Contingent Thinking and the Sure-Thing Principle: Revisiting Classic Anomalies in the Laboratory*

Ignacio Esponda (UCSB)

Emanuel Vespa (UCSB)

August 30, 2019

Abstract

Failure of contingent thinking has been increasingly offered as a possible explanation for behavior that deviates from standard rationality postulates. The concept of contingent thinking, however, remains a fairly elusive one. We study a specific aspect of contingent thinking that can be formally defined and tested in the laboratory, and show that it potentially underlies many of the classic anomalies in decision and game theory. We perform an experiment and find that failure of this type of contingent thinking can in large part explain some of the most common anomalies found in the laboratory, including overbidding in auctions, naive voting in elections, and both Ellsberg and Allais types of paradoxes.

*We thank Eduardo Azevedo, Ted Bergstrom, Gary Charness, Benjamin Enke, Larry Epstein, Erik Eyster, Drew Fudenberg, Xavier Gabaix, Itzhak Gilboa, Daniel Gottlieb, Yoram Halevy, Paul J. Healy, Shachar Kariv, Shengwu Li, Bart Lipman, Mark Machina, Paulo Natenzon, Muriel Niederle, Ryan Oprea, Pietro Ortoleva, Matthew Rabin, Collin Raymond, Ariel Rubinstein, Marciano Siniscalchi, Joel Sobel, Charlie Sprenger, Tomasz Strzalecki, Severine Toussaert, Georg Weizsäcker, Leeat Yariv, and Sevgi Yuksel for helpful comments. Esponda: Department of Economics, University of California at Santa Barbara, 2127 North Hall, University of California Santa Barbara, CA 93106, iesponda@ucsb.edu.; Vespa: Department of Economics, University of California at Santa Barbara, 2127 North Hall, University of California Santa Barbara, CA 93106, vespa@ucsb.edu.

Contents

1 Introduction

In many economic environments people need to anticipate the consequences of their actions under different contingencies before making a choice. Rational choice theory delivers predictions that rely on people considering such contingencies, but many experiments have documented behavior inconsistent with the predictions (for a survey see Eyster, 2019). While several explanations have been offered to rationalize deviations from standard theory, a recent literature explores to what extent discrepancies can be related to failures in contingent reasoning.[1]

The literature on contingent thinking, however, has made progress without an explicit definition of contingent reasoning. In this paper, we study a specific aspect of contingent thinking that can be formally defined and tested in the laboratory. An advantage of having an explicit definition of a particular type of contingent thinking is that it allows us to connect environments that have been studied in the contingent-thinking literature (e.g., auctions) to environments that have not (e.g., the Ellsberg paradox).

The type of contingent thinking that we study was informally introduced by Savage (1972) as the sure-thing principle (STP).[2] To illustrate the idea, consider a typical common-value election where, if a voter were pivotal, then she would maximize her payoff by voting against her private information. In the event she is not pivotal, her vote is irrelevant and so she would be indifferent. The principle of STP says that in this situation the voter would also vote against her private information even if she knew nothing about her pivotal status. Voting in line with her private information, however, would be a violation of STP. It also indicates a failure of contingent reasoning since she is not partitioning the events between states where her choice does matter and states where it does not.

Most of the literature on contingent thinking has focused on violations of *dominance*, where subjects make suboptimal choices and leave money on the table. Violations of STP can explain why dominance fails. In the common-value election discussed above, a voter who makes the right decision and votes against her private information when she knows her vote is pivotal is satisfying dominance. A failure to vote against her private information when she does not know her pivotal

[1] For recent experiments with a focus on contingent reasoning, see Esponda and Vespa (2014), Louis (2015), Dal Bó, Dal Bó, and Eyster (2016), Li (2017), Martínez-Marquina, Niederle, and Vespa (forthcoming), Ngangoue and Weizsäcker (2018), Bayona, Brandts, and Vives (2019), Martin and Munoz-Rodriguez (2019), and Moser (2019). Failures in contingent reasoning may create difficulties with understanding correlations (Eyster and Weizsäcker (2010; 2016), Enke and Zimmermann (2019) and Rees-Jones, Shorrer, and Tergiman (2019)) and/or selection effects (Esponda and Vespa (2018), Araujo, Wang, and Wilson (2019), Barron, Huck, and Jehiel (2019) and Enke (2019)).

[2] Savage did not formally introduce STP as a separate postulate, but only referred to it informally to motivate one of his main postulates. It is straightforward, however, to use his framework to define STP, as we show in Online Appendix A.

status is not a reflection of her desire to earn less money but rather an indication of a failure of contingent thinking, as embodied by STP.

An advantage of using the principle of STP to bring more discipline to the study of contingent thinking is that we can investigate the extent to which difficulty with contingent thinking is responsible for the failure of postulates other than dominance. In particular, we can study contingent thinking in settings beyond those that have been considered in the literature. For example, consider the classic one-urn Ellsberg problem, with three possible states of the world. The problem consists of two questions. For each question, the state is the color of a ball randomly selected from the urn and the subject needs to select one of two possible options without knowing the state. Crucially, the monetary payoffs in one state (say, for the blue ball) is the same for both options, so that states can be partitioned in two sets just as in our previous example. The two questions only differ in terms of the payoff the subject receives if the blue state results. For the remaining states, the two questions are identical and there is no dominant option. Typically, people revert their choices between these questions, which is a violation of a postulate that is central to expected utility theory and which we will call *separability*.[3] We will show that a large part of these reversals is due to a failure of contingent thinking as embodied by the violation of STP.

We study five problems in the laboratory, three of which correspond to previously documented failures of dominance (a second-price private-values auction, a common-value election, and the common-ratio Allais paradox) and two of which correspond to previously documented failures of separability (the Ellsberg paradox and the common-consequence Allais paradox). Following previous literature, we convert the last two game-theoretic problems into decision problems, thus abstracting from the possibility of incorrect beliefs about others' behavior.[4]

In each of these five problems, the states of the world can be partitioned into two non-empty sets: states where the subject's choice matters and states where it does not. For each problem, we elicit subjects' choices under two frames. The first frame (*noncontingent*) presents each problem in the standard way it appears in the literature, so that we start by replicating earlier experimental findings. The second frame (*contingent*) helps subjects focus on the set of states where her choice matters. For instance, in the case of the common-value election problem described above, we

[3] What we call separability in this paper is Savage's postulate P2.

[4] Previous experimental literature has found that game-theoretic anomalies persist when the problem is stripped away of strategic elements (guessing others' strategies), and it is now common to convert a game into a decision problem. This approach dates back to Roth and Murnighan (1978) and Walker, Smith, and Cox (1987), and recent work includes Neugebauer and Selten (2006), Charness and Levin (2009), Ivanov, Levin, and Niederle (2010), Esponda and Vespa (2014), and Agranov, Caplin, and Tergiman (2015). One advantage of this approach is that it allows one to test the extent to which deviations from equilibrium can be attributed to errors in optimization (or nonstandard preferences), as opposed to strategic features of the game such as incorrect beliefs about the opponents' behavior.

would describe the problem just as in the noncontingent frame but add that the subject's vote would only be used if it is needed to break a tie. The contingent version is simply a re-framing of the noncontingent version because when making her decision the subject does not know if her vote will be pivotal or not. Thus, if the subject were partitioning the states into the two sets when facing the noncontingent frame, we should observe the same choice in both frames.

We first successfully replicate standard findings in all the problems that we study, which represent anomalies in the sense that a standard postulate (dominance or separability) fails. We then document that a large part of the anomalies are driven by the failure of STP. In particular, we report a relatively large rate of violations of STP in *all* problems, ranging from 20% to 70%. Additionally, we find heterogeneity in subjects' response to the treatment. While most affected subjects go from being inconsistent to being consistent with the standard postulates of dominance and separability, several subjects are converted in the opposite direction, particularly for separability. The net effect of the treatment is that failures of dominance and separability drop by half in all problems, except in common-consequence Allais, where conversions in the opposite direction essentially cancel out.

We also find that focusing exclusively on the noncontingent version, as the previous literature has done, can obscure the connection between problems. For example, the correlation between consistency with separability in Ellsberg and consistency with separability in common-consequence Allais is negative in the *noncontingent* version, which would suggest, a bit surprisingly, that the same subject wants to satisfy separability in one problem but not in the other. We find, however, that this correlation is indeed positive in the contingent version. More generally, we find that the correlations in consistency across the problems are aligned with the two postulates, dominance and separability, for the case of the contingent versions of the problems, but not for the noncontingent versions.

Our results have several implications. The first implication is that, while anomalies in decision and strategic problems are typically studied separately, there is common ground between them. A failure of STP does not identify the mechanism behind anomalies in any specific problem that we study. In fact, the mechanisms examined in the literature (e.g., the thrill of winning, ambiguity and regret aversion, etc.) can be thought of as some of the factors that prevent people from engaging in contingent thinking and cause STP to fail in these problems. The reason why these problems have not been connected before is likely to be that the mechanisms are all very different. What we find in this paper is that regardless of the underlying mechanism, the consequence can be captured as a failure of STP. In other words, despite the fact that strategic and decision problems are usually rationalized with different psychological mechanisms, there is indeed common ground between these seemingly dissimilar settings, in the form of failure of a particular form of contingent

thinking.

Second, the finding that failure of STP underlies many of the classic anomalies suggests that models that allow STP to fail may better rationalize data across an array of environments. In this sense, our findings provide some guidance for future theoretical research and suggest paying more attention to theories of decision-making that do not rely on contingent thinking or more generally capture limited mental models, such as case-based decision theory (Gilboa and Schmeidler, 1995), the notion of obviously dominant strategies in games (Li, 2017), or models of local thinkers who focus on salient aspects of the decision problem (Gennaioli and Shleifer, 2010, Bordalo, Gennaioli, and Shleifer, 2012 and Gabaix, 2014).

A third implication is that our subjects are not good at thinking through the state space in the way modelers often assume, and there is indeed a very specific way in which this phenomenon can be formalized and tested. Our evidence corroborates the idea that incomplete preferences or anomalies may precisely stem from the fact that states are not naturally given, may be hard to construct, or some states may not be salient.[5] Moreover, our results caution us to draw inferences about people's preferences, particularly when people are not good at running through the state space.

A final implication speaks to welfare analysis. As the literature points out, it is reasonable that, for example, overbidding in auctions can be attributed to a preference for winning and that the Ellsberg paradox can be attributed to ambiguity aversion. From a welfare perspective, however, the question is the extent to which such behavior constitutes preferences or mistakes. While this is a difficult question to answer, the finding that overbidding and Ellsberg-type anomalies decrease by one-half in the contingent treatment suggests that difficulty with contingent thinking also plays an important role. On the other hand, there are also several subjects who, when facing a more transparent description of the state space, exhibit a preference to violate separability. Finally, the finding that anomalies decrease in the contingent treatment suggests that there is room for interventions that help people think hypothetically and be more consistent with postulates with normative appeal, such as dominance.

We conclude this introduction by discussing the related literature. Then, in Section 2, we use two examples to illustrate our experimental design. We describe the experiment in Section 3, and present the experimental results in Section 4. We conclude in Section 5.

[5]For more on these ideas, see, e.g., Gilboa, Postlewaite, and Schmeidler (2009), Bordalo, Gennaioli and Shleifer (2012, 2013).

Related literature

Our paper is related to several strands of the literature. Psychologists define hypothetical or contingent thinking as a form of "what–if" thinking that entails reasoning about events without knowing whether or not these events are true or will occur. A large literature in psychology finds that people have difficulty with various forms of hypothetical thinking.[6] The importance of contingent thinking has also been recognized in several economic environments. Early contributions include Shafir and Tversky (1992) and Croson (1999) in the context of the prisoner's dilemma and Charness and Levin (2009) in environments with adverse selection. A recent literature has expanded the scope of these ideas to several environments.

In an earlier paper (Esponda and Vespa, 2014), we distinguished between information extraction and contingent thinking in a voting context, but did not formalize the notion of contingent thinking. Moreover, that paper was closer to the literature on dynamic choice; its main goal was to compare behavior in the noncontingent treatment with behavior in a *sequential* treatment where, unlike the experiments in this paper, the subject was informed about the relevant event before making a decision.[7]

Li (2017) formally introduces the notion of an obviously dominant strategy as a strategy that a cognitively challenged player who cannot perform a certain kind of contingent thinking can recognize as dominant. He shows in an experiment that subjects play obviously dominant strategies at much higher rates than non-obviously dominant ones.[8] Zhang and Levin (2017) extend these results to allow for more general partitions of the state space.

An advantage of using STP to formalize a specific form of contingent thinking is that it allows us to link several seemingly unrelated anomalies with a common concept, and prescribes a very natural experimental test of that concept. A disadvantage is that it obviously does not necessarily encompass all possible forms of contingent thinking, such as difficulty in anticipating the equilibrium effects of new policies (Dal Bó, Dal Bó, and Eyster, 2016), in extracting information from market prices (Ngangoue and Weizsäcker, 2018, Bayona, Brandts, and Vives, 2019), in interpreting belief elicitation mechanisms (Martin and Munoz-Rodriguez, 2019), and in evaluating uncertain (vs. deterministic) events (Martínez-Marquina, Niederle, and Vespa, forthcoming).

Our paper is also connected to the literature on framing effects. The potential difference between the noncontingent and contingent treatments has been previously examined in the specific

[6] The early evidence dates back to Wason selection tasks (Wason 1966, 1968). See Evans (2007) and Nickerson (2015) for recent textbook treatments.

[7] For similar applications in other environments, see Louis (2015) and Moser (2019).

[8] Also see Glazer and Rubinstein (1996), who illustrate how the extensive form can facilitate contingent thinking relative to the normal form of a game.

context of the common-ratio Allais paradox. Tversky and Kahneman (1981) and Holler (1983) find significant differences between treatments, but Cubitt, Starmer, and Sugden (1998) find no significant differences in the marginal responses across treatments, like we do for this specific problem. Their focus, however, is different from ours and so they do not test if there is a reduction of reversals when both questions of the Allais paradox are asked using the standard (i.e., noncontingent) treatment (we find there is). To our knowledge, however, there has been no formalization of the form of hypothetical thinking that underlies this phenomenon nor any experimental work finding a common source linking mistakes in strategic environments such as auctions and elections with the classic anomalies in decision problems.[9] Moreover, while there is a large literature documenting the importance of framing effects (e.g., Tversky and Kahneman, 1986), we find a specific frame that has a systematic effect across a wide range of problems that seemed previously unrelated.

Our results are in line with the work of Halevy (2007), who in studying factors underlying the Ellsberg paradox found that there is a positive correlation between ambiguity neutrality and the ability to reduce compound lotteries. Our work supplements this earlier work in two directions. First, we find that the Ellsberg paradox is significantly mitigated once we help subjects to think contingently. Second, we find that contingent thinking fails systematically for a wide range of problems beyond the Ellsberg paradox.

There is a large experimental literature for each of the anomalies that we focus on. See, for example, Bazerman and Samuelson (1983), the survey by Kagel and Levin (2002), Charness and Levin (2009), Ivanov, Levin, and Niederle (2010), Esponda and Vespa (2014), and Levin, Peck, and Ivanov (2016) for mistakes in common value settings, such as auctions or elections, Kagel, Harstad, and Levin (1987), Kagel and Levin (1993), and Harstad (2000) for overbidding in second-price private value auctions; and MacCrimmon and Larsson (1979), the survey by Camerer (1995), Wakker (2001), Halevy (2007), Ahn, Choi, Gale, and Kariv (2014), Andreoni, Schmidt, and Sprenger (2014), Dean and Ortoleva (2015), and Kovářík, Levin, and Wang (2016) for a limited sample of the very large literature on the Ellsberg (1961) and Allais (1953) paradoxes. This experimental literature has also motivated a large theoretical literature for modeling the behavior of agents who make mistakes or have richer types of preferences.[10]

[9] In a theory paper, Eliaz, Ray, and Razin (2006) make a connection between anomalies in decision and game theory environments by establishing a formal equivalence between violations of expected utility theory and choice shifts in groups.

[10] For theoretical responses to mistakes, see Eyster and Rabin (2005), Crawford and Iriberri (2007), Jehiel and Koessler (2008), and Esponda (2008). For theoretical responses to the paradoxes, see the surveys by Machina (2008), Gilboa and Marinacci (2011) and Machina and Siniscalchi (2013). For a critical assessment of the ambiguity aversion literature, see Al-Najjar and Weinstein (2009) and Siniscalchi (2009).

		$\overbrace{\quad\quad A \quad\quad}$		$\overbrace{\quad A^c \quad}$
		red	$yellow$	$blue$
Question 1	f	\$10	\$0	\$10
	g	\$0	\$10	\$10
Question 2	f'	\$10	\$0	\$0
	g'	\$0	\$10	\$0

Figure 1: Example - Ellsberg problem

2 Illustrative examples

We illustrate the main experimental design using two examples, a second-price auction and the Ellsberg paradox.

ILLUSTRATIVE EXAMPLE 1: ELLSBERG. Consider Ellsberg's (1961) one-urn problem, where there is an urn with 90 balls, 30 of which are red and 60 of which are yellow or blue. The problem is depicted in Figure 1, where there are 3 states of the world, red, $yellow$, and $blue$. A decision maker faces choices between four options, f, g, f', and g', where each option maps the states of the world into monetary rewards.

In the standard experiment in the literature, which corresponds to what we call the *noncontingent treatment* in this paper, a subject is asked two questions. For each question, a ball is drawn from the urn with replacement, and the subject must make a choice before knowing the color of the drawn ball. In Question 1, the subject must choose between two options, f (win if red or blue) and g (win if yellow or blue). In Question 2, the subject is asked to choose between two other options, f' (win if red) and g' (win if yellow). A typical response pattern is to prefer g in the first question and f' in the second question. These choices are consistent with an aversion to ambiguity, because in each case the subject chooses the alternative with the known probability of success (2/3 in the first question and 1/3 in the second question). As is well known, these choices constitute a reversal and, therefore, violate one of Savage's central postulates of expected utility theory, which he called P2 and we refer to in this paper as separability and abbreviate by SEP.

Consider next a different treatment, which we call the *contingent treatment*, where we ask the same questions in a slightly different way. For the first question, we tell a subject that, if the ball is $blue$, then she will get \$10. And, if the ball is red or $yellow$, then she has a choice between the option that pays \$10 if the ball is red (f) and the option that pays \$10 if it is yellow (g). Crucially, we ask the subject to commit to a choice in case event $A = \{red, yellow\}$ is realized *before* she knows whether or not the state will be in A. Thus, the subject faces the same choice between f and g that is faced by a subject in the first question of the noncontingent treatment. For the second

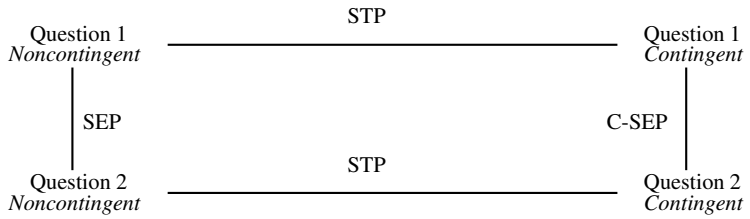

Figure 2: Experimental design and tests of STP, SEP, and C-SEP.

Notes: STP fails if there are differences between Question 1 noncontingent and Question 1 contingent, or differences between Question 2 noncontingent and Question 2 contingent. SEP fails if there are differences between Question 1 noncontingent and Question 2 noncontingent. C-SEP fails if there are differences between Question 1 contingent and Question 2 contingent.

question, we proceed in the same way except that we tell the subject that she will get \$0 if the ball is $blue$. Thus, the subject faces the same choice between f' and g' that is faced by a subject in the second question of the noncontingent treatment. (In particular, the subject also faces a static choice problem in the contingent treatment; for a discussion of a *dynamic* choice problem where the subject is asked to move *after* receiving information about an event, see Section 3.2.)

Because the two versions of the problem ask subjects to choose between the same options (in terms of having the same monetary payoffs for each possible state of the world), a subject's choice between f and g or between f' and g' should not depend on whether she faces the noncontingent or contingent version of the question, and the two treatments are simply different frames to describe the same options. But, in practice, one can imagine that the contingent treatment helps the subject to focus on the event A where the consequences differ.

There are two underlying principles behind the failure of separability (SEP). The first is a contingent version of SEP. For example, if the subject continues to choose g and f' in the contingent treatment, then we say that a contingent version of separability, which we call C-SEP, is violated.

The second principle, and the main focus of the paper, is what we refer to as the sure-thing principle. To see this principle intuitively, consider Question 1 and suppose that we ask the subject to commit to an option contingent on event A and to commit to an option contingent on the complementary event A^c, before knowing which event is realized. If the subject prefers f over g contingent on A and is indifferent between f and g contingent on A^c, then she should prefer f over g in the problem where the choice is not contingent. In the experiment, f and g give exactly the same monetary payoffs in A^c. To focus on the issue of contingent thinking and to avoid testing for indifference between two options that deliver the same consequences under A^c, we test a version of the sure-thing principle where we elicit choice contingent on A and the subject is told exactly what payoff she would get if a state in A^c is realized, a payoff that does not depend on her choice. Thus, we test a version of the sure-thing principle, which we abbreviate by STP, that says that if f is preferred to g contingent on A and f is *equal* to g outside of A, then f should be preferred

to g. In other words, if f is preferred to g in the contingent treatment, then f should be preferred to g in the noncontingent treatment.[11] The same is true regarding Question 2 between f' and g'. By comparing questions across treatments, we can assess the extent to which STP fails in our setting. In this example, an agent would violate STP if she were to prefer different alternatives across treatments for either Question 1 or 2. Figure 2 illustrates the connection between the questions and postulates for this example.

ILLUSTRATIVE EXAMPLE 2: SECOND-PRICE AUCTION. Consider a decision maker who participates in an auction where the highest bid wins and pays the bid of the second-highest bidder. There is one other bidder in this auction who is known to bid \$0.50, \$4.50, or \$8.50 with equal probability. The decision maker must choose an integer bid from \$1 to \$8. The decision maker gets \$5.50 if she wins the auction and an outside value of \$3 if she does not.[12]

We illustrate this problem in Figure 3. There are three states of the world, corresponding to the three possible bids of the competitor, and the auction problem is represented by Question 1, where the decision maker chooses one of two options. Option f' corresponds to a bid of \$1, \$2, \$3, or \$4 while option g' corresponds to a bid of \$5, \$6, \$7, or \$8. We also include a Question* that contrasts two options, f and g, that deliver constant payoffs of \$3 and \$1, respectively. We will not ask Question* in the experiment but will rather simply assume that more money is preferred to less, i.e., f is preferred to g.

		$\underbrace{\quad A \quad}$	$\underbrace{\quad A^c \quad}$	
		4.5	0.5	8.5
Question*	f	\$3	\$3	\$3
	g	\$1	\$1	\$1
Question 1	f'	\$3	\$5	\$3
	g'	\$1	\$5	\$3

Figure 3: Auction problem

In the auction, the choice of an integer bid from \$1 to \$8 is only relevant in the event that the other bidder bids \$4.50. In this case, however, it is optimal to lose the auction by submitting a bid of \$1, \$2, \$3 or \$4, which is represented by f'. The paradox in this problem is that, in a treatment where subjects in the role of the decision maker are asked to submit an integer bid from \$1 to

[11] Note also that, as mentioned earlier, the version of STP that we test is one that essentially amounts to a reframing of the problem, since the agent faces the same options in both treatments, in the sense that the options deliver the same state-contingent payoffs. Thus, one could argue on normative grounds that subjects should want to satisfy STP, and that a failure to satisfy it suggests a cognitive limitation.

[12] While it is standard to normalize the outside to be zero, here we illustrate the ideas using the payoffs from the actual experiment, where an outside value of \$3 guarantees that subjects do not incur losses.

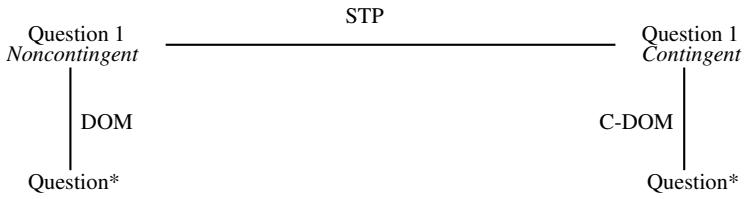

Figure 4: Experimental design and tests of STP, DOM, and C-DOM.

Notes: STP fails if there are differences between Question 1 noncontingent and Question 1 contingent, or differences between Question 2 noncontingent and Question 2 contingent. DOM fails if there are differences between Question 1 noncontingent and Question* noncontingent. C-DOM fails if there are differences between Question 1 contingent and Question* contingent. In AUCT and ELECT, Question* was not presented to subjects as we directly assumed that subjects prefer more money to less. In CR ALLAIS, Question* corresponds to Question 2 in Figure 7.

\$8, a significant fraction of subjects end up overbidding and choose g' over f'. Note that, under the assumption that f is preferred to g in Question*, which is the assumption that more money is preferred to less money, then this choice constitutes a reversal: f is preferred to g but g' is preferred to f'. This reversal is a violation of dominance (henceforth abbreviated by DOM). We refer to this benchmark treatment as the *noncontingent treatment.*

Consider next a different treatment, which we call the *contingent treatment*, where we describe the same problem in a slightly different way. We now tell the subject that she will get \$5 if the other bidder bids \$0.50 and she will get \$3 if the other bidder bids \$8.50. If, however, the other bidder bids \$4.50, then the subject has to choose an integer bid from \$1 to \$8. Crucially, we ask the subject to commit to a choice in case event $A = \{\$4.50\}$ is realized *before* she knows whether or not event A is realized. Thus, the subject faces the same choice between f' and g' that is faced by a subject in the noncontingent treatment.

Once again, the subject faces the same options (in terms of having the same monetary payoffs for each possible state of the world) in the two treatments. But, in practice, one can imagine that the contingent treatment helps subjects focus on the event A where the consequences differ.

There are two underlying principles behind the failure of dominance (DOM). The first is a contingent version of DOM. In particular, if the subject continues to choose g' in the contingent treatment, then we say that a contingent version of dominance, which we call C-DOM, is violated.

The second principle, which is the principle that, as we emphasize, underlies problems from both decision and game theory is, once again, STP. In the example of Figure 3, STP says that if f' is preferred to g' in the contingent treatment (where the agent makes a choice contingent on event A, but before knowing which state will realize), then f' must be preferred to g' in the noncontingent treatment. By comparing questions across treatments, we can assess the extent to which STP fails in our setting. In this example, an agent would violate STP if she were to prefer different alternatives across treatments. Figure 4 illustrates the connection between the questions

and postulates for this example.

In Online Appendix A, we provide a simple formalization of the connections between the different concepts depicted in Figures 2 and 4. In particular, we show, as the informal discussion in this section suggests, that if STP and C-SEP (or C-DOM) hold, then SEP (or DOM) must also hold. These results formalize the claim that a very particular form of contingent thinking, embodied by the notion of STP, underlies behavior in classical environments from both decision and game theory. In the experiment, we will apply this simple framework to test to what extent it is failure of STP that explains the classic anomalies, as opposed to failures of C-DOM or C-SEP.

3 Experimental Design

We study five classic problems in the laboratory: Ellsberg (ELLS), common-consequence Allais (CC ALLAIS), a private-values second-price auction (AUCT), a common-value election (ELECT), and common-ratio Allais (CR ALLAIS). For each of these five problems, we conduct two treatments. The *noncontingent treatment* is the benchmark treatment and it is intended to replicate existing results in the literature. ELLS and CC ALLAIS constitute examples of violation of separability (SEP), while AUCT, ELECT and CR ALLAIS constitute examples of violation of dominance (DOM).

We also run a second treatment, the *contingent treatment*. For each of the five problems, we elicit choices contingent on the event for which the consequences of the two options differ, and we investigate if this manipulation has a systematic effect across problems. We evaluate the extent to which the sure-thing principle (STP) and either C-SEP (for ELLS and CC ALLAIS) or C-DOM (for AUCT, ELECT, and CR ALLAIS) are satisfied. A noteworthy feature of our design is that the subject faces a static choice problem in both the contingent and noncontingent treatments. We discuss the relationship to the dynamic choice literature in Section 3.2.

We conducted both *between-subjects* and *within-subjects* designs. There are well-known trade-offs between within- and between-subjects designs (e.g., Camerer (1995), pg. 633). In particular, the between-subjects design minimizes confounding effects, while the within-subjects design allows us to identify additional primitives such as the correlation between problems. In the between-subjects design, each subject participated in one of the five problems and in one of the two treatments only, where each problem is parameterized as described in Section 3.1.

We conducted *three* experiments with a within-subjects design. In the first version, which we refer to as the *within* experiment, each subject participated in both treatments for all of the five problems using the same parameterization of the between-subjects design. Subjects first go

over the noncontingent versions of all problems before going over the contingent versions of all problems. Provided we can replicate the between-subjects design, the within design provides valuable information to assess the effect of the treatment. For robustness purposes, we conducted a second version of a within-subjects design, which we refer to as as the *within+* experiment, where we exposed subjects to several parameterizations of each of the five problems; we discuss the additional parameterizations in Online Appendix D. Finally, in order to better understand the role of the contingent treatment in alleviating cognitive constraints, we conducted a third version that is identical to the first version except that the subjects were exposed to the contingent treatment first and the noncontingent treatment second. We refer to this version as the *withinCNC* experiment.

The experiments were conducted at the University of California, Santa Barbara and subjects were recruited using ORSEE (Greiner, 2015). There was a total of 1,093 subjects who participated only in one session: 742 participants in the between-subjects design and a total of 351 in the within-subjects designs (131 in the *within*, 119 in *within+*, and 101 in the *withinCNC*). The experiment was conducted using zTree (Fischbacher, 2007).[13] At the end of the experiment, each subject was asked additional questions to assess her level of risk or ambiguity aversion and cognitive ability. We describe these questions in further detail in Online Appendix B and in the Procedures Appendix.

We now describe the two treatments for each of the five problems.

3.1 Five experimental problems

3.1.1 Ellsberg problem

This problem was described in Section 2 and we refer the reader to Figure 1 in pg. 7 for reference. Recall that there is a jar with 90 balls. Of the 90 balls, 30 are red and 60 are yellow or blue. A subject must answer two questions, Q1 and Q2, in sequential order and she does not know the second question when answering the first. For each of the two questions, a ball is randomly drawn (with replacement).

Noncontingent treatment. In Q1, the subject must choose between f and g, while in Q2 the subject must choose between f' and g', as described in Figure 1. As mentioned earlier, the typical paradox is that a significant number of subjects choose g in the first question and f' in the second question. This is a violation of SEP.[14] For this treatment and for all other treatments and problems,

[13] A session in the between-subjects (*within*, *within+*, *withinCNC*) design lasted approximately 30 (90, 120, 90) minutes and on average subjects received \$9.50 (\$19.10, \$27.10, \$18.60) in compensation.

[14] In accordance with the literature, we take these choices to reflect strict preferences. The underlying assumption is that a subject who is in fact indifferent in *both* questions chooses the first of the two options with the same probability

subjects did not receive feedback until the end of the experiment. Moreover, the written instructions did not include a table such as Figure 1 in this example.

Contingent treatment. In Q1, we tell the subject that if the drawn ball is blue, the problem ends there and the decision-maker gets \$10. But, if the drawn ball is red or yellow, the decision maker has to make a choice between two options: (1) get \$10 if the ball is red, and (2) get \$10 if it is yellow. Thus, the subject makes a choice contingent on the event in which the ball is red or yellow, without yet knowing if this is the event that will happen. Note that a choice between (1) and (2) corresponds to a choice between f and g in Figure 1. Q2 is identical to the first question except that the decision maker gets \$0 if the drawn ball is blue. Thus, the subject is in fact facing a choice between f' and g' in Figure 1.

Testable hypotheses. In both the noncontingent and contingent treatments, the subject faces the same choices. In practice, however, one may see a difference in behavior between the two treatments to the extent that the second treatment helps the subject focus on the event that is payoff relevant for her decision, $\{red, yellow\}$. Figure 2 (reported earlier in pg. 8) illustrates the testable hypotheses in the context of this problem. A comparison between Q1 and Q2 in the noncontingent treatment is the standard way of testing for SEP in the literature. A comparison between Q1 and Q2 in the contingent treatment provides a test of C-SEP. A comparison between Q1 in the noncontingent treatment and Q1 in the contingent treatment provides a test of STP; the same is true for the comparison between Q2 across treatments.

3.1.2 Common-consequence Allais problem

There is a jar with 100 balls. Of the 100 balls, 1 is red (R), 10 are yellow (Y), and 89 are blue (B). For each of the two questions, Q1 and Q2, a ball is randomly drawn (with replacement).

Noncontingent treatment. In Q1, the subject must choose between f, which gives \$100 million for sure, and g, which gives \$500 million if the ball is yellow and \$100 million if it is blue.[15] In Q2, the subject must choose between f', which gives \$100 million if the ball is red or yellow, and g', which gives \$500 million if the ball is yellow.

in each question. The same point holds for all other problems.

[15]This is the only decision problem for which we use hypothetical payoffs. Huck and Müller (2012) study three alternative ways of implementing the common-consequence Allais problem. They find few violations of SEP with small payoffs, but that violations become more prevalent with hypothetical payoffs expressed in millions of euros. We also conducted a version of this problem with real payoffs; see the discussion after Finding #3 in Section 4.1.

		R (1)	Y (10)	B (89)
Question 1	f	\$100m	\$100m	\$100m
	g	\$0	\$500m	\$100m
Question 2	f'	\$100m	\$100m	\$0
	g'	\$0	\$500m	\$0

Figure 5: Common-consequence Allais problem

Figure 5 depicts the choices in each question. The paradox is that a significant number of subjects choose f (safe option) in Q1 and g' (risky option) in Q2. This is a violation of SEP.

Contingent treatment. In Q1, we tell the subject that if the drawn ball is blue, the problem ends there and the decision maker gets \$100 million. But, if the drawn ball is red or yellow, the decision maker has to make a choice between: (1) get \$100 million if the ball is red or yellow, and (2) get \$500 million if the ball is yellow. We then ask them to make a choice contingent on the event in which the ball is red or yellow, without yet knowing if this is the event that will happen. Note that a choice between (1) and (2) corresponds to a choice between f and g in Figure 5. Q2 is identical except that the decision maker gets \$0 if the drawn ball is blue. Thus, the subject is in fact facing a choice between f' and g' in Figure 5.

Testable hypotheses. This problem has the same structure of ELLS, and so it is straightforward to see that the previous discussion applies here. In particular, Figure 2 (reported earlier in pg. 8) also describes the hypotheses for CC ALLAIS.

3.1.3 Auction problem

This problem was described in Section 2 and we refer the reader to Figure 3 in pg. 9 for reference. In Section 2, we used the wording of auction theory to describe the problem, but, in the experiment, we used a more neutral language. In particular, we did not make an explicit reference to an auction environment partly due to the concern that subjects may mistakenly interpret it as a first-price auction rather than the less familiar second-price auction (see, e.g., Cason and Plott, 2014).[16]

There are three cards, numbered 4.5, 0.5, and 8.5, and one card is randomly drawn. There is only one question in each treatment.

[16]This is not an issue in second-price auction experiments, where subjects face multiple rounds. While it is typical in experiments of strategic behavior to repeat the same or similar questions many times to test for experience effects, here we decided to only ask each question once, in order to make the results comparable to the standard decision experiments (Ellsberg and Allais). The literature documents that the strategic biases that we replicate here are actually robust to experience (see, for example, Kagel and Levin (1993) for the Auction problem and Esponda and Vespa (2014) for the Election problem).

Noncontingent treatment. Without knowing the drawn card, the subject must choose an integer between 1 and 8. If the number she chooses is higher than the number on the drawn card, her payoff is \$5.5 minus the number on the card (in dollars). If the number she chooses is lower than the number on the card, her payoff is \$3. In this problem, one's choice is only relevant if the card is 4.5, and, in that case, it is optimal to choose 1, 2, 3, or 4.

Figure 3 in pg. 9 depicts the problem faced by the agent. As mentioned earlier, we simplify the exposition of the results and codify an optimal choice of 1, 2, 3, or 4 as option f' and a suboptimal choice of 5, 6, 7, or 8 (overbidding) as option g'. We also define options f (payoff of \$3 in all states) and g (payoff of \$1 in all states). The paradox in this problem is that a significant fraction of subjects choose g' over f'. This choice violates DOM under the reasonable assumption that subjects prefer more money to less, i.e., f is preferred to g.

Contingent treatment. We tell the subject that she will get \$5 if the drawn card is 0.5, \$3 if it is 8.5, and we ask her to make a contingent choice for the event in which the card is 4.5, without her knowing if this is the event that will happen.

Testable hypotheses. Figure 4 (reported earlier in pg. 10) describes the hypotheses for AUCT. In the figure, Question 1 corresponds to the question we asked and Question* corresponds to the choice between options f and g in Figure 3. Therefore, a comparison between Q1 and Q* in the noncontingent treatment tests DOM. Because Q* seems trivial, we did not ask it; instead, we simply assume that, if Q* were asked, a subject would prefer more to less money, i.e, f over g. Therefore, DOM is violated in this example provided that a subject prefers g' to f' in the non-contingent treatment. Similarly, C-DOM is violated if a subject prefers g' to f' in the contingent treatment. Finally, a comparison between Q1 in the noncontingent treatment and Q1 in the contingent treatment tests STP.

3.1.4 Election problem

There is a jar with 7 white balls and 3 black balls, and one ball is randomly drawn. There are two computers. If the drawn ball is white (w), both computers vote White (WW). If the drawn ball is black (b), computers vote for different colors (WB). We focus on one question in each treatment, represented in Figure 6 as Question 1.

Noncontingent treatment. Without observing either the color of the drawn ball or the votes of the computers, the subject must choose between voting for Black and voting for White. If the color chosen by the majority matches the color of the drawn ball, the subject gets \$5; otherwise, she gets \$0. It is optimal for the subject to vote for Black, since, if her vote matters, it must be that the ball is indeed black.

		pivotal		not pivotal			
		bWB	wWB$^{\oslash}$	bWW$^{\oslash}$	wBB$^{\oslash}$	bBB$^{\oslash}$	**wWW**
Question*	f	**$5**	$5	$5	$5	$5	**$5**
	g	**$0**	$0	$0	$0	$0	**$0**
Question 1	f'	**$5**	$0	$0	$0	$5	**$5**
	g'	**$0**	$5	$0	$0	$5	**$5**

Figure 6: Election problem. States marked by $^{\oslash}$ have zero probability.

Figure 6 represents the payoffs. Voting for Black is represented by f' in Figure 6 and voting for White is represented by g'. We also define options f (a sure payment of \$5) and g (a sure payment of \$0) in Figure 6, and assume that f is preferred to g. The paradox in this problem is that a significant fraction of subjects choose g' over f'. Whether or not this choice violates DOM depends on the subject's subjective perception of state wWB. If the subject correctly believes that this state has zero probability, then choosing g' over f' violates DOM. Otherwise, all we can say is that an *objective version* of DOM is violated. It is in principle possible for a subject to subjectively believe that wWB has positive probability and to, therefore, choose g' over f' without violating DOM. We will be content with either interpretation (DOM or objective DOM) since our main focus will be on STP.

Contingent treatment. The question is the same, except that we tell subjects that, if both computers vote for the same color, they will receive \$5 if the color matches the color of the drawn ball, and \$0 otherwise. We then ask them to make a choice contingent on the event that computers vote for different colors, without knowing if this is the event that will happen.

Testable hypotheses. Figure 4 (reported earlier in pg. 10) can also be used to describe the hypotheses for ELECT. In the figure, Question 1 corresponds to the question we asked and Question* corresponds to the choice between options f and g in Figure 6. Therefore, a comparison between Q1 and Q* in the noncontingent treatment tests either DOM or an objective version of DOM, as explained earlier. Because Q* seems trivial, we did not ask it; instead, we simply assume that, if Q* were asked, a subject would prefer more to less money, i.e, f over g. Therefore, DOM (or its objective version) is violated in this example provided that a subject prefers g' to f' in the noncontingent treatment. Similarly, C-DOM (or its objective version) is violated if a subject prefers g' to f' in the contingent treatment. Finally, a comparison between Q1 in the noncontingent treatment and Q1 in the contingent treatment tests STP. Note that STP is tested irrespective of whether or not the subject understands that state wWB has zero probability. For example, a subject may choose g' over f' because she thinks that state wWB is very likely. But the same subject who then prefers f' over g' in the contingent treatment would be violating STP.

		RY	B
Question 2	f	x	x
	g	y	y
Question 1	f'	x	\$0
	g'	y	\$0

Figure 7: Common-ratio Allais problem

3.1.5 Common-ratio Allais problem

There is a jar with 100 balls. In each treatment, the subject answers two questions, Q1 and Q2.

Noncontingent problem. In Q1, the jar has 12 red, 3 yellow, and 85 blue balls. The subject must choose an option that gives \$4 if the drawn ball is red or yellow, and an option that gives \$5.30 if it is red. In Q2, the jar has 80 red balls and 20 yellow balls. The subject must choose between an option that gives \$4 for sure and an option that gives \$5.30 if it is red. Note that the ratio of red to yellow balls is the same in both jars, which explains the term "common-ratio" in this experiment.

To depict this problem in Savage's framework, let the space of consequences be given by $Z = \{x, y, 0\}$, where x is a lottery that gives a sure payoff of \$4, y is a lottery that gives \$5.30 with probability .8 and nothing otherwise, and 0 is a lottery that pays \$0 for sure. The set of states is $S = \{RY, B\}$, where RY is the state where the ball drawn from the urn is red or yellow and B is the state where it is blue. Figure 7 depicts the choices faced by the subject in each question. In Q1, the subject must choose between f' and g'. In Q2, the subject must choose between f (a sure payoff of \$4, represented by x) and g (a lottery that pays \$5.30 with probability .8, represented by y). The typical paradox is that many subjects choose g' in Q1 and f in Q2. This is a violation of DOM.

Contingent treatment. In Q1, the question is the same as in the noncontingent treatment, except that if the ball is blue, the decision maker gets \$0, and she has to decide what to do in the event that the ball is red or yellow, before knowing if this is the event that will happen. Note that a choice between the options is equivalent to a choice between f' and g' in Figure 7. In Q2, the question is *identical* to Q2 in the noncontingent treatment, and so the choice is between f and g in Figure 7.

Testable hypotheses. Figure 4 (reported earlier in pg. 10) also depicts the hypotheses for CR ALLAIS, where, in this particular case, Q2 (which is labelled as Question* in the figure) is the *same* question for the contingent and noncontingent treatments. A comparison between Q1 and Q2 in the noncontingent treatment tests DOM. This is the standard test in the literature. A comparison between Q1 and Q2 in the contingent treatment tests C-DOM. Finally, a comparison between Q1 in the noncontingent treatment and Q1 in the contingent treatment tests STP.

3.2 Comparison to the dynamic choice literature

One feature that distinguishes our paper from previous work is that we focus on *static* contexts, in the sense that the subject does not receive any information about the realized state. There is a literature that instead studies decision making in *dynamic* contexts. This literature defines a conditional preference relation as the preference relation that applies *after* the agent observes the realization of an event. Two of the main postulates studied in dynamic choice settings include dynamic consistency and consequentialism (e.g., Hammond (1988), Machina (1989), Cubitt (1996)). Dynamic consistency corresponds to comparing our contingent treatment, which elicits a plan of action from the subject, to a sequential treatment in which the agent is informed, before making her choice, that an event has actually occurred.[17] The notion of dynamic consistency is also related to the strategy method in experimental work (for a survey, see Brandts and Charness, 2011). Under the strategy method, subjects are asked what they would hypothetically do at every contingency. Experiments test if the strategy method introduces a bias by comparing the contingent treatment to a sequential treatment, in which subjects are told the specific contingency that occurred. Consequentialism, on the other hand, corresponds to comparing two sequential treatments where the agent is informed that an event has actually occurred but where the forgone payoffs under the complement of the event are different. In contrast, we focus on the comparison between the noncontingent and contingent treatments, both of which correspond to static choice situations.

4 Results

4.1 Between-subjects results

Table 1 summarizes the findings for the between-subjects design. For each of the five problems and for each treatment (noncontingent, NC, and contingent, C), we report the number of observations and the percent of subjects making each of *four* (in ELLS, CC ALLAIS and CR ALLAIS, where subjects face two questions) or *two* (in AUCT and ELECT, where subjects face one question) possible choices. Based on these choices, we compute and report the percent of subjects failing SEP or DOM (in the noncontingent treatment) and C-SEP or C-DOM (in the contingent treatment). In the last row, we report, separately for each problem, the p-value of a test of the null hypothesis that percentage failure of SEP or DOM equals percentage failure of C-SEP or C-DOM. There are three

[17] Because most of the literature implicitly views the noncontingent and contingent treatments as equivalent, some tests of dynamic consistency compare behavior in the noncontingent and sequential treatments. For experiments on dynamic consistency and consequentialism, see Cohen, Gilboa, Jaffray, and Schmeidler (2000), Dominiak, Dürsch, and Lefort (2012) and Nebout and Dubois (2014).

	ELLS		CC ALLAIS		AUCT		ELECT		CR ALLAIS	
treatment	NC	C	NC	C	NC	C	NC	C	NC	C
# of observ.	59	61	63	63	62	62	66	63	62	63
% (f, f')	18.6	42.6	22.2	34.9	67.7	87.1	15.2	54.0	45.2	55.6
% (g, g')	23.7	29.5	58.7	36.5	—	—	—	—	4.8	19.0
% (f, g')	6.8	9.8	12.7	12.7	32.3	12.9	84.8	46.0	32.3	15.9
% (g, f')	50.9	18.0	6.4	15.9	—	—	—	—	17.7	9.5
% fail SEP/C-SEP	57.7	27.8	19.1	28.6	—	—	—	—	—	—
% fail DOM/C-DOM	—	—	—	—	32.3	12.9	84.8	46.0	50.0	25.4
p-value	.001		.213		.010		.000		.004	

Table 1: Between-subjects design: summary of results

Notes: 1) In ELLS and CC ALLAIS, (f, f') indicates the proportion of subjects who selected f in Q1 and f' in Q2. In AUCT and ELECT, (f, f') indicates choices of f in Q* and f' in Q1. We assume that all subjects prefer more money to less and so we impute that all subjects select f in Q*. In CR ALLAIS, (f, f') indicates choices of f in Q2 and f' in Q1.
2) % fail SEP/C-SEP and % fail DOM/C-DOM presents the addition of subjects who chose (f, g') and subjects who chose (g, f').
3) The reported p-value results from a regression in which the unit of observation is a subject. The dependent variable is a dummy that takes value 1 if the subject's choices fail to satisfy the corresponding postulate, and the right-hand side includes a constant and a treatment dummy that takes value 1 if the subject participated in the contingent treatment. The p-values we report correspond to the coefficient estimated for the treatment dummy.

main takeaways from this table.

Finding #1. *We replicate the anomalies pointed out in the literature for all noncontingent versions of the problems*. This finding is observed by looking at each column labeled NC (non-contingent treatment). Consider first the case of ELLS. Most subjects (50.9%) select g in Q1 and f' in Q2.[18] These choices are consistent with the heuristic of ambiguity aversion (Ellsberg, 1961): subjects in this group prefer g in the first question (where the probability of receiving \$10 is known to be 2/3) and f' in the second question (where the probability of receiving \$10 is known to be 1/3). There is also a 6.8% of subjects who fail SEP by selecting f in Q1, but g' in Q2. Overall, the percent of subjects with choices that are inconsistent with SEP in ELLS is 57.7%. In the case of CC ALLAIS, in line with the literature, we find that the most common violation of SEP occurs when subjects select f in Q1 and g' in Q2. These choices are consistent, for example, with the heuristic of regret aversion (Loomes and Sugden, 1982) or with saliency of payoffs (Bordalo, Gennaioli, and Shleifer, 2012). Overall, the percent of subjects with choices that are inconsistent with SEP is 19.1%.

Choices of g' in Q1 of AUCT and ELECT are inconsistent with DOM (or objective DOM, in the case of ELECT) under the assumption that subjects prefer more money to less. In Table 1, we force this assumption by imposing that all subjects would select f in Q*. We find that 32.3% and 84.8% of subjects select g' in AUCT and ELECT, respectively.[19] In AUCT, the observed overbidding

[18]For previous results that are qualitatively in line with our findings see Camerer (1995).

[19]For previous evidence on overbidding in second-price auctions see Kagel (1995) and for non-pivotal voting see

is consistent with the illusion that it increases the chance of winning with little cost because the winner pays the second highest bid (Kagel, Harstad, and Levin, 1987). In ELECT, the mistake is consistent with the heuristic that subjects choose the color that is more prevalent in the jar.

Finally, in CR ALLAIS, DOM is violated if subjects' choices are (f, g') or (g, f'). Table 1 shows that 50% of subjects make choices inconsistent with DOM, qualitatively in line with previous findings; see Camerer (1995). As in ELLS and CC ALLAIS, the most prevalent anomaly in CR ALLAIS coincides with the one found in the literature, which corresponds to choosing (f, g'). These choices are consistent with the certainty effect heuristic (Tversky and Kahneman, 1986).

Finding #2. *In all problems except* CC ALLAIS*, the anomalies drop by about half in the contingent treatment; in* CC ALLAIS*, anomalies increase but the difference is not statistically significant.* In particular, reversals decrease from 57.7% to 27.8% (p-value of .001) in ELLS and from 50.0% to 25.4% (p-value of .004) in CR ALLAIS. Inconsistent choices decrease from 32.3% to 12.9% in AUCT and from 84.8% to 46% in ELECT, with both differences being statistically significant at the 1% level.[20] Finally, the proportion of subjects making inconsistent choices in CC ALLAIS increases from 19.1% to 28.6%, but this difference is not statistically significant (p-value of .213).

Finding #3. *We find a treatment effect in all problems. In particular, by comparing the joint distribution of responses across treatments, we can reject the hypothesis that STP holds in all of the five problems.* We can also look at the extent to which differences in Q1 or Q2 are driving the failures of STP. In Q1 of ELLS, f is preferred by 25.4% of subjects (this is the sum of $(f, f')\%$ and $(f, g')\%$) in the noncontingent treatment and by 52.4% of subjects in the contingent treatment; the difference is statistically significant (p-value of .002). We also find significant differences for Q1 of ELECT (15.2% vs. 54%; p-value of .000) and Q1 of AUCT (67.7% vs. 87.1%; p-value of .010).[21] For Q1 of CC ALLAIS, there is a difference but it is marginally not significant (34.9% vs 47.6%; p-value of .15). Moreover, there is a significant difference for Q2 of CC ALLAIS (28.6% vs. 50.8%; p-value of .011). Finally, we do not find a significant difference for either Q1 or Q2 of CR ALLAIS. The result for Q2 is reassuring, because Q2 is the same question in the noncontingent and contingent treatments for this problem. Interestingly, despite no difference in the marginal distributions, the joint distribution for CR ALLAIS is significantly different, as evidenced by the drop in half in anomalies.

Esponda and Vespa (2014; 2018).

[20]We conduct a regression in which the unit of observation is a subject. The dependent variable is a dummy that takes value 1 if the subject's choices are inconsistent with the corresponding postulate, and the right-hand side includes a constant and a treatment dummy that takes value 1 if the subject participated in the contingent treatment. The p-values we report correspond to the coefficient estimated for the treatment dummy. A similar approach is followed for the other tests reported in the paper.

[21]For ELECT and AUCT, the test for STP is numerically the same as the test presented in Finding #2 to assess the difference between DOM and C-DOM.

The reader may wonder if the low rate of violations of SEP and the lack of statistical significance between SEP and C-SEP for CC ALLAIS could be due to the fact that we used hypothetical payoffs (in millions of dollars) for this particular problem. To answer this question, we conducted a new experiment where we used monetary payoffs for CC ALLAIS. We replaced \$100m with 10 dollars and \$500m with \$20 dollars in Table 5. We recruited 58 subjects for the noncontingent treatment and 60 subjects for the contingent treatment. We found that 39.6% of subjects violated SEP in the noncontingent treatment.[22] In the contingent treatment, violations of C-SEP are almost identical, at 41.7%, and the difference is not statistically significant (p-value .826). Moreover, the three findings listed above are robust to using hypothetical or real payoffs in CC ALLAIS.[23]

4.2 Within-subjects results

Table 2 summarizes the findings for the *within* experiment. Recall that this is an experiment with 131 subjects, where each subject participates first in the noncontingent version of all problems and then in the contingent version of all problems. For each problem, there are four possible outcomes regarding consistency with the relevant postulate: always consistent (i.e., satisfies the relevant postulate in both treatments; e.g., satisfies SEP and C-SEP in ELLS), always not consistent, consistent in the noncontingent treatment but inconsistent in the contingent treatment ($NOT \rightarrow Cons$ in the table) and the other way around ($Cons \rightarrow NOT$ in the table). Based on these four outcomes, we compute and report the percent of subjects failing SEP/DOM and C-SEP/C-DOM. In particular, failure of SEP/DOM is given by the sum of $\%Always\,NOT$ and $\%NOT \rightarrow Cons$, since these were the subjects who were not consistent in the noncontingent treatment. Similarly, the failure of C-SEP/C-DOM (i.e., the failure of the relevant postulate in the contingent treatment) is given by the sum of $\%Always\,NOT$ and $\%Cons \rightarrow NOT$. We also report the p-value from a test of the null hypothesis of an equal proportion of failures of SEP/DOM and C-SEP/C-DOM.

Towards the bottom of the table, we report the percent of STP violations for each problem. For the problems that have two questions to test STP (ELLS and CC ALLAIS), we say that the subject fails STP if it fails for at least one of the questions. The bottom three rows in the table will be explained later. There are four main takeaways from Table 2.

Finding #4. *The findings of the within experiment are in line with the previous findings from the*

[22] Of all the anomalies we replicated, the one that is documented to be the hardest to find is the paradox in CC ALLAIS (see Huck and Müller, 2012 and Blavatskyy, Ortmann, and Panchenko, 2015).

[23] The percentages of choices between the noncontingent and contingent treatments were 19.0 vs 28.3 for (f, f'), 41.4 vs 30.0 for (g, g'), 29.3 vs 10.0 for (f, g'), and 10.3 vs 31.7 for (g, f'). We find no significant difference in the choice of f in Q1 (48.3% vs. 38.3%; p-value of .280), but we find a significant difference in the choice of f' in Q2 (29.3% vs. 60.0%; p-value of .001).

	ELLS	CC ALLAIS	AUCT	ELECT	CR ALLAIS
% Always Cons	33.6	54.2	75.6	17.6	60.3
% Always NOT	16.8	15.3	4.6	36.6	12.2
% NOT $\rightarrow$ Cons	36.6	19.9	18.3	42.0	22.9
% Cons $\rightarrow$ NOT	13.0	10.7	1.5	3.8	4.6
% fail SEP/DOM	53.4	35.1	22.9	78.6	35.1
% fail C-SEP/C-DOM	29.8	25.9	6.1	40.5	16.8
($\circledast$) p-value	.000	.058	.000	.000	.000
% fail STP	72.5	43.5	19.9	45.8	25.2
$q_{C\|I}$	.69	.57	.80	.53	.65
$q_{I\|C}$	.28	.16	.02	.18	.07
($\otimes$) p-value	.000	.000	.000	.000	.000

Table 2: *Within* experiment: summary of results

Notes: 131 Observations. Let $j \in \{$ELLS, CC ALLAIS, AUCT, ELECT, CR ALLAIS$\}$ capture each of the problems. ($\circledast$) For each subject D_{Inc} is a dummy variable that takes value 1 if the subject is not consistent with the corresponding postulate (SEP/DOM in the noncontingent versions and C-SEP/C-DOM in the contingent versions). D^C is a dummy that takes value 1 if the observation is from a Contingent problem and the dummy variable D_j takes value 1 if the answers are for problem j. We run the following regression, in which for each subject we have ten observations (corresponding to the contingent and noncontigent answers to each of the five problems): $D_{Inc} = \sum_j \left(\delta_j D_j + \phi_j (D^C \times D_j)\right) + \upsilon$, where υ is an error term. The null hypothesis of interest is that there is no effect of the contingent treatment for each j. The p-value reported in the table correspond to the null hypothesis that $\phi_j = 0$.

($\otimes$)For each subject in the contingent (noncontingent) treatment of each problem D^C_{Inc} (D^{NC}_{Inc}) is a dummy that takes value 1 if the subject is not consistent. We run the following regression, in which for each subject we have five observations (corresponding to the five problems): $D^C_{Inc} = \sum_j \left(\alpha_j D_j + \beta_j (D^{NC}_{Inc} \times D_j)\right) + \epsilon$, where ϵ is an error term. The null hypothesis $q_{C|I} = q_{I|C}$ implies that $2\alpha_j = 1 - \beta_j$. The reported p-values in the last row of the table correspond to a Wald test of this equality for the corresponding j. In the estimation we cluster standard errors by subject.

between-subjects design. Inspection of Tables 1 and 2 shows that the levels of failure of SEP/DOM and C-SEP/C-DOM are comparable to the levels observed in the between-subjects design. In particular, failures of consistency drop by about one-half in all problems except CC ALLAIS, where the difference continues to be not significant.[24] Based on this evidence, we feel more confident in taking advantage of the additional information that the *within* experiment provides.

Finding #5. *There is a large treatment effect across all problems, as evidenced by the number of failures of STP documented.* The range of STP violations goes from a low of 19.9% in AUCT to a high of 72.5% in ELLS. Note that a lower bound for failure of STP is given by those subjects whose consistency is affected by the treatment (either $NOT \rightarrow Cons$ or $Cons \rightarrow NOT$).[25] This bound is tight in the AUCT and ELECT problems. But for ELLS and CC ALLAIS, where there are two questions to test for STP, a subject who is either always or never consistent may still provide different responses to these questions and, therefore, fail STP.

Finding #6. *Subject heterogeneity: In all problems, most subjects whose consistency is affected by the treatment go from being inconsistent to consistent, but there is also a considerable number of subjects who go from consistent to inconsistent in the case of SEP.* To the extent that SEP is a postulate describing the preferences of a subject, there is no a priori reason to believe that every subject should be more likely to satisfy this postulate once faced with the contingent treatment. Indeed, while 36.6% of subjects go from inconsistent to consistent in ELLS, there are also 13% of subjects who go from consistent to inconsistent.

By definition, the difference between $\%NOT \rightarrow Cons$ and $\%Cons \rightarrow NOT$ is exactly the difference between subjects failing the noncontingent and contingent versions of the postulate. For example, in the case of CC ALLAIS, 19.9% of subjects change consistency in one direction and 10.7% in the other direction; these effects essentially cancel out and the difference of 9.2 percentage points, which is exactly the difference between %fail SEP (35.1%) and %fail C-SEP (25.9%), is not statistically significant. But, despite these effects cancelling out, there are still a total of 30.6% of subjects in CC ALLAIS whose consistency is affected by the treatment. Interestingly, when the relevant postulate is dominance (as in AUCT, ELECT, and CR ALLAIS), there are essentially no switches in the direction $Cons \rightarrow NOT$.

Another way to interpret these numbers is to note that, if subjects' actual preferences were more clearly elicited by the contingent treatment, then we would be incorrectly misclassifying a large fraction of subjects using the noncontingent treatment, where this large fraction is equal to the sum of $\%NOT \rightarrow Cons$ and $\%Cons \rightarrow NOT$, a figure that is as high as 49.5% in ELLS.

[24] See Online Appendix C for a more detailed analysis of Findings 1-3 using the data from the *within* experiment. In addition, Table 17 of Online Appendix C provides detailed comparison of the between-design and *within* experiment.

[25] This is true except in CR ALLAIS, where only Q1, and not Q2, can be used to test for STP.

% of subjects (**fictional**)

		Contingent NOT	*Contingent* consistent	
Noncontingent	NOT	19.0	59.6	78.6
	consistent	21.4	0	21.4
		40.4	59.6	

% of subjects (**true data**)

		Contingent NOT	*Contingent* consistent	
Noncontingent	NOT	36.6	42.0	78.6
	consistent	3.8	17.6	21.4
		40.4	59.6	

Figure 8: Illustration: fictional and true data for ELECT.

Taken together, the two previous findings indicate that, while the treatment can move subjects in either direction, in all problems except CC ALLAIS, the net effect is to decrease inconsistencies.

The data analyzed so far, however, does not account for base rates of consistency and, in particular, does not yet prove that the treatment is more likely to convert an inconsistent subject into consistent than vice versa. The following example illustrates this point. The table on the top of Figure 8 shows a fictional joint distribution of consistent and inconsistent behavior across treatments for ELECT. The table on the bottom of Figure 8 shows the true distribution using our data. The marginals are the same in both tables. Let $q_{C|I}$ be the estimated probability of switching to being consistent (contingent treatment) conditional on being inconsistent in the noncontingent treatment. Let $q_{I|C}$ be the estimated probability of switching in the opposite direction: becoming inconsistent conditional on being consistent in the original treatment. In the top table of Figure 8, every subject who is consistent in the noncontingent treatment later becomes inconsistent in the contingent treatment. In fact, based on this fictional data,

$$q_{C|I} = \frac{59.6}{19 + 59.6} = .76 < 1 = \frac{21.4}{21.4} = q_{I|C}.$$

Therefore, even though the percent of inconsistent subjects goes down from 78.6% to 40.4%, the contingent treatment in this fictional example is more likely to turn people from consistent to inconsistent than vice versa!

When computing these fractions using the true data (bottom table of Figure 8), we obtain a very different picture:

$$q_{C|I} = \frac{42}{36.6 + 42} = .53 > .18 = \frac{3.8}{3.8 + 17.6} = q_{I|C}.$$

In other words, the estimated conditional probability of switching from inconsistent to consistent is about three times higher than the estimated conditional probability of switching from consistent to inconsistent. This example illustrates that focusing on treatment effect without examining switching behavior may lead to misleading conclusions. Therefore, we now turn to examining switching behavior.

Finding #7. *In all problems, the estimated probability of switching to being consistent conditional on being inconsistent ($q_{C|I}$) is significantly higher than the estimated probability of switching to being inconsistent conditional on being consistent ($q_{I|C}$).* These probabilities are reported in the last rows of Table 2, pg. 22, together with the p-value for the test of the null hypothesis that $q_{C|I} = q_{I|C}$. For ELLS, CC ALLAIS, and ELECT, $q_{C|I}$ is about three times higher than $q_{I|C}$. For AUCT and CR ALLAIS, the differences are even larger, as almost all subjects who are originally consistent remain consistent in the contingent treatment.

The calculation of $q_{C|I}$ and $q_{I|C}$ can also help us understand why there is no (significant) net effect on consistency for CC ALLAIS, despite large changes in response to the treatment, as described earlier. In particular, the difference between $q_{C|I} = .57$ and $q_{I|C} = .17$, while considerable, is not sufficiently large to lead to a statistically significant reduction in inconsistencies given that the baseline rate of inconsistency in the noncontingent treatment is 35.1%. It is interesting to compare these figures to the ones for AUCT, where the difference between $q_{C|I} = .80$ and $q_{I|C} = .02$ is sufficiently large to lead to a significant reduction in inconsistencies, despite a baseline rate of inconsistency of only 22.9% in the noncontingent treatment.

4.3 Robustness and extensions

The previous results were obtained for specific parameterizations of five classical problems. Each of these parameterizations was chosen based on previous literature, but the reader may wonder if the results would have been different if we had chosen different parameterizations. For this reason, we conducted an additional, within-subjects experiment where subjects were exposed to both the noncontingent and contingent versions of *multiple* (rather than just one, as before) parameterizations for each of the five problems. In particular, we considered five parameterizations for ELLS, three for CC ALLAIS, ten for AUCT, ten for ELECT, and four for CR ALLAIS; for details, see Online Appendix D. To compare with our previous findings, we include the benchmark parameterizations examined in our earlier results as the first parameterization encountered by the subject in each of the problems.

Following the *within* experiment reported earlier, each subject in this new experiment participated in all parameterizations of all five problems, with all noncontingent versions taking place

before all the contingent versions. We refer to this experiment as the *within+* experiment.

There are two main purposes of this experiment. The first is to examine robustness of our results to different parameterizations of each problem. The second is to have a more reliable measure of behavior for individual subjects that can be used to obtain additional information about correlations across problems.

Table 3 presents the results for the *within+* experiment. For each problem, we report a column with the findings for the benchmark parameterization, which is the parameterization we used to derive our previous findings, #1-7, and a second column with the average of all parameterizations.

Finding #8. *The findings of the within+ experiment are qualitatively in line with the previous findings from the within experiment.* In the *within* experiment, inconsistencies in the contingent version of the problem dropped by about one half for all problems except CC ALLAIS. As reported in the first columns of Table 3 for each problem, for this benchmark parameterization inconsistencies also drop, but not as dramatically as before. Presumably, this is driven by the fact that the *within+* experiment is much longer, lasting 2 hours. The direction of the findings, however, are all in line with findings from the *within* experiment. A similar pattern arises when looking at the second columns of Table 3 for each problem, which reports the average over all parameterizations. We report findings separately for each parameterization and conduct a more detailed analysis in Online Appendix D. The most notable difference arises for CC ALLAIS, where there is now a significant drop in inconsistencies in the contingent treatment driven by our third parameterization, where we used real payoffs as high as $50 to incentivize subjects.

So far, we have focused on each problem individually. One advantage of having multiple measures of consistency for each subject and problem is that we can obtain reliable measures of correlations across problems. Table 4 shows correlations of consistency both for the noncontingent version (SEP or DOM) and for the contingent version (C-SEP or C-DOM) of the problems.

Finding #9. *The correlations in consistency across the problems are aligned with the two postulates, separability and dominance, for the case of the contingent versions of the problems, but not for the noncontingent versions.* For ELLS and CC ALLAIS, consistency with (contingent) separability in one problem is correlated with consistency in the other problem. The correlation is .482. For AUCT, ELECT, and CR ALLAIS, consistency with (contingent) dominance in one problem is correlated with consistency in each of the other problems. In particular, the correlation between AUCT and ELECT is .419, the correlation between ELECT and CR ALLAIS is .389, and the correlation between AUCT and CR ALLAIS is .415. Across postulates, the correlation is very close to zero or very small, with the exception of the correlation between CC ALLAIS and CR ALLAIS of .692.

Interestingly, the correlations are very different for the noncontingent treatment. For example,

	ELLS		CC ALLAIS		AUCT		ELECT		CR ALLAIS	
	Benchmark	All	Benchmark	All	Benchmark	All	Benchmark	All	Benchmark	All
% Always Consistent	26.9	23.7	42.9	46.2	47.9	49.6	13.4	24.1	51.3	47.5
% Always NOT Consistent	21.0	24.9	13.5	14.6	19.3	16.1	59.7	53.8	12.6	13.2
% From NOT to Consistent	35.3	35.0	22.7	25.5	21.9	23.0	24.4	19.4	25.2	25.2
% From Consistent to NOT	16.8	16.5	21.0	13.7	10.9	11.3	2.5	2.6	10.9	14.1
% fail SEP or DOM	56.3	59.8	36.1	41.7	41.2	39.2	84.0	73.2	37.8	38.4
% fail C-SEP or C-DOM	37.8	41.3	34.4	28.3	30.2	27.4	62.2	56.4	23.5	27.3
p-value	.001	.000	.762	.001	.049	.000	.000	.000	.010	.000
% fail STP	66.4	69.9	49.6	49.9	32.8	34.3	26.9	22.1	31.9	36.1
q_N	.63	.58	.63	.64	.53	.58	.29	.27	.67	.66
q_C	.38	.42	.33	.23	.19	.19	.16	.09	.18	.23
p-value	.001	.000	.000	.000	.000	.000	.294	.003	.000	.000

Table 3: All problems: *Within* and *within+* experiments

Notes: 119 Observations. For detailed information on all parameterizations, see Tables 25, 26 and 27 of Online Appendix D. Online Appendix D also provides details on how p-values are computed.

	CC ALLAIS		AUCT		ELECT		CR ALLAIS	
ELLS	*-.560*	**.482**	*-.006*	**-.060**	*.111*	**-.030**	*.298*	**.086**
CC ALLAIS	—		*.254*	**-.070**	*-.250*	**.047**	*-.248*	**.692**
AUCT	—		—		*.198*	**.419**	*.006*	**.415**
ELECT	—		—		—		*-.429*	**.389**

Table 4: Correlations across problems

Notes: Correlations in Italics are for noncontingent versions. Correlations in Bold are for contingent versions. Correlations are computed using standard corrections for multiple measures; see Online Appendix E for details on how correlations are computed.

ELLS and CC ALLAIS are negatively correlated, even though separability underlies both problems, and several other correlations are negative or have no clear pattern. This finding shows that focusing on the noncontingent treatment, as the previous literature has done, obscures the connection between these problems.

The previous findings suggest some alignment of the five problems in terms of their underlying postulate, SEP or DOM. In particular, we showed that SEP is a feature that some people appear to be prefer to violate once the problem is framed contingently, but this is much less likely to be the case for DOM. To further investigate this matter, we use data from the Cognitive Reflection Test (CRT, see Frederick, 2005), which we administered at the end of the experiment. This test is composed of three questions that are intended to measure a subject's tendency to override their gut response and engage in further reflection. This test appears appropriate in our setting, where the gut response may be to follow some heuristic (ambiguity aversion, overbidding, etc.) and this heuristic may be overridden by those who further reflect on the nature of the state space.

For each of the five problems, we regress a dummy for the treatment (where the contingent treatment equals 1), a dummy for CRT (where a high CRT score of 2 or 3 correct answers equals 1), an interaction effect of the treatment and CRT, and a constant on the proportion of consistent responses for all questions in the problem. The results appear in Table 5.

Finding #10. *The CRT is positively correlated with consistency in the problems where the underlying postulate is DOM (*ELECT*,* AUCT *and* CR ALLAIS*), but there is no correlation when the underlying postulate is SEP (*ELLS *and* CC ALLAIS*).* For ELECT*,* a high CRT score is positively associated with consistency within a treatment and, moreover, subjects with a high CRT score respond more to the contingent treatment, in the direction of being more consistent, compared to those with a low CRT score. For AUCT*,* the first of these effects is present, while, for CR ALLAIS, the second effect is present. These findings are consistent with the idea that subjects desire to be consistent, that cognitive limitations make it more difficult to be consistent, and that the contingent treatment alleviates these limitations particularly for subjects who are less cognitively challenged, as measured by the CRT. In contrast, for ELLS *and* CC ALLAIS, neither the CRT nor its interaction

	ELLS	CC ALLAIS	AUCT	ELECT	CR ALLAIS
Contingent	0.188***	0.087**	0.043**	0.091***	0.054
	(0.051)	(0.037)	(0.021)	(0.029)	(0.038)
CRT	0.018	−0.022	0.083**	0.246***	−0.053
	(0.052)	(0.054)	(0.037)	(0.055)	(0.052)
Contingent × CRT	−0.008	0.073	0.041	0.113**	0.136**
	(0.073)	(0.057)	(0.036)	(0.049)	(0.065)
Constant	0.394***	0.609***	0.681***	0.338***	0.638***
	(0.034)	(0.034)	(0.024)	(0.028)	(0.033)
Observations	238	238	238	238	238

Table 5: CRT and consistency (*Within*+ All Questions)

Notes: Results from a regression in which the dependent variable is the proportion of answers consistent with the postulate within a problem. The right-hand side includes a treatment dummy (1=Contingent), a dummy variable for CRT answers (0 indicates 0 or 1 correct CRT answer, 1 indicates 2 or 3 correct CRT answers), and the interaction between the treatment dummy and the CRT score. Standard errors (between parentheses) are clustered by subject. Significant at 10% (*), 5% (**), 1% (***).

with the contingent treatment are significant.[26]

An alternative to using the CRT is to first expose the subjects to the contingent treatment and then assess their responses in the noncontingent treatment. If the contingent treatment is alleviating the subjects' cognitive limitations, then we might expect to see that behavior in the noncontingent treatment is closer to the contingent treatment under this order (compared to the between-subjects design or compared to the previous within-subjects designs where subjects were first exposed to the noncontingent treatment and subsequently to the contingent treatment). For this purpose, we conducted an additional within-subjects experiment with a pool of 101 subjects. The experiment is just like the *within* experiment, except that subjects were first exposed to the contingent versions of the problems and later exposed to the noncontingent versions. We refer to this experiment as the *withinCNC* experiment.

Table 6 shows the results of the *withinCNC* experiment and compares them to the previous between-subjects design and *within* experiment.[27] For each problem, the top of the table shows the percentage of violations of consistency with the corresponding postulate for each of the experiments and each of the treatments. The bottom part of the table shows the p-values for tests where the null hypothesis is the equality across experimental designs.

Finding #11. *Behavior in the contingent treatment is similar across all experiments for all five problems. Behavior in the noncontingent treatment is relatively similar across all experimental designs for* ELLS *and* CC ALLAIS, but changing the order to contingent and then noncontingent

[26]In Online Appendix D, we show that the results are robust if we use a sample that includes subjects who participated in the within-subjects design and subjects who participated in the *within*+.

[27]Further details of the *withinCNC* experiment are provided in Online Appendix F.

	ELLS			CC ALLAIS			AUCT			ELECT			CR ALLAIS		
treatment	NC	C	Δ	NC	C	Δ	NC	C	Δ	NC	C	Δ	NC	C	Δ
Between	57.7	27.8	-29.9***	19.1	28.6	9.5	32.3	12.9	-19.4***	84.8	46.0	-38.8***	50.0	25.4	-24.6***
Within	53.4	29.8	-23.6***	35.1	25.9	-9.2*	22.9	6.1	-16.5***	78.6	40.5	-38.1***	35.1	16.8	-18.3***
WithinCNC	50.5	38.6	-11.9*	28.7	35.7	7.0	17.8	20.8	3.0	69.3	47.5	-21.8***	28.7	24.7	-4.0
p-values															
Bet=With	.592	.787	.560	.014	.705	.038	.184	.156	.755	.278	.467	.944	.053	.182	.508
Bet=WithCNC	.385	.157	.109	.153	.345	.793	.043	.183	.011	.016	.853	.085	.007	.927	.011
With=WithCNC	.659	.163	.195	.301	.116	.044	.341	.002	.001	.113	.286	.038	.301	.144	.048

Table 6: All treatments: % fail the corresponding postulate

Notes: The p-values report on the null hypothesis that the failures in each pair-wise comparison {*Between* and *Within*, *Between* and *WithinCNC*, *Within* and *WithinCNC*}. For each problem we run a regression in which the left-hand side is a dummy that takes value 1 if the subject's choices were inconsistent with the postulate. The right hand side includes four dummies that capture whether the observation belongs to one of four possible groups that result from interacting {C, NC} and the two treatments in the corresponding pair-wise comparison. The p-values correspond to a Wald test for the null hypothesis of no difference in the comparison of interest. For example, the .592 p-value corresponds to the null hypothesis that the coefficient for the between NC is equal to the coefficient for the within NC in the Ellsberg problem. The column indicated with Δ computes the difference between failures in NC and failures in C. The stars are computed using a Wald test on the same regression described earlier, but in the comparison of the coefficients for a fixed experimental design. The stars indicate that each corresponding comparison between NC and C is significant at the 1%(***), 5%(**) and 10%(*) level. Standard errors are clustered by subject.

introduces significant differences in ELECT, AUCT *and* CR ALLAIS. *In particular, subjects in those problems are significantly more likely to satisfy DOM in the noncontingent treatment when they are first exposed to the contingent treatment.*

As shown by the table, we can reject the hypotheses that the differences between the noncontingent and contingent versions of the problems are the same in the *withinCNC* experiment compared to the between-subjects design and the *within* experiment. In addition, in AUCT and CR ALLAIS, the difference between the noncontingent and contingent versions of the problems is essentially zero for the *withinCNC* experiment (3.0 and -4.0, respectively, not significantly different from zero). In ELECT, which is the hardest problem for subjects, being first exposed to the contingent version provides a significant improvement later in the noncontingent version, but there are still significant differences across these versions. In contrast, for ELLS and CC ALLAIS, there is little to no effect of being first exposed to the contingent versions.

Overall, the evidence suggests that being exposed to the contingent treatment alleviates cognitive constrains in the problems where DOM matters. In problems where SEP matters, however, choices are very much tied to the frame that is used, and the contingent treatment, while having a large effect on behavior, has little to no effect on a later noncontingent version.

5 Conclusion

We document relative large rates of failure of a particular type of contingent thinking in classic decision problems and in game-theoretic environments. The environments we study have the property that the state-space can be partitioned into two non-empty sets, such that payoffs between the alternatives only differ in one of these sets. If a subject prefers one alternative over the other in the contingent version of the problem where the set of states where payoffs differ is emphasized, then the sure-thing principle (STP) requires that she prefers the same alternative in the noncontingent version where no states are emphasized. A violation of STP suggests a failure of contingent reasoning since the subject is not partitioning the events between states where her choice matters and states where it does not.

We find that some of the most common anomalies uncovered in laboratory experiments, including overbidding in auctions, naive voting in elections, and Ellsberg and Allais types of paradoxes, spring, at least in large part, from the failure of the type of contingent thinking embodied by STP. Our strategy is to run subjects through standard versions of each of these canonical problems (noncontingent frame), and then we run subjects through slight alterations of each problem (contingent frame), which focus on the subset of states where their choices affect payoffs. STP would be

satisfied if choices were the same across the two frames, but we find that many subjects behave differently in the two versions of the problem, and that this difference explains about half of the anomalies in many classic problems.

Despite the fact that strategic and decision problems are usually rationalized with different psychological mechanisms, our findings indicate that there is actually common empirical ground between them in the form of a failure of contingent thinking, as captured by STP. This result is useful because it suggests that theories that explain anomalies should incorporate the possibility of failure of STP. More generally, the finding that anomalies in a set of seemingly dissimilar settings originate from a common source suggests that finding other root phenomena underlying failures of classical assumptions may be an important and fruitful avenue for future research.

References

AGRANOV, M., A. CAPLIN, AND C. TERGIMAN (2015): "Naive play and the process of choice in guessing games," *Journal of the Economic Science Association*, 1, 146–157.

AHN, D., S. CHOI, D. GALE, AND S. KARIV (2014): "Estimating ambiguity aversion in a portfolio choice experiment," *Quantitative Economics*, 5, 195–223.

AHN, D. S. AND H. ERGIN (2010): "Framing contingencies," *Econometrica*, 78, 655–695.

AL-NAJJAR, N. I. AND J. WEINSTEIN (2009): "The ambiguity aversion literature: a critical assessment," *Economics and Philosophy*, 25, 249–284.

ALLAIS, M. (1953): "Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine," *Econometrica: Journal of the Econometric Society*, 503–546.

ANDREONI, J., T. SCHMIDT, AND C. SPRENGER (2014): "Measuring Ambiguity Aversion: Experimental Tests of Subjective Expected Utility," *mimeo*.

ARAUJO, F., S. WANG, AND A. WILSON (2019): "The times they are a-Changing: Dynamic Adverse Selection in the Laboratory," *Working Paper*.

BARRON, K., S. HUCK, AND P. JEHIEL (2019): "Everyday econometricians: Selection neglect and overoptimism when learning from others," *Working Paper*.

BAYONA, A., J. BRANDTS, AND X. VIVES (2019): "Information Frictions and Market Power: A Laboratory Study," .

BAZERMAN, M. H. AND W. F. SAMUELSON (1983): "I won the auction but don't want the prize," *Journal of Conflict Resolution*, 27, 618–634.

BLAVATSKYY, P. R., A. ORTMANN, AND V. PANCHENKO (2015): "Now you see it, now you don't: How to make the Allais Paradox appear, disappear, or reverse," *UNSW Business School Research Paper*.

BORDALO, P., N. GENNAIOLI, AND A. SHLEIFER (2012): "Salience theory of choice under risk," *The Quarterly journal of economics*, 127, 1243–1285.

——— (2013): "Salience and consumer choice," *Journal of Political Economy*, 121, 803–843.

BRANDTS, J. AND G. CHARNESS (2011): "The strategy versus the direct-response method: a first survey of experimental comparisons," *Experimental Economics*, 14, 375–398.

CAMERER, C. (1995): "Individual Decision Making. Chapter 8 in John Kagel and Al Roth (eds.), The Handbook of Experimental Economics," .

CASON, T. N. AND C. R. PLOTT (2014): "Misconceptions and game form recognition: Challenges to theories of revealed preference and framing," *Journal of Political Economy*, 122, 1235–1270.

CHARNESS, G. AND D. LEVIN (2009): "The origin of the winner's curse: a laboratory study," *American Economic Journal: Microeconomics*, 1, 207–236.

CHEW, S. H. AND W. S. WALLER (1986): "Empirical tests of weighted utility theory," *Journal of Mathematical Psychology*, 30, 55–72.

COHEN, M., I. GILBOA, J.-Y. JAFFRAY, AND D. SCHMEIDLER (2000): "An experimental study of updating ambiguous beliefs," *Risk Decision and Policy*, 5, 123–133.

CRAWFORD, V. AND N. IRIBERRI (2007): "Level-k Auctions: Can a Nonequilibrium Model of Strategic Thinking Explain the Winner's Curse and Overbidding in Private-Value Auctions?" *Econometrica*, 75, 1721–1770.

CROSETTO, P. AND A. FILIPPIN (2013): "The 'bomb' risk elicitation task," *Journal of Risk and Uncertainty*, 47, 31–65.

CROSON, R. T. (1999): "The disjunction effect and reason-based choice in games," *Organizational Behavior and Human Decision Processes*, 80, 118–133.

CUBITT, R., C. STARMER, AND R. SUGDEN (1998): "Dynamic choice and the common ratio effect: An experimental investigation," *The Economic Journal*, 108, 1362–1380.

CUBITT, R. P. (1996): "Rational dynamic choice and expected utility theory," *Oxford Economic Papers*, 48, 1–19.

DAL BÓ, E., P. DAL BÓ, AND E. EYSTER (2016): "The Demand for Bad Policy When Voters Underappreciate Equilibrium Effects," .

DEAN, M. AND P. ORTOLEVA (2015): "Is it all connected? a testing ground for unified theories of behavioral economics phenomena," *mimeo*.

DOMINIAK, A., P. DÜRSCH, AND J.-P. LEFORT (2012): "A dynamic Ellsberg urn experiment," *Games and Economic Behavior*, 75, 625–638.

ELIAZ, K., D. RAY, AND R. RAZIN (2006): "Choice shifts in groups: A decision-theoretic basis," *The American economic review*, 96, 1321–1332.

ELLSBERG, D. (1961): "Risk, ambiguity, and the Savage axioms," *The quarterly journal of economics*, 643–669.

ENKE, B. (2019): "What you see is all there is," *Working Paper*.

ENKE, B. AND F. ZIMMERMANN (2019): "Correlation neglect in belief formation," *The Review of Economic Studies*, 86, 313–332.

ESPONDA, I. (2008): "Behavioral equilibrium in economies with adverse selection," *The American Economic Review*, 98, 1269–1291.

ESPONDA, I. AND E. VESPA (2014): "Hypothetical Thinking and Information Extraction in the Laboratory," *American Economic Journal: Microeconomics*, 6, 180–202.

——— (2018): "Endogenous sample selection: A laboratory study," *Quantitative Economics*, 9, 183–216.

EVANS, J. S. B. (2007): *Hypothetical thinking: Dual processes in reasoning and judgement*, vol. 3, Psychology Press.

EYSTER, E. (2019): "Errors in Strategic Reasoning. D. Bernheim, S. DellaVigna, D. Laibson, eds., Handbook of Behavioral Economics - Foundations and Applications 2, Volume 2," .

EYSTER, E. AND M. RABIN (2005): "Cursed equilibrium," *Econometrica*, 73, 1623–1672.

EYSTER, E. AND G. WEIZSÄCKER (2010): "Correlation neglect in financial decision-making," *Working Paper*.

——— (2016): "Correlation neglect in portfolio choice: Lab evidence," *Available at SSRN 2914526*.

FISCHBACHER, U. (2007): "z-Tree: Zurich toolbox for ready-made economic experiments," *Experimental Economics*, 10, 171–178.

FREDERICK, S. (2005): "Cognitive reflection and decision making," *The Journal of Economic Perspectives*, 19, 25–42.

GABAIX, X. (2014): "A sparsity-based model of bounded rationality," *The Quarterly Journal of Economics*, 129, 1661–1710.

GENNAIOLI, N. AND A. SHLEIFER (2010): "What comes to mind," *The Quarterly journal of economics*, 125, 1399–1433.

GHIRARDATO, P. (2002): "Revisiting Savage in a conditional world," *Economic theory*, 20, 83–92.

GILBOA, I. (2009): *Theory of decision under uncertainty*, vol. 1, Cambridge university press.

GILBOA, I. AND M. MARINACCI (2011): "Ambiguity and the Bayesian paradigm," *Chapter*, 7, 179–242.

GILBOA, I., A. POSTLEWAITE, AND D. SCHMEIDLER (2009): "Is it always rational to satisfy Savage's axioms?" *Economics and Philosophy*, 25, 285–296.

GILBOA, I. AND D. SCHMEIDLER (1995): "Case-based decision theory," *The Quarterly Journal of Economics*, 605–639.

GLAZER, J. AND A. RUBINSTEIN (1996): "An Extensive Game as a Guide for Solving a Normal Game," *Journal of Economic Theory*, 1, 32–42.

GREINER, B. (2015): "Subject pool recruitment procedures: organizing experiments with ORSEE," *Journal of the Economic Science Association*, 1, 114–125.

HALEVY, Y. (2007): "Ellsberg revisited: An experimental study," *Econometrica*, 75, 503–536.

HAMMOND, P. J. (1988): "Consequentialist foundations for expected utility," *Theory and decision*, 25, 25–78.

HARSTAD, R. M. (2000): "Dominant strategy adoption and bidders' experience with pricing rules," *Experimental economics*, 3, 261–280.

HOLLER, M. J. (1983): "Do economics students choose rationally? A research note." *Social Science Information/sur les sciences sociales*.

HUCK, S. AND W. MÜLLER (2012): "Allais for all: Revisiting the paradox in a large representative sample," *Journal of Risk and Uncertainty*, 44, 261–293.

IVANOV, A., D. LEVIN, AND M. NIEDERLE (2010): "Can relaxation of beliefs rationalize the winner's curse?: an experimental study," *Econometrica*, 78, 1435–1452.

JEHIEL, P. AND F. KOESSLER (2008): "Revisiting games of incomplete information with analogy-based expectations," *Games and Economic Behavior*, 62, 533–557.

KAGEL, J., R. HARSTAD, AND D. LEVIN (1987): "Information impact and allocation rules in auctions with affiliated private values: A laboratory study," *Econometrica*, 1275–1304.

KAGEL, J. AND D. LEVIN (2002): *Common value auctions and the winner's curse*, Princeton Univ Pr.

KAGEL, J. H. (1995): "Auctions: A survey of experimental research. John H. Kagel, Alvin E. Roth, eds., The Handbook of Experimental Economics," .

KAGEL, J. H. AND D. LEVIN (1993): "Independent private value auctions: Bidder behaviour in first-, second-and third-price auctions with varying numbers of bidders," *The Economic Journal*, 103, 868–879.

KOVÁŘÍK, J., D. LEVIN, AND T. WANG (2016): "Ellsberg paradox: Ambiguity and complexity aversions compared," *Journal of Risk and Uncertainty*, 52, 47–64.

LEVIN, D., J. PECK, AND A. IVANOV (2016): "Separating bayesian updating from non-probabilistic reasoning: An experimental investigation," *American Economic Journal: Microeconomics*, 8, 39–60.

LI, S. (2017): "Obviously strategy-proof mechanisms," *American Economic Review*, 107, 3257–87.

LOOMES, G. AND R. SUGDEN (1982): "Regret theory: An alternative theory of rational choice under uncertainty," *The economic journal*, 92, 805–824.

LOUIS, P. (2015): "The barrel of apples game: Contingent thinking, learning from observed actions, and strategic heterogeneity," .

MACCRIMMON, K. R. AND S. LARSSON (1979): "Utility theory: Axioms versus paradoxes," in *Expected utility hypotheses and the Allais paradox*, Springer, 333–409.

MACHINA, M. (2008): "Non-Expected Utility Theory. In Steven Durlauf and Lawrence Blume (eds.), The New Palgrave Dictionary of Economics," .

MACHINA, M. J. (1989): "Dynamic consistency and non-expected utility models of choice under uncertainty," *Journal of Economic Literature*, 27, 1622–1668.

MACHINA, M. J. AND M. SINISCALCHI (2013): "Ambiguity and ambiguity aversion," *Economics of Risk and Uncertainty, Handbook in Economics, North Holland, Machina and Viscusi eds*.

MARTIN, D. AND E. MUNOZ-RODRIGUEZ (2019): "Misperceiving Mechanisms: Imperfect Perception and the Failure to Recognize Dominant Strategies," .

MARTÍNEZ-MARQUINA, A., M. NIEDERLE, AND E. VESPA (forthcoming): "Failures in Contingent Reasoning: The Role of Uncertainty," *American Economic Review*.

MOSER, J. (2019): "Hypothetical thinking and the winner's curse: an experimental investigation," *Theory and Decision*, 87, 17–56.

NEBOUT, A. AND D. DUBOIS (2014): "When Allais meets Ulysses: Dynamic axioms and the common ratio effect," *Journal of Risk and Uncertainty*, 48, 19–49.

NEUGEBAUER, T. AND R. SELTEN (2006): "Individual behavior of first-price auctions: The importance of information feedback in computerized experimental markets," *Games and Economic Behavior*, 54, 183–204.

NGANGOUE, K. AND G. WEIZSÄCKER (2018): "Learning from Unrealized versus Realized Prices," .

NICKERSON, R. (2015): *Conditional Reasoning: The Unruly Syntactics, Semantics, Thematics, and Pragmatics of "If"*, Oxford University Press.

REES-JONES, A., R. I. SHORRER, AND C. TERGIMAN (2019): "Correlation Neglect in Student-to-School Matching," *Available at SSRN 3434662*.

ROTH, A. AND J. MURNIGHAN (1978): "Equilibrium behavior and repeated play of the prisoner's dilemma," *Journal of Mathematical Psychology*, 17, 189–198.

SAMET, D. (2015): "The sure-thing principle in epistemic terms," .

SAVAGE, L. J. (1972): *The foundations of statistics*, Courier Corporation.

SHAFIR, E. AND A. TVERSKY (1992): "Thinking through uncertainty: Nonconsequential reasoning and choice," *Cognitive psychology*, 24, 449–474.

SINISCALCHI, M. (2009): "Two out of three ain't bad: A comment on 'the ambiguity aversion literature: A critical assessment'," *Economics and Philosophy*, 25, 335–356.

TVERSKY, A. AND D. KAHNEMAN (1981): "The framing of decisions and the psychology of choice," *Science*, 211, 453–458.

——— (1986): "Rational choice and the framing of decisions," *Journal of business*, S251–S278.

WAKKER, P. P. (2001): "Testing and characterizing properties of nonadditive measures through violations of the sure-thing principle," *Econometrica*, 69, 1039–1059.

WALKER, J., V. SMITH, AND J. COX (1987): "Bidding behavior in first price sealed bid auctions: Use of computerized Nash competitors," *Economics Letters*, 23, 239–244.

WASON, P. C. (1966): "Reasoning," *New horizons in psychology, Vol. 1, BM Foss ed.*

——— (1968): "Reasoning about a rule," *The Quarterly journal of experimental psychology*, 20, 273–281.

ZHANG, L. AND D. LEVIN (2017): "Partition Obvious Preference and Mechanism Design: Theory and Experiment," *Available at SSRN: https://ssrn.com/abstract=2927190*.

Online Appendices

A Theoretical framework

We now use Savage's framework to formalize the connections between the different postulates that form the main basis of our experiment. The main advantages of this approach are that it allows us to link several seemingly unrelated anomalies with a common concept and that it prescribes a very natural experimental test of that concept.[28]

Let Z be a set of consequences and S a set of states. We assume Z and S are finite for simplicity. An act (which we refer to as an option in the main text) $f : S \to Z$ maps states into consequences, and so $f(s)$ is the consequence of choosing act f if the state of the world is s. Let F denote the set of all acts, i.e., all functions from S to Z. We assume that the agent is characterized by a (complete and transitive) relation $\succsim \subseteq F \times F$ over the set of acts, in the sense that the choices that the agent makes in the noncontingent version of the problem can be rationalized by this relation. As usual, we denote the symmetric and asymmetric parts of this relation by $\sim$ and $\succ$, respectively.

We denote the complement of a set A by A^c. Given two acts f, g, the act equal to f for all states in an event A and equal to g in the complement of A is denoted by $f_A g$. Moreover, degenerate acts (acts with $f(s) = f(s')$ for all $s \in S$) are identified with consequences $x, y \in Z$.

We focus on two of Savage's main postulates, separability and dominance. The next postulate corresponds to postulate P2 in Savage (1972) and is one of the main (and most controversial) postulates of subjective expected utility theory.

Separability (SEP; P2 in Savage). For all $A \subseteq S$ and acts $\hat{f}, \hat{g}, h, h' \in F$, $\hat{f}_A h \succsim \hat{g}_A h$ implies $\hat{f}_A h' \succsim \hat{g}_A h'$.

SEP is one of the postulates of Savage on which we will focus attention. To illustrate SEP, consider the following example, where $S = \{s_1, s_2, s_3\}$, $A = \{s_1, s_2\}$, $Z = \{0, 1\}$, and the relevant acts are given by:

[28] Samet (2015) provides a different formalization of STP based on a formal definition of knowledge.

	$\overbrace{\quad}^{A}$		$\overbrace{\quad}^{A^c}$
	s_1	s_2	s_3
f	1	0	1
g	0	1	1
f'	1	0	0
g'	0	1	0

In particular, SEP says that $f \succsim g$ if and only if $f' \succsim g'$. The Ellsberg and Common-Consequence Allais paradoxes are classic examples where this relationship does not hold and, therefore, SEP is violated.

Dominance (DOM) is the second postulate implied by Savage's framework on which we will focus attention.[29]

Dominance (DOM) For all $A \subseteq S$ and all $x, y \in Z$ and $h \in F$, if $x \succsim y$ then $x_A h \succsim y_A h$.

To illustrate DOM, consider the following example, where $S = \{s_1, s_2\}$, $A = \{s_1\}$, $Z = \{x, y, z\}$ are monetary payoffs, and the relevant acts are given by:

	s_1	s_2
f	x	x
g	y	y
f'	x	z
g'	y	z

Suppose that $x \geq y$. It is reasonable for agents to prefer more money to less, so that $f \succsim g$. DOM then implies that $f' \succsim g'$. We will show that this is the kind of dominance relationship that fails in strategic environments, such as auctions and elections. Finally, by taking the space of consequences Z to be a specific set of lotteries, we will show that the common-ratio Allais paradox constitutes a violation of DOM.[30]

Contingent choices

We interpret the relation $\succsim$ as capturing the agent's choices over acts before any uncertainty about the state is realized. For clarity, we will refer to $\succsim$ as the *noncontingent* relation. We also introduce

[29] It is easy to see that DOM is implied by Savage's SEP (P2) and a postulate he calls P3. We use DOM and not P3 in our experimental application because P3 would require us to assume or test that a particular event is non-null.

[30] When consequences are not monetary payoffs, DOM can be interpreted as requiring utility to be state independent; see, e.g., Gilboa (2009).

as a new *primitive*, for each $A \subseteq S$, a (complete and transitive) relation pair $(\succsim_A, \succsim_{A^c})$, with corresponding symmetric and asymmetric parts $\sim_A$ and $\succ_A$, and similarly for event A^c. We refer to the collection $(\succsim_A, \succsim_{A^c})_{A \subseteq S}$ as the *contingent relation*.

There are several interpretations of the contingent relation, and the subsequent theory holds for any of these interpretations. But, for the purposes of this paper, we view $(\succsim_A, \succsim_{A^c})$ as capturing the choices a subject would make if they were asked to commit to a choice conditional on each of the events A and A^c, without knowing which event will realize; this type of elicitation is known in the experimental literature as the strategy method (for a survey, see Brandts and Charness, 2011).[31],[32]

The sure-thing principle places important restrictions on the connections between the noncontingent and contingent relations.

Sure-thing principle (STP*) For all $A \subseteq S$ and acts $f, g \in F$: If $f \succsim_A g$ and $f \succsim_{A^c} g$, then $f \succsim g$. If, in addition, either $f \succ_A g$ or $f \succ_{A^c} g$, then $f \succ g$.

For the purposes of our experiment, the important implication of STP* is that, if f is preferred to g contingent on A and if f is indifferent to g contingent on A^c, then f must be preferred to g according to the noncontingent preference relation. In the experiment, f and g give exactly the same outcome under event A^c. To focus on the issue of contingent thinking and to avoid testing for indifference of two acts that deliver the same consequence under event A^c, we test a version of STP* where subjects are told that both acts yield the same payoff in the event A^c. This version, which we denote by STP, is formalized as follows.

Sure-thing principle (STP) For all $A \subseteq S$ and acts $f, g, h \in F$: If $f_A h \succsim_A g_A h$, then $f_A h \succsim g_A h$. If, in addition, $f_A h \succ_A g_A h$, then $f_A h \succ g_A h$.

STP says that if there is an act that is preferred to another act contingent on A, and if the two acts are equal to each other in the complement of A, then the first act is preferred to the second act

[31] Alternatively, the contingent relation could represent hypothetical or real choices in a sequential problem, after the agent finds out that event A has actually realized. Savage defines a conditional relation that relates to hypothetical choices and is derived from the standard preference relation and the assumption of certain postulates. Ghirardato (2002) introduces the notion of a conditional preference relation as a primitive of its own in the context of sequential decision making, where the decision maker acts after receiving information about the realized state. He shows that the Bayesian model of sequential decision making, where the prior is updated according to Bayes' rule, can be obtained by weakening Savage's P2, provided that certain consistency conditions are imposed on the relation between the original and conditional preferences. Our approach follows this spirit of distinguishing between noncontingent and contingent preferences, although we propose a different weakening of Savage's main postulates that has the sure-thing principle as a central postulate.

[32] Ahn and Ergin (2010) formalize a different kind of framing effect where the primitive is a family of preference relations indexed by partitions, rather than subsets, of the state space. They interpret a partition as a particular description of the contingencies facing the decision maker.

according to the noncontingent preference. Note that, under the assumption that two acts that are equal to each other in all states in A^c satisfy $f \sim_{A^c} g$, it follows that STP is implied by STP*, so a failure of STP also implies a failure of STP*.

Next, we turn to the contingent version of separability.

Contingent separability (C-SEP) For all $A \subseteq S$ and acts $f, g, h, h' \in F$, $f_A h \succsim_A g_A h$ implies $f_A h' \succsim_A g_A h'$.

While we naturally view C-SEP as corresponding to a particular version of SEP in the domain of contingent choices, it is important to highlight the following difference: The definition of SEP fixes the noncontingent relation $\succsim$ and restricts it to satisfy certain conditions for *all* subsets $A \subseteq S$. In contrast, in the definition of C-SEP, the relation $\succsim_A$ changes as a function of the set A.[33]

Our first main result is as follows:

Proposition 1. *If the relations $(\succsim_A, \succsim_{A^c})_{A \subseteq S}$ and $\succsim$ satisfy STP and C-SEP, then $\succsim$ satisfies SEP.*

Proof. Fix $A \subseteq S$ and acts $f, g, h, h' \in F$ such that $f_A h \succsim g_A h$. By STP, $f_A h \succsim_A g_A h$ (otherwise, we would have $g_A h \succ f_A h$, a contradiction of the fact that $f_A h \succsim g_A h$). By C-SEP, $f_A h' \succsim_A g_A h'$. By applying STP once again, $f_A h' \succsim g_A h'$, thus implying SEP. □

Proposition 1 shows that violations of SEP in Savage's setting can be attributed to violations of two different postulates defined over the noncontingent and contingent relations.

We now state a third postulate on the noncontingent and contingent relations.

Contingent dominance (C-DOM) For all $A \subseteq S$ and all $x, y \in Z$ and $h \in F$, if $x \succsim y$ then $x_A h \succsim_A y_A h$.

Our second main result is as follows:

Proposition 2. *If the relations $(\succsim_A, \succsim_{A^c})_{A \subseteq S}$ and $\succsim$ satisfy STP and C-DOM, then $\succsim$ satisfies DOM.*

Proof. Fix $A \subseteq S$ and $x, y \in Z$ and $h \in F$ such that $x \succsim y$. By C-DOM, $x_A h \succsim_A y_A h$. Then STP implies $x_A h \succsim y_A h$, so that DOM is satisfied. □

[33] In particular, it is not true that C-SEP by itself implies SEP.

Proposition 2 shows that violations of DOM in Savage's setting can be attributed to violations of two different postulates defined over the noncontingent and contingent relations.[34]

B Further details on the between-subjects design

Table 7 shows for each question the proportion of subjects who selected f or f', depending on the question. Table 8 reproduces the information presented in Table 1 of the main text, but it also adds the information pertaining to the CC ALLAIS ($) treatment.

		Noncontingent	Contingent	*p-value*
ELLS	% select f in Q1	25.4	52.5	*.002*
	% select f' in Q2	69.5	60.7	*.314*
	# of Participants	59	61	-
CC ALLAIS	% select f in Q1	34.9	47.6	*.150*
	% select f' in Q2	28.6	50.8	*.011*
	# of Participants	63	63	-
CC ALLAIS ($)	% select f in Q1	48.3	38.3	*.280*
	% select f' in Q2	29.3	60.0	*.001*
	# of Participants	58	60	-
AUCT	% select f' in Q1	67.7	87.1	*.010*
	# of Participants	62	62	-
ELECT	% select f' in Q1	15.2	54.0	*.000*
	# of Participants	66	63	-
CR ALLAIS	% select f' in Q1	62.9	65.1	*.802*
	% select f in Q2	77.4	71.4	-
	# of Participants	62	63	-

Table 7: Between-subjects design: Answers by question and testing for STP

Notes: The p-values are computed in the following manner. We run a regression where the left-hand side is a variable that takes value 1 if f (or f' depending on the question) is selected and the right-hand side variable is a treatment dummy (1=contingent). The reported p-value corresponds to the estimated coefficient for the treatment dummy. For ELLS, CC ALLAIS and CC ALLAIS ($) we run one regression per question.

We now show that our treatment effects remain unchanged when we control for observables that we collected at the end of the experiment. We incentivized subjects to provide answers to the Cognitive Reflection Test (CRT, see Frederick, 2005). This test involves three questions (see

[34]These propositions can be extended to the case where preferences are neither complete nor transitive. First, it should be clear that transitivity was not used in either proposition. Second, we used completeness in proposition 1 for convenience; we could drop completeness and modify the definitions of STP and C-SEP accordingly.

	ELLS		CC ALLAIS		CC ALLAIS ($)		AUCT		ELECT		CR ALLAIS	
treatment	NC	C	NC	C	NC	C	NC	C	NC	C	NC	C
# of observ.	59	61	63	63	58	60	62	62	66	63	62	63
% (f, f')	18.6	42.6	22.2	34.9	19.0	28.3	67.7	87.1	15.2	54.0	45.2	55.6
% (g, g')	23.7	29.5	58.7	36.5	41.4	30.0	0	0	0	0	4.9	19.1
% (f, g')	6.8	9.8	12.7	12.7	29.3	10.0	32.3	12.9	84.8	46.0	32.2	15.9
% (g, f')	50.9	18	6.4	15.9	10.3	31.7	0	0	0	0	17.7	9.5
% fail SEP/C-SEP	57.7	27.8	19.1	28.6	39.6	41.7	—	—	—	—	—	—
% fail DOM/C-DOM	—	—	—	—	—	—	32.3	12.9	84.8	46.0	49.9	25.4
p-value	.001		.213		.826		.010		.000		.004	

Table 8: Between-subjects design: summary of results (including CC ALLAIS ($))

Notes: 1) In ELLS and CC ALLAIS, (f, f') indicates the proportion of subjects who selected f in Q1 and f' in Q2. In AUCT and ELECT, (f, f') indicates choices of f in Q* and f' in Q1. We assume that all subjects prefer more money to less and so we impute that all subjects select f in Q*. In CR ALLAIS, (f, f') indicates choices of f in Q2 and f' in Q1.
2) % fail SEP/C-SEP and % fail DOM/C-DOM presents the addition of subjects who chose (f, g') and subjects who chose (g, f').
3) The reported p-value results from a regression in which the unit of observation is a subject. The dependent variable is a dummy that takes value 1 if the subject's choices fail to satisfy the corresponding postulate, and the right-hand side includes a constant and a treatment dummy that takes value 1 if the subject participated in the contingent treatment. The p-values we report correspond to the coefficient estimated for the treatment dummy.

the Procedures Appendix) and is intended to measure the extent to which the respondent gives spontaneous (System 1) vs. reasoned (System 2) answers. Total CRT is a variable that captures how many of the three CRT questions the subject answered correctly; possible values it can take are 0, 1, 2 or 3. We also control for other factors that might affect behavior by eliciting additional information for each subject, including a measure of risk aversion (or ambiguity aversion for the Ellsberg problem) and demographic information, including gender, major (economics related or not), and their year of study (freshman, sophomore, junior, senior, or graduate student). The measure of risk or ambiguity aversion is obtained using the bomb risk elicitation test (BRET, see Crosetto and Filippin, 2013).[35] Table 9 shows descriptive statistics. The measures of risk aversion and cognitive ability are in line with previous literature.[36]

[35]For ambiguity aversion, we modify the experiment by creating ambiguity about the location of the bomb; see the Online Appendix for details.

[36]Frederick (2005) finds a mean of 1.24 for Total CRT and Crosetto and Filippin (2013) find a mean of 46.5 for BRET (Risk). There is no comparable exercise for ambiguity aversion.

Variable	Mean	Median	Std. dev.	Observations
Total CRT	1.44	1.00	1.11	742
BRET (Ambiguity)	40.45	40.00	15.50	120
BRET (Risk)	38.57	39.50	14.53	622
Econ Major	0.202	-	-	742
Female	0.579	-	-	742
Junior or older	0.536	-	-	742

Table 9: Between-subjects design: Summary statistics of observables

Table 10 shows the results of regressions using data from the between-subjects design, where the rand-hand side is a dummy variable that takes value 1 if the subject chooses f (or f' depending on the question). The regression controls for the treatment and the observables described above. As shown by the table, the treatment effect highlighted in Finding #3 continues to hold (the results are also in line with the p-values reported in Table 7), but the observables are mostly statistically not significant. Although the measure of risk or ambiguity aversion tends to have the expected sign, i.e., a lower aversion measure -higher value for the BRET variable- is associated with less risky choice the coefficient is not significant.[37] Table 11 shows the results of regressions of whether or not behavior across Q1 and Q2 is inconsistent for each problem, controlling for the treatment and the observables described above. Once again, the treatment effects reported in Finding #2 continue to hold.

[37]Note that in AUCT and ELECT, there is an optimal choice and there is no reason why risk aversion should explain who is sophisticated and makes the right choice.

	ELLS		CC ALLAIS		CC ALLAIS($)		AUCT	ELECT	CR ALLAIS
	Q1	Q2	Q1	Q2	Q1	Q2	Q1	Q1	Q1
Contingent	0.243***	−0.116	0.104	0.220**	−0.033	0.326***	0.193**	0.389***	0.034
	(0.086)	(0.089)	(0.090)	(0.087)	(0.090)	(0.090)	(0.075)	(0.077)	(0.088)
Total CRT	−0.096**	−0.058	0.015	−0.016	−0.029	0.011	0.037	0.002	−0.049
	(0.044)	(0.046)	(0.042)	(0.041)	(0.042)	(0.042)	(0.036)	(0.038)	(0.041)
BRET	−0.003	−0.003	−0.000	−0.002	0.001	−0.006*	−0.003	0.002	−0.002
	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.004)	(0.002)	(0.003)	(0.003)
Econ Major	0.258**	0.030	−0.059	0.040	0.035	−0.011	−0.079	−0.160	−0.118
	(0.107)	(0.111)	(0.104)	(0.101)	(0.117)	(0.118)	(0.096)	(0.105)	(0.119)
Female	−0.121	−0.056	0.095	0.068	0.154	0.134	−0.064	−0.113	0.096
	(0.091)	(0.095)	(0.094)	(0.091)	(0.107)	(0.107)	(0.087)	(0.087)	(0.098)
Junior or Older	0.018	−0.106	−0.138	−0.186**	−0.307***	−0.076	0.068	0.008	−0.029
	(0.103)	(0.107)	(0.091)	(0.088)	(0.092)	(0.093)	(0.078)	(0.081)	(0.088)
Constant	0.485***	0.990***	0.372**	0.403**	0.477***	0.437**	0.760***	0.144	0.762***
	(0.168)	(0.175)	(0.162)	(0.157)	(0.182)	(0.182)	(0.165)	(0.150)	(0.173)
Observations	120	120	126	126	118	118	124	129	125

Table 10: Between-subjects design: Linear regression results that test for STP

Note: These regressions use data from the between-subjects design. Left-hand side variable: We run a regression where the left-hand side is a dummy variable that takes value 1 if the subject chooses f (or f' depending on the question). Right-hand side variables: Contingent (1=Contingent Treatment), Total CRT (number of correct answers in CRT test; it takes integer values from 0 to 3), BRET (Box for which subjects stopped the collecting process, a higher number indicates willingness to take more risk/ambiguity), Econ Major (1=Economics/Accounting/Business Economics Majors), Female (1=Subject is a female), Junior or older (1=Subject is a junior, senior or graduate student). Standard errors between parentheses. (*) Significant at 10% level, (**) Significant at 5% level, (***) Significant at 1% level.

	ELLS	CC ALLAIS	CC ALLAIS($)	CR ALLAIS
Contingent	-0.297^{***}	0.088	0.071	-0.284^{***}
	(0.088)	(0.077)	(0.092)	(0.087)
Total CRT	0.078^{*}	0.007	-0.095^{**}	-0.010
	(0.046)	(0.036)	(0.043)	(0.040)
BRET	-0.002	-0.003	0.005	0.005
	(0.003)	(0.003)	(0.004)	(0.003)
Econ Major	-0.110	0.128	-0.202^{*}	-0.229^{*}
	(0.110)	(0.089)	(0.120)	(0.117)
Female	0.144	0.178^{**}	-0.039	0.028
	(0.094)	(0.080)	(0.109)	(0.096)
Junior or Older	0.013	0.033	-0.050	-0.056
	(0.106)	(0.078)	(0.094)	(0.087)
Constant	0.506^{***}	0.153	0.402^{**}	0.404^{**}
	(0.174)	(0.139)	(0.185)	(0.170)
Observations	120	126	118	125

Table 11: Between-subjects design: Linear regression results for inconsistencies

Note: These regressions use data from the between-subjects design. Left-hand side variable: Dummy that takes value 1 if choice is inconsistent. The choice is consistent if the subject selects (f, g') or (g, f'). Right-hand side variables: Contingent (1=Contingent Treatment), Total CRT (number of correct answers in CRT test; it takes integer values from 0 to 3), BRET (Box for which subjects stopped the collecting process, a higher number indicates willingness to take more risk/ambiguity), Econ Major (1=Economics/Accounting/Business Economics Majors), Female (1=Subject is a female), Junior or older (1=Subject is a junior, senior or graduate student). Standard errors between parentheses. (*) Significant at 10% level, (**) Significant at 5% level, (***) Significant at 1% level.

C Further details on the within experiment

A total of 131 subjects participated in the *within* experiment. Subjects first faced noncontingent treatments and afterwards the corresponding contingent version. The order in which problems were faced was as follows: ELLS, ELECT, CC ALLAIS, AUCT, CR ALLAIS. The only criterion to select the order was to collate decision-theory based and game-theory based problems. The treatment consisted of ten parts plus a bonus part (Part 11) that involves the cognitive reflexion test, a risk aversion task and answers to demographic questions. For details on the instructions and other procedures see the Procedures Appendix.

Table 12 reproduces the information of Table 1, using answers from the within-subjects design. The information on choices inconsistent with SEP/C-SEP and DOM/C-DOM are informed on Table 2 of the main text. Table 12 adds the decomposition of the frequency of pair of choices.

	ELLS		CC ALLAIS		AUCT		ELECT		CR ALLAIS	
treatment	NC	C	NC	C	NC	C	NC	C	NC	C
# of observ.	131	131	131	131	131	131	131	131	131	131
% (f, f')	19.9	35.9	19.1	29.7	77.1	93.9	21.4	59.5	55.7	69.5
% (g, g')	26.7	34.4	45.8	44.3	0.0	0.0	0.0	0.0	9.2	13.7
% (f, g')	9.9	13.7	24.4	14.5	22.9	6.1	78.6	40.5	12.2	11.5
% (g, f')	43.5	16.0	10.7	11.5	0.0	0.0	0.0	0.0	22.9	5.3
% fail SEP/C-SEP	53.4	29.7	35.1	26.0	—	—	—	—	—	—
% fail DOM/C-DOM	—	—	—	—	22.9	6.1	78.6	40.5	35.1	16.8
p-value	.000		.058		.000		.000		.000	

Table 12: *Within* experiment: summary of results

Notes: 1) In ELLS and CC ALLAIS, (f, f') indicates the proportion of subjects who selected f in Q1 and f' in Q2. In AUCT and ELECT, (f, f') indicates choices of f in Q* and f' in Q1. We assume that all subjects prefer more money to less and so we impute that all subjects select f in Q*. In CR ALLAIS, (f, f') indicates choices of f in Q2 and f' in Q1.
2) % fail SEP/C-SEP and % fail DOM/C-DOM presents the addition of subjects who chose (f, g') and subjects who chose (g, f').
3) Column (1) of Table 16 reports a linear regression in which the unit of observation is a subject's answer in the contingent or noncontingent version and the right-hand side includes a set of controls. Among the controls there is a variable that results from the interaction of the problem dummy and the treatment dummy (contingent=1). The reported p-value for each problem corresponds to the p-value of the coefficient for this interaction variable.

Table 13 reproduces the information of Table 7 using answers in the within-subjects design that allow us to test directly for failures of STP.

		Noncontingent	Contingent	*p-value*
ELLS	% select f in Q1	29.8	49.6	*.001*
	% select f' in Q2	63.4	51.9	*.051*
	# of Participants	131	131	-
CC ALLAIS	% select f in Q1	43.5	44.3	*.854*
	% select f' in Q2	29.8	41.2	*.025*
	# of Participants	131	131	-
AUCT	% select f' in Q1	77.1	93.9	*.000*
	# of Participants	131	131	-
ELECT	% select f' in Q1	21.4	59.5	*.000*
	# of Participants	131	131	-
CR ALLAIS	% select f' in Q1	67.9	80.9	*.003*
	% select f in Q2	78.6	74.8	-
	# of Participants	131	131	-

Table 13: *Within* experiment: Answers by question and testing for STP

Notes: The p-values are computed in the following manner. We run a linear regression where the left-hand side is a variable that takes value 1 if f (or f' depending on the question) is selected and the right-hand side variable is a treatment dummy (1=contingent). The reported p-value corresponds to the estimated coefficient for the treatment dummy. For ELLS and CC ALLAIS we run one regression per question. Standard errors are clustered by subject.

Table 14 provides summary statistics on the observables we collected at the end of the session in the *within* experiment. The statistics are comparable to those of the between-subjects population reported in Table 9 of Online Appendix B.

Variable	Mean	Median	Std. dev.	Observations
Total CRT	1.44	1.0	1.2	131
BRET (Risk)	43.8	44.0	17.1	131
Econ Major	0.191	-	-	131
Female	0.603	-	-	131
Junior or older	0.466	-	-	131

Table 14: *Within* experiment: Summary statistics on observables

Table 15 reproduces the regressions reported in Table 10, but now using data from the within-subjects design. In both cases we find that controlling for other observables does not change the findings that we report. Table 15 shows that the documented treatment effects of inconsistencies hold if we control for observables. In both regressions the left-hand side variable is a dummy that

captures whether the subject was consistent (=1) or not (=0) in a given contingent or noncontingent problem. The right-hand side includes a dummy for the problem (the dummy for Ellsberg is excluded) and dummies that interact the problem with a dummy that takes value 1 if the observation is from a contingent treatment. Column (2), in addition, controls for other observables we collected. The interaction dummies are all negative, indicating that inconsistencies go down in all problems when the treatment is contingent. The main point is that these treatment effects are virtually unchanged if we control for other observables.

	ELLS		CC ALLAIS		AUCT	ELECT	CR ALLAIS
	Q1	Q2	Q1	Q2	Q1	Q1	Q1
Contingent	0.198^{***}	-0.115^{*}	0.008	0.115^{**}	0.168^{***}	0.382^{***}	0.130^{***}
	(0.060)	(0.059)	(0.042)	(0.051)	(0.037)	(0.050)	(0.043)
Total CRT	-0.027	0.027	-0.006	-0.041	0.028	0.085^{***}	-0.079^{***}
	(0.024)	(0.026)	(0.033)	(0.028)	(0.023)	(0.024)	(0.025)
BRET	0.001	-0.004^{**}	-0.001	-0.003^{*}	-0.001	-0.004^{**}	-0.002
	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)
Econ Major	-0.037	0.118	-0.242^{***}	-0.157^{**}	0.043	-0.110	0.132^{*}
	(0.063)	(0.078)	(0.085)	(0.073)	(0.057)	(0.080)	(0.069)
Female	-0.026	0.039	0.142^{*}	0.129^{**}	0.014	-0.189^{***}	-0.137^{**}
	(0.065)	(0.066)	(0.076)	(0.065)	(0.049)	(0.067)	(0.060)
Junior or Older	-0.010	-0.059	-0.060	-0.056	0.059	0.018	0.019
	(0.059)	(0.063)	(0.076)	(0.063)	(0.048)	(0.055)	(0.060)
Constant	0.339^{***}	0.767^{***}	0.483^{***}	0.464^{***}	0.725^{***}	0.393^{***}	0.947^{***}
	(0.125)	(0.118)	(0.143)	(0.112)	(0.088)	(0.118)	(0.105)
Observations	262	262	262	262	262	262	262

Table 15: *Within* experiment: Linear regression results that test for STP

Note: These regressions use data from the within-subjects design. Left-hand side variable: We run a regression where the left-hand side is a dummy variable that takes value 1 if the subject chooses f (or f' depending on the question). Right-hand side variables: Contingent (1=Contingent Treatment), Total CRT (number of correct answers in CRT test; it takes integer values from 0 to 3), BRET (Box for which subjects stopped the collecting process, a higher number indicates willingness to take more risk/ambiguity), Econ Major (1=Economics/Accounting/Business Economics Majors), Female (1=Subject is a female), Junior or older (1=Subject is a junior, senior or graduate student). Standard errors between parentheses are clustered by subject. (*) Significant at 10% level, (**) Significant at 5% level, (***) Significant at 1% level.

	(1) inconsistent	(2) inconsistent
Contingent × ELLS	-0.237^{***}	-0.237^{***}
	(0.058)	(0.059)
Contingent × CC ALLAIS	-0.092^{*}	-0.092^{*}
	(0.048)	(0.048)
Contingent × AUCT	-0.168^{***}	-0.168^{***}
	(0.036)	(0.036)
Contingent × ELECT	-0.382^{***}	-0.382^{***}
	(0.049)	(0.049)
Contingent × CR ALLAIS	-0.183^{***}	-0.183^{***}
	(0.043)	(0.043)
CC ALLAIS	-0.183^{***}	-0.183^{***}
	(0.060)	(0.060)
AUCT	-0.305^{***}	-0.305^{***}
	(0.054)	(0.054)
ELECT	0.252^{***}	0.252^{***}
	(0.056)	(0.056)
CR ALLAIS	-0.183^{***}	-0.183^{***}
	(0.056)	(0.056)
Total CRT		-0.021^{*}
		(0.011)
BRET		0.001
		(0.001)
Econ Major		-0.064^{**}
		(0.029)
Female		0.047
		(0.030)
Junior or Older		-0.048^{*}
		(0.026)
Constant	0.534^{***}	0.536^{***}
	(0.044)	(0.069)
Observations	1310	1310

Table 16: *Within* experiment:Linear regression results for inconsistencies

Note: These regressions use data from the within-subjects design. Left-hand side variable: Dummy that takes value 1 if choice is inconsistent. The choice is consistent if the subject selects (f, g') or (g, f'). Right-hand side variables: Contingent (1=Contingent Treatment), Total CRT (number of correct answers in CRT test; it takes integer values from 0 to 3), BRET (Box for which subjects stopped the collecting process, a higher number indicates willingness to take more risk/ambiguity), Econ Major (1=Economics/Accounting/Business Economics Majors), Female (1=Subject is a female), Junior or older (1=Subject is a junior, senior or graduate student). Standard errors between parentheses are clustered by subject. (*) Significant at 10% level, (**) Significant at 5% level, (***) Significant at 1% level.

Finally, we provide a comparison across the between-subjects and the *within* experiment. Table 17 compares answers in the between-subjects design and the *within* experiment question by question. We find no statistical difference in the comparison question by question with one exception. In the case of the contingent CR ALLAIS treatment, we find that more subjects select f' in question 1 of the the *within* experiment relative to the between-subjects design.

		Noncontingent			Contingent		
		Between	Within	*p-value*	Between	Within	*p-value*
ELLS	% select f in Q1	25.4	29.8	*.541*	52.5	49.6	*.716*
	% select f' in Q2	69.5	63.4	*.414*	60.7	51.9	*.259*
	# of Participants	59	131		61	131	-
CC ALLAIS	% select f in Q1	34.9	43.5	*.256*	47.6	44.3	*.663*
	% select f' in Q2	28.6	29.8	*.864*	50.8	41.2	*.211*
	# of Participants	63	131		63	131	-
AUCT	% select f' in Q1	67.7	77.1	*.168*	87.1	93.9	*.111*
	# of Participants	62	131		62	131	-
ELECT	% select f' in Q1	15.2	21.4	*.299*	54.0	59.5	*.464*
	# of Participants	66	131		63	131	-
CR ALLAIS	% select f' in Q1	62.9	67.9	*.492*	65.1	80.9	*.016*
	% select f in Q2	77.4	78.6	*.851*	71.4	74.8	*.619*
	# of Participants	62	131		63	131	-

Table 17: Comparing answers across the between-design and *within* experiment

Notes: The p-values are computed in the following manner. We run regression where the left-hand side is a variable that takes value 1 if f (or f' depending on the question) is selected and the right-hand side variable is a treatment dummy (1=subject participated in the between-subjects design). The reported p-value corresponds to the estimated coefficient for the design dummy. For ELLS and CC ALLAIS and CR ALLAIS we run one regression per question.

D Further details on the within+ experiment

In this appendix we provide further details on results for the *within+* experiment. In the *within+* experiment, subjects faced several parameterizations of each of the five problems. The experiment was divided into eight parts plus a bonus part. Parts 1-4 capture noncontingent treatments, and parts 5-8 consist of the corresponding contingent treatment. That is, for each noncontingent version of the question in parts 1-4 there is a contingent version in parts 5-8. The number of parameterizations in each part is selected so that each comparison of noncontingent parts generates ten tests of STP. For example, parts 1 and 5 involve five parameterizations of the Ellsberg problem, which allow for ten STP tests (one for each of two questions of each parameterization). Parts 2 and 8 consist of ten parameterizations of the Election problem, parts 3 and 9 include three parameterizations of the CC Allais problem and four parameterizations of the CR Allais problem, and parts 4 and 8 are built using ten parameterizations of the Auction problem. Four of the eight parts were randomly selected for payment. In each part selected for payment one question was randomly selected for payment. The Bonus part includes cognitive reflexion test (CRT) questions, and the bomb risk-elicitation task. Detailed procedures are provided in the Procedures Appendix. There were 119 participants in total, the average payment was $27.1, and each session lasted about two hours.

Parameterizations

Ellsberg problems

Three of the five parameterizations are based on the one-jar Ellsberg problem and the remaining two use the two-jar Ellsberg problem. The three parameterizations of the one-jar Ellsberg problem are presented in Figure 9. For each parameterization there is a jar that contains Red, Yellow and Blue balls. The number of balls for one of the three colors is set to 30, and the remaining 60 balls are distributed among the other two colors depending on a number provided by a monitor, which is not revealed to participants until the end of the session. For example, Parameterization 1 coincides with the case used in the between- and within-subjects design and involves 30 Red balls, m_1 Yellow balls and $60 - m_1$ Blue balls, where m_1 is a number between 0 and 60 provided by a monitor.[38] A ball is then randomly selected from the jar, with all balls having an equal chance

[38] Specifically, at the beginning of each session five subjects were selected as monitors. Each monitor wrote five numbers in a green sheet of paper and five numbers in a pink sheet of paper. Each number is an integer in $[0, 60]$. Monitors did not know what these numbers would be used for, were paid after they wrote the numbers and left the session. We used the numbers that the monitors provided to generate the composition of the jar in each parameterization. In noncontingent (contingent) versions we used the numbers in the green (pink) sheet of paper. For example, m_3 (which is used in Parameterization 3) represents the number written by monitor #3 (the one written in the green sheet would

Parameterization			Primitives		
1	Jar		R(30)	Y(m_1)	B(60-m_1)
	States		R	Y	B
	Q1	f	\$10	0	\$10
		g	0	\$10	\$10
	Q2	f'	\$10	0	0
		g'	0	\$10	0
2	Jar		R(60-m_2)	Y(30)	B(m_2)
	States		R	Y	B
	Q3	f	\$5	\$5	0
		g	\$5	0	\$5
	Q4	f'	0	\$5	0
		g'	0	0	\$5
3	Jar		R(m_3)	Y(60-m_3)	B(30)
	States		R	Y	B
	Q5	f	0	\$15	\$15
		g	\$15	\$15	0
	Q6	f'	0	0	\$15
		g'	\$15	0	0

Figure 9: *Within*+ experiment: Ellsberg Problem Parameterizations (One Jar)

Notes: (1) For each parameterization the row entitled 'Jar' describes the composition of the jar for the two questions. There are (R)ed, (Y)ellow and (B)lue balls in the jar and between parentheses we provide the number of balls of each color. The parameters m_1,..., m_5 were provided by 'monitors' before the session started and were not revealed to participants. Specifically, at the beginning of each session five subjects were selected as monitors. Each monitor wrote five numbers in a green sheet of paper and five numbers in a pink sheet of paper. Each number is an integer in $[0, 60]$. Monitors did not know what these numbers would be used for, were paid after they wrote the numbers and left the session. We used the numbers that the monitors provided to generate the composition of the jar in each parameterization. In noncontingent (contingent) versions we used the numbers in the green (pink) sheet of paper. For example, m_3 represents the number written by monitor #3 (the one written in the green sheet would be used for the noncontingent version and the one in the pink sheet for the contingent version). For further details see the Procedures Appendix.
(2) The row 'States' includes the possible states of the world for these problems. The state of the world is determined by one randomly drawn ball from the jar.

Parameterization		Primitives				
4		Jar 1	R(30), B(30)			
		Jar 2	R(m_4), B(60-m_4)			
		States	RR	RB	BR	BB
	Q7	f: Get \$10 if Red from Jar 1	\$10	\$10	0	0
		g: Get \$10 if Red from Jar 2	10	0	\$10	0
	Q8	f': Get \$10 if Blue from Jar 2	0	\$10	0	\$10
		g': Get \$10 if Blue from Jar 1	0	0	\$10	\$10
5		Jar 1	Y(30), B(30)			
		Jar 2	Y(60-m_5), B(m_5)			
		States	YY	YB	BY	BB
	Q9	f: Get \$15 if Blue from Jar 2	0	\$15	0	\$15
		g: Get \$15 if Yellow from Jar 1	\$15	\$15	0	0
	Q10	f': Get \$15 if Blue from Jar 1	0	0	\$15	\$15
		g': Get \$15 if Yellow from Jar 2	\$15	0	\$15	0

Figure 10: *Within*+ experiment: Ellsberg Problem Parameterizations (Two Jars)

Notes: (1) For each parameterization the rows entitled 'Jar 1' and 'Jar 2' describes the compositions of the jars for the two questions. The number of (R)ed, (Y)ellow and (B)lue balls in each jar is provided between parentheses. The parameters m_1,..., m_5 were provided by 'monitors' before the session started and were not revealed to participants. Specifically, at the beginning of each session five subjects were selected as monitors. Each monitor wrote five numbers in a green sheet of paper and five numbers in a pink sheet of paper. Each number is an integer in $[0, 60]$. Monitors did not know what these numbers would be used for, were paid after they wrote the numbers and left the session. We used the numbers that the monitors provided to generate the composition of the jar in each parameterization. In noncontingent (contingent) versions we used the numbers in the green (pink) sheet of paper. For example, m_4 represents the number written by monitor #4 (the one written in the green sheet would be used for the noncontingent version and the one in the pink sheet for the contingent version). For further details see the Procedures Appendix.
(2) The row 'States' includes the possible states of the world for these problems. The state of the world is determined by one randomly drawn ball from each jar. For example, if the state is 'YB' it means that the ball drawn from Jar 1 is Yellow and the ball drawn from Jar 2 is Blue.

of being selected. The possible colors of the selected ball are referred to in the table as possible 'States,' and each parameterization consists of two questions. For example, parameterization 1 consists of Question 1 (Q1) and Question 2 (Q2), and each question involves two alternatives as described in the table. In the laboratory subjects were presented the questions in the same order that questions are presented in the table. Parameterizations 2 and 3 are similar to parameterization 1 except that we switch the colors that are affected by the monitor's choice and we change the maximum amount of money that can be earned.

After subjects finished with the one-jar Ellsberg problems, instructions were read for the two-jar Ellsberg problem described in parameterization 4 of Figure 10. In the two-jar problem, both jars are composed of 60 balls that could be of two possible colors. Jar 1 is known to contain an equal amount of balls of each possible color. The composition of Jar 2 is determined by a number submitted by the monitor. For example, in parameterization 4, Jar 2 contains m_4 Red balls

be used for the noncontingent version and the one in the pink sheet for the contingent version). For further details see the Procedures Appendix.

Parameterization	Jar	Computer 1's voting rule	Computer 2's voting rule
1	7W, 3B	Votes White	Votes for Color of Ball
2	8W, 2B	Votes White	Votes for Color of Ball
3	9W, 1B	Votes for Color of Ball	Votes White
4	2W, 8B	Votes for Color of Ball	Votes Black
5	3W, 7B	Votes Black	Votes for Color of Ball
6	6W, 4B	Votes for Color of Ball	Votes White
7	4W, 6B	Votes for Color of Ball	Votes Black
8	3W, 7B	Votes White	Votes for Color of Ball
9	8W, 2B	Votes for Color of Ball	Votes Black
10	1W, 9B	Votes for Color of Ball	Votes White

Figure 11: *Within*+ experiment: Election Problem Parameterizations

Notes: (1) The 'Jar' column describes the composition of the jar, where W and B represent White and Black. For example, '7W, 3B' indicates that the ten-ball jar is composed of seven white and three black balls.
(2) In parameterizations 1 through 7 it is optimal to vote for the color with the least number of balls in the jar. In parameterizations 8-10, it is optimal to vote for the color that represents most balls in the jar.

and $60 - m_4$ Blue balls, where m_4 is an integer between 0 and 60 written down by a monitor. Subjects are told that one ball is independently selected from each jar. In the first question of parameterization 4 (displayed as Question 7) subjects have to select between: (f) get \$10 if a Red ball is selected from Jar 1 and (g) get \$10 if a Red ball is selected from Jar 2. Figure 10 presents payoffs in the state space for this problem. A state involves a ball from Jar 1 and a ball from Jar 2. For parameterization 4, state RB captures the case in which a Red ball is selected from Jar 1 and a Blue ball is selected from Jar 2. Thus, a subject who selects f would get paid \$10 if the state is RR or RB (the Red ball is selected from Jar 1). In the second question (displayed as Question 8) the options are: (f') get \$10 if a Blue ball is selected from Jar 2 and (g') get \$10 if a Blue ball is selected from Jar 1. A subject who selects f in Q7 presumably believes that there are fewer than 30 Blue balls in Jar 2, so that when faced with Q8 her choice would be consistent with SEP if she selects f'. It is common, however, to observe choices of f in Q7 and g' in Q8, which are consistent with ambiguity aversion, but inconsistent with SEP. Parameterization 5 also involves a two-jar Ellsberg problem and subjects face it after parameterization 4. In this case jars can have Yellow or Blue balls, the maximum amount that can be earned is \$15 but the structure of the test of SEP is similar to that of parameterization 4.

Election problems

Figure 11 shows the ten parameterizations that correspond to parts 2 and 6. Parameterization 1 coincides with the parameterization used in the between- and within-subjects designs. In parame-

Parameterization			Primitives		
1	Jar		R(1)	Y(10)	B(89)
	States		R	Y	B
	Q1	f	\$100M	\$100M	\$100M
		g	0	\$500M	\$100M
	Q2	f'	\$100M	\$100M	0
		g'	0	\$500M	0
2	Jar		R(5)	Y(20)	B(75)
	States		R	Y	B
	Q3	f	\$10	\$10	\$20
		g	0	\$20	\$20
	Q4	f'	\$10	\$10	0
		g'	0	\$20	0
3	Jar		R(89)	Y(1)	B(10)
	States		R	Y	B
	Q5	f	\$20	\$20	\$20
		g	\$20	0	\$50
	Q6	f'	0	\$20	\$20
		g'	0	0	\$50

Figure 12: *Within*+ experiment: CC Allais Problem Parameterizations

Notes: (1) For each parameterization the row entitled 'Jar' describes the composition of the jar for the two questions. There are (R)ed, (Y)ellow and (B)lue balls in the jar and between parentheses we provide the number of balls of each color.
(2) The row 'States' includes the possible states of the world for these problems. The state of the world is determined by one randomly drawn ball from the jar.
(3) M stands for Millions of dollars. In parameterization 1 payoffs are hypothetical. In parameterizations 2 and 3 payoffs are in dollars.

terizations 1 through 7 it is optimal to vote for the minority color, that is, it is optimal to vote for the color with fewer balls in the jar. These are the cases for which naive and optimal voting differ. A naive voter would always vote for the color with more balls in the jar. We also include three parameterizations (8, 9 and 10) for which optimal and naive behavior coincide. Subjects faced the parameterizations in the following order. They were first presented with one parameterization in which it is optimal to vote for the minority color (parameterization 1) and then with one for which it is optimal to vote for the majority color (parameterization 8). The order in which the remaining eight parameterizations were presented was randomized at the subject level.

Allais problems

From the perspective of explaining a task the instructions for CC Allais and CR Allais problems are the same. In each question of each parameterization, subjects are first presented with a jar of a known composition from which a ball will be selected. They are then presented with two

options to choose from and each option describes how payoffs depend on the selected ball. Given that instructions are the same in all Allais problems we merged both type of problems into a single part of our within+ treatment (implemented as part 3 in the noncontingent, and part 7 in the contingent version). Taken together there are ten tests of STP in Allais problems: six in CC Allais parameterizations (one from each question) and four from CR Allais parameterizations (one from each parameterization).

Figure12 presents all parameterizations for CC Allais problems. Parameterization 1 is the only parameterization that involves hypothetical payoffs, and coincides with the parameterizations in the between- and within-subjects designs. Parameterization 2 is based on the CC Allais test of Chew and Waller (1986). Parameterization 3 uses the same jar composition of parameterization 1 (albeit switching colors) but increases monetary payoffs relative to Parameterization 2.

Figure 13 includes all parameterizations of CR Allais problems. Parameterization 1 coincides with the parameterization used in between- and within-subjects designs. Parameterization 4 involves similar probabilities (switching colors) relative to parameterization 1, but there is a change in monetary payoffs used. Parameterizations 2 and 3 are based on the CR Allais parameterization of Chew and Waller (1986).

Overall there are fourteen questions related to Allais problems, of which only two involve hypothetical payments. To avoid any possible confusion, parameterization 1 of the CC Allais question, which involves hypothetical payments, was always presented last. That is, the 13th and 14th question that subjects faced in the Allais problems part where Q1 and Q2 of the CC Allais parameterization 1. The first twelve questions in this part involved real monetary payoffs. The order was fixed for the first two and the last parameterization. That is, subjects first faced parameterization 2 of the CC Allais problems first, then parameterization 1 of the CR Allais problems and parameterization 1 of the CC Allais problems last. The remaining four parameterizations (one of CC Allais and three of CR Allais problems) were faced in a random order. That is, the order was randomized at the parameterization level but not at the question level. For example, assume that parameterization 4 of the CC Allais problems was randomly selected to be third in the order (recall that the order of the first two and the last parameterization was not randomized, but fixed), then after subjects faced the first two parameterizations they would be presented with Q7 and Q8 (in that order) of Figure 13.

Parameterization		Primitives			
1	Q1	Jar	R(12)	Y(3)	B(85)
		States	R	Y	B
		f'	\$10	\$10	0
		g'	0	\$20	0
	Q2	Jar	R(80)	Y(20)	
		States	R	Y	
		f	\$4	\$4	
		g	0	\$5.3	
2	Q3	Jar	R(75)	Y(5)	B(20)
		States	R	Y	B
		f'	\$10	\$5	\$5
		g'	\$10	0	\$10
	Q4	Jar		Y(20)	B(80)
		States		Y	B
		f		\$5	\$5
		g		0	\$10
3	Q5	Jar	R(20)	Y(75)	B(5)
		States	R	Y	B
		f'	\$7	\$14	\$7
		g'	\$14	\$14	0
	Q6	Jar	R(80)		B(20)
		States	R		B
		f	\$7		\$7
		g	\$14		0
4	Q7	Jar	R(85)	Y(12)	B(3)
		States	R	Y	B
		f'	0	\$8	\$8
		g'	0	\$11	0
	Q8	Jar		Y(80)	B(20)
		States		Y	B
		f		\$8	\$8
		g		\$11	0

Figure 13: *Within*+ experiment: CR Allais Problem Parameterizations

Notes: (1) For each question the row entitled 'Jar' describes the composition of the jar. There can be (R)ed, (Y)ellow and (B)lue balls in each jar and between parentheses we provide the number of balls of each color.
(2) The row 'States' includes the possible states of the world for each question. The state of the world is determined by one randomly drawn ball from the jar.

Auction Problems

Parameterization	X	Y
1	5.5	3
2	9	5
3	8.25	4
4	7.5	4
5	7.75	4
6	5	0
7	4.75	0
8	6.25	1
9	5.5	0
10	8.5	2

Figure 14: *Within*+ experiment: Auction Problems Parameterizations

Notes: (1) X refers to the following parameter. If the number chosen by the subject is higher than the number on the drawn card, the payoff is X minus the number on the card.
(2) Y refers to the following parameter. If the number chosen by the subject is lower than the number on the drawn card, the payoff is Y.

Parts 4 and 8 involve ten parameterizations of the auction problem, which are shown in Figure 14. Each case is parameterized by two numbers, X and Y. Recall that a first step in each auction problem is that the computer randomly selects a card out of three possible cards, where the numbers on each card are 8.5, 4.5 and 0.5. The subject is not told which card was randomly selected but has to select an integer between 1 and 8. If the subject's choice is higher than the number on the drawn card, the subject's payoff is X minus the number on the card. If the subject's choice is lower than the number on the drawn card, the subject's payoff is Y.

In the first five parameterizations it is optimal to select a number lower than or equal to four (underbidding is optimal). In parameterizations 6-10, it is optimal to select a number higher than or equal to five (overbidding is optimal). Parameterization 1 corresponds to the parameterization that subjects face in the between- and within-subjects designs and was the first parameterization that all subjects faced in the within+ design. The order of the remaining nine parameterizations was randomly determined at the subject level.

Results

Tables 18, 19, 20, and 21 show for each parameterization of each problem the proportion of subjects who select the first alternative (f or f') in each question.

Parameterization		Noncontingent	Contingent
1	% select f in Q1	19.3	49.6
	% select f' in Q2	63.9	58.8
2	% select f in Q3	31.9	44.5
	% select f' in Q4	52.1	60.5
3	% select f in Q5	23.5	46.2
	% select f' in Q6	58.0	56.3
4	% select f in Q7	79.8	64.7
	% select f' in Q8	37.0	52.9
5	% select f in Q9	27.7	59.7
	% select f' in Q10	84.0	56.3

Table 18: *Within*+ experiment: Answers Contingent and Noncontingent of ELLS problems

Notes: For a description of each parameterization, see figures 9 and 10.

Parameterization		Noncontingent	Contingent
1	% select f in Q1	49.6	45.4
	% select f' in Q2	30.3	49.6
2	% select f in Q1	49.6	44.5
	% select f' in Q2	40.3	51.3
3	% select f in Q1	32.8	14.3
	% select f' in Q2	10.1	24.4

Table 19: *Within*+ experiment: Answers Contingent and Noncontingent of CC ALLAIS problems

Notes: For a description of each parameterization, see Figure 12.

Parameterization		Noncontingent	Contingent
1	% select f' in Q1	68.1	74.8
	% select f in Q2	85.7	74.8
2	% select f' in Q3	47.1	42.0
	% select f in Q4	42.9	43.7
3	% select f' in Q5	33.6	43.7
	% select f in Q6	43.7	49.6
4	% select f' in Q7	66.4	73.1
	% select f in Q8	82.4	77.3

Table 20: *Within*+ experiment: Answers Contingent and Noncontingent of CR ALLAIS problems

Notes: For a description of each parameterization, see Figure 13.

Parameterization	AUCT		ELECT	
	Noncontingent	Contingent	Noncontingent	Contingent
1	58.8	69.8	16.0	37.8
2	57.1	72.3	25.2	44.5
3	57.1	71.4	25.2	40.3
4	63.0	75.6	27.7	45.4
5	68.1	74.0	26.9	45.4
6	79.0$^{\circledast}$	79.8$^{\circledast}$	32.8	46.2
7	82.4$^{\circledast}$	80.7$^{\circledast}$	33.6	45.4
8	85.7$^{\circledast}$	82.4$^{\circledast}$	80.7*	91.6*
9	84.0$^{\circledast}$	84.0$^{\circledast}$	83.2*	88.2*
10	80.7$^{\circledast}$	86.6$^{\circledast}$	89.9*	95.0*

Table 21: *Within*+ experiment: % of f'Answers in AUCT and ELEC problems

Notes: *Question in which selecting f' implies voting for the color with a majority of balls in the jar. $^{\circledast}$Question in which overbidding (selecting f') is optimal. For details on the parameterizations see Figure 14 (Auction Problems) and Figure 11 (Election Problems).

Tables 22, 23, and 24 show information for problems that involve two questions (ELLS, CC ALLAIS and CR ALLAIS, respectively) on the joint distribution of choices. The tables also present the p-value of a test on whether the proportion of failures is equal across noncontingent and contingent treatments, and we explain how we compute these p-values later in this section.

parameterization	1		2		3		4		5	
treatment	NC	C	NC	C	NC	C	NC	C	NC	C
(f, f')	13.5	35.3	11.8	30.3	12.6	31.9	31.1	37.8	20.2	36.1
(g, g')	30.2	26.9	27.7	25.2	31.1	29.4	14.3	20.2	8.4	20.2
(f, g')	5.9	14.3	20.2	14.3	10.9	14.3	48.7	26.9	7.6	23.5
(g, f')	50.4	23.5	40.3	30.3	45.4	24.4	5.9	15.1	63.9	20.2
% fail SEP/C-SEP	*55.3*	*37.8*	*60.5*	*44.6*	56.3	*28.7*	*54.6*	*42.0*	*71.5*	*43.7*
$(\circledast)$ p-value	.005		.009		.006		.059		.000	

Table 22: Within+design: Proportion of subjects by choices across ELLS questions (in %)

Notes: (f, f') indicates the proportion of subjects who selected f in the first question and f' in the second question of each parameterization. % fail SEP/C-SEP presents the addition of subjects who chose (f, g') and subjects who chose (g, f').
$(\circledast)$ p-value on the null hypothesis that the % of failures in NC are equal to those in C. An explanation on the computation of the p-values is provided in the text.

parameterization	1		2		3	
treatment	NC	C	NC	C	NC	C
(f, f')	21.9	30.3	19.3	31.9	5.0	10.1
(g, g')	42.0	35.3	29.4	36.1	62.2	71.4
(f, g')	27.7	15.1	30.3	12.6	27.8	4.2
(g, f')	8.4	19.3	21.0	19.4	5.0	14.3
% fail SEP/C-SEP	*36.1*	*34.4*	*56.3*	*32.0*	*32.8*	*18.5*
(⊛) p-value	.784		.001		.006	

Table 23: Within+ design: Proportion of subjects by choices across CC ALLAIS questions (in %)

Notes: (f, f') indicates the proportion of subjects who selected f in the first question and f' in the second question of each parameterization. % fail SEP/C-SEP presents the addition of subjects who chose (f, g') and subjects who chose (g, f').
(⊛) p-value on the null hypothesis that the % of failures in NC are equal to those in C. An explanation on the computation of the p-values is provided in the text.

parameterization	1		2		3		4	
treatment	NC	C	NC	C	NC	C	NC	C
(f', f)	58.0	63.0	26.9	28.6	17.6	31.1	55.5	62.2
(g', g)	4.2	13.4	37.0	42.9	40.3	37.8	6.7	11.8
(f', g)	10.1	11.8	20.2	13.4	16.0	12.6	10.9	10.9
(g', f)	27.7	11.8	15.9	15.1	26.1	18.5	26.9	15.1
% fail DOM/C-DOM	*37.8*	*23.6*	*36.1*	*28.5*	42.1	*31.1*	*37.8*	*26.0*
(⊛) p-value	.010		.172		.049		.034	

Table 24: Within+ design: Proportion of subjects by choices across CR ALLAIS questions (in %)

Notes: (f', f) indicates choices of f' in the first question and f in the second question of each parameterization. % fail DOM/C-DOM presents the addition of subjects who chose (f, g') and subjects who chose (g, f').
(⊛) p-value on the null hypothesis that the % of failures in NC are equal to those in C. An explanation on the computation of the p-values is provided in the text.

Tables 25, 26 and 27 have the same structure as Table 3, except that tables provide detailed results for all specifications within each problem. The first specification for each problem corresponds to the benchmark column in Table 3. Columns entitled 'All' compute the average for all specifications, which is also reported in Table 3.

We now explain how we compute the p-values reported in Tables 22, 23, 24, 25, 26 and 27. Let $j \in \{\text{ELLS, CC ALLAIS, AUCT, ELECT, CR ALLAIS}\}$ capture each of the problems, and let k_j index parameterizations within problem j.

We first address the significance test on the reduction of failures between noncontingent and contingent versions, as reported in rows indicated with (⊛). For each subject and each parameterization, D_{Inc} is a dummy variable that takes value 1 if the subject is not consistent with the corresponding postulate. D^C is a dummy that takes value 1 if the observation is from a contingent version and the dummy variable D_{k_j} takes value 1 if the answers are for parameteri-

zation k_j. For each subject we have 64 answers (five specifications of ELLS, three of CC ALLAIS, ten of AUCT, ten of ELECT and four of CR ALLAIS, each in a contingent and noncontingent version). We run the following regression (in which we cluster standard errors by subject) $D_{Inc} = \sum_j \sum_{k_j} \left(\delta_{k_j} D_{k_j} + \phi_{k_j}(D^C \times D_{k_j})\right) + \upsilon$, where υ is an error term. The null hypothesis of interest is that there is no effect of the contingent treatment for each k_j. The p-value reported for each parameterization in row ($\circledast$) corresponds to the null hypothesis that $\phi_{k_j} = 0$. The p-values in the columns entitled 'All' result from running the following regression (in which we cluster standard errors by subject): $D_{Inc} = \sum_j \left(\delta_j D_j + \phi_j(D^C \times D_j)\right) + \upsilon$, where υ is an error term and D_j is a dummy variable that takes value 1 if the answers are for problem j. The p-value reported under the 'All' columns corresponds to the null hypothesis that $\phi_j = 0$.[39]

Now we explain how we compute the p-values for the null hypothesis that $q_C = q_I$, reported in rows ($\otimes$) of Tables 25, 26 and 27. For each subject in the contingent (noncontingent) version of each parameterization of each problem D^C_{Inc} (D^{NC}_{Inc}) is a dummy that takes value 1 if the subject's answer is not consistent with the corresponding postulate. We run the following regression, in which standard errors are clustered by subject and for each subject we have 32 observations: $D^C_{Inc} = \sum_j \sum_{k_j} \left(\alpha_{k_j} D_{k_j} + \beta_{k_j}(D^{NC}_{Inc} \times D_{k_j})\right) + \epsilon$, where ϵ is an error term. The null hypothesis $q_C = q_I$ implies that $2\alpha_{k_j} = 1 - \beta_{k_j}$. The reported p-values in the last row of the tables correspond to a Wald test of this equality for the respective parameterization k_j. To compute p-values for the 'All' columns we proceed as in described in the previous paragraph, by running the same regression without distinguishing between parameterizations of the same problem.

[39] To report p-values for each of the two groups of parameterizations in ELECT and AUCT we run the same regression except excluding questions that are not in the classification. For example, the p-value reported under 'All' for parameterizations AUCT in which underbidding is optimal is the coefficient of the interaction $D^C \times D_{\text{AUCT}}$ in a regression that excludes parameterizations in which overbidding is optimal.

Problem	ELLS						CC ALLAIS				CR ALLAIS				
Parameterization	1	2	3	4	5	All	1	2	3	All	1	2	3	4	All
% Always Consistent	26.9	25.2	27.7	25.2	13.5	23.7	42.9	37.8	58.0	46.2	51.3	45.4	41.2	52.1	47.5
% Always NOT Consistent	21.0	30.3	22.7	21.9	28.6	24.9	13.5	21.0	9.2	14.6	12.6	10.1	14.3	16.0	13.2
% From NOT to Consistent	35.3	30.3	33.6	32.8	42.9	35.0	22.7	30.3	23.5	25.5	25.2	26.1	27.7	21.8	25.2
% From Consistent to NOT	16.8	14.3	16.0	20.2	15.1	16.5	21.0	10.9	9.2	13.7	10.9	18.5	16.8	10.1	14.1
% fail SEP or DOM	56.3	60.5	56.3	54.6	71.5	59.8	36.1	51.3	32.8	40.0	37.8	36.1	42.0	37.8	38.4
% fail C-SEP or C-DOM	37.8	44.5	38.6	42.0	43.7	41.3	34.4	31.9	18.5	28.3	23.5	28.6	31.1	26.0	27.3
(⊛) p-value	.005	.009	.006	.059	.000	.000	.784	.001	.006	.001	.010	.172	.049	.034	.000
% fail STP	66.4	61.3	67.2	77.3	77.3	69.9	49.6	58.0	42.0	49.9	31.9	35.3	43.7	33.6	36.1
q_I	.63	.50	.60	.60	.60	.58	.63	.59	.71	.64	.67	.72	.66	.58	.66
q_C	.38	.36	.37	.44	.53	.42	.33	.22	.14	.23	.18	.29	.29	.16	.23
(⊗) p-value	.009	.138	.012	.094	.490	.001	.002	.000	.000	.000	.000	.000	.000	.000	.000

Table 25: *Within*+ experiment: Switching behavior in ELLS, CC ALLAIS and CR ALLAIS

Notes: (⊛) p-value on the null hypothesis that the % of failures in NC are equal to those in C. (⊗) p-value on the null hypothesis that q_I is equal to q_C. An explanation on the computation of the p-values is provided in the text.

	Underbidding optimal						Overbidding optimal						All
Parameterization	1	2	3	4	5	All	6	7	8	9	10	All	
% Always Consistent	47.9	44.5	44.5	52.9	58.0	49.6	68.1	73.1	76.5	77.3	74.0	73.8	61.7
% Always NOT Consistent	19.3	15.3	16.0	14.3	16.0	16.1	9.2	10.0	8.4	9.2	6.7	8.7	12.4
% From NOT to Consistent	21.9	27.7	26.9	22.7	16.0	23.0	11.8	7.6	5.9	6.7	12.6	8.9	16.0
% From Consistent to NOT	10.9	12.6	12.6	10.1	10.0	11.3	10.9	9.2	9.2	6.7	6.7	8.6	9.9
% fail DOM	41.2	42.9	42.9	37.0	31.9	39.2	21.0	17.6	14.3	16.0	19.3	17.6	28.4
% fail C-DOM	30.2	27.7	28.6	24.4	26.0	27.4	20.2	19.3	17.6	16.0	13.5	17.3	22.4
($\circledast$) p-value	.038	.009	.013	.016	.212	.000	.849	.658	.350	1.00	.147	.891	.001
% fail STP	32.8	40.3	39.5	32.8	26.0	34.3	22.7	16.8	15.1	13.5	19.3	17.5	25.9
q_I	.53	.65	.63	.61	.50	.58	.56	.43	.41	.42	.65	.49	.54
q_C	.19	.22	.22	.16	.15	.19	.14	.11	.11	.08	.08	.10	.15
($\otimes$) p-value	.000	.000	.000	.000	.000	.000	.000	.006	.016	.005	.000	.000	.000

Table 26: *Within*+ experiment: Switching behavior in AUCT

Notes: ($\circledast$) p-value on the null hypothesis that the % of failures in NC are equal to those in C. ($\otimes$) p-value on the null hypothesis that q_I is equal to q_C. An explanation on the computation of the p-values is provided in the text.

	Minority color optimal								Majority color optimal				All
Parameterization	1	2	3	4	5	6	7	All	8	9	10	All	
% Always Consistent	13.4	25.2	23.5	26.0	26.0	26.9	27.7	24.1	77.3	79.0	88.2	81.5	41.3
% Always NOT Consistent	59.7	55.5	58.0	52.9	53.8	47.9	48.7	53.8	5.0	7.6	3.4	5.3	39.2
% From NOT to Consistent	24.4	19.3	16.8	19.3	19.3	19.3	17.7	19.4	14.3	9.2	6.7	10.1	16.6
% From Consistent to NOT	2.5	0.0	1.7	1.7	0.8	5.9	5.9	2.6	3.4	4.2	1.7	3.1	2.7
% fail DOM	84.0	74.8	74.8	72.3	73.1	67.2	66.4	73.2	19.3	16.8	10.1	15.4	55.9
% fail C-DOM	62.2	55.5	59.7	54.6	54.6	53.8	54.6	56.4	8.4	11.8	5.0	8.4	42.0
($\circledast$) p-value	.000	.004	.000	.001	.001	.003	.008	.000	.004	.136	.058	.027	.000
% fail STP	26.9	19.3	18.5	21.0	20.2	25.2	23.5	22.1	17.7	13.5	8.4	13.2	19.4
q_I	.29	.26	.22	.27	.26	.29	.27	.27	.74	.55	.67	.65	.38
q_C	.16	.00	.07	.06	.03	.18	.17	.09	.04	.05	.02	.04	.08
($\otimes$) p-value	.173	.000	.015	.002	.000	.182	.252	.002	.000	.000	.000	.000	.000

Table 27: *Within+* experiment: Switching behavior in ELECT

Notes: ($\circledast$) p-value on the null hypothesis that the % of failures in NC are equal to those in C. ($\otimes$) p-value on the null hypothesis that q_I is equal to q_C. An explanation on the computation of the p-values is provided in the text.

Table 28 provides summary statistics on the observables we collected at the end of the session in the within+ design. The statistics are comparable to those of the between-subjects (*within* experiment) population reported in Table 9 (Table 14) of Online Appendix B (Online Appendix C).

Variable	Mean	Median	Std. dev.	Observations
Total CRT	1.26	1.00	1.14	119
BRET (Risk)	39.98	40.00	16.40	119
Econ Major	0.260	-	-	119
Female	0.605	-	-	119
Junior or older	0.445	-	-	119

Table 28: Within+ design: Summary statistics on observables

Finally, Table 29 provides a robustness test of the results described in Table 5. The regressions presented in the main text used data from all questions of the within+ design. In Table 29 we use data from the within+ design only for the parameterization that coincides with the *within* experiment, but we include 119+131=250 subjects in the analysis.

	ELLS	CC ALLAIS	AUCT	ELECT	CR ALLAIS
Contingent	0.213^{***}	0.078	0.071	0.199^{***}	0.085^{*}
	(0.059)	(0.050)	(0.044)	(0.042)	(0.046)
CRT	-0.021	0.013	0.056	0.171^{***}	-0.119^{*}
	(0.064)	(0.061)	(0.059)	(0.051)	(0.062)
Contingent $\times$ CRT	-0.002	-0.050	0.158^{***}	0.242^{***}	0.181^{***}
	(0.087)	(0.078)	(0.061)	(0.066)	(0.068)
Constant	0.461^{***}	0.638^{***}	0.660^{***}	0.113^{***}	0.688^{***}
	(0.042)	(0.041)	(0.040)	(0.027)	(0.039)
Observations	500	500	500	500	500

Table 29: CRT and consistency (*Within* & *Within*+ Parameterization 1)

Notes: Results from a regression in which the dependent variable takes value 1 if the subject's choice in the problem is consistent with the postulate and 0 otherwise. The right-hand side includes a treatment dummy (1=Contingent), a dummy variable for CRT answers (0 indicates 0 or 1 correct CRT answer, 1 indicates 2 or 3 correct CRT answers) Subjects from the *within* experiment and answers from subjects in the *within+* questions that coincides with the *within* experiment are included in the sample, for a total of 250 subjects. Standard errors (between parentheses) are clustered by subject. Significant at 10% (*), 5% (**), 1% (***).

E Computation of correlations with multiple observations per subject

Let X_j be a measure of a variable X, and let $X_j = \alpha + \eta_j^X + \epsilon^X$, where $\eta_j^X \sim F(\overline{\eta}_j^X, \text{Var}(\eta_j^X))$ and $\epsilon^X \sim F(\overline{\epsilon}^X, \text{Var}(\epsilon^X))$. There is also a second variable, Y for which we assume the same structure. That is, $Y_j = \alpha + \eta_j^Y + \epsilon^Y$, where $\eta_j^Y \sim F(\overline{\eta}_j^Y, \text{Var}(\eta_j^Y))$ and $\epsilon^Y \sim F(\overline{\epsilon}^Y, \text{Var}(\epsilon^Y))$. We assume that we have k^X measures of X and k^Y measures of Y.[40] For each subject i we observe one draw of X_j and one draw of Y_j, which we refer to as X_{ij} and Y_{ij}. Our objective is to compute the correlation between ϵ^X and ϵ^Y.

We first compute the mean of X and Y for each subject (averaging across j). The mean by subject is given by: $\overline{X}_i = \frac{1}{k^X} \sum_{j=1}^{k^X} X_{ij}$, and $\overline{Y}_i = \frac{1}{k^Y} \sum_{j=1}^{k^Y} Y_{ij}$. The overall mean given m total subjects is given by: $\overline{X} = \frac{1}{m} \sum_{i=1}^{m} \overline{X}_i$, and $\overline{Y} = \frac{1}{m} \sum_{i=1}^{m} Y_i$. We define standardized measures in the following way: $\overline{X}_i^* = \frac{\overline{X}_i - \overline{X}}{(\text{Var}(\overline{X}_i))^{1/2}}$ and $\overline{Y}_i^* = \frac{\overline{Y}_i - \overline{Y}}{(\text{Var}(\overline{Y}_i))^{1/2}}$.

By construction we have that

$$\text{Cov}(\overline{X}_i^*, \overline{Y}_i^*) = \text{Cov}\left(\frac{\overline{X}_i - \overline{X}}{(\text{Var}(\overline{X}_i))^{1/2}}, \frac{\overline{Y}_i - \overline{Y}}{(\text{Var}(\overline{Y}_i))^{1/2}}\right) = \frac{\text{Cov}\left(\overline{X}_i, \overline{Y}_i\right)}{(\text{Var}(\overline{X}_i)\text{Var}(\overline{Y}_i))^{1/2}} = \text{Corr}\left(\overline{X}_i, \overline{Y}_i\right).$$

Now, $\text{Cov}\left(\overline{X}_i, \overline{Y}_i\right)$ is given by: $\text{Cov}\left(\overline{X}_i, \overline{Y}_i\right) = \text{Cov}\left(\frac{1}{k^X} \sum_{j=1}^{k^X} X_{ij}, \frac{1}{k^Y} \sum_{j=1}^{k^Y} Y_{ij}\right) = \text{Cov}(\epsilon^X, \epsilon^Y)$. Therefore, we obtain

$$\text{Corr}(\epsilon^X, \epsilon^Y) = \text{Corr}(\overline{X}_i, \overline{Y}_i)\Delta,$$

where $\Delta = \left(\frac{\text{Var}(\overline{X}_i)\text{Var}(\overline{Y}_i)}{\text{Var}(\epsilon^X)\text{Var}(\epsilon^Y)}\right)^{1/2}$. We conclude by showing how to express $\text{Var}(\epsilon^X)$ and $\text{Var}(\epsilon^Y)$ in terms of the data.

We know that:

$$\text{Var}(\overline{X}_i) = \frac{1}{(k^X)^2}\left(\text{Var}\left(\sum_{j=1}^{k^X} \eta_j^X\right) + \left(k^X\right)^2 \text{Var}(\epsilon_{ij}^X)\right) = \frac{1}{(k^X)^2}\left(\sum_{j=1}^{k^X} \text{Var}\left(\eta_j^X\right)\right) + \text{Var}(\epsilon^X) \tag{1}$$

The average variance of the different measures of X is given by:

[40] As an example, let variable X be "the subject is inconsistent with SEP in the Ellsberg problem" and variable Y can be "the subject is inconsistent with SEP in the CC Allais problem." In the within+ design we have $k^X = 5$ measures of X and $k^Y = 3$ measures of Y for each subject.

$$\overline{\text{Var}\left(X_j\right)} = \frac{1}{k^X}\sum_{j=1}^{k^X}\text{Var}(X_j) = \frac{1}{k^X}\sum_{j=1}^{k^X}\text{Var}\left(\eta_j^X + \epsilon^X\right) = \frac{1}{k^X}\sum_{j=1}^{k^X}\text{Var}\left(\eta_j^X\right) + \text{Var}(\epsilon^X) \qquad (2)$$

Using (1) and (2) to solve for Var $\left(\epsilon^X\right)$,we get that:

$$\text{Var}\left(\epsilon^X\right) = \frac{k^X\text{Var}(\overline{X}_i) - \overline{\text{Var}\left(X_j\right)}}{k^X - 1}.$$

Similarly, for Var $\left(\epsilon^Y\right)$,we get that:

$$\text{Var}\left(\epsilon^Y\right) = \frac{k^Y\text{Var}(\overline{Y}_i) - \overline{\text{Var}\left(Y_j\right)}}{k^Y - 1}.$$

F Further details on the withinCNC experiment

A total of 101 subjects participated in the *withinCNC* experiment. Subjects first faced contingent treatments and afterwards the corresponding noncontingent version. The order in which problems were faced was as follows: ELLS, ELECT, CC ALLAIS, AUCT, CR ALLAIS. The treatment consisted of ten parts plus a bonus part (Part 11) that involves the cognitive reflexion test, a risk aversion task and answers to demographic questions. The instructions are the same as in the *within* experiment except for the change in order. For details on the instructions, see the Procedures Appendix.

Table 30 shows information on choices inconsistent with SEP/C-SEP and DOM/C-DOM are informed on Table 6 of the main text.

	ELLS		CC ALLAIS		AUCT		ELECT		CR ALLAIS	
treatment	NC	C	NC	C	NC	C	NC	C	NC	C
# of observ.	101	101	101	101	101	101	101	101	101	101
% (f, f')	24.8	34.7	20.8	33.7	82.2	79.2	32.3	54.8	64.4	62.4
% (g, g')	24.8	26.7	50.5	30.7	—	—	—	—	6.9	12.9
% (f, g')	6.9	20.8	11.9	14.9	17.8	20.8	67.7	45.2	22.8	7.9
% (g, f')	43.6	17.8	16.8	20.8	—	—	—	—	5.9	16.8
% fail SEP/C-SEP	50.5	38.6	28.7	35.7	—	—	—	—	—	—
% fail DOM/C-DOM	—	—	—	—	17.8	20.8	69.3	47.5	28.7	24.7
p-value	.091		.279		.517		.001		.497	

Table 30: *WithinCNC* experiment: summary of results

Notes: 1) In ELLS and CC ALLAIS, (f, f') indicates the proportion of subjects who selected f in Q1 and f' in Q2. In AUCT and ELECT, (f, f') indicates choices of f in Q* and f' in Q1. We assume that all subjects prefer more money to less and so we impute that all subjects select f in Q*. In CR ALLAIS, (f, f') indicates choices of f in Q2 and f' in Q1.
2) % fail SEP/C-SEP and % fail DOM/C-DOM presents the addition of subjects who chose (f, g') and subjects who chose (g, f').
3) The p-value is obtained by conducting a regression similar to that of Table 16. The reported p-value for each problem corresponds to the p-value of the coefficient for this interaction variable.

Table 31 shows the answers question by question.

		Noncontingent	Contingent	*p-value*
ELLS	% select f in Q1	31.7	55.5	*.000*
	% select f' in Q2	68.3	52.5	*.018*
	# of Participants	101	101	-
CC ALLAIS	% select f in Q1	32.7	48.5	*.013*
	% select f' in Q2	37.6	54.5	*.006*
	# of Participants	101	101	-
AUCT	% select f' in Q1	82.2	79.2	*.516*
	# of Participants	101	101	-
ELECT	% select f' in Q1	30.7	52.5	*.001*
	# of Participants	101	101	-
CR ALLAIS	% select f' in Q1	70.3	79.2	*.119*
	% select f in Q2	87.1	70.3	-
	# of Participants	101	101	-

Table 31: *WithinCNC* experiment: Answers by question

Notes: The p-values are computed in the following manner. We run a linear regression where the left-hand side is a variable that takes value 1 if f (or f' depending on the question) is selected and the right-hand side variable is a treatment dummy (1=contingent). The reported p-value corresponds to the estimated coefficient for the treatment dummy. For ELLS and CC ALLAIS we run one regression per question. Standard errors are clustered by subject.