one-box. *Synthese, 187,* 95–122.
Steele, K. (2006). What can we rationally value? Mimeo: University of Sydney.
Steele, K. (2010). What are the minimal requirements of rational choice? Arguments from the sequential-decision setting. *Theory and Decision, 68,* 463–487.
Sugden, R. (1991). Rational choice: a survey of contributions from economics and philosophy. *The Economic Journal*, *101*, 751–785.
Tenenbaum, S. (2018). Reconsidering intentions. *Nous, 52,* 443–472.
Thoma, J. (2017). Temptation and preference-based instrumental rationality. In J. Bermudez (Ed.), *Self-control, decision theory, and rationality*. Cambridge: Cambridge University Press.
Thoma, J. (2020). Instrumental rationality without separability. *Erkenntnis*, *85*, 1219–1240.
Turri, J., Rose, D., & Buckwalter, W. (2018). Choosing and refusing: doxastic voluntarism and folk psychology. *Philosophical Studies, 175,* 2507–2537.
Van Hees, M., & Roy, O. (2009). Intentions, decisions and rationality. In T. Boylan & R. Gekker (Eds.), *Economics, rational choice and normative philosophy* (pp. 56–72). London: Routledge.
Velleman, D. (1989). *Practical reflection*. Princeton: Princeton University Press.
Velleman, D. (2007). What good is a will? In A. Leist & H. Baumann (Eds.), *Action in context* (pp. 193–215). Berlin: de Gruyter.
Williams, B. (1981). *Moral luck*. Cambridge: Cambridge University Press.