Cognitive Psychology 77 (2015) 42–76

Contents lists available at ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

Do the right thing: The assumption of optimality in lay decision theory and causal judgment

Samuel G.B. Johnson [a,*], Lance J. Rips [b]

[a] Department of Psychology, Yale University, United States
[b] Department of Psychology, Northwestern University, United States

ARTICLE INFO

Article history:
Accepted 28 January 2015

Keywords:
Lay decision theory
Causal attribution
Rationality
Decision-making
Theory of mind
Behavioral game theory

ABSTRACT

Human decision-making is often characterized as irrational and suboptimal. Here we ask whether people nonetheless assume optimal choices from *other* decision-makers: Are people intuitive classical economists? In seven experiments, we show that an agent's perceived optimality in choice affects attributions of responsibility and causation for the outcomes of their actions. We use this paradigm to examine several issues in lay decision theory, including how responsibility judgments depend on the efficacy of the agent's actual and counterfactual choices (Experiments 1–3), individual differences in responsibility assignment strategies (Experiment 4), and how people conceptualize decisions involving trade-offs among multiple goals (Experiments 5–6). We also find similar results using everyday decision problems (Experiment 7). Taken together, these experiments show that attributions of responsibility depend not only on what decision-makers *do*, but also on the quality of the options they choose *not* to take.

1. Introduction

Psychologists, economists, and philosophers are united in their disagreements over the question of human rationality. Some psychologists focus on the fallibility of the heuristics we use and the systematic biases that result (Kahneman & Tversky, 1996), while others are impressed by the excellent

* Corresponding author at: 2 Hillhouse Ave., New Haven, CT 06520, United States.
E-mail address: samuel.johnson@yale.edu (S.G.B. Johnson).

http://dx.doi.org/10.1016/j.cogpsych.2015.01.003
0010-0285/

performance of heuristics in the right environment (Gigerenzer & Goldstein, 1996). Economists spar over the appropriateness of rationality assumptions in economic models, with favorable views among classically-oriented economists (Friedman, 1953) and unfavorable views among behavioral theorists (Simon, 1986). Meanwhile, philosophers studying decision theory struggle to characterize what kind of behavior *is* rational, given multifaceted priorities, indeterminate probabilities, and pervasive ignorance (Jeffrey, 1965).

Although decision scientists have debated sophisticated theories of rationality, less is known about people's lay theories of decision-making. Understanding how people predict and make sense of *others'* decision-making has both basic and applied value, just as research on lay theories of biology (e.g., Shtulman, 2006), psychiatry (e.g., Ahn, Proctor, & Flanagan, 2009), and personality (e.g., Haslam, Bastian, & Bissett, 2004) has led to both theoretical and practical progress. The study of lay decision theory can illuminate aspects of our social cognition and reveal the assumptions we make when interacting with others.

In this article, we argue that people use an *optimality* theory in thinking about others' behavior, and we show that this optimality assumption guides the attribution of causal responsibility. In the remainder of this introduction, we first describe game theory research on optimality assumptions, then lay out the connections to causal attribution research. Finally, we derive predictions for several competing theoretical views, and preview our empirical strategy.

1.1. Optimality assumptions in strategic interaction

Psychologists are well-versed in the evidence against human rationality (e.g., Shafir & LeBoeuf, 2002; the collected works of Kahneman and Tversky). Nonetheless, optimality assumptions have a venerable pedigree in economics (Friedman, 1953; Muth, 1961; Smith, 1982/1776), and are incorporated into some game-theoretic models. In fact, classical game theory assumes not only first-order optimality (i.e., behaving optimally relative to one's self-interest) but also second-order optimality (assuming that others will behave optimally relative to their own self-interest), third-order optimality (assuming that others will assume that others will behave optimally), and so on *ad infinitum* (von Neumann & Morgenstern, 1944). Understanding the nature of our assumptions about others' decision-making is thus a foundational issue in *behavioral game theory*—the empirical study of strategic interaction (Camerer, 2003; Colman, 2003).

Because people are neither infinitely wise nor infinitely selfish, rational self-interest models of economic behavior break down even in simple experimental settings (Camerer & Fehr, 2006). For example, in the *beauty contest game* (Ho, Camerer, & Weigelt, 1998; Moulin, 1986; Nagel, 1995), a group of players each picks a number between 0 and 100, with the player choosing the number closest to 2/3 of the average winning a fixed monetary payoff. The Nash Equilibrium for this game is that every player chooses 0 (i.e., only if every player chooses 0 is it the case that no player can benefit by changing strategy). If others played the game without any guidance from rationality, choosing randomly, then their mean choice would be 50, so the best response would be around 33. But if others followed that exact reasoning, then their average response would be 33, and the best response to 33 is about 22. Applying this same logic repeatedly leads us to the conclusion that the equilibrium guess should be 0. Yet average guesses are between 20 and 40, depending on the subject pool, with more analytic populations (such as Caltech undergraduates) tending to give lower guesses (Camerer, 2003). Which assumption or assumptions of classical game theory are being violated here? Are people miscalculating the equilibrium? Are they assuming that others will miscalculate, or assuming that others will assume miscalculations from others? Are they making a perspective-taking error, or assuming that others will make perspective-taking errors?

One approach toward answering such questions is to build an econometric model of each player's behavior, interpreting the parameter estimates as evidence concerning the players' underlying psychology (e.g., Camerer, Ho, & Chong, 2004; Stahl & Wilson, 1995). This approach has led to important advances, but the mathematical models often underdetermine the players' thinking, because a variety of mental representations and cognitive failures can often produce identical behavior. In this paper, we approach the problem of what assumptions people make about others' behavior using a different set of tools—those of experimental psychology.

1.2. An optimality assumption in lay theories of decision-making?

Two key assumptions of mathematical game theory—perfect self-interest and perfect rationality—are not empirically plausible (Camerer, 2003). However, a third assumption—that people assume (first-order) optimality in others' decision-making—may be more plausible. To test this possibility, we studied how people assign causal responsibility to agents for the outcomes of their decisions: How do people evaluate Angie's responsibility for an outcome, given Angie's choice of a means for achieving it? Our key prediction is that if people use an optimality theory, agents should be seen as more responsible for outcomes flowing from their actions when those actions led optimally to the outcome.

We hypothesized this connection between lay decision theory and perceived responsibility because (a) rational behavior is a cue to agency (Gao & Scholl, 2011; Gergely & Csibra, 2003), and (b) agents are perceived as more responsible than non-agents (Alicke, 1992; Hart & Honoré, 1959; Hilton, McClure, & Sutton, 2010; Lagnado & Channon, 2008). Putting these two findings together, a lay decision theorist should assign higher responsibility to others to the extent that those others conform to her theory of rational decision-making (see Gerstenberg, Ullman, Kleiman-Weiner, Lagnado, & Tenenbaum, 2014, for related computational work). Conversely, decision-making that contradicts her theory could result in attenuated responsibility assignment, on the grounds that the decision-maker is not operating in a fully rational way. In extreme cases, murderers may even be acquitted on grounds of mental defect when their decision-making mechanism is perceived as wildly discrepant from rational behavior (see Sinnott-Armstrong & Levy, 2011), overriding the strong motivation to punish morally objectionable actions (Alicke, 2000).

Studying attributions of responsibility also has methodological and practical advantages. Responsibility attributions can be used to test inferences not only about agents' actual choices, but also about their *counterfactual* choices—the options that were available but not taken. Intuitively, responsibility attributions are a way of assigning "ownership" of an outcome to one or more individuals after a fully specified outcome has occurred (Hart & Honoré, 1959; Zultan, Gerstenberg, & Lagnado, 2012). This method allows us to independently vary the quality of the actual and counterfactual decision options. Further, attributions of causal responsibility have real-life consequences. They affect our willingness to cooperate (Falk & Fischbacher, 2006), our predictions about behavior (McArthur, 1972; Meyer, 1980), and our moral evaluations (Cushman, 2008). For this reason, understanding how people assign responsibility for outcomes has been a recurring theme in social cognition research (e.g., Heider, 1958; Kelley, 1967; Weiner, 1995; Zultan et al., 2012).

1.3. Strategies for assigning responsibility

In this article, we argue that perceived responsibility depends on the *optimality* of an action—that people behave like lay classical economists in the tradition of Adam Smith. People believe a decision maker is responsible for an outcome if the decision maker's choice is the best of all available options. However, optimality is not the only rule people could adopt in evaluating decisions, and in this section, we compare optimality to other strategies.

To compare the alternative strategies, suppose that Angie wants the flowers of her cherished shrub to turn red, and faces a decision as to which fertilizer to purchase—*Ever-Gro* or *Green-Scream*. Suppose she purchases *Ever-Gro*, which has a 50% chance of making her flowers turn red. We abbreviate this probability as P_{ACT}, where $P_{ACT} = P(\text{Outcome} \mid \text{Actual Choice})$. In this case, $P_{ACT} = .5$. Suppose, too, that the rejected option, *Green-Scream*, has a 30% chance of making her flowers turn red; we abbreviate this as $P_{ALT} = P(\text{Outcome} \mid \text{Alternative Choice})$. Since $P_{ACT} > P_{ALT}$, Angie's choice was optimal. However, if the rejected option, *Green-Scream*, had instead had a 70% chance of making the flowers turn red, then $P_{ALT} > P_{ACT}$, and Angie's choice of *Ever-Gro* would have been suboptimal. Finally, if both fertilizers had a 50% chance of producing red flowers, then $P_{ACT} = P_{ALT}$, and there would have been no uniquely optimal decision. Supposing that the fertilizer of her choice does cause the flowers to turn red, is Angie responsible for the successful completion of her goal—for the flowers turning red?

One possibility is that the quality of the rejected options is not relevant to Angie's responsibility. What does it matter if Angie *might* have made a choice more likely to fulfill her goal, given that she actually did fulfill it? People are ordinarily more likely to generate "upward" counterfactuals in cases of failure than "downward" counterfactuals in cases of success (e.g., Mandel & Lehman, 1996), and on some accounts, the primary function of counterfactual reasoning is to elicit corrective thinking in response to negative episodes (Roese, 1997). So people may not deem counterfactual actions relevant if the actual choice led to a success (see Belnap, Perloff, & Xu, 2001 for a different rationale for such a pattern). If people do not view Angie's rejected options as relevant to evaluating her actual (successful) decision, then they would follow a strategy we call *alternative-insensitive*: For a given value of P_{ACT}, there would be no relationship between attributions of responsibility and P_{ALT}. Table 1 summarizes this possibility by showing that this view predicts that people will assign Angie responsibility (indicated by + in the table) as long as (a) she chooses an option that has a nonzero probability of leading to the desired outcome and (b) that outcome actually occurs.

A quite different pattern would appear if people assume that agents are *optimizers*. Although much of the time people do not themselves behave optimally (e.g., Simon, 1956), the assumption of optimal decision-making might be useful for predicting and explaining behavior (Davidson, 1967; Dennett, 1987) and is built into game theory models of strategic interaction (e.g., Von Neumann & Morgenstern, 1944). If optimality of this sort underlies our lay decision theories, the perceived responsibility of other decision-makers should depend on whether they select the highest quality option available (i.e., on whether $P_{ACT} > P_{ALT}$). For example, given Angie's choice of *Ever-Gro* ($P_{ACT} = .5$), Angie might be seen as more responsible for the flowers turning red if the rejected option of *Green Scream* is inferior ($P_{ALT} = .3$) than if it is superior ($P_{ALT} = .7$). According to this account, the *size of the difference* between P_{ACT} and P_{ALT} should have little impact on responsibility ratings. That is, if $P_{ACT} = .5$, Angie would be seen as equally responsible for the flowers turning red, regardless of whether the rejected option is only somewhat worse ($P_{ALT} = .3$) or is much worse ($P_{ALT} = .1$), because she chose *optimally* either way. Likewise, Angie's (non-)responsibility for the outcome would be similar whether the rejected option is only somewhat better ($P_{ALT} = .7$) or much better ($P_{ALT} = .9$), because she chose *suboptimally* either way.

The prediction that responsibility ratings would be insensitive to the magnitude of $[P_{ACT} - P_{ALT}]$ is an especially strong test of optimality, because in other contexts, people often judge the strength of a cause to be proportional to the size of the difference the cause made to the probability of the outcome (Cheng & Novick, 1992; Spellman, 1997). The canonical measure of probabilistic difference-making is ΔP (Allan, 1980), which is equal to $[P(\text{Effect} \mid \text{Cause}) - P(\text{Effect} \mid \sim\text{Cause})]$. One might expect, based on those previous results, that responsibility ratings would be sensitive to the magnitude of $[P_{ACT} - P_{ALT}]$, which is equivalent to ΔP if one interprets the actual decision as the cause and the rejected option as the absence of the cause (i.e., ∼Cause). We refer to this strategy as ΔP *dependence*.

The final strategy we consider is *positive difference-making*. If more than two alternatives are available, the difference-making status of any one of them is best evaluated against a common baseline. For example, if Angie can choose to apply *Ever-Gro*, *Green-Scream*, or neither, then we can calculate ΔP separately for each fertilizer relative to the do-nothing option. Suppose, for example, that Angie's plant

Table 1
Distinctions among four ways of assigning responsibility for the outcome of an action.

Assignment strategy	Responsibility judgment		
	Option A ($P_A = .5$)	Option B ($P_B = .3$)	Option C ($P_C = .1$)
Optimality	+	–	–
Positive difference-making (positive ΔP)	+	+	–
ΔP dependence (linear ΔP)	+.4	+.2	–
Alternative-insensitive	+	+	+

Note: A plus sign indicates that an agent is responsible for an outcome under the named strategy, and a minus sign indicates that the agent is not responsible. Numbers following the plus sign represent degrees of responsibility. The probabilities associated with the options are $P_X = P(\text{outcome} \mid \text{choice of } X)$, and the responsibility assignments assume that the outcome was actually realized.

has a 10% chance of developing red flowers even if she does not add any fertilizer, and that Angie's choice of *Ever-Gro* was suboptimal ($P_{ACT} = .5$) relative to the rejected choice of *Green-Scream* ($P_{ALT} = .7$). Now, ΔP is positive both for her actual (suboptimal) choice ($\Delta P_{ACT} = .4$) and for the rejected option ($\Delta P_{ALT} = .6$). If people simply assign higher responsibility ratings when $\Delta P > 0$ than when $\Delta P < 0$—in contrast to both the ΔP dependence and the optimality strategies—then Angie would be seen as highly causal, despite her suboptimal choice.

Table 1 compares these four methods of assigning responsibility. Suppose the decision-maker has three options, A, B, and C. For illustration, we will assume that P_A [= P(outcome | choice of A)] = .5, $P_B = .3$, and $P_C = .1$. Then, as Table 1 shows, optimizing implies that the decision-maker is responsible (indicated by a +) only if she chooses A, whereas positive difference-making implies that she is responsible if she chooses either A or B (assuming that ΔP is calculated relative to the worst option, C). A pure ΔP strategy (i.e., responsibility is directly proportional to ΔP) also assigns responsibility to A and B, but more strongly for the former. Finally, if people are insensitive to alternative choices, then so long as a positive outcome occurred, the decision-maker would be credited with responsibility if she chooses any of A, B, or C.

1.4. Overview

These issues are explored in seven experiments. Experiments 1 and 2 distinguish the predictions of the four accounts summarized in Table 1 by varying the quality of the decision-makers' rejected options (P_{ALT}). Experiment 3 then turns to how people combine information about the quality of both the actual and rejected options (P_{ACT} and P_{ALT}) in forming responsibility judgments, and Experiment 4 looks at individual differences in assignment strategies. Experiments 5 and 6 then examine how people conceptualize trade-offs among multiple goals, testing whether perceived responsibility for a goal tracks optimality for that goal or optimality relative to the agents' overall utility. Finally, Experiment 7 uses more naturalistic decision problems to see how people spontaneously assign responsibility when the probabilities are supplied by background knowledge rather than by the experimenter (Experiment 7).

2. Experiment 1: The influence of rejected options

In Experiment 1, we ask whether people typically use an optimality assumption to guide their attributions of responsibility, or whether instead they follow a linear ΔP or alternative-insensitive strategy (see Table 1). To do so, we examine how agents' perceived responsibility for a desired outcome depends on the quality of a counterfactual choice—that is, an option they rejected. Participants read about agents who made decisions leading to an outcome with probability P_{ACT} (always .5), but could have made an alternative decision that would have led to that outcome with probability P_{ALT} (which varied between .1 and .9 across conditions). Table 2 exhibits the full set of combinations of P_{ALT} and P_{ACT}, which were varied across vignettes such as the following:

> *Angie has a shrub, and wants the shrub's flowers to turn red. She is considering two brands of fertilizer to apply:*
> *If she applies Formula PTY, there is a 50% chance that the flowers will turn red.*
> *If she applies Formula NRW, there is a 10% chance that the flowers will turn red.*

Table 2
Design of Experiment 1.

	Cond. 1	Cond. 2	Cond. 3	Cond. 4	Cond. 5
Choice A (P_{ACT})	**.5**	**.5**	**.5**	**.5**	**.5**
Choice B (P_{ALT})	*.1*	*.3*	*.5*	*.7*	*.9*

Note: Entries denote the probability of the goal outcome occurring given each choice. The bold entries indicate probabilities given the agent's actual choice, whereas the italicized entries indicate probabilities given the agent's rejected options.

Angie chooses Formula PTY, and the flowers turn red.

To assess the consistency of these effects, some participants were asked about responsibility and others about causation. Although judgments of social causation and judgments of responsibility are often treated similarly in social cognition research (Shaver, 1985; Weiner, 1995), we thought that the more moral character of the term *responsibility* could produce a different pattern of results than the more neutral term *cause*. In all other experiments (except as noted for Experiment 3), only the responsibility question was asked, because wording did not interact with the variables of interest.

Because P_{ACT} is fixed at .5 across all versions of the items (see Table 2), a lay theory of decision-making that is insensitive to counterfactuals should predict that P_{ALT} will produce no differences in responsibility judgments. A theory that assumes optimizing, in contrast, *should* distinguish between cases for which $P_{ACT} > P_{ALT}$ (the actual decision was optimal) and those for which $P_{ACT} < P_{ALT}$ (the actual decision was suboptimal). But an optimizing theory would be less likely to discriminate between different values of P_{ALT} as long as they are on the same side of P_{ACT}. That is, responsibility should show a *qualitative dependence* on P_{ALT}, a step or sigmoid function with a steep drop at the value of P_{ALT} that makes $[P_{ACT} - P_{ALT}] = 0$. Because P_{ACT} is always .5 in these items, the drop would occur when $P_{ALT} = .5$. Both possibilities (optimality and alternative insensitivity) can be distinguished from the *linear dependence* of responsibility on $[P_{ACT} - P_{ALT}]$ that might be expected on the basis of the causal attribution literature (Cheng & Novick, 1992; Spellman, 1997).

2.1. Method

Participants read five vignettes similar to the text above. All vignettes involved an agent with a goal who is deciding between two options for achieving that goal. They covered a variety of different types of decisions: choices between various kinds of fertilizers, flour, shampoo, clothing as a gift, and bicycles. In these vignettes, P_{ACT} was fixed at .5, while P_{ALT} was varied across vignettes (at .1, .3, .5, .7, and .9), as shown in Table 2. A participant read the five cover stories, each paired with a different value of P_{ALT}. However, across participants, the five stories appeared about equally often with the five P_{ALT} values, as determined by a Latin square. Participants rated their agreement with either a responsibility statement (e.g., "Angie is responsible for the flowers turning red"), or with a causal statement ("Angie caused the flowers to turn red"), on an 11-point scale (0: "disagree"; 5: "neither agree nor disagree"; 10: "agree"). The phrasing (responsibility or causality) was varied between-subjects. The vignettes appeared in a new random order for each participant.

One hundred participants were recruited from Amazon Mechanical Turk for Experiment 1. In each experiment reported in this paper, participants answered a series of multiple-choice check questions after the main experiment, and any participant answering more than 33% of these questions incorrectly was excluded from data analysis. For Experiment 1, ten check questions were asked, and five participants were excluded from analysis for answering more than three questions incorrectly.

2.2. Results and discussion

Ratings of both the agents' responsibility and causality depended qualitatively on the rejected choice. The conditional probability of the desired outcome given the actual choice was always .5, but participants' evaluation of this choice varied with the conditional probability of the same outcome given the *rejected* option. Responsibility and causality ratings were highest ($M = 6.18$ on a 0-to-10 scale) when the probability of the actual choice was greater than that of the rejected choice ($P_{ACT} > P_{ALT}$) and lowest ($M = 4.90$) when the probability of the actual choice was less than that of the rejected choice ($P_{ACT} < P_{ALT}$). However, the magnitude of the difference between P_{ALT} and P_{ACT} had no further effect on judgments, consistent with an optimality strategy.

Because there was no interaction between question type and P_{ALT} [$F(4,372) = 0.76$, $p = .55$, $\eta_p^2 < .01$], we collapsed across responsibility and causality judgments in conducting pairwise comparisons for adjacent P_{ALT} values. Fig. 1 graphs these mean ratings and their standard errors. These comparisons revealed a significant difference between the .3 condition (where Angie's actual decision was optimal) and the .5 condition (where no uniquely optimal decision existed, since P_{ACT} also was .5)

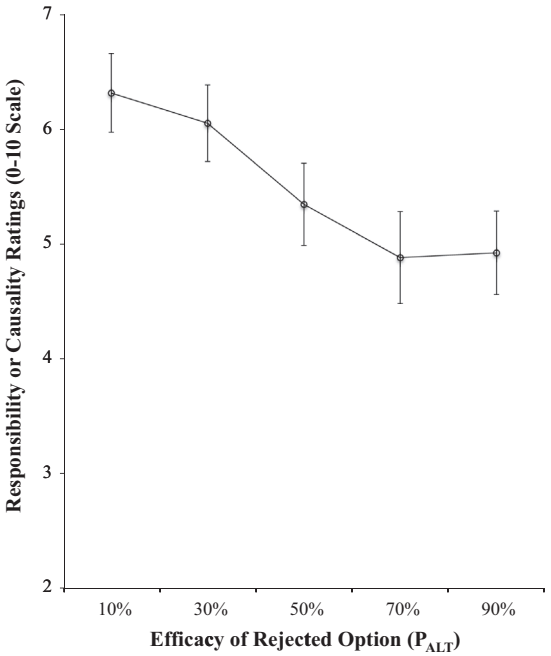

Fig. 1. Results of Experiment 1. Error bars represent ±1 standard error.

[$t(94) = -2.74$, $p = .007$, $d = -0.28$, $BF_{10} = 2.9$].[1] However, there was evidence *against* a difference between adjacent optimal conditions [for the .1 condition vs. the .3 condition, $t(94) = 1.21$, $p = .23$, $d = 0.12$, $BF_{01} = 5.99$] and between adjacent suboptimal conditions [for the .7 vs. .9 conditions, $t(94) = -0.04$, $p = .86$, $d = -0.02$, $BF_{01} = 12.16$]. Finally, the evidence was equivocal with respect to the difference between the .5 and .7 conditions [$t(94) = 2.03$, $p = .029$, $d = 0.23$, $BF_{01} = 1.14$]. Because $P_{ACT} = .5$, there was no optimal choice in the .5 condition, and the agent chose *suboptimally* in the .7 condition. Thus, if only *optimality* matters, one would expect no difference between these conditions, but if *suboptimality* matters, one would expect a difference. Given the lack of firm evidence either way, we reserve judgment on this question until more definitive evidence can adjudicate this issue.

The effect of P_{ALT} on responsibility ratings occurred only because judgments depended on the ordinal difference between P_{ACT} and P_{ALT} (i.e., the sign of ΔP). The magnitude of the difference did not affect judgments. This is most consistent with the idea that causal attribution involves assessing the optimality of the decision: Decision-makers acted optimally if $P_{ACT} > P_{ALT}$ and suboptimally if $P_{ACT} < P_{ALT}$. In addition, because the lack of magnitude-dependence held for both attributions of responsibility and of causation, these results are not due to idiosyncratic properties of either phrasing, and in particular, the more moral character of the term *responsibility* did not drive judgments.

We take these results as support for the optimality model, for cases where outcomes are well-defined. That is, most participants in Experiment 1 appear to have assigned increased responsibility to Angie when her choice was optimal relative to her achieved goal. However, the nature of her intention is crucial. If her goal were to make her plant grow in general, then a choice that caused the plant to

[1] Because null effects were predicted for some comparisons, all *t*-tests in this paper are supplemented with Bayes Factors (BFs), computed using a default Jeffrey–Zellner–Siow (JZS) prior with a scaling factor of 1, as recommended by Rouder, Speckman, Sun, Morey, and Iverson (2009). Unlike *p*-values, BFs quantify evidence either against or in favor of a null hypothesis. A BF favoring the null hypothesis is denoted 'BF_{01}'; e.g., $BF_{01} = 3.0$ means that the data is three times likelier under the null hypothesis. Conversely, a BF favoring the alternative hypothesis is denoted 'BF_{10}'; e.g., $BF_{10} = 5.0$ means that the data is five times likelier under the alternative hypothesis. For a conceptual comparison of Bayesian vs. null hypothesis significance testing, see Dienes (2011), and for computational details, see Rouder et al. (2009).

grow to 8 in. rather than 4 in. is unlikely to yield higher responsibility ratings, because people would probably view Angie as responsible in both cases—just for different outcomes. But if her goal were to make the plant grow to 8 in. and she chose an option that made it more likely to grow to 4 in. rather than to 8 in., then her responsibility might indeed be discounted because she behaved suboptimally relative to her particular goal. Similarly, if Angie had instead desired the plant *not* to turn red, then she would likely be viewed as more responsible for it not turning red if she *minimized* the probability of it turning red. The current results, then, are consistent with the idea that most people use optimality relative to a specified, desired outcome for judging responsibility (see Johnson & Rips, 2014, for the effects of manipulating goals on perceived responsibility).

2.3. Replication experiment

Given the intuitive appeal of a linear relationship between ΔP and responsibility, a skeptical reader may not be fully convinced by the evidence in favor of the null effects between the .1 and .3 conditions and the .7 and .9 conditions in Fig. 1. (Note, however, that a lack of power could not account for the substantial Bayes Factors favoring the null effects.) An additional concern about Experiment 1 is that the vignettes did not specify whether or not the agent knew the probabilities of the outcome. If participants had inferred that the agent knew the probabilities when she chose optimally but did not know the probabilities when she chose suboptimally, then the greater responsibility judgments could be due to the agent's greater knowledge in the optimal conditions rather than to her optimal behavior *per se*.

To address these concerns, we performed a near-exact replication of Experiment 1 that differed only in specifying that the agent knew the probabilities of the outcome for each choice. We recruited 199 participants from Amazon Mechanical Turk, and 29 were excluded because they incorrectly answered more than 33% of the 15 check questions. (Additional check questions were added to ensure that participants understood that the agent knew the probabilities.) The results of this replication experiment were similar to those of Experiment 1, as shown in Fig. 2. These results corroborated the null differences between the .1 and .3 conditions [$t(169) = 1.38$, $p = .17$, $d = 0.11$, $BF_{01} = 6.42$] and between the .7 and .9 conditions [$t(169) = 1.67$, $p = .10$, $d = 0.13$, $BF_{01} = 4.13$], while revealing strong evidence for a difference between the .3 and .5 conditions [$t(169) = 3.80$, $p < .001$, $d = 0.29$, $BF_{10} = 59.7$]. Unlike Experiment 1, however, the evidence *against* a difference between the .5 and .7 conditions was substantial [$t(169) = 1.12$, $p = .26$, $d = 0.09$, $BF_{01} = 8.78$]. Thus, it appears that for most participants, it is the *optimal* decision-makers who are assigned increased responsibility rather than the *suboptimal* decision-makers who are assigned diminished responsibility (see Section 9.1 for discussion).

To further address the concern that these findings could reflect a linear rather than sigmoid (step) pattern, we fit linear and sigmoid functions to the data from (a) Experiment 1, (b) the Experiment 1 replication, (c) Experiment 4 (which used a similar procedure), and (d) these three experiments combined. The linear function had two free parameters ($y = a + b\ P_{ALT}$), and the sigmoid function could have either three ($y = a + (b(P_{ALT} - c))/\text{sqrt}(1 + (b(P_{ALT} - c))^2)$) or two ($y = a + (b(P_{ALT} - 0.5))/\text{sqrt}(1 + (b(P_{ALT} - 0.5))^2)$) free parameters, if we assume in the latter case that the inflection point occurs at $P_{ALT} = 0.5$. (This sigmoid function was chosen rather than a logistic function because the logistic function was difficult to specify uniquely with low-noise data.) For all four data sets, the two-parameter sigmoid model was a better fit than the linear model (R^2s = .956 vs. .915; .971 vs. .963; .893 vs. .845; and .966 vs. .936, for the four data sets, respectively). Further, the three-parameter sigmoid model was a better fit than the linear model, using AIC to penalize the sigmoid model for the additional parameter (*AIC* = −3.90 vs. 2.46; −3.97 vs. −0.64; 4.65 vs. 4.80; −3.47 vs. 1.30, for the four data sets, respectively, where low values of AIC indicate a better fit). Thus, the results of these three experiments consistently reveal a sigmoid dependence of responsibility judgments on P_{ALT}, consistent with the analyses presented above, finding a qualitative dependence of responsibility on P_{ALT}.

3. Experiments 2A and 2B: Optimizing vs. positive difference making

Although the results of Experiment 1 are consistent with optimality, they could also be explained by participants attributing responsibility to the agent whenever $\Delta P > 0$. This is the positive difference-

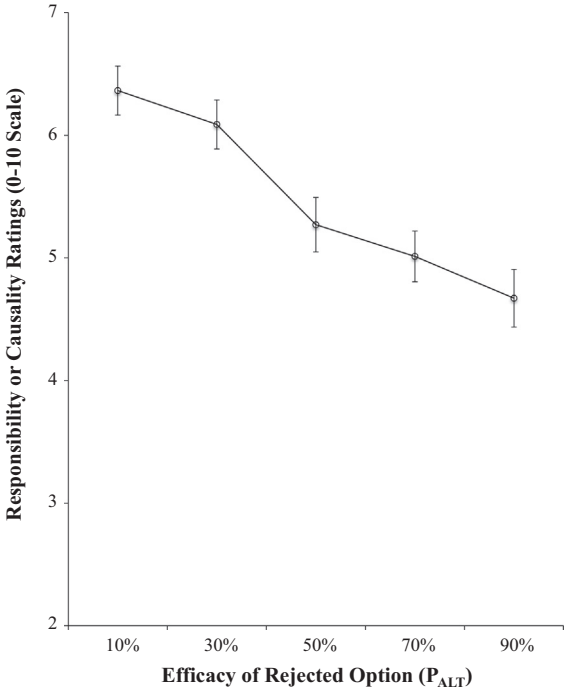

Fig. 2. Results of Experiment 1 replication. Error bars represent ±1 standard error.

making strategy of Table 1. The $\Delta P > 0$ relationship occurs when the decision makes a positive difference to the outcome, relative to some reference point. If participants were computing ΔP relative to the worst available option, then the optimal choice in every condition of Experiment 1 was also the only option for which $\Delta P > 0$. This possibility is distinct from the prediction that judgments would be linearly dependent on ΔP, which Experiment 1 disconfirmed.

To examine whether the response pattern in Experiment 1 was based on optimality or on positivity, Experiment 2 used vignettes in which agents faced three options: a first option with a high probability of leading to the goal, a second option with a moderate probability of leading to the goal, and a third option with a low probability of leading to the goal, as shown in Table 3. In the first condition, item *A* (the actual choice) is the optimal choice with $P_{ACT} = .5$, but both *A* and *B* have positive ΔP relative to *C*, the worst choice (as in Table 1). If judgments are based on optimality, then an agent choosing *A* should be rated more responsible than an agent choosing *B*, since *A* is optimal but *B* and *C* are not. However, if people are merely sensitive to ΔP's positivity, they should rate agents choosing *A* and *B* equally highly, since $\Delta P > 0$ for both. In the second condition, option *A* (again the actual choice with $P_{ACT} = .5$) is suboptimal relative to B ($P_{ALT} = .7$). So optimality predicts that an agent choosing *A* should be less responsible than an agent choosing *B*, but positivity once again predicts that they are equally responsible.

A related hypothesis is that people use a positive difference-making strategy, but rather than computing ΔP values relative to the worst option, they compute these values relative to the average option. This *average* version of the positive-difference strategy is also addressed by our experimental design (see Table 3). In the first ($P_{ALT} = .3$) condition, the average efficacy among all the options is .3, and in the second ($P_{ALT} = .7$) condition, the average efficacy among all the options is .43. Both of these averages are less than P_{ACT} (=.5). Thus, a positive difference-making strategy that computes ΔP relative to the average option should predict that an agent choosing A is responsible in both conditions. (Note that a *linear* ΔP strategy that computes ΔP relative to the average option would have produced a strictly increasing pattern in Experiment 1, contrary to the results.)

Table 3A
Design of Experiment 2A.

	Cond. 1	Cond. 2
Choice A (P_{ACT})	**.5**	**.5**
Choice B (P_{ALT})	*.3*	*.7*
Choice C (P_{BR})	*.1*	*.1*

Table 3B
Design of Experiment 2B.

	Cond. 1	Cond. 2
Choice A (P_{ACT})	**.5**	**.5**
Choice B (P_{ALT})	*.3*	*.7*
Doing nothing (P_{BR})	*.1*	*.1*

Note: Entries denote the probability of the goal outcome occurring given each choice. The bold entries indicate probabilities given the agent's actual choice, whereas the italicized entries indicate probabilities given the agent's rejected options.

We used two framings of the "least optimal" alternative (which always led to the outcome with $P_{BR} = .1$). In Experiment 2A, all three options were described as alternatives with varying probabilities of success (e.g., three types of fertilizers). In Experiment 2B, the "least optimal" alternative was described as a base rate—the probability of the goal occurring in the absence of any action. This difference in framing is relevant to a test of the average positive-difference strategy (see Section 3.2).

3.1. Method

In Experiment 2A, participants read two vignettes based on those of Experiment 1. For example, the fertilizer vignette read:

Angie has a shrub, and wants the shrub's flowers to turn red. She is thinking about applying a fertilizer, and has three options:
If she applies Formula LPN, there is a 10% chance that the flowers will turn red.
If she applies Formula PTY, there is a 50% chance that the flowers will turn red.
If she applies Formula NRW, there is a [30/70]% chance that the flowers will turn red.
Angie chooses Formula PTY, and the flowers turn red.

For Experiment 2B, the phrase "if she applies Formula LPN" was replaced by the phrase "if she applies nothing." In both experiments, whether Formula NRW had a .3 or .7 chance of leading to the goal (P_{ALT}) was manipulated within-subjects. In the former case, the actual choice was optimal, while in the latter case, the actual choice was suboptimal. The assignment of P_{ALT} to vignette was counterbalanced across participants. Participants rated the agent's responsibility for the outcome on the same 11-point scale as Experiment 1A. The order of the vignettes was randomized separately for each participant.

One hundred participants were recruited from Amazon Mechanical Turk for Experiment 2A, and a different group of 100 participants for Experiment 2B. Four participants from Experiment 2A and seven participants from Experiment 2B were excluded because they answered more than 33% of the four check questions incorrectly. Each experiment was conducted as part of a session that included additional experiments; the order of the experiments was counterbalanced.

3.2. Results and discussion

In all conditions of this experiment, participants rated agents' responsibility for an outcome, where the agents' choice had a constant probability of success ($P_{ACT} = .5$). However, as Fig. 3 shows, agents

were viewed as more responsible when their choice was optimal than when it was suboptimal. This result held no matter whether the "least optimal" choice (with $P_{BR} = .1$) was described as an alternative (Experiment 2A) or as a base rate (Experiment 2B). The agents' choice had positive ΔP, both when this actual choice was optimal and when it was second best. So a strategy that assigns equivalent responsibility whenever $\Delta P > 0$ cannot explain the difference between conditions, whether ΔP is defined relative to the worst option or to the average option.

Over Experiments 2A and 2B, responsibility was higher when $P_{ALT} = .3$ ($M = 6.87$, $SD = 2.02$) than when $P_{ALT} = .7$ ($M = 6.27$, $SD = 2.15$), $t(188) = 4.39$, $p < .001$, $d = 0.32$, $BF_{10} = 531.2$. This result contradicts the positive ΔP strategy. Further, responsibility ratings were higher for optimal decisions regardless of whether the "least optimal" choice was phrased as an alternative or as a base rate: The difference between conditions was about equally large for the two framings of the options, $t(187) = 0.30$, $p = .77$, $d = 0.04$, $BF_{01} = 8.42$. Finally, Fig. 3 shows a main effect of this phrasing, because judgments were higher with the base-rate framing of Experiment 2B ($M = 7.02$, $SD = 1.53$) than with the alternative framing of Experiment 2A ($M = 6.11$, $SD = 2.05$), $t(187) = 3.48$, $p = .001$, $d = 0.51$, $BF_{10} = 33.5$. This may have occurred because describing the least effective option as an omission rather than as an action created a clearer contrast with the remaining options.

A positive difference strategy has difficulty explaining these results when ΔP is calculated relative to the worst option or to the options' average probability. But could the strategy be salvaged if ΔP is calculated in some other way? One further possibility is defining ΔP relative to the average of only the most salient alternatives rather than to the average of all alternatives. Because omissions are treated as less salient causes than actions (Ritov & Baron, 1995) and decision-makers are thought less likely to choose suboptimal omissions than suboptimal actions (Johnson & Rips, 2014), one would expect the ΔP calculations to include the worst option only when it is framed as an action. If so, then ΔP would be positive in both conditions of Experiment 2A, since P_{ACT} (=.5) is greater than the average probabilities of .30 and .43 in the two conditions. In Experiment 2B, however, ΔP would be positive in the $P_{ALT} = .3$ condition but *negative* in the $P_{ALT} = .7$ condition, since P_{ACT} is greater than the average of the non-omission probabilities (.40) in the first but less than the average (.60) in the second. Thus, if people are using only the salient options in their ΔP calculations, one would expect the difference between conditions to be larger for Experiment 2B (where ΔP is positive for one condition but negative for the other) than in Experiment 2A (where ΔP is positive in both conditions). We have seen that no such interaction emerges from the data, and Fig. 3 shows that the mean ratings are in the opposite direction. So positive difference-making delivers the wrong predictions no matter whether ΔP is calculated relative to the probability of the worst alternative, of the average alterative, or the average of the salient alternatives.

Although these results show that optimality plays a role above positive difference-making, even the suboptimal decision-makers were rated above the scale midpoint (5) in their responsibility ratings. Likewise, the responsibility ratings in Experiment 1 were above the midpoint even for suboptimal decision-makers. This may seem to be in tension with our claim that P_{ALT} affects responsibility judgments in a qualitative manner. However, other factors are likely to affect responsibility ratings. For example, the value of P_{ACT} is likely to have an effect: If your decision guaranteed that an outcome occurs ($P_{ACT} = 1.0$), you will likely be judged more responsible than if your decision was the best available but nonetheless had a low probability of leading to the outcome (e.g., $P_{ACT} = .1$). Thus, even though P_{ALT} appears to exert a qualitative effect on responsibility judgments, enhancing responsibility for outcomes caused by optimal decisions, optimality can certainly combine with other factors to affect responsibility in a more graded way. Toward an integrated account of how the efficacies of decision options impact responsibility judgment, we turn to the effects of P_{ACT} in Experiment 3, before exploring the optimality assumption in greater detail.

4. Experiments 3A and 3B: Varying the quality of the actual choice

Causes with higher probabilities of bringing about their effects are ordinarily assigned higher causal strength than causes with lower probabilities (e.g., Cheng, 1997). Therefore, holding optimality constant, one might expect a positive relationship between responsibility judgments and P_{ACT}. In

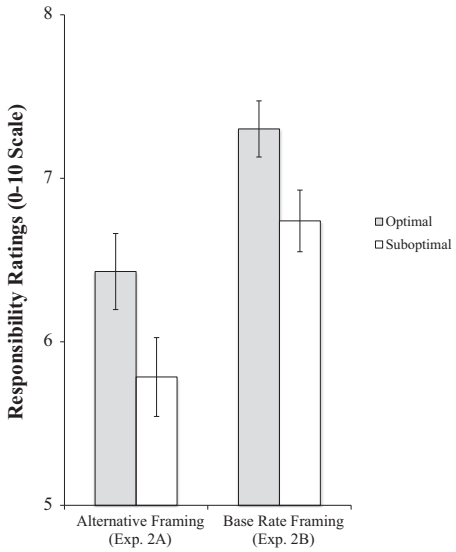

Fig. 3. Results of Experiment 2. Error bars represent ±1 standard error.

Experiment 3, we measure the effect of P_{ACT}, both for decisions where $P_{ALT} < P_{ACT}$ and the decision was therefore optimal (Experiment 3A) and for decisions where $P_{ALT} = P_{ACT}$ and the decision was not optimal (Experiment 3B). In addition to quantifying the effect of P_{ACT} overall, this study also provides the opportunity to replicate the effect of optimality in a between-subjects design—we would expect higher overall responsibility ratings in Experiment 3A than in Experiment 3B, since only in the former cases were the decisions optimal.

4.1. Method

The materials and procedure for Experiment 3 were the same as those of Experiment 1, except that rather than manipulating P_{ALT} across vignettes, we manipulated P_{ACT} (at .3, .4, .5, .6, or .7), as shown in Table 4. We also manipulated optimality across experiments: In Experiment 3A, P_{ALT} was always .1 lower than P_{ACT}, and in Experiment 3B, P_{ALT} was always equal to P_{ACT}. Thus, in Experiment 3A, the agent's decision was optimal, while in Experiment 3B, the decision was non-optimal (i.e., no uniquely optimal choice existed). Half of participants made responsibility judgments and half made causality judgments, as in Experiment 1. However, because wording did not interact with the effects of interest, this variable is not discussed further.

One hundred participants were recruited from Amazon Mechanical Turk for Experiment 3A, and a different group of 100 participants for Experiment 3B. Ten participants from Experiment 3A and four participants from Experiment 3B were excluded because they answered more than 33% of 10 check questions incorrectly.

4.2. Results and discussion

Participants used P_{ACT} in assigning responsibility for the outcomes, both when the decision was optimal (in Experiment 3A) and when it was not optimal (in Experiment 3B), as shown in Fig. 4. However, ratings were higher overall in Experiment 3A than in Experiment 3B, reflecting the optimality strategy found in Experiments 1 and 2.

As suggested by Fig. 4, P_{ACT} produced significant linear trends both for the optimal decisions in Experiment 3A [$t(89) = 3.59$, $p = .001$, $d = 0.38$, $BF_{10} = 32.1$] and the non-optimal decisions in

Table 4A
Design of Experiment 3A.

	Cond. 1	Cond. 2	Cond. 3	Cond. 4	Cond. 5
Choice A (P_{ACT})	**.3**	**.4**	**.5**	**.6**	**.7**
Choice B (P_{ALT})	*.2*	*.3*	*.4*	*.5*	*.6*

Table 4B
Design of Experiment 3B.

	Cond. 1	Cond. 2	Cond. 3	Cond. 4	Cond. 5
Choice A (P_{ACT})	**.3**	**.4**	**.5**	**.6**	**.7**
Choice B (P_{ALT})	*.3*	*.4*	*.5*	*.6*	*.7*

Note: Entries denote the probability of the goal outcome occurring given each choice. The bold entries indicate probabilities given the agent's actual choice, whereas the italicized entries indicate probabilities given the agent's rejected options.

Experiment 3B [$t(95) = 5.53$, $p < .001$, $d = 0.56$, $BF_{10} > 1000$]. As the probability of the outcome given the agent's actual choice increases, so does the agent's perceived responsibility for the outcome. The effect size on the linear contrast was somewhat larger for the non-optimal decision-makers than for the optimal decision-makers ($d = 0.56$ vs. 0.38), although this difference did not reach significance, $t(184) = 1.56$, $p = .121$, $d = 0.23$, $BF_{01} = 2.71$. The effect of P_{ACT} for optimal decision-makers also appeared to be less consistent than for suboptimal decision-makers, as shown in Fig. 4, and this more consistent linear effect in the non-optimal condition led to a significant interaction between P_{ACT} and optimality, $F(4,736) = 2.53$, $p = .040$, $\eta_p^2 = .01$. (However, Bayes Factor analyses revealed that the linear trends are more informative than comparisons between adjacent conditions in this experiment, so the apparent non-monotonicity in the optimal condition is unlikely to be reliable.)

Finally, the main effect of optimality was significant, $t(184) = 2.31$, $p = .022$, $d = 0.34$, $BF_{10} = 1.5$. Just as in Experiments 1 and 2, agents were perceived as more responsible when they acted optimally ($M = 5.89$, $SD = 2.12$) compared to when there was no optimal choice available ($M = 5.12$, $SD = 2.44$). This result is particularly striking in light of the small size of the difference between choices ($P_{ACT} - P_{ALT} = .1$ in Experiment 3A) and the between-subjects design, although the modest Bayes Factor suggests that this result should be interpreted cautiously.

These results begin to paint a picture of how optimality and probability are used together in assigning responsibility. For both optimal and non-optimal decisions, higher values of P_{ACT} led to higher ratings of perceived responsibility. Whereas Experiment 1 revealed a nonlinear effect of ΔP ($P_{ALT} - P_{ACT}$) on responsibility judgments as a result of optimality, P_{ACT} appears to have a more linear effect (though Fig. 4 hints at less stable effects of P_{ACT} for optimal decision-makers). People appear to determine their responsibility judgments based on whether the actual choice is optimal, and adjust upward or downward depending on the efficacy of that choice. However, the probability of the rejected choice, P_{ALT}, does not appear to have any further influence on responsibility judgments, as shown in Experiment 1. Next, we examine how stable these uses of P_{ACT} and P_{ALT} are across different individuals.

5. Experiment 4: Individual differences in responsibility assignment

We have so far described our findings at the group level, averaging across participants and comparing means across conditions. The possibility remains, however, that some of our findings reflect a mix of strategies at the individual level. For example, a subset of participants who were insensitive to counterfactuals and a subset of participants using the optimality principle could lead to a pattern of group means like that in Fig. 1. This possibility is particularly plausible in light of findings in behavioral game theory that people use a variety of strategies in game settings, even differing in how rationally they expect their opponents to behave (e.g., Camerer et al., 2004; Stahl & Wilson, 1995). In Experiment 4, we therefore assess what fraction of participants follow an optimizing strategy and what fraction follow the other strategies listed in Table 1.

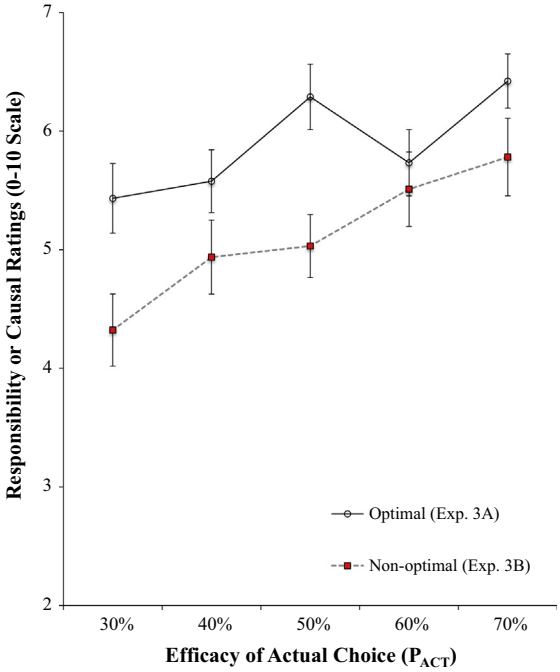

Fig. 4. Results of Experiment 3. Error bars represent ±1 standard error.

A methodological strength of Experiments 1–3 was a weakness for assessing individual differences. In those experiments, we counterbalanced across several cover stories. This allowed us to make the manipulations of P_{ACT} and P_{ALT} less transparent, avoiding concerns about demand characteristics. However, these stories are also likely to have elicited different baseline levels of responsibility. Indeed, very few participants in Experiment 1 had classifiable response patterns across conditions, probably due to the balancing of condition with vignette. Here, we use the same story—the flower story—across nine conditions. We vary only the names of the products (e.g., formulas PTY and NRW) and agents (e.g., Angie or Matt) across vignettes, along with the values of P_{ACT} and P_{ALT}.

To lessen the possibility of experimental demand that could accompany repetitions of very similar vignettes, we varied both P_{ACT} and P_{ALT}, as summarized in Table 5. In five conditions, P_{ACT} was held constant at .5, while P_{ALT} was varied at .1, .3, .5, .7, and .9. In four other conditions, P_{ACT} was equal to P_{ALT} (at .1, .3, .7, and .9), in addition to the condition where $P_{ACT} = P_{ALT} = .5$. This design allows us to assess the possibility of individual differences for both the P_{ALT} manipulation (analogous to Experiment 1) and the P_{ACT} manipulation (analogous to Experiment 3). It also allows us to see whether individuals who use different strategies for assigning responsibility in light of P_{ALT} also use systematically different strategies for assigning responsibility in light of P_{ACT}.

5.1. Method

The materials were similar to those for Experiments 1 and 3, except that only the flower cover story was used. The probabilities were varied across conditions, as specified in Table 5, and the names of the agents and products were also varied across stories to reduce carry-over effects. All participants made responsibility judgments on the same scale used in Experiments 1–3, and items were presented in a random order.

One hundred participants were recruited from Amazon Mechanical Turk for Experiment 4. A series of 18 check questions was included at the end of the study. However, these questions were more

Table 5
Design of Experiment 4.

	Cond. 1	Cond. 2	Cond. 3	Cond. 4	Cond. 5	Cond. 6	Cond. 7	Cond. 8	Cond. 9
Choice A (P_{ACT})	**.5**	**.5**	**.5**	**.5**	**.5**	**.1**	**.3**	**.7**	**.9**
Choice B (P_{ALT})	*.1*	*.3*	*.5*	*.7*	*.9*	*.1*	*.3*	*.7*	*.9*

Note: Entries denote the probability of the goal outcome occurring given each choice. The bold entries indicate probabilities given the agent's actual choice, whereas the italicized entries indicate probabilities given the agent's rejected options.

difficult than those used in previous studies (listing the names of the characters in the vignettes, rather than the contents of the vignettes, as this was the most salient feature that varied), so excluding participants with a greater than 33% error rate would lead to an unacceptable exclusion rate (indeed, the median error rate was 33.3%). We instead included all participants in data analysis but used performance on these questions to assess attentiveness.

5.2. Results

5.2.1. Replications of Experiments 1 and 3

Before categorizing participants into strategy groups, we checked that the overall results of Experiments 1 and 3 were replicated. This was not a foregone conclusion, as the use of very similar vignettes across condition may have created twin demand characteristics—a pressure to respond linearly to a salient manipulation and a pressure to give consistent responses across items—which both work against finding an optimality strategy (favoring a linear ΔP strategy and an alternative-insensitive strategy, respectively). Nonetheless, both the optimality pattern found in Experiment 1 for P_{ALT} and the linear pattern found for P_{ACT} in Experiment 3 were replicated, as shown in Fig. 5.

First, we compared the five conditions where P_{ACT} was set at .5, and P_{ALT} varied between .1, .3, .5, .7, and .9 (the solid line in Fig. 5). Thus, the decision was optimal in the .1 and .3 conditions, suboptimal in the .7 and .9 conditions, and neither optimal nor suboptimal in the .5 condition. Similar to Experiment 1, the .3 and .5 conditions significantly differed, with attributions of responsibility higher when the decision was optimal at $P_{ALT} = .3$ than when neither decision was optimal at $P_{ALT} = .5$, $t(99) = 3.36$, $p = .001$, $d = 0.34$, $BF_{10} = 15.8$. However, consistent with the step function predicted by an optimality strategy and found in Experiment 1, the .1 and .3 conditions did not differ [$t(99) = -0.92$, $p = .36$, $d = -0.09$, $BF_{01} = 8.33$] nor did the .7 and .9 conditions [$t(99) = 1.11$, $p = .27$, $d = 0.11$, $BF_{01} = 6.94$]. Whereas in Experiment 1, there was equivocal evidence for a difference between the .5 and .7 conditions (with the Bayes factor slightly favoring the null hypothesis), the Bayes factor here clearly favors the null hypothesis at the group level, $t(99) = 0.92$, $p = .36$, $d = 0.09$, $BF_{01} = 8.37$.

Second, we compared the five conditions where $P_{ACT} = P_{ALT}$, where neither decision was optimal and where these probabilities varied between .1, .3, .5, .7, and .9. The linear trend found in Experiment 3 was replicated, with a highly significant linear contrast, $t(99) = 5.91$, $p < .001$, $BF_{10} > 1000$. That is, just as in Experiment 3, higher values of P_{ACT} were associated with higher ratings of responsibility, even though P_{ALT} was always equal to P_{ACT}. When $P_{ACT} = P_{ALT} = .1$, responsibility ratings were lower than the scale midpoint, whereas when $P_{ACT} = P_{ALT} = .9$, responsibility ratings were higher than the midpoint.

5.2.2. Individual differences in P_{ALT} strategies

The main goal of Experiment 4 was to assess what proportion of participants were driving the group-level optimality strategy, and what proportion followed other strategies. To categorize participants into strategy groups, we computed four difference scores for each participant, reflecting their responses to the five items for which $P_{ACT} = .5$ and P_{ALT} varied (at .1, .3, .5, .7, and .9). These difference scores were: (a) the participant's responsibility rating at $P_{ALT} = .1$ – the same participant's responsibility rating at $P_{ALT} = .3$; (b) the responsibility rating for $P_{ALT} = .3$ – responsibility rating at $P_{ALT} = .5$; and so on. Participants are broken down by P_{ACT} and P_{ALT} strategy groups in Table 6. We first consider how variations in P_{ALT} affected the full set of participants, as represented by the final column in the table.

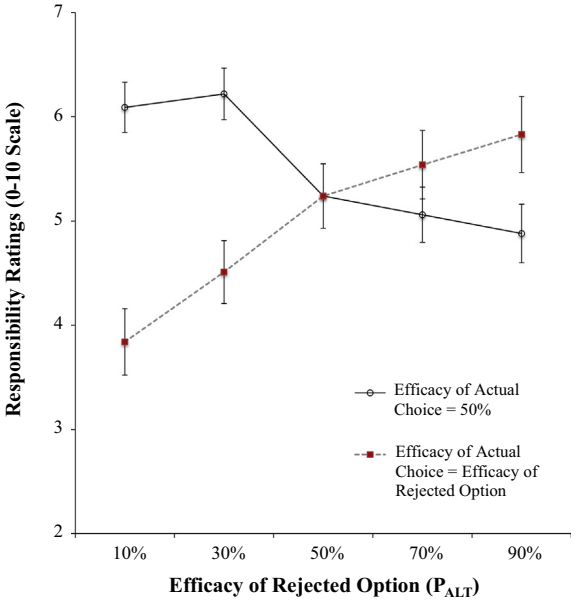

Fig. 5. Overall means from Experiment 4. Error bars represent ±1 standard error.

A participant following a *linear* ΔP strategy would have positive difference scores between the .1 and .3, the .3 and .5, the .5 and .7, and the .7 and .9 conditions. This is because the probability of success given the actual choice (relative to the alternative choice) decrease as P_{ALT} changes from .1 to .3 to .5, and so on (see Table 5). Given the step function in Fig. 5 and the lack of any difference in the group-level analysis between the .1 and .3 conditions and between the .7 and .9 conditions, it may not be surprising that few participants appear to have followed a linear ΔP strategy. In fact, no participant in our sample had positive difference scores for all four of these differences. Even on a looser criterion of having 3 out of 4 positive difference scores, only 6 out of 100 participants could be categorized as following a ΔP strategy.

A participant following an *alternative-insensitive* strategy would give similar responsibility ratings regardless of the value of P_{ALT}, and would hence have difference scores of 0 for the .1–.3, .3–.5, .5–.7, and .7–.9 comparisons. Of the 100 participants in our sample, 30 followed this pattern. This number may overestimate the proportion who would follow this strategy in a design with less pressure for response consistency. But the error rates on the check questions were similar for this group and for the sample as a whole (both medians were 33.3%, as were the medians for the linear ΔP group and for the optimizing group described below), suggesting that these participants were not merely perseverating out of fatigue or inattentiveness. Overall, this suggests that a sizable minority uses an alternative-insensitive assignment strategy.

A few participants (8 out of 100) exhibited an "anti-optimality" pattern, with a positive difference score for either the .1–.3 or the .7–.9 comparison but for neither the .3–.5 or .5–.7 comparison. Since this pattern was relatively rare and did not conform to any of the predicted strategies, we suspect this pattern may reflect inattentiveness rather than anything deeper about responsibility assignment (indeed, half of these participants responded at chance levels on the check questions, with a median error rate of 41.7%).

The remaining 56 participants followed an optimality strategy, with no more than two positive difference scores, of which at least one was for the .3–.5 or the .5–.7 comparison. As shown in Fig. 6, these participants tended to give similar ratings in the .1 and .3 conditions and in the .7 and .9 conditions. However, this group of participants distinguished strongly between the .3 and .5 conditions, giving

Table 6
Number of participants classified as following each strategy in Experiment 4.

P_{ALT} strategy	Positive sensitivity to P_{ACT}	Non-positive sensitivity to P_{ACT}	Total
Optimality	32	24	**56**
Linear ΔP	4	2	**6**
Alternative-insensitive	17	13	**30**
Total	**53**	**39**	**92**

much higher responsibility ratings when a decision was optimal than when it was not. This pattern is consistent with the aggregate trends found in Experiments 1 and 4, where participants were sensitive to *whether* a decision was optimal, but not to the *degree* to which it was better or worse than the alternative.

Thus, of the 92 participants who used one of the responsibility assignment strategies summarized in Table 1, most (about 61%) followed an optimality strategy while a large minority (about 33%) followed an alternative-insensitive strategy. Very few participants (about 7%) followed a linear ΔP strategy—an especially striking finding in light of the demand characteristics to respond in proportion to the salient manipulation of P_{ALT}.

5.2.3. Individual differences in P_{ACT} strategies

Our design also allowed us to look for individual differences in the use of P_{ACT} for assigning responsibility, across the five conditions where $P_{ACT} = P_{ALT}$, varying between .1, .3, .5, .7, and .9 (see Table 5). Using our analysis strategy in Experiment 2, we computed a single linear contrast for each participant across the P_{ACT} conditions. These contrasts were much more frequently positive than negative (58% vs. 12%), suggesting that the effect of P_{ACT} was relatively consistent across participants. However, as we found in the preceding section for the P_{ALT} manipulation, a sizable minority (30% of participants) were not sensitive to P_{ACT}.

Interestingly, the participants who were insensitive to P_{ACT} were not the same as those who were insensitive to P_{ALT}. The first two columns in Table 6 divide participants into those who were sensitive to P_{ACT} and those who were not, and they show that the distribution of P_{ALT} strategies was similar for both P_{ACT} groups. That is, having an insensitivity strategy for P_{ACT} was not associated with having an insensitivity strategy for P_{ALT}. Indeed, the linear effect of P_{ACT} was larger among those *insensitive* to P_{ALT} than among those using an optimality strategy, $t(84) = 2.25$, $p = .027$, $BF_{10} = 1.70$, although we interpret this result cautiously given the modest Bayes factor. It is nonetheless consistent with our finding in Experiment 3 that the linear effect of P_{ACT} was somewhat weaker when agents behaved optimally (i.e., when $P_{ACT} > P_{ALT}$), in that people who used an optimality strategy here were somewhat less sensitive to absolute levels of P_{ACT}.

5.3. Discussion

Taken together, these results replicate the overall findings of Experiments 1 and 3 in showing an optimality pattern in responding to changes in P_{ALT} and a linear pattern in responding to P_{ACT}. They also testify to the robustness of these patterns over participants. Although a sizable minority of about 30% of participants were insensitive to P_{ALT} and a (different) minority of about 30% of participants were insensitive to P_{ACT}, both effects were robust over the majority of participants. These results are especially interesting in light of findings from behavioral game theory (e.g., Camerer et al., 2004; Stahl & Wilson, 1995) that people differ in their assumptions about other players' rationality. One possible direction for future research would be to see to what extent these individual differences are stable across tasks, and whether they are associated with other individual differences (such as individualist/collectivist cognitive style, intelligence, or personality factors).

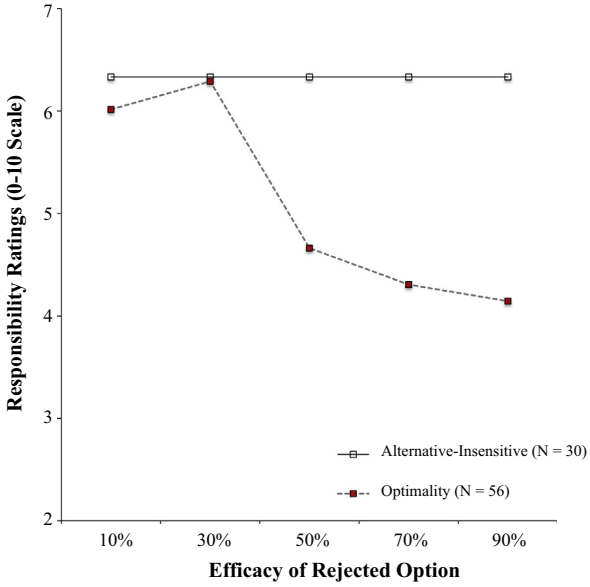

Fig. 6. Means broken down by predominant strategy groups in Experiment 4.

6. Experiment 5: Local and global optimality

Sometimes an agent has multifaceted priorities, and the optimal means toward some particular end may not maximize the agent's overall utility. An example from Audi (1993) illustrates this point:

> Suppose I want to save money, purely as a means to furthering my daughter's education. . .. I might discover that I can save money by not buying her certain books which are available at a library, and then, in order to save money, and with no thought of my ultimate reason for wanting to do so, decline to buy the books. By doing so, I might act rationally in a narrow instrumentalist sense.... If, however, the damage to her education from not owning the books obviously outweighs—and should have been seen by me to outweigh—the benefits to her of the saving, the action is not rational from the broad instrumentalist point of view. .. . (p. 289).

Not buying the books is an efficient action with respect to the goal of saving money, but not a rational action on the part of the agent because it hampered the agent's broader set of goals, in particular his daughter's education. In cases where local and global optimality come into conflict, which is the critical factor for assigning responsibility?

Toward answering this question, participants in Experiment 5 read vignettes such as the following (see Table 7):

> *Jill is shopping for a new shampoo, and wants her hair both to smell like apples, and to curl up. Both of these goals are equally important to her. She is considering three brands of shampoo to use:*
> *If she uses Variety JLR, there is a 70% chance that her hair will smell like apples and a 70% chance that her hair will curl up.*
> *If she uses Variety WYZ, there is an 80% chance that her hair will smell like apples and a 40% chance that her hair will curl up.*
> *If she uses Variety HPN, there is a 40% chance that her hair will smell like apples and an 80% chance that her hair will curl up.*
> *Jill chooses Variety JLR; then, her hair smells like apples and her hair curls up.*

In this situation, Jill's locally optimal choice for making her hair smell like apples is Variety WYZ, while the locally optimal option for making her hair curl up is Variety HPN, because those choices maximize the probability of their respective goals being satisfied. However, because Jill is said to weight the goals equally, the *globally* optimal choice that maximizes her overall utility would be Variety JLR.

Given that we know from Experiments 1–4 that most people use a decision-maker's optimality in assigning responsibility, what should we predict about responsibility judgments in this situation? Consider the assignment of responsibility to Jill for her hair smelling like apples. From the results of our previous experiments, we would expect her to have the highest responsibility when she chooses Variety WYZ, because it maximizes the probability of that goal. However, given Jill's overall preference set, the optimal action would be to choose Variety JLR. This leads to two competing predictions: People might assign responsibility for goal *X* according to the local optimality or efficiency of the action for reaching the goal *X* (Baker, Saxe, & Tenenbaum, 2009; Csibra, Gergely, Bíró, Koós, & Brockbank, 1999); or according to the action's global optimality for satisfying the agent's overall goals, even if it is not locally optimal for goal *X*.

6.1. Method

Each participant read three vignettes similar to the shampoo example, in which agents simultaneously attempted to satisfy two goals of equal priority. Table 7 summarizes the probabilities in each condition. Each participant was randomly assigned one of the two goals as the *focal goal* for which they completed ratings (e.g., "How responsible is Jill for her hair smelling like apples?" vs. "How responsible is Jill for her hair curling up?"). We varied which choice the agent made (globally optimal, optimal for focal goal, or optimal for non-focal goal) as a within-subject factor. The content of the vignette and the agent's choice was balanced using a Latin square. Participants completed responsibility ratings using the same scale as in previous experiments. Vignettes appeared in a random order.

One hundred and forty-two participants were recruited from Amazon Mechanical Turk for this experiment. The experiment was conducted as part of a session that included an additional experiment, and the order of the experiments was balanced. Thirty-one participants were excluded because they incorrectly answered more than 33% of the 12 check questions. Despite this somewhat high exclusion rate, the results remain the same if all participants are included in the analyses.

6.2. Results and discussion

Participants' responsibility ratings differed for the choices that were globally optimal, optimal for a focal goal, or optimal for a non-focal goal. Ratings were higher for the focal goal when agents made the locally optimal choice for that goal than when they made the locally optimal choice for the alternative, equally valued goal [$M = 6.87$, $SD = 2.45$ vs. $M = 5.93$, $SD = 2.45$; $t(110) = 3.73$, $p < .001$, $d = 0.35$, $BF_{10} = 49.7$]. For example, participants rated Jill in our earlier example as more responsible for her hair smelling like apples when she chose the shampoo (WYZ) with the best chance of producing that effect than when she chose the shampoo (HPN) with the best chance of curling her hair. Further, agents were actually rated *more* responsible for the focal goal after making the globally optimal choice than after making the locally optimal choice for the focal goal [$M = 7.59$, $SD = 2.03$ vs. $M = 6.87$, $SD = 2.45$; $t(110) = 2.81$, $p = .006$, $d = 0.27$, $BF_{10} = 3.3$]. This was true even though the globally optimal choice was a less efficient means to the focal goal compared with the locally optimal choice. Jill was rated more responsible for her hair smelling like apples when she chose the shampoo (JLR) that best fulfilled both her goals as when she chose the shampoo (WYZ) that specifically maximized the apple smell. Thus, participants took global optimality, as well as local optimality, into account when making responsibility attributions.

These results suggest that optimality expectations are *disjunctive*: Most people appear to compute optimality both relative to the agent's global utility, and relative to whichever particular goal is under consideration. The agent's responsibility was somewhat lower for a focal goal if the agent acted optimally for the focal goal but in a globally suboptimal way, and was much lower if the agent acted optimally for the non-focal goal.

Table 7
Design of Experiment 5.

	Cond. 1		Cond. 2		Cond. 3	
	Focal goal	Other goal	Focal goal	Other goal	Focal goal	Other goal
Choice A	**.7**	**.7**	*.7*	*.7*	*.7*	*.7*
Choice B	*.8*	*.4*	**.8**	**.4**	*.8*	*.4*
Choice C	*.4*	*.8*	*.4*	*.8*	**.4**	**.8**

Note: Entries denote the probability of each goal outcome occurring given each choice. The bold entries indicate probabilities given the agent's actual choice, whereas the italicized entries indicate probabilities given the agent's rejected options.

Sensitivity to global optimality strongly distinguishes the optimality strategy from other ways of assigning responsibility. Participants seem to be taking into account the agent's overall rationality in evaluating their responsibility for achieving a goal, not merely their efficiency, narrowly construed, toward achieving that particular goal. An alternative-insensitive strategy would have particular difficulty accommodating such a finding, since this demonstrates sensitivity to *two* different aspects of the decision alternatives: Global optimality is dependent not only on the quality of available alternatives relative to the focal goal, but also on the quality of these alternatives relative to the agent's overall set of preferences. Further, this result is consistent with our explanation for why optimality influences responsibility. Since rationality is used as an agency cue, a globally optimal person should seem most responsible for the intended outcomes if global optimality is the critical factor for determining a person's degree of rationality.

However, a variant on the linear ΔP strategy could potentially accommodate these results if the additional assumption is made that both the focal and non-focal goals are used, but the focal goal is weighted more heavily. For example, if the weight on the focal goal was .6 and the weight on the non-focal goal was .4, then the globally optimal choice would have the highest score on this measure ($.7 \times .6 + .7 \times .4 = .70$), the locally optimal choice for the focal goal would have the next highest score ($.8 \times .6 + .4 \times .4 = .64$), and the locally optimal choice for the non-focal goal would have the lowest score ($.4 \times .6 + .8 \times .4 = .56$). This strategy would be similar to responding on the basis of expected utility, but with greater weight on the focal goal. We addressed this possibility in Experiment 6.

7. Experiment 6: Varying the probability of the non-focal goal

In the present experiment, we manipulate global optimality in a more subtle way than we did in the previous experiment, this time asking whether responsibility for a focal goal is sensitive to manipulations of the choice's efficacy for *non-focal* goals. More concretely, consider how Jill's response to her newest conundrum could affect her responsibility for her hair smelling like apples (where P_{ACT} of the non-focal goal is varied across conditions as indicated in brackets):

> Jill is shopping for a new shampoo, and wants her hair both to smell like apples, and to curl up. Both of these goals are equally important to her. She is considering three brands of shampoo to use:
> If she uses Variety JLR, there is a 70% chance that her hair will smell like apples and a [70/50/40]% chance that her hair will curl up.
> If she uses Variety WYZ, there is an 80% chance that her hair will smell like apples and a 40% chance that her hair will curl up.
> Jill chooses Variety JLR; then, her hair smells like apples and her hair curls up.

Table 8 lists the relevant probabilities in the three conditions of this experiment. In all three conditions, Variety JLR is locally inefficient for making her hair smell like apples (the focal choice), because the alternative choice (Variety WYZ) has a higher probability of making her hair smell like apples (i.e., $P_{ACT} = .7$ and $P_{ALT} = .8$ for the focal goal). However, the *global* optimality of Variety JLR varies depending on the probability of the non-focal goal. When the non-focal goal has $P_{ACT} = .7$, then her choice is globally optimal, because her overall utility (across both equally weighted goals) is maximized by her

actual choice. When the non-focal goal has P_{ACT} = .5, then both choices are globally equivalent, but differ in local efficiencies (the alternative choice is locally efficient for the focal goal). Finally, when the non-focal goal has P_{ACT} = .4, then her actual choice is both globally suboptimal and locally suboptimal for the focal goal.

On the basis of Experiment 5, we would expect Jill to be rated more responsible for the focal goal when her choice was globally optimal (i.e., P_{ACT} = .7 for the non-focal goal) than when her choice was globally suboptimal (i.e., P_{ACT} = .4 for the non-focal goal). Less clear, however, is where the P_{ACT} = .5 condition would fall. On the one hand, the difference in expected utility is greater between the .7 and .5 conditions than between the .5 and .4 conditions, so any strategy responding linearly to expected utility or a weighted combination of the focal and non-focal probabilities would predict a larger gap between the .7 and .5 conditions. However, if agency is discounted in light of a decision-maker's suboptimal behavior, then we might expect a *bigger* gap between the .5 and .4 conditions than between the .7 and .5 conditions, because only in the .4 condition is the actual choice globally suboptimal (despite being locally inefficient for the focal goal in all conditions).

7.1. Method

Each participant read three vignettes similar to the above example, where P_{ACT} for the non-focal goal was varied within-subjects between .7, .5, and .4, as summarized in Table 8. The assignment of vignette content and P_{ACT} for the non-focal goal was balanced using a Latin square. Participants completed responsibility ratings using the same scale as previous experiments. The responsibility ratings concerned the focal goal only (e.g., "How responsible is Jill for her hair smelling like apples?" for the earlier example). Vignettes appeared in a random order.

Three hundred participants were recruited from Amazon Mechanical Turk for this experiment. The experiment was conducted as part of sessions that included additional experiments, and the order of the experiments was balanced. Fifty-four participants were excluded because they incorrectly answered more than 33% of 9 check questions. The results remain the same if all participants are included in the analyses.

7.2. Results and discussion

The agents' perceived responsibility varied, depending on P_{ACT} of the non-focal goal. Specifically, responsibility was lower when P_{ACT} = .4 than when P_{ACT} = .5 for the non-focal goal [M = 6.26, SD = 2.61 vs. M = 6.85, SD = 2.38; $t(245) = -4.10$, $p < .001$, $d = -0.26$, BF_{10} = 166.5]. That is, when Jill's choice was globally suboptimal, her responsibility for the focal goal was discounted, even though the efficiency of achieving the focal goal (her hair smelling like apples) was the same in both conditions. However, the difference between the P_{ACT} = .5 and P_{ACT} = .7 conditions reached only marginal significance [M = 6.85, SD = 2.38 vs. M = 7.09, SD = 2.33; $t(245) = -1.78$, p = .077, $d = -0.11$, BF_{01} = 4.14]. Even though the difference in global utility was larger between the .7 and .5 conditions than between the .5 and .4 conditions, only the latter difference led Jill's responsibility to be discounted reliably.

Experiment 6 reaffirms our finding in Experiment 5 that global optimality matters in assessing responsibility. Even when assessing an agent's responsibility for a focal goal (e.g., Jill's hair smelling like apples), people discounted the agent's responsibility for that goal if another option would have led to higher overall utility. This held despite the fact that the agent's actual choice was equally efficient for achieving the focal goal in all conditions. In addition, these results show that responsibility judgments are principally driven by optimality, not by linear responding to a weighted combination of the probabilities, because this strategy would lead to a large difference between the .7 and .5 conditions, but a smaller difference between the .5 and .4 conditions. For example, if participants placed a weight of .6 on the focal goal and .4 on the non-focal goal, then the .7 condition would have the largest score on this measure for the agent's actual choice ($.7 \times .6 + .7 \times .4 = .70$), followed by a substantial drop for the .5 condition ($.7 \times .6 + .5 \times .4 = .62$), followed by a smaller drop for the .4 condition ($.7 \times .6 + .4 \times .4 = .58$). Instead, we found the opposite—a substantial difference between the .5 and .4 conditions (because the agent's choice was suboptimal in the latter case) but little evidence for a

Table 8
Design of Experiment 6.

	Cond. 1		Cond. 2		Cond. 3	
	Focal goal	Other goal	Focal goal	Other goal	Focal goal	Other goal
Choice A (P_{ACT})	**.7**	**.7**	**.7**	**.5**	**.7**	**.4**
Choice B (P_{ALT})	*.8*	*.4*	*.8*	*.4*	*.8*	*.4*

Note: Entries denote the probability of each goal outcome occurring given each choice. The bold entries indicate probabilities given the agent's actual choice, whereas the italicized entries indicate probabilities given the agent's rejected options.

difference between the .7 and .5 conditions (where the agent's choice was always locally inefficient, but not globally suboptimal). A weighted linear ΔP strategy would be unable to explain this result.

However, this last finding raises an apparent inconsistency with our earlier experiments. In Experiments 1 and 4, most participants assigned higher responsibility when the agent's decision was optimal (i.e., $P_{ACT} > P_{ALT}$) than when it was suboptimal ($P_{ACT} < P_{ALT}$), but ratings were also relatively low when neither decision was optimal ($P_{ACT} = P_{ALT}$). Here, in contrast, when neither decision was globally optimal, responsibility ratings were similar to the globally optimal condition, and higher than the globally suboptimal condition. Although further study is needed to pinpoint the reason for this difference, we think this is likely due to the agent's perceived ability to act for reasons in deciding between the options in the globally equivalent condition of Experiment 6, but not in the $P_{ACT} = P_{ALT}$ condition of Experiments 1 and 4. That is, in Experiments 1 and 4, in the $P_{ACT} = P_{ALT}$ condition, there was no reason for agents to choose either option in particular, and hence the outcome may have been seen as outside their control (since not only their expected utility, but also the probability of the focal goal was the same regardless of what the agents did). But in Experiment 6, even in the .5 condition where both choices were globally equivalent, they differed with respect to their local efficiency. Thus, choosing one option over the other allowed the agents to steer their fate, even though it did not affect their overall expected utility. When both options are equivalent with respect to their overall utility, the agent could generate reasons for choosing either. It is only when the actual choice is globally suboptimal that the agent is not acting in accord with reason.

8. Experiments 7A and 7B: Tacit knowledge of conditional probabilities

We have so far explored lay decision theory using controlled vignettes that indicated optimal choice with exact probabilities. Unfortunately, life seldom wears probabilities on its sleeve: Would optimality assumptions also extend to more realistic decision environments that do not explicitly quantify uncertainty?

Experiment 7 addressed this question by using vignettes about decisions for which participants might have prior beliefs, omitting explicit mention of decision efficacies. We began by pretesting options for particular goals to determine participants' degree of agreement on these options. For example, we asked participants in the pretest to rate the likelihood that liquid glue would hold a science project together and to rate the likelihood that a glue stick would hold a science project together. In Experiment 7A, we used decisions for which the pretest revealed a consensus on which option was superior. For instance, participants believed that liquid glue is more likely to keep a science project intact than is a glue stick, so we hypothesized that a decision-maker who used liquid glue would be seen as more responsible when her project in fact stayed intact, compared to a decision-maker who used a glue stick.

In Experiment 7B, we used decisions for which the pretest revealed a lack of consensus. For instance, participants were divided on whether root beer was a more likely beverage than Dr. Pepper to be enjoyed by someone new to soft drinks. Suppose Mark gives his friend a Dr. Pepper (rather than root beer) and Mark's friend enjoys it. Then people should assign Mark relatively high responsibility for his friend enjoying the soft drink if they think Dr. Pepper is the superior beverage. However, they should assign him relatively low responsibility if they do not believe that Dr. Pepper is superior.

8.1. Pretest

Each participant read 40 decision situations, each involving a choice between two or three potential options (e.g., "Mark is recommending a soda for his best friend, who has never tried soft drinks before. He is deciding between two beverages"). Participants rated the probability of an outcome occurring given each option (e.g., "Suppose Mark gives his friend **Dr. Pepper**. In your opinion, what is the probability that his friend will enjoy the taste of the beverage?" [emphasis in original]). The order of the items and choices within each item were randomized. Thirty participants were recruited from Amazon Mechanical Turk for the pretest.

From this set of 40 items, we selected two sets of ten items—a *High Consensus* set and a *No Consensus* set. The High Consensus items were those on which participants tended to agree that one option led to the outcome with greater probability than the other did. Averaged across items, the mean difference in probability between the optimal and suboptimal options was 29.22 (SD = 29.82), measured on a 0-to-100 scale. The No Consensus items were those on which participants were in disagreement about which option was more effective. The mean difference in probability between the higher- and lower-rated options was 5.66 (SD = 28.92) on a 0-to-100 scale.

8.2. Method

In Experiment 7A, participants made judgments about modified versions of the High Consensus items that emerged from the pretest. Table 9 lists the full set of items. We presented the problems in the following format:

> *Mary is making a model solar system for her school science fair. She is deciding what kind of glue to use to hold the pieces of her project together:*
>
> *The first option is to use* ***liquid glue****.*
>
> *The second option is to use* ***a glue stick****.*
>
> *Mary decides to use [**liquid glue**/**a glue stick**]. As a result, all of her project stays intact.*

In the optimal version of the item, Mary used liquid glue, and in the suboptimal version, Mary used a glue stick. Which option (e.g., liquid glue or glue stick) the vignette described as the first option was randomized. Participants then rated their agreement with the statement "Mary is responsible for all of her project staying intact" on the same 11-point scale used in the other experiments. Each participant saw five items in which the character acted optimally and five items in which the character acted suboptimally, according to the pre-test ratings. Across participants, each vignette appeared equally often with the optimal and with the suboptimal choice.

Participants in Experiment 7B made judgments about modified versions of the ten No Consensus items from the pretest. Table 10 lists the problems based on these items, and their format was the same as that in Experiment 7A. Because we did not know in advance which participants would believe a given choice to be optimal for these items, we randomized each agent's choice in each vignette, as well as the order in which the options were listed. After making responsibility judgments as in Experiment 7A, participants completed a *choice evaluation* task, in which they were asked their opinion about the efficacy of each option in the following way:

> *Suppose a person is recommending a soda for their best friend, who has never tried soft drinks before, and has the goal of choosing a soda that their friend will enjoy.*
>
> *Which of the following courses of action is most likely to lead to the person's friend enjoying their first soft drink?*
>
> *(A) The person chooses Dr. Pepper.*
>
> *(B) The person chooses root beer.*

Participants answered this question on an 11-point scale (0: "Option A much more likely to lead to goal"; 10: "Option B much more likely to lead to goal"). Participants always completed these ratings

Table 9
Stimuli and results from Experiment 7A.

Decision (decision-maker underlined)	Goal/outcome	Options	Pretest	Resp. rating
Mary is making a model solar system for her school science fair. She is deciding what type of glue to use to hold the pieces of her project together	All of her project stays in tact	Using liquid glue	76.17 (14.77)	7.47 (2.18)
		Using a glue stick	47.13 (23.00)	5.87 (2.09)
Kevin is filling up his water thermos while getting ready for work in the morning. He is deciding between two sources of water, making his choice only based on taste	The water in his thermos tastes clean	Using bottled water	89.40 (10.53)	6.19 (3.28)
		Using tap water from his kitchen sink	52.27 (27.29)	4.33 (3.27)
Steven has just finished buying groceries. He would like to be able to carry all of his groceries in one trip, and the cashier is deciding between two ways of bagging them	Steven is able to carry all of his groceries in one trip	Using plastic bags	75.63 (16.91)	5.59 (2.96)
		Using paper bags.	44.23 (22.76)	5.01 (2.84)
Along with her normal diet and exercise, Anna is planning to skip one meal per day to lose weight. Anna's friend is deciding which meal Anna should skip to try and accelerate her weight loss	Anna's weight loss accelerates	Skipping dinner	56.13 (29.11)	4.32 (2.83)
		Skipping breakfast	20.80 (18.92)	3.95 (2.39)
Jonathan's feet hurt after each day at work. He is hoping that purchasing a new pair of shoes will help solve this problem. He is deciding between two types of shoes	Jonathan's feet no longer hurt after work	Running shoes	71.93 (13.95)	7.78 (1.97)
		Loafers	49.63 (18.59)	6.95 (2.21)
Paul wants to lose weight, and is considering a new diet to [try] along with his workout regimen. Paul's friend is deciding which diet program to recommend to Paul	Paul lost ten pounds	Mostly proteins/few carbohydrates	72.80 (19.67)	4.84 (2.51)
		Mostly carbohydrates/few proteins	25.60 (20.70)	3.76 (2.55)
Noah asked to borrow a pencil from his friend for an essay he has to write. His friend wants to make sure the lead does not break in the middle of writing, and is deciding between two pencils to give Noah	Noah's lead did not break while writing	Giving Noah a wooden pencil	75.50 (17.06)	4.17 (2.89)
		Giving Noah a mechanical pencil	55.20 (21.27)	4.01 (2.76)
Lauren is mailing a package, and wants to make sure it is securely sealed for shipping. She is deciding between two different types of tape to seal the box	The box stayed sealed	Duct tape	84.93 (18.28)	5.78 (2.98)
		Masking tape	56.90 (25.95)	5.76 (2.45)

(*continued on next page*)

Table 9 (*continued*)

Decision (decision-maker underlined)	Goal/outcome	Options	Pretest	Resp. rating
Sam's nose is bleeding, and he would like it to stop. He is deciding between two different methods of treatment	Sam's nosebleed stopped	Tilting his head slightly back	71.67 (19.16)	6.54 (2.99)
		Tilting his head slightly forward	46.97 (30.38)	6.39 (2.69)
Taylor is purchasing a new smartphone. She is deciding between two different brands, with battery life being the most important quality	Taylor is able to go through a day of normal use without recharging her phone	Samsung Galaxy S4	77.57 (13.66)	5.24 (2.83)
		iPhone 5	60.80 (26.77)	5.62 (2.98)
Mean		Optimal	75.17	5.79
		Suboptimal	45.95	5.17

Note: For each item, the optimal choice is given in the first row and the suboptimal choice in the second row. The *Pretest* column lists the mean estimate of *P*(Goal | Choice) for each decision option, as estimated by participants in the pretest. The *Resp. Rating* column lists the mean responsibility rating in Experiment 7A. (SDs in parentheses.)

Table 10
Stimuli and results from each item/choice combination in Experiment 7B.

Decision (decision-maker underlined)	Goal/outcome	Options	Optimal participants		Nonoptimal participants	
			%	*M*	%	*M*
Angie has a shrub and wants the shrub's flowers to bloom. She is deciding between two brands of fertilizer to apply	The flowers bloom	Use a brand with nitrogen	40.6	6.41 (2.56)	59.4	5.48 (2.57)
		Use a brand with phosphorus	35.1	6.09 (2.45)	64.9	5.91 (2.07)
James is planning on buying a new pair of headphones to block out the sound of construction outside of the building he works in. He is deciding between two types of headphones	The sound of construction is blocked out	In-ear headphones	41.5	6.13 (3.27)	58.5	5.81 (2.68)
		Over-the-ear headphones	62.5	6.45 (2.55)	37.5	5.69 (2.22)
Mark is recommending a soda for his best friend, who has never tried soft drinks before. He is deciding between two beverages	His friend enjoys the taste of his first soft drink	Give his friend Dr. Pepper	37.9	5.89 (2.32)	62.1	5.68 (2.45)
		Give his friend root beer	33.7	6.15 (2.84)	66.3	5.51 (2.43)
Isaac's friend Bill wants to purchase a new car, and considers fuel efficiency to be a top priority. Bill asked Isaac for his advice about which car to purchase, and Isaac is deciding between two models to recommend	Bill only needs to fill up his gas tank once every two weeks	Recommend a Honda Civic	50.5	5.19 (2.58)	49.5	4.30 (2.45)
		Recommend a Mini Cooper	40.2	5.15 (2.55)	59.8	4.69 (2.02)
Claire is making coffee for Frank, and is unsure of what type of milk to use. She is deciding between two types of milk, and making the choice only based on taste	Frank is satisfied with the taste of his coffee	Use skim milk	42.7	6.51 (2.41)	57.3	6.41 (1.94)
		Use almond milk	36.2	6.60 (2.86)	63.8	6.64 (2.01)
Allen has a blister on his feet from breaking in new shoes. He is deciding between two methods of dealing with the blister	The blister heals within two weeks	Pop the blister	37.6	5.81 (2.59)	62.4	5.33 (2.88)
		Leave the blister alone	66.0	6.70 (2.53)	34.0	6.22 (1.62)
Ross just moved to his new home in a big city. He would like to be able to commute to work each morning in less than half an hour, and asks his wife for advice regarding which transportation method to use. Ross's wife is deciding between two modes of transportation to recommend	Ross's commute takes less than an hour	Recommend that Ross drive himself	57.9	5.38 (2.91)	42.1	5.39 (2.40)
		Recommend that Ross take the train	57.8	6.38 (2.00)	42.1	5.15 (2.38)

(*continued on next page*)

Table 10 (*continued*)

Decision (decision-maker underlined)	Goal/outcome	Options	Optimal participants		Nonoptimal participants	
			%	*M*	%	*M*
Jenna is switching cell phone carriers. She wants a carrier with a wide range of consistent coverage, as she lives and travels in a rural area. She is deciding between two different providers	She never has issues with dropped calls	Use Verizon	49.0	4.68 (2.82)	51.0	4.35 (2.63)
		Use AT&T	30.0	5.31 (2.73)	70.0	4.64 (2.37)
Kory wants to go camping in the mountains, and is concerned with going at a time when there are no mosquitoes. He is deciding between two seasons in which to go camping	There are no mosquitoes during his camping trip	Go camping during the spring	53.8	5.03 (3.02)	46.2	5.21 (2.44)
		Go camping during the fall	72.2	5.15 (3.01)	27.8	5.10 (2.96)
Susan and her friend are walking across campus to class when it begins to rain. Susan is holding her textbooks in her hands and does not have an umbrella. Susan does not want her textbooks to be damaged by the rain, and she asks her friend what to do. Susan's friend is deciding between two courses of action to recommend	Susan's textbooks are not damaged	Recommend that Susan start running to her class	75.6	5.60 (2.51)	24.4	5.31 (2.62)
		Recommend that Susan continue walking to her class	41.0	4.86 (2.54)	59.0	3.33 (2.61)
Mean			48.1	5.77	51.9	5.31

Note: Participants were divided for each item into those who indicated that the agent's choice was optimal and those who did not indicate that the agent's choice was optimal (according to the criterion described in the main text). The proportion of participants indicating that the agent's choice was optimal is given in the % column for each item/choice combination. The *M* column gives the mean (*SD* in parentheses) of responsibility ratings among optimal participants and among nonoptimal participants.

after the main phase of the experiment in order to prevent the participants' responsibility judgments from being contaminated by making their efficacy judgment.

Fifty participants were recruited from Amazon Mechanical Turk to participate in Experiment 7A, and 200 participants were recruited to participate in Experiment 7B. Zero participants from Experiment 7A and nine participants from Experiment 7B were excluded for incorrectly answering more than 33% of the 10 check questions.

8.3. Results

Participants in Experiment 7A believed the characters to be more responsible for the outcomes when the characters acted optimally, as defined by the pretest responses to each item. Participants gave a mean responsibility rating of 5.79 ($SD = 1.23$, calculated across items) to agents who picked the optimal choice, but a rating of 5.17 ($SD = 1.12$) to agents who picked the suboptimal choice. This difference is significant both by subject [$t(49) = 3.16$, $p = .003$, $d = 0.45$, $BF_{10} = 9.6$] and by item [$t(9) = 2.77$, $p = .022$, $d = 0.88$, $BF_{10} = 3.2$]. Summaries of the mean responses to each item appear in Table 9. These results accord with the optimality findings of the earlier experiments, but extend them to cases when participants did not see quantitative probabilities for each decision.

Experiment 7B looked at decisions where participants were split in their opinions about which choice was best. For our analysis, we categorized each participant as 'optimal' or 'non-optimal' for each item. A participant was designated as rating a choice as optimal if they indicated in the subsequent choice evaluation task that the agent's actual choice was the more likely of the options to lead to the goal, and the participant's judgment was at least two points beyond the midpoint of the scale (i.e., was lower than 3 if the character in the responsibility judgment task chose option *A*, or was higher than 7 if the character chose option *B*; this criterion was chosen *a priori*). Since there were two versions of each of the ten decision problems (one in which the agent chose option *A* and one in which the agent chose option *B*), the mean responsibility ratings for each of these 20 items was compared among the participants who rated that choice optimal and those who rated it non-optimal. As predicted, participants who rated the choice as optimal gave higher responsibility ratings than those who did not [$M = 5.77$, $SD = 0.64$ vs. $M = 5.31$, $SD = 0.78$; $t(19) = 4.73$, $p < .001$, $d = 1.06$, $BF_{10} = 206.6$].

8.4. Discussion

These results go beyond Experiments 1–6 in two main respects. First, this experiment shows that the assumption of optimality in causal judgment is not confined to relatively artificial settings, but also extends to more naturalistic situations. In our previous experiments, we eliminated any possible influence of prior beliefs by giving the decision options arbitrary labels (e.g., Formula PTY) and manipulating the efficacy of the options by tagging them with explicit probability values. Although consistent results under well-controlled conditions are strong evidence that people use optimality principles in causal judgment, it is less clear how well these results extend to everyday situations in which people *do* have prior beliefs and in which attention is not drawn to the efficacy of each option. Experiment 7 shows, however, that people use their prior beliefs about the efficacy of the options in assigning responsibility to decision-makers, in line with optimality.

Second, Experiment 7 helps to rule out two general methodological concerns about the previous experiments. One possibility is that the manipulations in the previous experiments were relatively transparent, leading to potential demand characteristics. We doubt that the results of the previous experiments can be explained this way, because (a) the content of the vignettes (e.g., fertilizer, shampoo, etc.) changed at the same time as decision efficacy, serving to mask the manipulation, and (b) experimenter demand cannot straightforwardly explain the step function pattern in Experiment 1. Nonetheless, Experiment 7 further rules out concerns about demand because the manipulation was highly opaque. Indeed, in Experiment 7B, rather than manipulating optimality between experimental conditions, we simply treated which decision alternative a given participant favored as an individual difference variable.

The second methodological concern is that some of the experiments might be explainable in terms of contrast effects. For example, in Experiment 2, participants judged two vignettes—one in which P_{ACT} was .5 and P_{ALT} was .3 (thus, the decision was optimal), and one in which P_{ACT} was .5 and P_{ALT} was .7 (thus, the decision was suboptimal). A contrast effect could have occurred if the psychological weight of P_{ACT} differed between conditions. P_{ACT} (.5) may have felt like a larger magnitude when compared to P_{ALT} = .3 than when the same probability was compared to P_{ALT} = .7, leading to higher ratings in the optimal condition. Although this sort of explanation is unlikely to account for the results of Experiment 1 because of the step function, a low-level contrast effect has special difficulty explaining the results of Experiment 7, in which numerical quantities never appeared.

9. General discussion

These experiments examined lay theories of decision-making, asking how the quality of decision-makers' actual and rejected options influences the perceived quality of their decisions, as measured by attributions of responsibility. Two main results emerged consistently across these studies.

First, an assumption of optimality guides attributions of responsibility. In Experiments 1 and 4 (as well as a near-exact replication of Experiment 1), attributions of responsibility depended *qualitatively* on the efficacy of the rejected options—agents were considered more responsible for positive outcomes when the actual choice was better than the rejected option, and less responsible when the actual choice was worse, but the *magnitude* of the difference in efficacy had no further effect. In Experiment 2, agents were considered more responsible when their choice was optimal rather than suboptimal, even if their suboptimal choice nonetheless raised the probability of a positive outcome. Varying both the quality of the actual choice (P_{ACT}) and the rejected option (P_{ALT}), Experiments 3 and 4 found that people base their responsibility judgments on the efficacy of the actual choice (P_{ACT}) in a linear way, and on whether the actual choice is better in this respect than the alternatives ($P_{ACT} > P_{ALT}$) in a qualitative way. That is, when assessing an agent's responsibility, people appear to consider the actual choice they made in a relatively nuanced way, and make an upward adjustment if the choice was optimal and a downward adjustment if the choice was suboptimal (though the *degree* of optimality is not important). Experiment 4 found that this strategy was relatively stable across participants, and Experiment 7 found a similar pattern in more naturalistic decision-making situations.

Second, this optimality assumption takes into account the agent's overall set of goals, not merely the action being evaluated. Experiment 5 looked at decision problems in which agents had to optimize along two dimensions simultaneously, with one option globally optimal (in that it maximized overall utility), another option locally optimal (in that it maximized the probability of a particular goal being accomplished), and a third option suboptimal in both respects. Agents were considered more responsible for achieving a particular goal when they made the best choice relative to their overall goals, not relative to the particular goal under consideration. Experiment 6 found that global *optimality*, rather than linear responses to expected utility, drive these judgments. Thus, people assume others should choose optimally relative to their overall set of goals, not just to their particular goals.

9.1. Optimality and responsibility judgments

We have used responsibility judgments as a way to probe our participants' lay theories of decision-making. Although it is fairly intuitive that optimality would affect other sorts of judgments such as predictions and explanations (see Johnson & Rips, 2014), it is less obvious why optimality should affect judgments of responsibility. Here, we consider two potential reasons for this link: (a) entities who are designated as rational agents are assigned higher responsibility than those who are not, and (b) responsibility judgments track predicted success in the future.

9.1.1. Rationality and responsibility

Philosophers have argued for a *Principle of Rationality*: Understanding others' behaviors and utterances critically depends on viewing these others as agents who take the rational course of action, relative to their beliefs and desires (Davidson, 1967; Dennett, 1987). This principle is used in two distinct

ways. First, we categorize as agents those entities that behave rationally, and subject them to the principles of folk-psychology (e.g., Gergely & Csibra, 2003; Gergely, Nádasdy, Csibra, & Bíró, 1995). Second, once we have designated an entity as an agent, we make inferences about the agent's mental states in accordance with his or her rationality (e.g., Baker et al., 2009). We call the first of these rationality functions *agent-designation* and the second *agent-based inference*.

The agent-designation function can account for why most people assigned higher responsibility to optimally behaving decision-makers. Those who behaved optimally were tacitly categorized as agents, and this led to greater perceived responsibility, because agents are assigned higher responsibility for outcomes than are non-agents (e.g., Hart & Honoré, 1959; Hilton et al., 2010; Lagnado & Channon, 2008). This sort of reasoning is even enshrined in legal systems. For example, many states in the U.S. subscribe to a Model Penal Code, according to which "A person is not responsible for criminal conduct if at the time of such conduct as a result of mental disease or defect he lacks substantial capacity either to appreciate the criminality (wrongfulness) of his conduct or to conform his conduct to the requirement of the law" (see Sinnott-Armstrong & Levy, 2011). Similarly, optimal decision-making in our experiments signaled that the character appreciated the nature of her actions, whereas suboptimal decision-making would signal the opposite.

This account also explains several related results. First, most participants treated the non-optimal decision-makers (those for whom $P_{ACT} = P_{ALT} = .5$) no differently than suboptimal decision-makers (Experiments 1 and 4). Although other people are likely assumed to be agents by default, rationality is used as an agency cue (Gergely & Csibra, 2003) and entities would therefore be categorized as more agentive when their behavior gives positive evidence for rationality. Second, the efficacy of the actual choice (i.e., P_{ACT}) had a somewhat weaker effect on responsibility for optimal rather than non-optimal decisions (Experiment 3) and among participants who followed an optimality strategy (Experiment 4). An optimality strategy, if based on agent-designation, could lead to less reliance on mechanistic factors such as probability because agents' actions are conceptualized in a more end-directed (or teleological) rather than means-directed (or mechanistic) way (Heider, 1958; Lombrozo, 2010). Finally, global optimality rather than local optimality (Experiment 5) or local expected utility (Experiment 6) was used for assigning responsibility. Global optimality (i.e., maximizing overall utility) is the best candidate for assessing the rationality of the *agent* as a whole, whereas local optimality may be a better measure of the rationality of a particular goal-directed action (construed in terms of one goal).

Could the other function of the Principle of Rationality—agent-based inferencing— explain our results in an alternative way? Perhaps people treat all decision-makers as agents to the same extent, but use rationality assumptions to make sense of their actions (Baker et al., 2009; Buchsbaum, Gopnik, Griffiths, & Shafto, 2011). For example, if participants attributed false beliefs to the suboptimal agents so that these agents erroneously believed their actions to be optimal, then this erroneous belief could have led to their discounted responsibility. However, our follow-up to Experiment 1 (see Section 1.3) showed that even knowledgeable actors were seen as less responsible after making suboptimal choices. Further, agent-based inferencing could have just as easily predicted the opposite result—given the agent's knowledge and clearly suboptimal choice (relative to the goals we know about), we could posit an additional goal to make rational sense of her actions, and this additional motivation could lead the agent to be seen as *more* responsible rather than less (see Johnson & Rips, 2014).

Although we favor the agent-designation account as an explanation for the relationship between optimality and responsibility, it is important to note that this link in any case provides clear evidence for the central role of optimality assumptions in lay decision theory, regardless of the mechanism. Either explanation accords a place for the Principle of Rationality in our interpretation of agents' decisions, and these experiments demonstrate that optimality, rather than other plausible conceptions of rationality, best characterizes the form of rationality assumed by most lay decision theorists.

9.1.2. The predictive role of responsibility judgments

Recent computational work suggests that responsibility judgments are a function of the extent to which an agent's action would change our predictions of her future behavior (Gerstenberg et al., 2014). This view can explain why intentional actions yield higher degrees of responsibility compared to accidental actions (Heider, 1958; Lagnado & Channon, 2008), and why outcomes requiring more skill yield higher degrees of responsibility compared to outcomes requiring less skill (Malle &

Knobe, 1997): Intentional and skilled actions carry more information about the agent's intrinsic character traits and causal powers, and these traits and powers are likely to carry over to later situations.

Gerstenberg et al. (2014) extended our optimality model (as reported in Johnson & Rips, 2013) in their Bayesian framework. Specifically, optimal actions provide information about whether the agent is likely to behave "reasonably," which Gerstenberg et al. define as choosing actions probabilistically in proportion to their expected utility. Although their model predicts a linear effect of P_{ALT} rather than the step function we found in the current studies, many of our findings are broadly consistent with this family of models. More generally, the idea that responsibility judgments depend on a difference in expectations about the future appears to be a promising direction, and we look forward to the possibility that additional modeling might illuminate phenomena like those reported in this paper.

9.2. Implications for social cognition

People use different assumptions for understanding the physical and the social worlds (e.g., Carey, 2009; Leslie, 1995; Lombrozo, 2010). The current research helps to clarify the principles that underlie social causation. Here, we sketch three examples of how these results illuminate issues in social cognition.

9.2.1. Moral psychology

Moral considerations affect a variety of apparently non-moral judgments, such as judgments of causality (Hitchcock & Knobe, 2009) and intentionality (Knobe, 2003). For instance, in Knobe's (2003) famous demonstration of the *side-effect effect*, participants read about a CEO who wished to maximize profits without regard for the environment. If the CEO authorized a program that he knew might harm the environment, most people said that he harmed the environment intentionally, while if the CEO authorized a program that he knew might *help* the environment, most people said that he did *not* help the environment intentionally. If moral judgments influence judgments of intentionality and causality, this poses a theoretical challenge, because intentionality and causality are precisely the factors that influence our *moral* judgments about blame and punishment (Cushman, 2008). Thus, an understanding of how attributions of causal responsibility work in non-moral contexts is needed to ground this circle. We take the current work as one step in this direction for grounding social cognition and moral psychology.

9.2.2. Theory of mind

Dennett (1987) suggested that rationality assumptions are the foundational assumption for understanding others. A rationality assumption is useful because it both constrains the space of possible motivations for any particular behavior (in making diagnostic or explanatory inferences), and because it constrains the space of plausible actions that an agent may take in the future (in making predictive inferences). The current results underscore previous empirical evidence for rationality assumptions (Baker et al., 2009; Gergely & Csibra, 2003), and go beyond those previous results in demonstrating a role not only for narrowly construed efficiency, but also for global optimality. In ongoing work, we are examining the possibility that optimality assumptions may be so pervasive in our social cognition that we apply them irresistibly to others, even in situations where an agent's ignorance may not warrant their use (de Freitas & Johnson, submitted for publication; Johnson & Rips, 2014).

9.2.3. Actor/observer asymmetries

We act the vast majority of the time in ways we *believe* to be rational; some have even argued that the very concept of acting irrationally is paradoxical (Davidson, 1982). Yet we seem to have little difficulty identifying irrationality in others. Kahneman (2011) went so far as to describe his book *Thinking, Fast and Slow* not as a guide for readers to act more rationally, but instead for readers to identify *other people's* errors while gossiping (p. 3). If people indeed perceive their own actions as optimal more often than they perceive another's actions as optimal, then our results would suggest that they should also attribute greater causal potency to their own actions in achieving their goals. Consistent with this idea, Pronin and Kugler (2010) found that people believe themselves to have more free will than others: A person's own behavior is seen as produced more by desires and intentions, while

others' behavior is determined more by personality and previous events. Self-perceptions of rational action may in part account for this phenomenon.

9.3. Implications for behavioral decision-making and game theory

Lay decision theory is a type of metacognition, since it concerns *beliefs* about decision-making, rather than how we make decisions. In some situations, however, lay decision theory might impact decision-making itself, and this is especially true when our decisions are contingent on the decisions of others, as in economic games (Camerer, 2003; Von Neumann & Morgenstern, 1944). As we explained in the introduction, achieving a more detailed understanding of lay decision theory may allow for more accurate game-theoretic predictions of group behavior. Here, we give one example of how the current results can illuminate debates in behavioral game theory.

A common approach to modeling strategic thinking in behavioral game theory is to measure performance in laboratory games, building an econometric model of each individual's behavior and interpreting the model's parameter estimates as estimates of participants' beliefs and ability. A particularly elegant example of this approach is the family of *cognitive hierarchy* models (e.g., Camerer et al., 2004; see also Stahl & Wilson, 1995). These models assume that each player belongs to a strategic *type* and that players act on their assumptions about the proportion of other players belonging to that type. The key idea is that players assume that their own strategy is the most sophisticated, and plan their strategies based on the assumed proportions of players of other types.

For example, we briefly described the beauty contest game (Nagel, 1995) in the introduction. In this game, a group of players choose numbers between 0 and 100, and the player who guesses nearest to two-thirds the average wins a monetary prize. The Nash equilibrium for this game is 0, yet average guesses are more typically in the range of 20–40. A cognitive hierarchy model defines a "step 0" type as a person who chooses a number randomly between 0 and 100. A "step 1" type would assume that all other players are "step 0," so the best response is 2/3 of 50, which is 33. A "step 2" type would assume that some proportion of players are "step 0" types (with a mean of 50) and some proportion are "step 1" types (with a mean of 33); for a "step 2" player, the best response depends on the assumed proportion of players in each group, but would be somewhere between 33 (2/3 of 50) and 22 (2/3 of 33). This same logic can be extended to higher-order types. The strategic behavior in a variety of games can be modeled in this way, and the median type is about 1.5—approximately halfway between "step 1" and "step 2"—for a remarkable range of games (Camerer et al., 2004). In other words, people seem to engage on average in about 1.5 steps of strategic thinking.

The indirect approach of inferring participants' beliefs from their strategic behavior has led to great advances in behavioral game theory (Camerer, 2003), but relatively little is known about what implicit theories guide these strategic choices. For example, do "step 1" thinkers actually believe that others will choose randomly, or do they simply fail to consider other players' strategic behavior altogether? The former would indicate use of a contentful (albeit odd) theory of others' decision-making, whereas the latter would indicate a failure of perspective-taking on the part of the player herself. The current results suggest that a perspective-taking error is more likely. If the player considers others' decision-making at all, she is likely to attribute optimal play to them, not random behavior.

More generally, we think that more interdisciplinary cross-talk between game theorists and psychologists would help to illuminate questions in both areas of behavioral science. The study of lay decision theory can profit by studying specific assumptions of econometric models of players' decision-making, with potential payoffs for fields such as social cognition and theory of mind. Likewise, game theorists may be able to build increasingly accurate models of players' behavior by incorporating more realistic assumptions about players' cognition.

9.4. The accuracy of lay decision theory

Classical economics assumes that economic agents act in accordance with their rational self-interest (e.g., Smith, 1982/1776). This assumption has been roundly criticized in light of well-documented biases in human decision-making (e.g., Shafir & LeBoeuf, 2002), and modern economics is divided over the value of rationality assumptions in economic models (e.g., Arrow, 1986; Kahneman, 1994; Simon,

1986). Arrow (1986) notes that these assumptions are most likely to be valid when applied to behavior at an aggregate level, but that individual economic agents will frequently behave in a suboptimal manner (see also Friedman, 1953). If this argument is correct, then the lay decision theorist's application of optimality assumptions at the level of individual behavior may be unreasonable; indeed, similar misunderstandings of emergent phenomena have been documented in domains such as biology and chemistry (e.g., Chi, Roscoe, Slotta, Roy, & Chase, 2012), with group-level emergent phenomena erroneously attributed to the intentions of individual agents.

However, even an individual-level optimality assumption may often be a useful heuristic. People certainly do behave in systematically biased ways under many circumstances (e.g., Kahneman, 1994; Shafir & LeBoeuf, 2002). But in other situations, such as those in our experiments, a systematic bias to behave suboptimally may be unlikely. Under many other circumstances, even if behavior is often suboptimal, the modal decision may still be the optimal one—that is, the uniquely optimal behavior may be more likely than a *particular* suboptimal behavior, even if each individual chooses *some* suboptimal behavior most of the time. Perhaps for this reason, humans seem to have evolved the use of rational behavior as an agency cue (Gergely & Csibra, 2003) and as an animacy cue (Gao & Scholl, 2011).

Ultimately, any assessment of the accuracy of lay decision theory will need to take into account the situation to which it is applied. A potential avenue for future research would be to test these boundaries, seeing for example whether situations that elicit biased behavior in actual decision-making lead lay decision theorists to make judgments that are less in keeping with the classical economist's assumption of optimality. The results of such inquiries may tell us a great deal about where our lay decision theories come from, and what effects they would have in the real world for thinking about and interacting with others.

Acknowledgments

This research was partially supported by funds awarded to the first author by the Yale University Department of Psychology. Experiments 1 and 2 were presented at the 35th Annual Meeting of the Cognitive Science Society. We thank the conference attendees and reviewers for their extremely helpful suggestions. We thank Andy Jin for assistance with stimuli development, Fabrizio Cariani, Winston Chang, Angie Johnston, Frank Keil, Doug Medin, Emily Morson, Axel Mueller, Eyal Sagi, and Laurie Santos for insightful comments, and the city of Nashville, TN for inspirational, sunny November weather.

References

Ahn, W., Proctor, C. C., & Flanagan, E. H. (2009). Mental health clinicians' beliefs about the biological, psychological, and environmental bases of mental disorders. *Cognitive Science, 33*, 147–182.
Alicke, M. D. (1992). Culpable causation. *Journal of Personality and Social Psychology, 63*, 368–378.
Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin, 126*, 556–574.
Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society, 15*, 147–149.
Arrow, K. J. (1986). Rationality of self and others in an economic system. *Journal of Business, 59*, S385–S399.
Audi, R. (1993). *Action, intention, and reason*. Ithaca, NY: Cornell University Press.
Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition, 113*, 329–349.
Belnap, N., Perloff, M., & Xu, M. (2001). *Facing the future: Agents and choices in our indeterminist world*. Oxford, UK: Oxford University Press.
Buchsbaum, D., Gopnik, A., Griffiths, T. L., & Shafto, P. (2011). Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence. *Cognition, 120*, 331–340.
Camerer, C. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton, NJ: Princeton University Press.
Camerer, C. F., & Fehr, E. (2006). When does 'Economic Man' dominate social behavior? *Science, 311*, 47–52.
Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2004). A cognitive hierarchy model of games. *Quarterly Journal of Economics, 119*, 861–898.
Carey, S. (2009). *The origin of concepts*. New York, NY: Oxford University Press.
Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*, 367–405.
Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review, 99*, 365–382.
Chi, M. T. H., Roscoe, R. D., Slotta, J. D., Roy, M., & Chase, C. C. (2012). Misconceived causal explanations for emergent processes. *Cognitive Science, 36*, 1–61.

Colman, A. W. (2003). Cooperation, psychological game theory, and limitations of rationality in social interaction. *Behavioral and Brain Sciences, 26*, 139–198.
Csibra, G., Gergely, G., Bíró, S., Koós, O., & Brockbank, M. (1999). Goal attribution without agency cues: The perception of "pure reason" in infancy. *Cognition, 72*, 237–267.
Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*, 353–380.
Davidson, D. (1967). Truth and meaning. *Synthese, 17*, 304–323.
Davidson, D. (1982). Paradoxes of irrationality. In R. Wollheim & J. Hopkins (Eds.), *Philosophical essays on Freud* (pp. 289–305). Cambridge, UK: Cambridge University Press.
De Freitas, J., & Johnson, S. G. B. (submitted for publication). Behaviorist thinking in judgments of wrongness, punishment, and blame.
Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.
Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6*, 274–290.
Falk, A., & Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior, 54*, 293–315.
Friedman, M. (1953). *Essays on positive economics*. Chicago, IL: University of Chicago Press.
Gao, T., & Scholl, B. J. (2011). Chasing vs. stalking: interrupting the perception of animacy. *Journal of Experimental Psychology: Human Perception and Performance, 37*, 669–684.
Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naïve theory of rational action. *Trends in Cognitive Sciences, 7*, 287–292.
Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition, 56*, 165–193.
Gerstenberg, T., Ullman, T. D., Kleiman-Weiner, M., Lagnado, D. A., & Tenenbaum, J. B. (2014). Wins above replacement: Responsibility attributions as counterfactual replacements. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.
Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103*, 650–669.
Hart, H. L. A., & Honoré, T. (1959). *Causation in the law*. Oxford: Clarendon Press.
Haslam, N., Bastian, B., & Bissett, M. (2004). Essentialist beliefs about personality and their implications. *Personality and Social Psychology Bulletin, 30*, 1661–1673.
Heider, F. (1958). *The psychology of interpersonal relations*. New York, NY: Wiley.
Hilton, D. J., McClure, J., & Sutton, R. M. (2010). Selecting explanations from causal chains: Do statistical principles explain preferences for voluntary causes? *European Journal of Social Psychology, 40*, 383–400.
Hitchcock, C., & Knobe, J. (2009). Cause and norm. *The Journal of Philosophy, 106*, 587–612.
Ho, T.-H., Camerer, C., & Weigelt, K. (1998). Iterated dominance and iterated best response in experimental 'p-beauty contests'. *American Economic Review, 88*, 947–969.
Jeffrey, R. C. (1965). *The logic of decision*. New York, NY: McGraw-Hill.
Johnson, S. G. B., & Rips, L. J. (2014). Predicting behavior from the world: Naïve behaviorism in lay decision theory. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.
Johnson, S. G. B., & Rips, L. J. (2013). Good decisions, good causes: Optimality as a constraint on attribution of causal responsibility. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual conference of the Cognitive Science Society* (pp. 2662–2667). Austin, TX: Cognitive Science Society.
Kahneman, D. (1994). New challenges to the rationality assumption. *Journal of Institutional and Theoretical Economics, 150*, 18–36.
Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.
Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review, 103*, 582–591.
Kelley, H. H. (1967). Attribution theory in social psychology. *Nebraska Symposium on Motivation, 15*, 192–238.
Knobe, J. (2003). Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology, 16*, 309–324.
Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition, 108*, 754–770.
Leslie, A. M. (1995). A theory of agency. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition* (pp. 121–149). New York, NY: Oxford University Press.
Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology, 61*, 303–332.
Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology, 33*, 101–121.
Mandel, D. R., & Lehman, D. R. (1996). Counterfactual thinking and ascriptions of cause and preventability. *Journal of Personality and Social Psychology, 71*, 450–463.
McArthur, L. A. (1972). The how and what of why: Some determinants and consequences of causal attribution. *Journal of Personality and Social Psychology, 22*, 171–193.
Meyer, J. P. (1980). Causal attribution for success and failure: A multivariate investigation of dimensionality, formation, and consequences. *Journal of Personality and Social Psychology, 38*, 704–718.
Moulin, H. (1986). *Game theory for the social sciences, second and* (revised ed.). New York, NY: New York University Press.
Muth, J. F. (1961). Rational expectations and the theory of price movements. *Econometrica, 29*, 315–335.
Nagel, R. (1995). Unraveling in guessing games: An experimental study. *American Economic Review, 85*, 1313–1326.
Pronin, E., & Kugler, M. B. (2010). People believe they have more free will than others. *Proceedings of the National Academy of Sciences of the United States of America, 107*, 22469–22474.
Ritov, I., & Baron, J. (1995). Outcome knowledge, regret, and omission bias. *Organizational Behavior and Human Decision Processes, 64*, 119–127.
Roese, N. J. (1997). Counterfactual thinking. *Psychological Bulletin, 121*, 133–148.
Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225–237.

Shafir, E., & LeBoeuf, R. A. (2002). Rationality. *Annual Review of Psychology, 53*, 491–517.
Shaver, K. G. (1985). *The attribution of blame: Causality, responsibility, and blameworthiness*. New York, NY: Springer-Verlag.
Shtulman, A. (2006). Qualitative differences between naïve and scientific theories of evolution. *Cognitive Psychology, 52*, 170–194.
Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review, 63*, 129–138.
Simon, H. A. (1986). Rationality in psychology and economics. *Journal of Business, 59*, S209–S224.
Sinnott-Armstrong, W., & Levy, K. (2011). Insanity defenses. In J. Deigh & D. Dolinko (Eds.), *Oxford handbook of philosophy of criminal law* (pp. 299–334). Oxford, UK: Oxford University Press.
Smith, A. (1982). *An inquiry into the nature and causes of the wealth of nations*. London, UK: Penguin (Original work published 1776).
Spellman, B. A. (1997). Crediting causality. *Journal of Experimental Psychology: General, 126*, 323–348.
Stahl, D. O., & Wilson, P. W. (1995). On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior, 10*, 218–254.
Von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.
Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct*. New York, NY: Guilford.
Zultan, R., Gerstenberg, T., & Lagnado, D. A. (2012). Finding fault: Causality and counterfactuals in group attributions. *Cognition, 125*, 429–440.