Spring 2020

Why Do We Follow Rules? An Exploration of Normativity and Possibility

Eliza Jane Shaeffer
Dartmouth College

Why do we follow rules?

An exploration of normativity and possibility

Author: Eliza Jane Schaeffer

Advisor: Jonathan Phillips

Dartmouth College, Cognitive Science

Senior Thesis, Winter and Spring 2020

Abstract

The way we interact with the world is governed by a body of rules, many of which are unspoken. What makes these rules so compelling? How do they interact with our decision-making infrastructure? In a series of three studies, this article explores the shared adaptive sampling model's ability to account for normative behaviors, using a time pressure paradigm in which subjects (N = 399) recruited via Amazon Mechanical Turk quickly judged whether actions of various degrees of moral permissibility and social acceptability were possible or impossible in a certain context (Phillips & Cushman, 2017; Phillips, Morris, & Cushman, 2019). When making intuitive judgements of possibility, participants regarded immoral actions as physically impossible, replicating findings from previous research (Phillips & Cushman, 2017). This effect was not found for actions that are simply abnormal, violating social norms rather than moral norms. These studies also provide evidence to suggest that beyond subjects' assessments of moral permissibility or impermissibility, it is the general perceived value (*good* vs. *bad*) of the proposed actions that drives the morality deliberation effect. This is consistent with a two-staged decision-making process, as described in the shared adaptive sampling model, in which the mind samples generally good and generally likely possibilities before using contextual information to further select amongst them. Overall, these studies elucidate the similarities and differences between various normative judgements and general assessments of value and probability, corroborating the shared adaptive sampling model, providing reason to believe that moral norms are distinct from social norms, and offering insight into how various sorts of norms may interact with possibility sampling and decision-making.

Why do we follow rules? An exploration of normativity and possibility

Introduction

Human societies are governed by rules that facilitate the smooth functioning of groups, small and large. Some rules, like "don't kill people," have clear moral valences while others, like "don't cut in line," are less obviously right or wrong; some rules, like "don't speed on the highway, are codified in law while others, like "don't hug strangers," are unspoken understandings. Rules, even arbitrary and implicit rules, are as ubiquitous as they are vital to the daily functioning of organizations and societies. Generally speaking, we eat with silverware, wash our hands, are silent in class, walk on the right side of the sidewalk, and show up to meetings. On a large scale, this level of cooperation is clearly beneficial. However, these rules are meaningless unless a critical mass chooses to follow them, and in any given moment, following a rule may involve incurring a personal cost for the benefit of another person or for the benefit of society more broadly. So, why is it that people follow rules?

This article reviews the literature on cooperative and normative behavior and relates existing theories to a unifying decision-making framework: the shared-adaptive sampling model, first proposed by (Phillips, Morris, & Cushman, 2019). Next, a series of three studies examines the shared adaptive sampling model and the extent to which it can account for intuitive rule-following behavior. Also discussed are the ways in which different categories of norms come into play at different points in the decision-making process.

Review of literature on human cooperation

Existing theories, and the research that backs them, seem to coalesce in general agreement around the proposition that cooperation is advantageous and that some sort of enforcement structure maintains the necessary levels of cooperative behavior (Alvard, 2013;

Baumard et al., 2013; Declerck et al., 2013, Fehr & Fischbacher, 2004; Gürerk et al., 2006; Martin et al., 2019; Rand, 2013; Rand et al., 2014; & Tomasello, 2019). These theories are broadly similar, though they take different positions on the question of how widespread human cooperation came to be.

Reciprocity and social punishment

By some accounts, it is internal incentives, arising over the course of evolution, that make mutual cooperation intrinsically rewarding (Baumard et al., 2013; Declerck et al., 2013). The condition of mutuality — or "reciprocity," as it is described in game theory — is an important caveat (Fehr & Fischbacher, 2004, Rand, 2013). Brain imaging and self-report measures indicate that people find reciprocated cooperation "personally satisfying," even when free-riding would be more personally advantageous (Kalish, 1998, p. 99).

This intrinsic valuing of reciprocity was solidified over the course of evolution as people learned how to avoid social punishments associated with violating reciprocity norms. Baumard et al. (2013) hold that moral systems are in essence an unspoken social contract enforced by "the risk of not being chosen as partners in future transactions" (p. 62). Within a reciprocity framework, ""free-riding" may turn out anything but free" (Baumard 2013, p. 66), while cooperation is likely to be reciprocated in the future (Declerck et al. 2013; Baumard et al. 2013).

Corrective punishments — whether they take the form of refusal to associate in future interactions or more explicit sanctions — function as a sort of "negative distribution" to restore fairness (Baumard et al. 2013). People punish even when they are a third party who receives no direct benefit from punishing and even when punishing is costly at a personal level (Declerck et al. 2013), though the likelihood of punishment decreases as associated costs increase (Baumard et al. 2013). These "altruistic" acts of punishment are associated with activity in brain areas

typically involved with the reward system, leading some to conclude that punishment is intrinsically rewarding (Declerck et al. 2013; Baumard et al. 2013). In the context of a web of interpersonal relations, direct reciprocity may be enough to account for the evolutionary rise of human cooperation at the macro level (Rand, 2013; Baumard et al., 2013).

Social expectations and group membership

Others maintain that an intrinsic valuing of reciprocity is not sufficient to maintain cooperation without additional enforcement from the external positive and negative incentives associated with group membership (Alvard, 2013; Declerck et al., 2013; & Tomasello 2019). People may cooperate because of the benefits — including stability, access to resources, belongingness, and social status — that naturally follow from group membership (Alvard, 2013; Declerck et al., 2013; Tomasello, 2019). They may also be motivated by the threat of punishment or sanctions, choosing to follow rules because they fear that fellow in-group members will enforce those rules in the case of a violation (Baumard et al., 2013; Fehr & Fichbacher, 2004; Rand, 2013; Tomasello, 2019).

Cooperation or defection in the context of one interaction may affect actors' reputations in a way that carries implications for future interactions (Baumard et al., 2013; Fehr & Fichbacher, 2004; Rand, 2013; Tomasello, 2019). People incorporate indirectly-observed interactions in their evaluations of others. For example, if X person slights Y person, and you observe this, you will be less willing to engage with X person in the future. This tendency is present early in development, with infants as young as 5 months old demonstrating preference for actors that display pro-social behaviors (Hamlin et al., 2011). The flip side of this is indirect reciprocity, which describes instances in which people cooperate with the expectation that others will hear of, and in turn reciprocate, their generosity (Rand, 2013).

Even if an actor is not personally involved in a dispute, their response, or lack thereof, can have a lasting effect on their reputation. Third-party punishment of defection "signal[s] prosociality and trustworthiness [and] a willingness to retaliate when harmed directly" (Martin et al., 2019, p. 11), further reinforcing expectations of cooperation (Fischbacher, 2004). To return to our earlier example, if X person slights Y person, and Z person doesn't condemn X's actions, and you observe this, you will be less likely to engage with Z in the future. In other words, it is expected that group members will practice, require, and enforce cooperation, even for interactions in which they are not directly involved.

Pro-social norms and conformity

When viewed in the context of increasingly higher-order cooperative behaviors, these 'expectations' begin to resemble social norms, which govern the rules under which agents affiliated with a particular group operate. Some argue that it is conformity to norms, rather than reciprocity, which explains the widespread nature of cooperation (Gürerk et al., 2006; Fehr & Fischbacher, 2004). Norms may be voluntarily obeyed if there is an alignment in incentives — Tomasello (2019) describes a joint "we" agent that is essentially an embodiment of value alignment.

A "we" implies a sense of belongingness. In following a rule, we do more than uphold social order — we choose a team. In experimental settings, following a single arbitrary rule has been enough to guarantee belongingness in a social group (Dunham, 2018). As evolutionary advances led to societies so large that one couldn't possibly know everyone, conformity to this joint "we" agent would have been an important means of signaling trustworthiness (Tomasello, 2019). In this way, normative behavior constructs and enforces social identities: conforming to a norm is a means of actively identifying with a group (Brennan et al., 2013).

A clarifying example: The Prisoner's Dilemma. To clarify the differences between these different, though not entirely incompatible, theories, consider the Prisoner's Dilemma (see Figure 1). From each players' perspective, the dominant strategy is defection; however, from a societal (collective) perspective, the dominant strategy is cooperation, as it results in a greater joint payoff.

Figure 1

Prisoner's Dilemma

		Player 2	
		Defect	Cooperate
Player 1	Defect	(4, 4)	(12, 0)
	Cooperate	(0, 12)	(7, 7)

What drives players to cooperate? The coordinating device at play may be an intrinsic valuing of mutual cooperation, providing additional individual incentive to cooperate even when the dominant strategy is defection and allowing partners to assume that their counterparts will also cooperate. Cooperation may also be attributable to external factors not captured by the individual payoff matrix — such as social affiliation, punishment, or reputational damage — and it may be these extrinsic incentives that push players to cooperate. Or, perhaps, in cases of joint agenthood, players respond to collective, not individual, payoffs. Finally, it could be the case that the specific actions under consideration (cooperate versus defect) are not particularly relevant, and it is the conformity to some external norm (in this case cooperation) that both players value.

If we take a step back and examine these theories' commonalities, we see that they each involve assigning value to actions. Behaviors are evaluated as valuable (they have historically positive social implications, they activate the reward pathway, etc.) or harmful (they have historically negative social implications, they cause personal discomfort, etc.). Some theories hold that this evaluation happens early in the decision-making process, at an intuitive level, and

other theories hold that this evaluation happens only as potential actions are explicitly considered. This article considers a broader, simpler model of rule-following behavior that describes how various means of assigning value to actions work in unison to inform decision-making.

Shared adaptive sampling model

First, consider that following a rule involves a deliberate choice. Making a deliberate choice requires two mental states: (1) the conceptualizing and entertaining of feasible possibilities and (2) selection amongst those possibilities (Byrne, 2007). Thus, in choosing to follow a rule, humans implicitly and explicitly consider and evaluate possible actions. Phillips, Morris, & Cushman (2019) outline a model that describes the means by which people reason across possibilities.

Their shared adaptive sampling model holds that the mind samples the space of possible actions within a task-specific partition, prioritizing possibilities with a high general cached value. The cached value is informed by past experience and learning, and it reflects historical value and probability of occurrence. More prevalent possibilities, or possibilities that the actor has encountered with greater frequency, would have a greater cached value, as would possibilities that the actor considers to be high-utility. In this first stage, possibilities are represented in a course-grain fashion. In the next stage, the brain selects amongst the adaptively sampled possibilities, ranking them according to context-specific value. A possibility may be generally valuable and common, but less advantageous given current circumstances. For example, someone may generally like eating soup, but not when it is hot outside. This latter stage involves finer-grain, context-specific representations that are accurate yet computationally expensive.

The shared adaptive sampling model is well-positioned to account for rule-following behavior. Rule-following is normative: rules are upheld by a critical mass of people that follow them, and any given actor's past experience with rules is likely largely characterized by compliance. Additionally — as was described in the earlier discussion of the positive and negative extrinsic and intrinsic incentive structures surrounding normative cooperation — rule-following is valuable. Rules maintain social cohesion. Following a rule can be an affiliative gesture, securing stability and belongingness, and it can be rewarding in of itself. Be it through direct experience, second-hand exposure, or internal conviction, we know that violating a rule is of low value, while following a rule is of high value.

Given that historical value and probability of occurrence receive priority in the formation of consideration sets, and assuming that rule violations are typically stored as low-probability and low-value options, it would follow that we consider rule-following possibilities to a greater degree than rule-violating possibilities. The priority received by representation of high-value, high-probability possible actions would be visible in actions taken, as we ultimately choose from the pool of possibilities represented in our consideration set (Hauser, 2014). In other words, we may follow a given rule simply because the high-value and high-probability nature of rule-following behaviors makes it easier to conceptualize them as possibilities.

Example: A dining hall

Say, for example, you are filling your plate at a buffet-style dining hall. You could start a food fight. You could make a salad at the salad bar, using any of the hundreds of permutations of ingredient combinations. You could wait in line for a slice of pizza. You could serve yourself chicken using tongs. You could serve yourself chicken using your hands. You could serve yourself chicken using your mouth. You could take a bite of your meal. You could rub your meal

in a stranger's face. You could eat at a table. You could eat seated on the floor. You could do yoga on a table. The array of possibly possible actions before you is mind-bogglingly expansive, though some of them are clearly more valuable and more common than others. The shared adaptive sampling model would hold that as you stand before the dining hall offerings, the possible courses of action that come to mind (that is, the adaptively-sampled initial consideration set) would be those that are high-value and high-probability.

Establishing the value of actions is an iterative process. Some may be intrinsically valuable — perhaps conformity, perhaps cooperation — while others may be learned. With experience, trial, and error, we update the relevant general cached value. For example, seeing a restaurant full of people eating at tables, rather than on the floor, reinforces this action's characterization as high-probability. Lack of success in attempts to serve yourself chicken with your mouth reinforces this action's characterization as low-value. In general, actions that comply with established rules and social norms are more likely to be high-value and high-probability; for example: waiting in line, serving yourself with utensils, eating at a table, etc.

This updating process, and the guiding effect it has on future behavior, is reminiscent of the Social Heuristic Hypothesis, which attributes normative cooperation to the internalization and generalization of learned social norms (Rand et al., 2014). It maintains that we intuitively apply historically-valuable norms like cooperation to novel situations, even when cooperation is not clearly advantageous, as is the case in artificial lab settings.

Advantages of the shared adaptive sampling model

Viewing rule-following behavior as an application of the shared adaptive sampling model offers several advantages. First, it is parsimonious. Other theories describe the independent evolution of cooperation, or higher-order punishment, or conformity. Recognizing rule-following

as one application of a more foundational cognitive mechanism allows us to acknowledge the contributions of these theories without endorsing one over another. Further, this structure has applications beyond just rule-following or decision-making; it describes the underpinnings of moral reasoning, rationality, judgements of force, socialization, and modal thought more broadly (Phillips & Cushman, 2017, Shtulman & Phillips, 2018; Rand et al., 2014).

Second, a shared adaptive sampling account of rule-following behavior allows for cultural variation in cooperative tendencies. If cooperation were intrinsically rewarding, we would expect to see universally high levels of cooperation; however, while it is widespread, cooperation is far from uniform across cultures (Ensminger 2014; Henrich 2000). It would be difficult to account for cross-cultural variation while arguing that humans consider cooperation to be intrinsically valuable; however, variation across cultures can readily be accounted for by the shared adaptive sampling model.

For example, a cross-cultural analysis of contributions in an ultimatum game — which included subject groups from western cities, the Peruvian Amazon, Japan, and Indonesia — found considerable variation in offerings. The sample group from a Peruvian tribe, the Machiguenga, had a mean and modal offering of 25 percent and 15 percent of the total pay-off, respectively; the Los Angeles sample group's mean and modal offerings were 48 percent and 50 percent, respectively (Henrich, 2000). But the Machiguenga live in small societies of 300 people and are largely self-sustaining at the family level. For them, it is not particularly common to cooperate with or act generously towards non-family members, and at the time of the study, what little experience they did have with non-kin cooperation was likely low-utility, due to the then-recent rise in unwelcome, foreigner-driven development and market integration (Henrich 2000). Thus, it is not likely that the possibility of issuing generous contributions would have appeared in

the Machiguengas' original consideration sets, and it is even less likely that the option issuing a generous contribution would have been selected after more deliberate consideration.

Criticisms of the shared adaptive sampling model

In spite of the advantages offered by the shared-adaptive sampling account of rule-following behavior, critics may find its implication — that people follow rules just because entertaining the possibility of violating them is computationally expensive — to be counterintuitive, as it is inconsistent with people's experience of reality. People tend to identify with a "conscious, reasoning self that has beliefs, makes choices, and decides what to think about and what to do" — this is our more intentioned, explicit system of thought, which Kahneman (2011) labels "System 2" (p. 21). However, it is the more automatic, intuitive system of thought ("System 1") that interprets our reality so that System 2 can operate with maximum efficiency. The two-step process modeled by the shared-adaptive sampling mechanism dovetails nicely with Kahneman's System 1 and System 2. Formation of a consideration set through adaptive sampling can be thought of as an operation of System 1, and explicit evaluation of the relative context-specific merits of options generated can be thought of as an operation of System 2. Because the intuitive System 1 systematically represents a normative and cohesive reality, it follows that rule-following behavior would be more easily identified as a possible course of action (Kahneman, 2011, p. 413, 424-6).

Importantly, the shared adaptive sampling model does not imply that weighting of value and probability happens only at the intuitive level, nor does it require that the original consideration set places permanent boundaries around the space of possibilities available for more explicit evaluation. Situations can be considered and evaluated and reconsidered and reevaluated in an iterative process. The shared adaptive sampling model just provides a means of

structuring that process, in a way that lends itself to accounting for the pervasiveness of rule-following behaviors.

Shared adaptive sampling and normative behavior

If we are to view rule-following behavior through the lens of the shared adaptive sampling model, the question becomes: At which stage of decision-making does the rule or norm become relevant? There are two basic positions that one may take:

1. Normative actions are more readily identified as possible and are selected for in the first stage of sampling, which occurs below conscious awareness.
2. Norms do not constrain intuitive possibility sampling and only become relevant in the second stage of sampling, when people determine which possibility or action is most appropriate or valuable in a given situation.
3. Moral norms constrain intuitive possibility sampling; other sorts of norms only become relevant in the second stage of sampling. This is a soft version of the first position.

Research by Phillips & Cushman (2017) offers us reason to reject the second position: They find that moral norms constrain intuitive assessments of possibility. Additionally, the literature on cooperation, reciprocity, and pro-sociality give us reason to expect that the same effect would be found for social norms. For these reasons, this article takes the first position.

Thesis Statement

If it is the case that actors conform to seemingly arbitrary rules in part because, on an intuitive level, they do not reflexively entertain the possibility of flouting those rules, we would expect that under time pressure, participants will judge all normative rule violations to be impossible to a higher extent than they would if given time to reflect. This finding, if it is indeed born out in the data, would offer evidence to further support a two-stage model of decision-

making in which social norm-compliant behaviors are prioritized in intuitive considerations of possible actions. It would also serve to further integrate current theories of justice, punishment, and cooperation, which focus on the positive and negative incentive structures surrounding rule-following behavior.

Study 1

The purpose of this study was to establish that participants' evaluations of each action item tracked with the a priori classification of those actions as abnormal, immoral, impossible, or ordinary. Below are the results from the final stimulus set, which first underwent two rounds of norming.

Participants

Fifty-one adult participants (M_{age} = 38, SD_{age} = 13.02; 17 females; 23 post-secondary degrees) were recruited through Amazon Mechanical Turk (www.mturk.com).

Methods

In an online survey, participants read a series of eight scenarios, each consisting of a one- to two-sentence description of a commonplace situation. For example, one scenario read: "As you enter a museum, the security guard informs you that you must leave your backpack at the coat check." Associated with each scenario was a list of 16 actions, coded as ordinary ("admire your favorite painting"), impossible ("transform into a very large seal"), abnormal ("search your backpack for a snack"), or immoral ("rip a painting into small pieces").

Ordinary actions were those that would conventionally be considered typical in the scenario at hand. Impossible actions were those that would violate the physical laws of space and time. Abnormal actions were those that would violate a context-specific amoral rule or social norm. Immoral actions were those that would be considered overtly wrong: for example, stealing, killing, or destructive behaviors.

In total, there were eight scenarios and 128 actions, with 16 actions associated with each scenario. In each scenario-specific grouping of actions, four were ordinary, four were impossible, four were abnormal, and four were immoral. Subjects were asked to rate each of the 128 actions

on a scale of 1 to 5, in terms of social acceptability or moral permissibility. For example, a subject might be asked, “Is it socially acceptable for you to … rip a painting into small pieces?” or “Is it morally permissible for you to … rip a painting into small pieces?” Subjects were randomly assigned to the social acceptability condition or the moral permissibility condition. The study was conducted using Testable.

Trials for which the response time was too fast, trials that timed out, and trials with *not applicable* responses were excluded from the analyses. The final dataset consisted of 4716 trials, with 2634 social acceptability ratings and 2082 moral permissibility ratings.

Results

A two-sample t-test confirmed the validity of the a priori classifications of actions as ordinary, immoral, abnormal, or impossible. The abnormal events were regarded as socially unacceptable (M = 4.41, SD = 1.023) more frequently than the ordinary events (M = 1.34, SD = 0.89), $t(1488.8) = 62.98$, $p < .001$. Abnormal events were also regarded as immoral (M = 3.77, SD = 1.30), more frequently than the ordinary events (M = 1.20, SD = 0.69), $t(853.15) = 42.51$, $p < .001$).

Immoral events were regarded as socially unacceptable (M = 4.82, SD = 0.64) more frequently than both the abnormal events (M = 4.41, SD = 1.02), $t(1252.6) = 9.241$, $p < .001$, and the ordinary events (M = 1.34, SD = 0.89), $t(1448.6) = 89.52$, $p < .001$. Immoral events were also regarded as immoral (M = 4.88, SD = 0.49) more frequently than the abnormal events (M = 3.77, SD = 1.30), $t(721.8) = 19.41$, $p < .001$, and the ordinary events (M = 1.20, SD = 0.69), $t(1186) = 111.12$, $p < .001$.

Finally, the immoral events were more immoral (M = 4.88, SD = 0.49) than they were socially unacceptable (M = 4.82, SD = 0.64), ($t(1435) = 2.10$, $p = .04$), and the abnormal events

were more socially unacceptable (M = 4.41, SD = 1.02) than they were immoral (M = 3.77, SD = 1.30), $t(1070.8) = 9.85$, $p < .001$. These general relationships are visible in Figure 2, which presents the acceptability ratings and morality ratings for each Event Type.

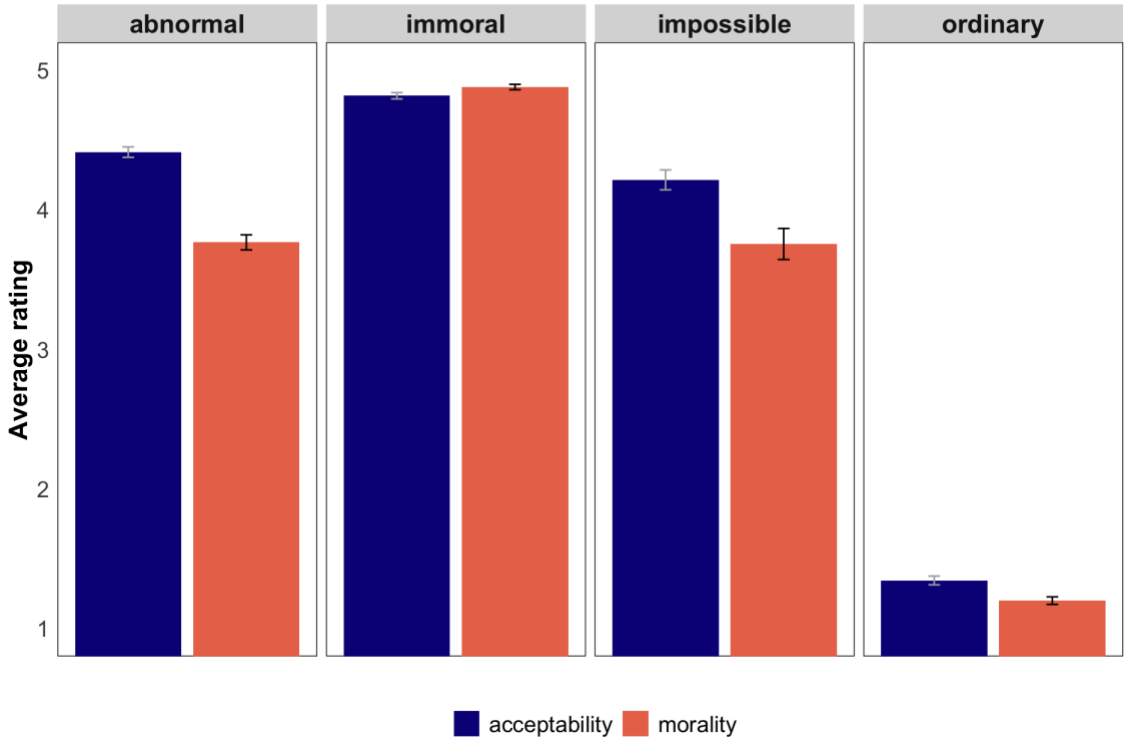

Figure 2. Average rating across Event Types. A rating of 5 has a strong negative valence (*not at all morally permissible, not at all socially acceptable*), while a rating of 1 has a strong positive valence (*very morally permissible, very socially acceptable*). Ratings of 3 are *in-between*. The bars reflect ratings' standard error.

Study 2

The purpose of this study was to measure speeded and reflective judgements of possibility for each action (Phillips & Cushman, 2017). Participants faced time constraints, which was not the case in Study 1. The time pressure paradigm allows us to identify which actions participants intuitively view as impossible. In the context of the shared adaptive sampling

model, these actions are the possibilities that participants would reject early in the possibly sampling process, perhaps without conscious awareness.

Participants

There were 186 participants (M_{age} = 36, SD_{age} = 10.02; 70 females; 73 post-secondary degrees) recruited through Amazon Mechanical Turk (www.mturk.com).

Methods

Subjects made judgements about the same 128 actions, labeling them either *possible* or *impossible* by pressing the corresponding key on their keyboard. Each subject judged four scenarios and their corresponding actions under time pressure (1.5 seconds) and judged the remaining four scenarios and their corresponding actions with no time pressure. In the slow trials, subjects were encouraged to reflect on their answers. The combination of scenarios represented in each fast and slow group were randomized using a Latin-square design. Subjects were randomly assigned to one of the eight conditions created by this randomization scheme.

Trials for which the response time was too fast, trials that timed out, and trials with *not applicable* responses were excluded. The final dataset consisted of 15,991 trials. There were 4157 abnormal trails, 4110 immoral trials, 3902 impossible trials, and 3822 ordinary trials, divided across 10,268 fast trials and 5723 slow trials.

Results

A generalized linear mixed model was used to understand how subjects' possibility assessments were affected by Event Type (immoral, impossible, ordinary, or abnormal) and Deliberation. The significance of an effect was computed by using an ANOVA to compare a model that included the factor of interest to a model that did not include that factor but was

otherwise identical. These models included random intercepts for both participants and scenarios.

There was no significant main effect of Event Type ($X_2(3) = 6225.7$, $p < .001$) or Deliberation ($X_2(1) = 0.024$, $p = .876$), though there was a significant interaction between Deliberation and Event Type ($X_2(3) = 204.58$, $p < .001$). Figure 3 presents the spread of impossibility ratings for each Event Type under the speeded and reflective conditions.

To further explore the relationship between Event Type and Deliberation, a series of generalized linear mixed models were used to predict the effect of Deliberation on possibility judgements for each Event Type in isolation. An ANOVA test revealed no significant main effect of Deliberation for abnormal events ($z = 0.495$, $p = .621$), though it did reveal a significant main effect of Deliberation for immoral events ($z = -5.621$, $p < .001$), impossible events ($z = 10.82$, $p < .001$), and ordinary events ($z = -6.591$, $p < .001$). Because judgements for impossible events and ordinary events fall so close to the ends of the possibility spectrum, it is likely that their respective interactions with Deliberation are attributable to a regression to the mean.

Finally, it is noteworthy that, even given time to reflect, the average impossibility rating for abnormal events (M = .30, SD = .46) and immoral events (M = .40, SD = .49) was far above the impossibility ratings for ordinary events (M = .04, SD = .19).

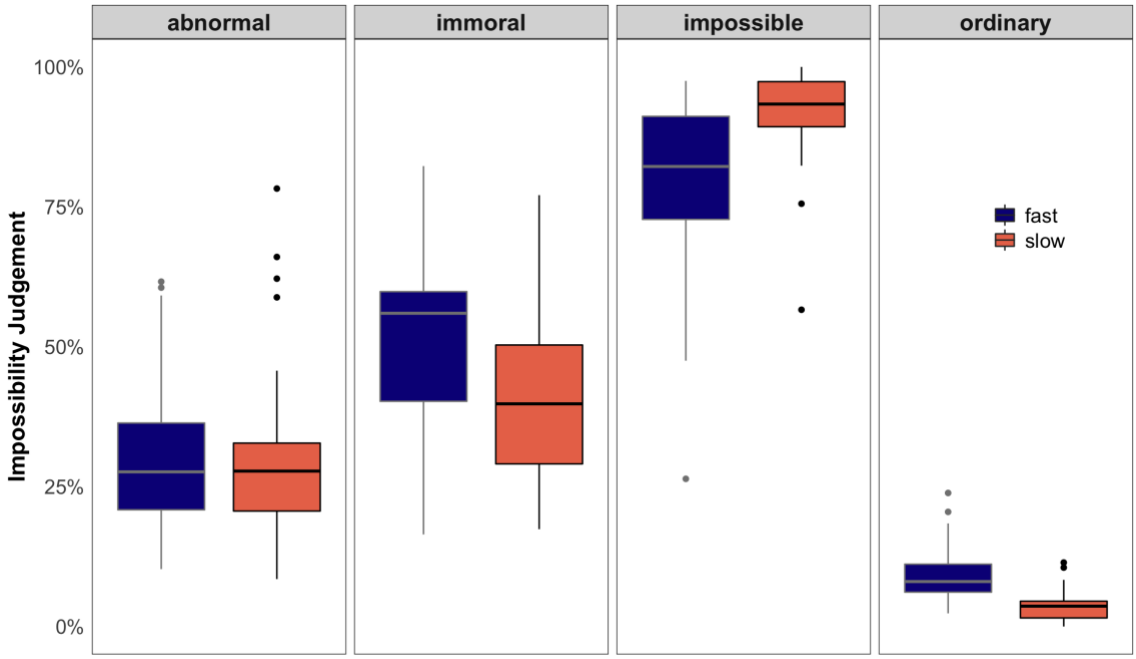

Figure 3. Percent of events judged impossible. Figure 3 displays the spread of impossibility ratings for each a priori Event Type, across fast and slow conditions. Fast judgements were made under time pressure; slow judgements were not. The analysis was performed at the level of each action (N = 128). The boxes reflect the upper and lower quartiles of trials' average impossibility ratings for each Event Type. The vertical lines represent the spread of the data beyond the middle two quartiles, and the dots represent outliers.

For a more detailed view, see Figure 4, which regresses morality judgements against the difference between fast and slow judgements of possibility for each Event Type. Because impossible events were excluded from the dataset used to make this map, any y-axis value greater than 0 reflects an increase in mistaken judgements of impossibility under time pressure, whereas y-axis values less than 0 reflect an increase in correct judgements of possibility under time pressure. Of the three categories, immoral events clearly have the steepest slope.

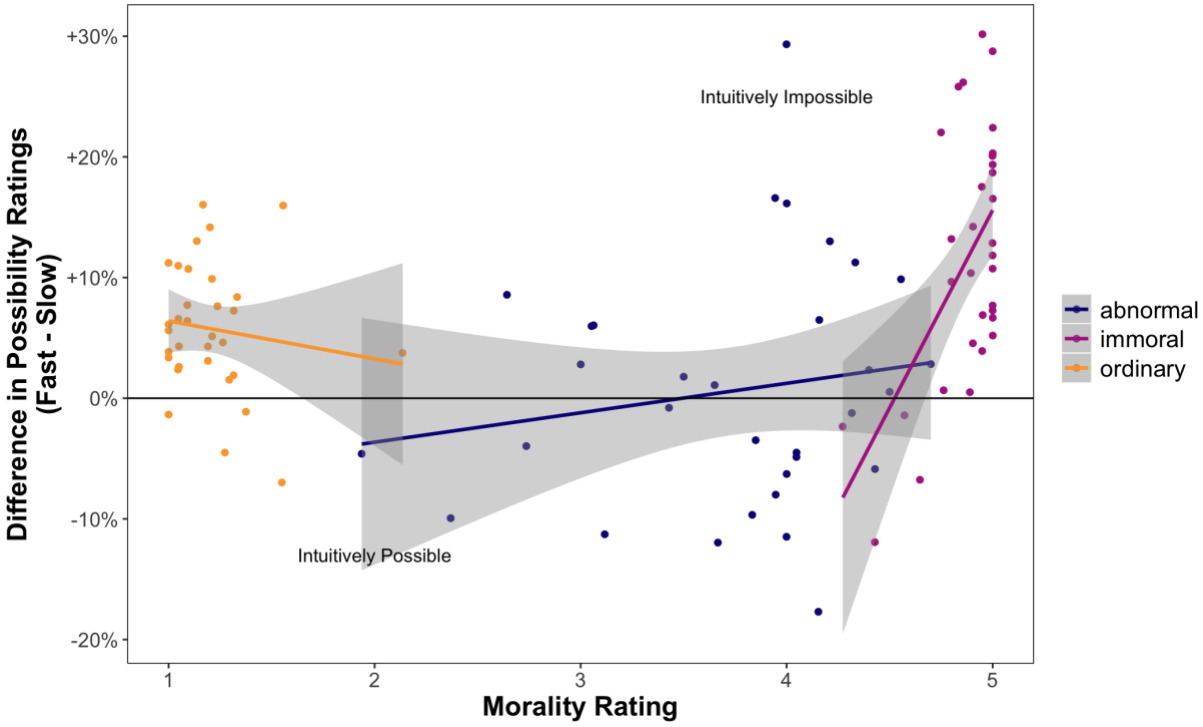

Figure 4. Deliberation effect by Event Type. The y-axis of Figure 4 gives the difference between fast and slow judgements of possibility. A point falling above y = 0 represents an action that was judged to be impossible to a greater extent under time pressure than with deliberation. Morality ratings are on the x-axis. A rating of 5 refers to *not at all morally permissible*, while a rating of 1 refers to *very morally permissible.*

This sharp uptick in impossibility judgements is also visible in Figure 5, which does not differentiate between a priori Event Types. Instead, Figure 5 features just two linear regressions: one for events that had a morality judgement less than 4.5, and one for events that had a morality judgement of greater than 4.5. This is a somewhat arbitrary number, lying halfway in between 4 (*somewhat not morally permissible*) and 5 (*not at all morally permissible*). It was chosen because it reflects a value that might be analogous to "not morally permissible" and because it cleanly partitioned the data into one linear regression that is largely flat and one linear regression that is sharply sloped. Thus, it is clear that at the higher end of the immorality spectrum, the gap

between fast and slow assessments of actions' possibility widens even when the priori classifications of those actions are not taken into account.

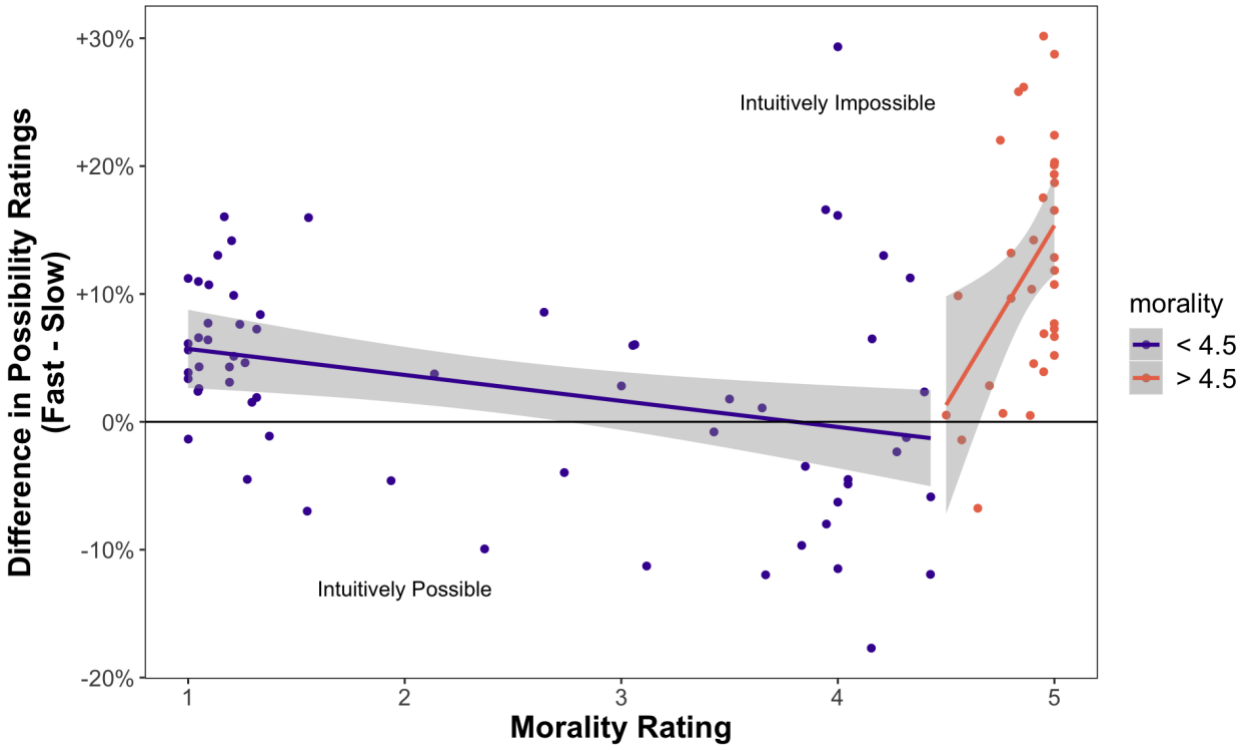

Figure 5. Deliberation Effect Across Morality Ratings. The y-axis of Figure 5 gives the difference between fast and slow judgements of possibility. A point falling above y = 0 represents an action that was judged to be impossible to a greater extent under time pressure than with deliberation. A rating of 5 refers to *not at all morally permissible*, while a rating of 1 refers to *very morally permissible.* The figure presents two linear regressions: one for events that had a morality judgement less than 4.5, and one for events that had a morality judgement of greater than 4.5.

Study 3

The purpose of this study was to measure context-independent judgements of the value and probability of each action. These ratings reflect the cached value of each actions, as described by the shared adaptive sampling model.

Participants

There were 62 participants (M_{age} = 33, SD_{age} = 12.42; 19 females; 13 post-secondary degrees) recruited through Amazon Mechanical Turk (www.mturk.com).

Methods

Subjects rated each of the 128 actions on a scale of 1 to 5, in terms of value or probability. For example, a subject might be asked, "Is it good or bad for you to … rip a painting into small pieces?" or "How likely is it that you … rip a painting into small pieces?" For both conditions, actions were presented *without* context. This is at difference with the prior two studies, in which subjects read a blurb about the context in which they would hypothetically perform the proposed actions.

Subjects were randomly assigned to the value condition or the probability condition. Trials for which the response time was too fast, trials that timed out, and trials with *not applicable* responses were all excluded. Trial 97 was also excluded from all analyses, due to experimenter error. The final dataset consisted of 6704 trials, with 3444 for the value condition and 3260 for the probability condition.

Results

Colinearity of Judgements.

It proved difficult to pull apart assessments of general value (study 3), general probability (study 3), situation-specific acceptability (study 1), and situation-specific morality (study 1), as

the four are largely co-linear. Each bar in Figure 6 represents an action (N = 127), presented in ascending order according to the average of all four judgements for each action. There is variation across judgements, but on the whole, they trend in the same direction at a similar rate. Actions that were less valuable were deemed less moral, less acceptable, and less likely.

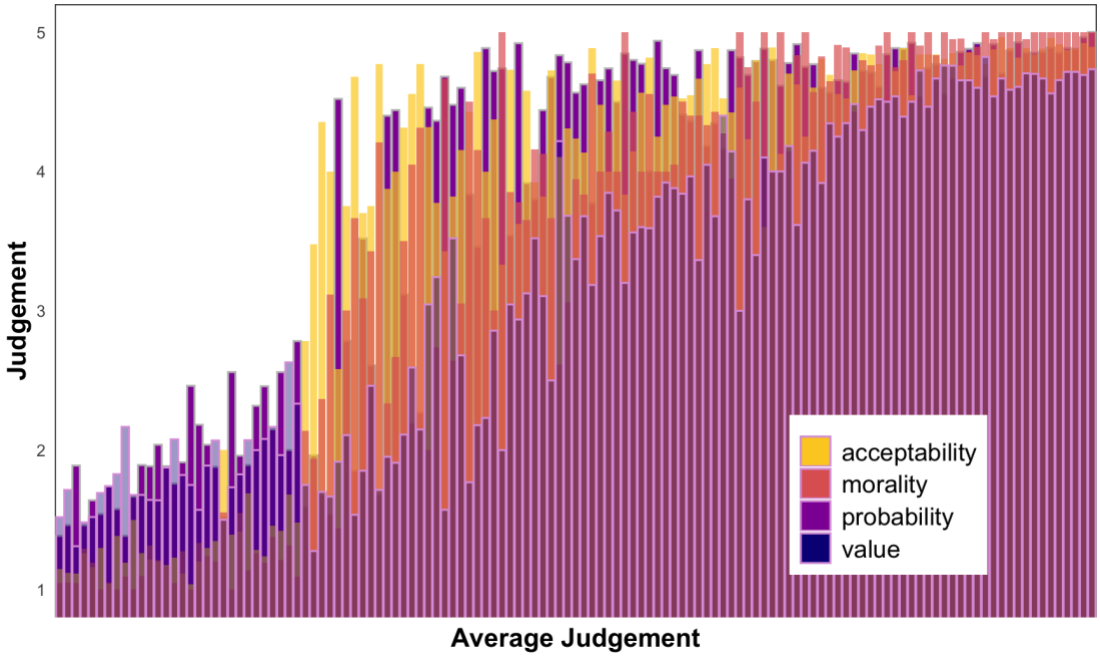

Figure 6. Colinearity of judgements. Each bar represents an action (N = 127), presented in ascending order according to the average of all four judgements for each action. A rating of 5 carries a negative valence (*not at all morally permissible, not at all socially acceptable*), while a rating of 1 carries a positive valence (*very morally permissible, very socially acceptable*).

Probability and value by Event Type

Figure 7 displays the value and probability ratings of each action, coded by Event Type. In order to aid visibility, the scatterplot is divided into 9 sections, reflecting various levels of value and likelihood. The ordinary events are clustered around the "good and likely" section, whereas the immoral events are entirely "bad and unlikely." The impossible events are almost all "unlikely," though they cover the range of values pretty evenly. Abnormal events are the only

ones that are scattered throughout the plot. The sections surrounding “bad and likely” are noticeable empty. This figure clarifies the relationship between value, probability, acceptability, and morality as they relate to the a priori Event Types.

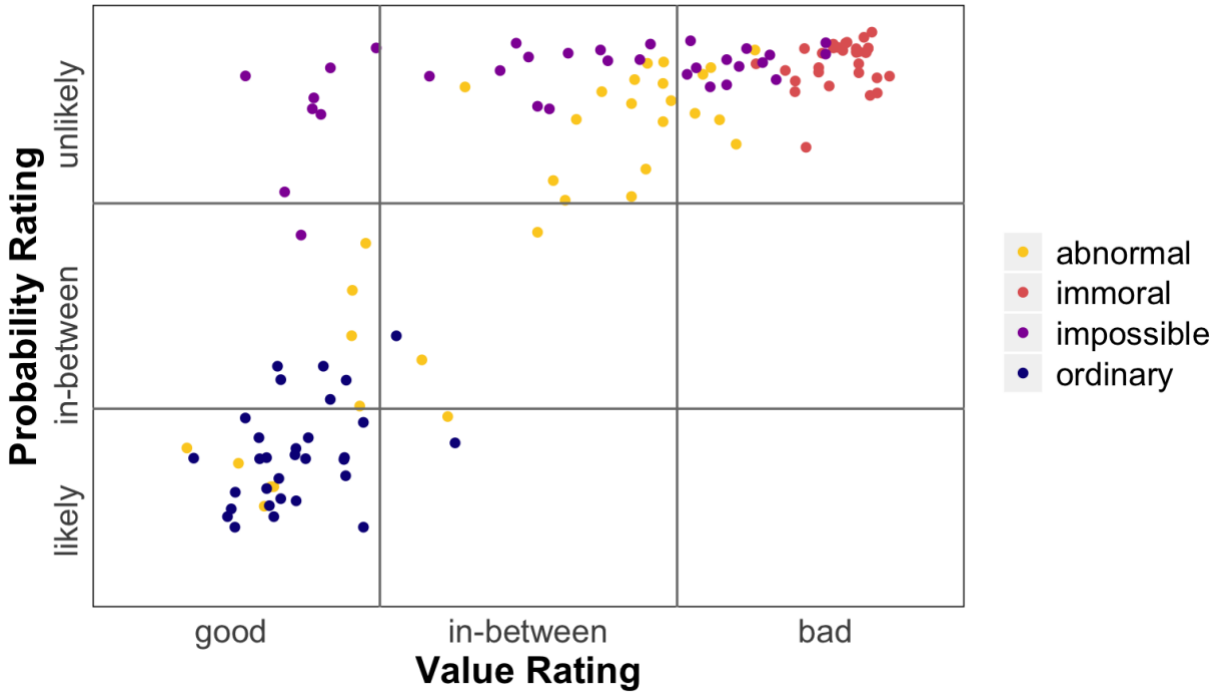

Figure 7. Event value and probability. Figure 7 displays the value and probability of each action, coded by Event Type. In order to aid visibility, the scatterplot is divided into 9 sections reflection various levels of value and likelihood.

Deliberation and value.

The deliberation data, the two situation-specific judgements, and the two general judgements were combined for a final round of modeling. In these analyses, impossible events were excluded, narrowing the scope of the analysis to focus on the mistake participants made in judging unacceptable or impermissible, yet possible, events to be impossible.

Deliberation effect for possible events. A linear mixed effect model was used to predict the difference between fast and slow responses from the variables value, probability,

acceptability, and morality. The data was analyzed at the level of the 95 possible actions, and the model included a random intercept for the 8 scenarios. The significance of an effect was computed by using an ANOVA to compare a model that included the factor of interest to a model that did not include that factor but was otherwise identical.

There was a significant main effect of value ($X^2(1) = 6.58$; p = .01), such that actions with a higher value rating (more negative valence) were judged to be impossible to a greater extent under time pressure. There was also a main effect of acceptability ($X^2(1) = 10.91$; $p < .001$), but in the opposite direction: actions with a higher acceptability rating (more negative valence) were judged to be impossible to a lesser extent under time pressure. There was no main effect for probability ($X^2(1) = 0.3226$; p=.5701) or morality ($X^2(1) = 3.1604$; $p = .07545$). The effect of value is in the positive direction, whereas the effect of acceptability is in the negative direction; thus, under this model, value drives the positive Deliberation interaction while acceptability judgements act as a moderating force in order to accommodate the difference in fast and slow possibility judgements between acceptable/good events (ordinary) and somewhat unacceptable/bad events (abnormal).

Ordinary events ($M_{Acceptability} = 1.34$, $SD_{Acceptability} = 0.89$) were judged to be impossible to a greater extent under time pressure, $t(47.62) = 5.29$, $p < .001$, while abnormal events ($M_{Acceptability} = 4.41$, $SD_{Acceptability} = 1.02$) were not, $t(61.30) = 0.16$, $p = .88$. The deliberation effect in ordinary events is attributable to a regression to the mean; the lack of a deliberation effect in abnormal events is attributable to their middling value ratings (see Figure 8). Acceptability ratings function as a moderating force due to the disconnect between perceptions of acceptability and value for abnormal events (see Figures 8).

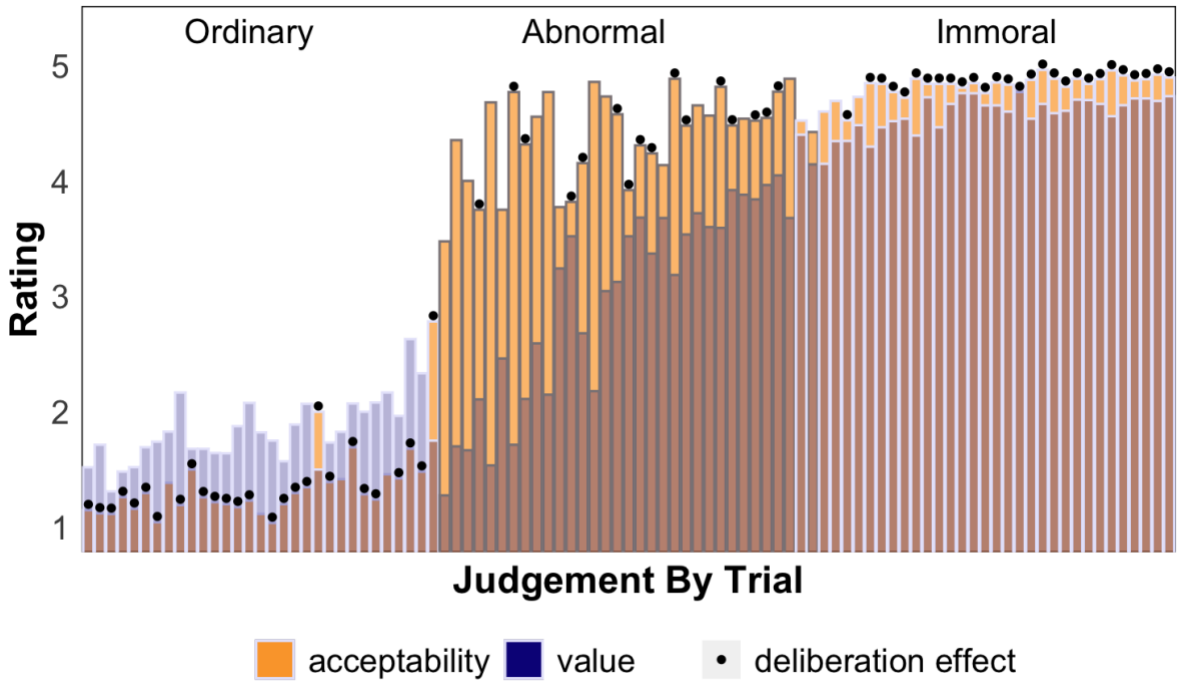

Figure 8. Acceptability and value judgements. Each bar represents an action (*N* = 95), presented in ascending order according to the average of all four judgements for each action. A rating of 5 carries a negative valence (*very bad, not at all socially acceptable*), while a rating of 1 carries a positive valence (*very good, very socially acceptable*).

Drivers of the deliberation effect. Among events for which there was a deliberation effect, which variables predict differences in fast and slow judgements of possibility? A linear mixed effects model was used to predict the difference between fast and slow possibility ratings from the variables value, probability, acceptability, and morality, considering only possible events for which there were a deliberation effect (*N* = 75). The model included a random intercept for the 8 scenarios.

The significance of an effect was computed by using an ANOVA to compare a model that included the factor of interest to a model that did not include that factor but was otherwise identical. There was a main effect for value ($X_2(1) = 4.03$; p=.04), such that actions with a higher

value rating (more negative valence) were judged to be impossible to a greater extent under time pressure. There was no main effect for probability ($X_2(1) = 0.07$; $p = .79$), morality ($X_2(1) = 0.08$; $p = .7741$), or acceptability ($X_2(1) = 0.01$; $p = .94$).

Discussion

The three studies above use the shared adaptive sampling model as a framework for considering how social norms interact with the decision-making process, hypothesizing that they, like moral norms, would constrain intuitive assessments of possibility. Had this finding been born out in the data, it would have provided evidence to suggest that we follow rules and social norms because the possibility of violating them does not occur to us. However, at difference with this prediction, abnormal actions — those which violate general social norms, like crab-walking in a public space — were the only actions to produce no significant difference between speeded and reflective judgements of possibility (see Figure 2). To better understand the disconnect between the hypothesis and the results, this article now turns to a deeper exploration of the shared adaptive sampling model and of moral and social norms.

Shared adaptive sampling and cached value

To review, the shared adaptive sampling model holds that decision-making occurs via a two-step, iterative process. In the first step, the mind subconsciously forms a consideration set, informed by the historic value and probability of non-actual possibilities. In the second step, the mind reasons over those possibilities more explicitly, dismissing those that would not be situationally appropriate or possible (Phillips, Morris, & Cushman, 2019).

Under the shared adaptive sampling model, we would expect to see that actions that were considered bad and unlikely would produce a larger deliberation effect, even when presented without context. With this in mind, in Study 3 subjects were asked to make context-free judgements of the probability and value of the 128 actions proposed in the first two studies. These assessments reflect the general cached probability (from *very likely* to *very unlikely*) and general cached value (from *very good* to *very bad*) of each action.

Though broadly speaking, worse events were judged to be impossible to a greater extent under time pressure, this effect is specifically traceable to the context-free value of those events, which interacted with deliberation above and beyond other judgements' interactions. That value was a strong and singular predictor of the deliberation effect provides new evidence to suggest that, consistent with the shared adaptive sampling model, context-general value plays a key role early in the possibility sampling process.

The same effect was not found for probability; this is at odds with past theoretical and empirical work on possibility sampling (Kahneman, 2011; Lieder, Hsu, & Griffiths, 2014; Phillips, Morris, & Cushman, 2019; Bear et al., 2020). However, this finding, or lack thereof, should be taken with a grain of salt, as the time pressure paradigm used to measure intuitive judgements of possibility (Phillips & Cushman, 2017) successfully identifies the possibilities that participants' minds instinctively reject but fails to replicate the process of possibility generation. Because possibility generation must necessarily occur before pruning, the time pressure paradigm falls short of fully capturing the first stage of decision-making, as described in the shared adaptive sampling.

It may be the case that probability is particularly relevant during the initial, pre-pruning round of possibility generation. This would be consistent with the availability heuristic, a "mental shortcut" through which exposure and ease of access influence assessments of probability or likelihood (Kahneman, 2011; Lieder, Hsu, & Griffiths, 2014). It would also be consistent with research on spontaneous generation of possibilities (Bear et al, 2020). Without further research, though, it is difficult to make any definitive claims.

Situation-specific unacceptability

Determining the general cached value for each action is helpful in that it provides some structure to the results obtained in the previous two studies. We would not expect to see a deliberation effect for actions that are generally acceptable but situationally maladaptive, as was the case for many of the abnormal actions. Indeed, abnormal events were consistently judged to be more specifically unacceptable than generally bad ($t(87.39) = 11.72$, $p < .001$). The variability in the context-free value and probability of abnormal events (see Figure 7) provides context for the lack of a clear deliberation effect for these events. On the whole, immoral events are also more specifically-unacceptable than they are generally-bad ($t(111.26) = 9.90$, $p < .001$), though to a significantly lesser degree than is the case with the abnormal events ($t(65.95) = -10.62$, $p < .001$).

Moral norms vs. social norms. This is consistent with the argument that moral norms are different from social norms in that they hold more weight across contexts. As a class, norms involve "a right to expect and demand" certain things of group members (Brennan et al., 2013). The key difference between moral norms and social norms is the audience to which each applies. Regardless of whether you believe in a universal moral code, those who subscribe to moral norms generally consider them to apply to the human species as a whole (Brennan et al., 2013). They are treated as universal and non-arbitrary (Brennan et al., 2013; Lewis, 1969). Moral norm violations are generally bad as well as specifically unacceptable. For example, skinning a cat for personal enjoyment is bad (Value = 4.73), regardless of whether or not you do it during your professor's lecture (Acceptability = 4.89).

Social norms, on the other hand, may only apply to a certain group and in certain situations (Brennan et al., 2013). While moral norms hold weight across contexts, social norms

require context: certain actions may be inappropriate in certain settings but fine in other settings. For example, it is generally good to play boardgames (Value = 1.54), but it is far from socially acceptable to play a boardgame during your professor's lecture (Acceptability = 4.68).

Comparing the relative morality of abnormal and immoral events, we see that immoral events were almost universally considered to be more immoral than abnormal events (see Figure 4). Further, abnormal events fall within the range of morality judgements for which fast and slow possibility judgements were largely similar; however, immoral events, which fall at the higher end of the moral impermissibility spectrum, saw a sharp increase in reflective impossibility judgements (See Figures 4 and 5). Perhaps more personally salient social norm violations would have been capable of producing a deliberation effect: future research should explore this possibility. But regardless of whether or not that is the case, the data suggest a clear division between moral and social norms.

Normativity and what's worth considering

Deliberative norms vs. practice-based norms

Brennan et al. (2013) distinguish between deliberative norms and practice-based norms. Deliberative norms govern the acceptability of certain thoughts, whereas practice-based norms govern the acceptability of certain behaviors. Building a bonfire in your office breakroom (Value = 4.6, Probability = 4.92) or hiding a dead body in a public library's bathroom (Value = 4.67, Probability = 4.7) are also not options that many would act on.

The difference, however, is that you might be less likely to admit to seriously entertaining the latter two possibilities (Brennan et al., 2013). There's a sense in which it feels inappropriate to even consider burning down your office or murdering someone and then placing the evidence in a building filled with children. These actions violate engrained values about respect for others'

property and wellbeing. They also feel less arbitrary and more inherently wrong, unlike crab-walking, which is a harmless activity. One could conceive of a possible world in which walking around on four legs is the normal, even respectful, thing to do. In this world, we walk on two legs. Public buildings would be chaotic and possibly dangerous places with some people crab-walking and others travelling bipedally. Our walking norms are not entirely arbitrary (walking on two legs is faster), but they don't restrict our imagination in the same way that norms against bodily harm do.

Deliberative norms and morality

As they are more engrained, and more universal, we might expect moral norms to constrain thought in addition to constraining practice. This is not to suggest that moral and deliberative norms are one in the same. Rather, moral norms may be a subset of deliberative norms. Particularly potent social norms — for example, social norms governing the operation of a group that is central to one's social identity — may also function as deliberative norms, constraining thought in addition to constraining action (Brennan et al., 2013). Less potent social norms may become relevant later in the deliberation process; these may include the generally-good, situationally-unacceptable norms discussed earlier.

Consider Figure 9, which displays a ceiling effect wherein people consistently report that it would be *very unlikely* that they would do things falling above a certain "bad" value. The vertical line here is drawn at 3.75, after which the linear regression becomes nearly flat. Thus, actions falling below a certain value (3.75 out of 5, slightly better than *somewhat bad*) were almost uniformly judged as unlikely; people were not willing to consider that it would be "likely" that they do something that is generally "bad." Some of the abnormal events were

judged to be just as "bad" as were the immoral events, and the ceiling (above 3.75) consists of both immoral and abnormal events.

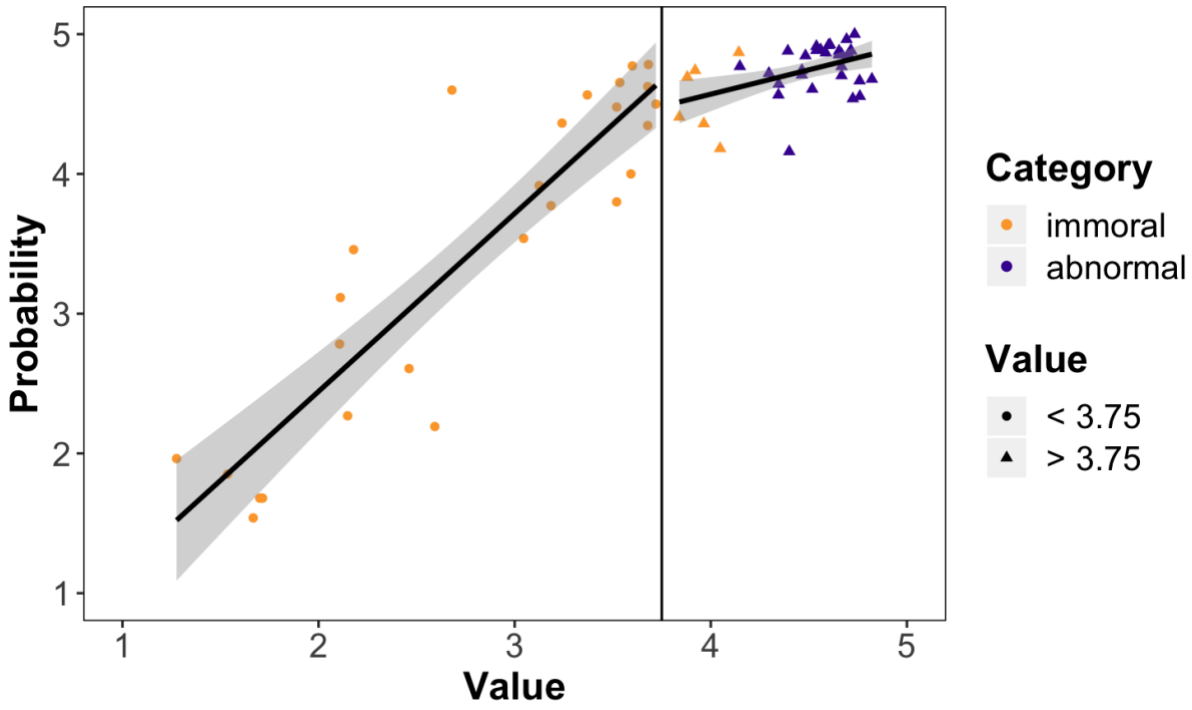

Figure 9. Probability and value of immoral and abnormal events. The points in Figure 9 represent immoral and abnormal actions. The x-axis reflects value ratings, and the y-axis reflects probability ratings. A rating of 5 refers to *very bad* or *very unlikely*, while a rating of 1 refers to *very good* or *very unlikely*. The figure presents two linear regressions: one for one for events that had a value judgement less than 3.75, and one for events that had a value judgement of greater than 3.75.

Tying it together: Deliberation, norms, and the shared adaptive sampling model

Having explored all the above, we can now return to the original research question. Previously, we considered two main positions on normative behavior and the shared adaptive sampling model: (1) that social norms, like moral norms, constrain possibility sampling in the first, intuitive stage of decision-making and (2) that social norms constrain possibility sampling

in the second, deliberative stage of decision-making. The results from these three studies replicated the morality effect reported by Phillips & Cushman (2017) but failed to extend this finding to social norm violations.

However, to say that decision-making is a clean, two-step process, and that the time pressure paradigm accurately assesses judgements made in the first step of that process, may be misleading. Intuitive judgements of pre-written actions are hardly comparable to possibilities generated internally. It may be more appropriate to model the initial stage of possibility sampling — the formation of the initial consideration set — as a two-part process: first possibility generation, then instinctive pruning. Both processes may be informed by historic value and probability, and both may occur below conscious awareness.

And, if possibility sampling is an iterative process, deliberative norms likely come into play in the earlier rounds of sampling, "removing some alternatives from the agenda of available options" in the initial round of intuitive, context-free pruning (Brennan et al., 2013, p. 251). A practice-based norm, however, might not become relevant until later in the decision-making process, when contextual factors like environment, actors, and mood are considered. Thus, when it comes to the shared adaptive sampling model, the more relevant distinction may be between deliberative and practice-based norms, rather than moral and social norms. In this way, we can understand deliberative norms as those norms which constrain thought in the first stage of possibility sampling (intuitive pruning of the original consideration set) and practice-based norms as those norms which constrain thought in the second stage of possibility sampling (the explicit deliberation of the consideration set).

As a class, the abnormal actions used in this study were not sufficiently "bad" or salient to qualify as violations of deliberative norms; thus, they did not receive higher impossibility ratings under time pressure.

Limitations and Implications for Future Research

Population Studied

In the case of this study, a diverse sample, drawn from across education backgrounds, genders, and geographies, may have been a barrier rather than a boon. Social norms are highly specific to certain groups of people (Brennan et al., 2013; Lewis, 1969). It would have been more appropriate to study the effects of norms on decision-making for a narrowly defined social group, like a college campus or a workplace. It is difficult to convey the social meaning of certain actions via an online survey that, by design, must be broadly interpretable to a wide range of participants. The need for interpretability limited the sorts of social norms that could be included in the survey; narrowing the population studied would have made it possible to target more specific and salient social norms.

Motivations for Conforming to Norms

This study offers little insight into the reasons why participants issued the judgements they did. Did they view certain actions as unacceptable because they were intrinsically wrong, because they would feel guilty for doing them, or because they feared backlash from others? Research suggests that social punishment and social expectations are critical in enforcing norms; this study does not involve either (Declerck et al. 2013; Baumard et al. 2013; Rand et al., 2014; Brennan et al., 2013; Lewis, 1969; Gürerk et al., 2006; Tomasello, 2019).

First-Person Questionnaire

A key limitation of the study is that its subject is social norms, yet its questions do not at all involve judgements about other people: all questions were asked in the first-person (ie, how good or bad would it be for you to …). First, this is problematic because we are often not faithful judges of ourselves. People tend to overestimate the degree to which they personally comply with positive social and moral norms (Alicke & Govorun, 2005; Tappin & McKay, 2016). Future research might compare these first-person judgements with similar third-person judgements, to measure and control for any sort of self-serving bias, particularly with regard to judgements of probability. Second, first-person judgements isolate the participant from considerations of social accountability. Some research suggests that some degree of observation is required for people to act on social norms (Brennan et al., 2013).

Possibility Generation

The time pressure paradigm measures what possibilities people are quick to rule out — perhaps reflecting the first round of revisions in the adaptive sampling process — but it doesn't measure what initially comes to mind. It is possible that at this stage, probability becomes more relevant, as we would more readily generate or conceptualize a possible action that we have frequently encountered in the past.

Future Research

Future empirical research on intuition and empirical norms might focus on a specific demographic group, beginning with an ethnographic study of the group's normative behaviors and then following up with a survey similar to the one used in this study, using understanding gained in the ethnography to inform the content of the survey. Alternatively, future research could use a time pressure paradigm with possible actions that violate norms or social

expectations that apply specifically to a certain group but not others, focusing, for example, on gender norms or religious practices.

In order to better understand people's motivations for conforming to norms, researchers might ask 'how mad would people be if you …' rather than 'how socially acceptable is it for you to …' This would focus attention on the social consequences of norm violations. An alternative approach — one that would avoid the potential confound of any self-flattering biases — would be to ask participants to make social judgements about third parties by presenting actions in the third person. Finally, future research might present participants with a scenario, ask them to quickly list a series of possible actions they might take in that scenario, and then ask them to revisit and edit that list. This approach would better capture the process of possibility generation.

Conclusion

When making intuitive judgements of possibility, participants regarded immoral actions as impossible; the same effect was absent in evaluations of actions that are simply abnormal, violating social norms rather than moral norms. However, assessments of moral permissibility are ultimately not responsible for this effect; the general perceived value (*good* vs. *bad*) of proposed actions is what best predicts increased error in speeded vs. reflective judgements of possibility. This is consistent with a two-staged decision-making process, as described in the shared adaptive sampling model, in which the mind samples generally good and generally likely possibilities before giving greater consideration to contextually appropriate possibilities. Future research, however, should explore possibility generation, an underexplored prerequisite step to the first stage of intuitive possibility sampling.

Overall, these studies elucidate the similarities and differences between various normative judgements (moral permissibility, social acceptability) and general assessments of value and probability, corroborating the shared adaptive sampling model, providing reason to believe that moral norms are wholly distinct from social norms, and offering insight into how deliberative and practice-based norms may interact with possibility sampling and decision-making.

Different norms operate at different levels of the possibility sampling process: Some may govern what we do when others are watching, some govern what we are willing to consider, and some may constrain possibilities that comfortably come to mind, without governing our thoughts or actions. The time pressure paradigm used in this project is best suited to study the sorts of norms that determine which possibilities are appropriate to seriously entertain, with near-universal relevance. Future research may consider norms affecting possibility generation, group-

specific social norms as identified through ethnographic research, or practice-based norms enforced by social punishment.

References

Alicke, M. & Govorun, O. (2005) The better-than-average effect. In Alicke, M., Dunning, D., & Krueger, J. (Eds), *The Self and Social Judgement* (pp. 85-106). New York, NY: Psychology Press.

Alvard, M. (2013). Partner selection, coordination games, and group selection. *Behavioral and Brain Sciences,* 36(1), 80-81. doi:10.1017/S0140525X12000702

Baumard, N., André, J., & Sperberg, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences*, 36, 59-122.

Bear, A., Bensinger, S., Jara-Ettinger, J., Knobe, J., & Cushman, F. (2020). What comes to mind? *Cognition*, 194, 104057.

Brennan, G., Eriksson, L., Goodin, R., & Southwood, N. (2013). *Explaining Norms*. Oxford: Oxford University Press.

Byrne, R. M. (2007). Précis of *The Rational Imagination: How People Create Alternatives to Reality. Behavioral and Brain Sciences*, 30, 439-480.

Declerck, C.H., Boone, C., & Emonds, G. (2013). When do people cooperate? The neuroeconomics of prosocial decision making. *Brain and Cognition*, 81, 85-117.

Dunham, Y. (2018). Mere membership. *Trends in Cognitive Sciences,* 22, 780-793.

Emonds, G. (2013). *Cooperation in social dilemmas: Using fMRI to investigate the influence of extrinsic incentives and intrinsic social preferences on social decision making* (Doctoral dissertation). Retrieved from UMI (3538745).

Ensminger, J. (2004). Market integration and fairness: Evidence from ultimatum, dictator, and public goods experiments in East Africa. In Henrich, J. et al. (Ed.), *Foundations of Human Sociality* (pp. 356-381). New York, NY: Oxford University Press.

Fehr, E., & Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Science*, 8(4), 185-190.

Gürerk, Ö., Irlenbusch, B., & Rockenbach, B. (2006). The competitive advantage of sanctioning institutions. *Science*, 312, 108-110.

Hamlin, J.K. et al. (2011) How infants and toddlers react to antisocial others. Proc. Natl. Acad. Sci. U.S.A. 108, 19931–19936.

Hauser, J.R. (2014). Consideration-set heuristics. *Journal of Business Research*, 67(8), 1688-1699.

Henrich, J. (2000). Does culture matter in economic behavior? Ultimatum game bargaining among the Machiguenga of the Peruvian Amazon. *The American Economic Review*, 90(4), 973-979.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Kalish, C. (1998). Reasons and Causes: Children's Understanding of Conformity to Social Rules and Physical Laws. *Child Development*, 69(3), 706-720.

Lewis, D. (1969). *Convention*. Cambridge, MA: Harvard University Press.

Lieder, F., Hsu, M., & Griffiths, T. L. (2014). The high availability of extreme events serves resource-rational decision-making. In *Proceedings of the annual meeting of the cognitive science society*, 36(36).

Martin, J.W., Jordan, J.J., Rand, D.G., & Cushman, F. (2019). When do we punish people who don't? *Cognition*, 193, 1-13.

Phillips, J., Cushman, F. (2017). Morality Constrains the Default Representation of What Is Possible. *Proceedings of the National Academy of Sciences*, 114 (18), 4649 -4654.

Phillips, J., Morris, A, & Cushman, F. How we know what not to think. Unpublished manuscript, Program in Cognitive Science, Dartmouth College & Department of Psychology, Harvard University.

Rand, D.G., Peysakhovich, A., Kraft-Todd, G.T., Newman, G.E., Wurzbacher, O., Nowak, M.A., & Greene, J.D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications*, 5(3677), 1-12.

Rand, D.G., Nowak, M.A. (2013). Human cooperation. *Trends in Cognitive Science*, 17(8), 413-425.

Shtulman, A., & Phillips, J. (2018). Differentiating "could" from "should": Developmental changes in modal cognition. *Journal of Experimental Child Psychology*, 165, 161-182.

Smith, A (1759). *The Theory of Moral Sentiments*. London: Printed for A. Millar, and A. Kincaid and J. Bell.

Tappin, B. & McKay, R. (2016). "The Illusion of Moral Superiority" *Social Psychology and Personality Science*, 8(6), 623-631.

Tomasello, M. (2019). The Moral Psychology of Obligation. *Behavioral and Brain Sciences*, p. 1-33, forthcoming.

Tyler, T. (2008). Psychology and Institutional Design. *Review of Law and Economics*, 4 (3), 801-885.