Maxim Consequentialism for Bounded Agents

Mayank Agrawal
Princeton University

David Danks
UC San Diego

Abstract

Normative moral theories are frequently invoked to serve one of two distinct purposes: (1) explicate a *criterion of rightness*, or (2) provide an ethical *decision-making procedure*. Although a criterion of rightness provides a valuable theoretical ideal, proposed criteria rarely can be (nor are they intended to be) directly translated into a feasible decision-making procedure. This paper applies the computational framework of bounded rationality to moral decision-making to ask: how ought a bounded human agent make ethical decisions? We suggest agents ought to follow moral *maxims*: principles that approximate rightness in many situations, but that can be overridden in specific, precisely describable circumstances. While this intuitive idea has been proposed many times before, we provide a precise model of how *maxim consequentialism* functions as an approximation to an act-consequentialist criterion of rightness, while maintaining the flexibility and defeasibility that has eluded most forms of rule consequentialism. Furthermore, while our overarching aim is to propose a new normative standard of moral decision-making, we demonstrate how maxim consequentialism can also function as a descriptive account of human behavior. We conclude by noting that different criteria of rightness may lead to different maxim-based ethics.

Keywords: consequentialism, bounded rationality, cognitive science

Maxim Consequentialism for Bounded Agents

I. Introduction

Normative moral theories are frequently invoked to serve one of two distinct, separable purposes: (1) explicate a *criterion of rightness*, or (2) provide an ethical *decision-making procedure* (Bales, 1971; Adams, 1976). These are clearly distinct: a characterization of rightness does not necessarily provide a tractable way to achieve it, while a defensible decision-making procedure may sometimes yield actions that fail to be right. One must be clear about the goal of a particular normative ethical theory, else inapt objections may be levied (Hare, 1981).

In particular, although a criterion of rightness provides a valuable theoretical ideal, proposed criteria rarely can be (nor are they intended to be) directly translated into a feasible decision-making procedure (Smart, 1956; Railton, 1984). Humans are epistemically bounded, cognitively limited agents. The criterion of rightness may involve information that we cannot know in the moment, or require inferences and calculations that we cannot perform, or otherwise describe a computationally intractable ideal that is unrealistic (and perhaps even self-defeating) in everyday situations.

A clear articulation of this gap can be found in the two-level utilitarian theory of Hare (1981), though the general distinction is widespread in moral theories.[1] He distinguishes between a 'critical' and 'intuitive' level of utilitarian thinking, where the former provides for the selection of moral principles and the latter for application of them to real-world situations.[2] He concedes that the computationally unbounded *archangel* can (and ought to) exclusively use the critical level of thought, while the *prole* who is incapable of critical thought should instead rely solely on intuitive reasoning. Hare argues that we humans lie between these two extremes, and so our utilitarian thinking should be some sort of rational "blend," where the exact details are ultimately

[1] Hare himself cites Plato, Aristotle, Mill, and Rawls as precursors to his approach, but considered the proposed distinction to be largely neglected by the philosophical audience in his day.

[2] Hare's main justification for this distinction is cognitive costs (rather than epistemic bounds).

a psychological question. However, Hare (1981) lacked the psychological data and formal frameworks to show how this might work (in addition to being committed to a more narrow form of utilitarianism).

We aim in this paper to return to this general idea in light of forty additional years of (computational) cognitive science. Human limitations have been extensively catalogued in cognitive psychology, largely discrediting the normative *homo economicus* assumption that humans ought to be perfectly rational cognitive agents (Kahneman & Tversky, 1979). At the same time, other work has provided reinterpretations of those limitations in an attempt to save rationality (Lewis, Howes, & Singh, 2014; Gershman, Horvitz, & Tenenbaum, 2015; Lieder & Griffiths, 2020). The core idea is that people are optimizing (i.e., behaving rationally) *relative to their cognitive bounds*, even if they cannot optimize *simpliciter*. Moreover, this work has led to precise mathematical frameworks that capture those bounds, and rational behaviors within them, so we can now often derive the rational action or cognitive process for a computationally bounded human.

In this paper, we apply these computational frameworks to moral decision-making to ask: how ought a bounded human agent make ethical decisions? For expository purposes, we will assume some form of act consequentialism as the criterion of rightness. Act consequentialism has often been dismissed as computationally impossible for bounded humans, and so it is a particularly appropriate place to apply our approach. Having said that, our approach is relatively modular in the sense that other criteria of rightness could be used instead.[3] The view that we develop bears many similarities (in both argument and substance) with rule consequentialism. Section II thus provides a high-level sketch of traditional motivations for rule consequentialism, as well as standard objections against it. Section III then introduces a computational framework for bounded rationality, outlining both the mathematical formalization as well as some of the core results from this paradigm. Section IV provides our answer to the focal question of this paper: We

[3] At the very least, one could consequentialize an alternative moral theory or criterion of rightness (e.g. Portmore, 2007, 2009) and then derive the boundedly rational decision procedure for those consequences.

show that boundedly rational agents ought to make moral decisions by applying moral *maxims* (at least, in many situations). This maxim consequentialism is both rationally justifiable for agents such as us, and also avoids the standard objections to rule consequentialism. We conclude by considering possible extensions to our analysis, as well as opportunities for other types of maxim-based ethics.

II. Rule Consequentialism

Rule consequentialism has often been defended on similar grounds to our approach—namely, as the proper decision procedure for computationally bounded ethical agents. In its strongest form, rule consequentialism (Harrod, 1936; Rawls, 1955; Harsanyi, 1977; Brandt, 1984; Hooker, 1990; Parfit, 2011) combines elements from three of the main families of normative ethics (consequentialism, Kantian deontology, and contractualism), and so has also attracted interest from those seeking a convergent solution to moral action (Hare, 1981; Parfit, 2011; Awad et al., 2022).

The easiest path to rule consequentialism is arguably as a response to act consequentialism, as illustrated by the following dilemma from Ross (1930, pp. 34-35):

> Suppose, to simplify the case by abstraction, that the fulfilment of a promise to A would produce 1,000 units of good for him, but that by doing some other act I could produce 1,001 units of good for B, to whom I have made no promise, the other consequences of the two acts being of equal value; should we really think it self-evident that it was our duty to do the second act and not the first? I think not. We should, I fancy, hold that only a much greater disparity of value between the total consequences would justify us in failing to discharge our *prima facie* duty to A.

Here, the act consequentialist prescribes the agent to break their promise in order to increase utility. The objector disagrees, arguing that while promise-breaking in this individual scenario increases utility, a society with promise-keeping as a norm will overall perform better (since we typically cannot know the exact act consequentialist verdict). Generalizing, the rule consequentialist considers an act to be right if it results from a right rule, and a rule is right if, when universally adopted, it increases utility.[4] In the above scenario, rule consequentialism prescribes the agent to keep their promise to A, despite the assumed fact that the (local) act-consequentialist criterion of rightness favors the other action.

Despite a measure of intuitive appeal, rule consequentialism has failed to achieve broad acceptance. One persistent issue is its perceived instability (Scanlon, 1982; Arneson, 2005): rule consequentialism is often thought to be inconsistent in some way. Two common objections are:

1. RULE WORSHIP: Rule consequentialism prescribes an agent to act in accordance with an ideal set of rules even if an alternative act is, by the agent's own lights, more beneficial and this fact is known to the agent.

2. COLLAPSE: The precision needed to identify a set of ideal rules will cause rule consequentialism to collapse into act consequentialism in practical scenarios[5].

Rebuttals to these objections (most notably by Brad Hooker, 1990, 2002) frequently invoke the flexibility of human cognition: humans are not automatons blindly following rules, but instead are dynamic agents that adapt and act accordingly. As a result (continues the rebuttals), people need not blindly follow rules nor fully specify them *a priori*, but rather can develop or adapt rules as appropriate. Of course, the natural reply to these rebuttals is to question how this flexibility can be captured (in a defensible way) without collapsing back either into act consequentialism or rule worship.

Our proposal answers these concerns using frameworks from (computational) cognitive science. We suggest agents ought to follow moral *maxims*—principles that approximate rightness

[4]As one can see, this proposal contains aspects of consequentialism (utility maximization), deontology (rules), and contractualism (universalization).

[5]As Hare (1981) noted: "By the time we have been in, or even considered without actually being in them, a few such dilemmas, we shall be getting very long principles indeed. Very early on we shall get principles like 'One ought never to do an act which is G, except that one may when it is necessary in order to avoid an act which is F, and the act is also H; but if the act is not H, one may not' (43 words)."

in many situations—but they should override those maxims in specific, precisely describable circumstances. This form of "rule" consequentialism is more in line with a classical conceptualization that emphasizes the need for "rules of thumb" when approximating an act-consequentialist criterion of rightness (Mill, 1861; Sidgwick, 1913; Urmson, 1953; Smart, 1956), and provides responses to the standard objections levied against rule consequentialism.

We motivate these responses by way of an analogy with chess (and then provide a more formal, decision-theoretic characterization in Section IV). Chess has long been of interest to the psychology and artificial intelligence communities (Chase & Simon, 1973; de Groot, 1978; Silver et al., 2018; Russek, Acosta-Kane, van Opheusden, Mattar, & Griffiths, 2022) because it requires agents to perform a well-defined objective (checkmate the opponent's king) that is almost always computationally intractable. How ought a bounded cognitive agent play chess? Chess players cannot evaluate exhaustive search trajectories over all possible scenarios. Rather, players *ought* to combine short-term search trajectories with general principles of play: 'control the center', 'don't double pawns', 'castle your king', etc. These principles identify common motifs and thus attenuate redundant computation. Successfully using these principles is considered a hallmark of human intelligence, as demonstrated when Garry Kasparov, a human chess grandmaster, adequately competed with Deep Blue, an artificially intelligent chess-playing system capable of searching over two million positions per second (Campbell, Hoane Jr, & Hsu, 2002).

Chess players manage to use these principles without falling prey to RULE WORSHIP or COLLAPSE. While novices may be in danger of RULE WORSHIP, even a little bit of training enables chess players to recognize conditions in which these general principles should be violated. One may consider sacrificing control over the center if they see a way to force their opponent to be checkmated in three moves. Center control is a useful principle, but the overarching objective is to checkmate the opponent, and so the chess player ought to (and in fact, does with experience) override the control-the-center principle when it conflicts with the checkmating-the-king objective.[6] COLLAPSE is also easily resolved. Chess players cannot (and are thus not expected to)

[6]A typical hallmark of (chess) expertise is precisely such violation of a very useful principle in order to win the

perform exhaustive deliberation at every opportunity, nor use hyper-detailed principles. Rather, their experience and available time determine the extent to which they (ought to) use fine- and coarse-grained principles to navigate the complex problem space.

We suggest that ethical maxims should be regarded as similar to chess principles, thereby yielding a maxim consequentialism that provides a flexible, decision-making procedure for choices that can be evaluated by an act-consequentialist criterion of rightness.[7] Flexibility and defeasibility emerge automatically from the computational constraint arguments below, and so this approach can navigate the standard objections levied against rule consequentialism. We turn now to formalization of this analysis.

III. Bounded Rationality[8]

Classical notions of rationality propose agents select the action a^* that corresponds to maximizing expected utility (Morgenstern & Von Neumann, 1953):

$$a^* = \arg\max_{a \in A} \int u(o)p(o|a)do \tag{1}$$

where A is the set of actions, o is a potential outcome, $u(\cdot)$ is a function mapping outcome to utility, and $p(o|a)$ is the probability of realizing outcome o given action a.

This rational ideal was quickly seen to not accurately describe human behavior. Researchers in what we now call the "heuristics and biases" program measured participants' behavior on simple economic decisions (*e.g.* risky choice; Edwards, 1954; Kahneman & Tversky, 1979; Peterson, Bourgin, Agrawal, Reichman, & Griffiths, 2021) and found that humans

game, e.g. when a player sacrifices their queen in order to set up an eventual checkmate.

[7] We note again that the analysis below is modular, so could be replicated for other criteria of rightness. For example, "maxim virtue theory" would result from using a virtue-theoretic criterion instead of a utility function. And so on for almost any other criterion of rightness, though we do not (for reasons of space) explore these other possibilities.

[8] For reasons of space, we provide only a high-level explanation of relevant computational cognitive science frameworks. See Lieder and Griffiths (2020) for an extensive overview.

systematically deviated from the rational ideal. Phenomena such as loss aversion (Kahneman & Tversky, 1979), base rate neglect (Kahneman & Tversky, 1973), and anchoring (Tversky & Kahneman, 1974) were identified and conceptualizated as hallmarks of human irrationality (Ariely & Jones, 2008; Kahneman, 2011; Thaler & Ganser, 2015).[9] In retrospect, perhaps the most important contribution of this program was the demonstration, not that humans are irrational, but that they are *predictably* irrational. This predictability suggests that people might exhibit procedural rationality (Simon, 1955), which focuses on the decision-making process as opposed to the final outcome. That is, people's seemingly irrational choices might be the product of rational cognitive processes that implement sophisticated (computational) tradeoffs.

Recent work in the cognitive sciences (Sims, 2003; Lewis et al., 2014; Gershman et al., 2015; Lieder & Griffiths, 2020) has started to revive and make precise this idea of procedural rationality. The insight here is that humans are bounded agents who do not have the resources to compute the classical economic ideal action. The correct normative standard ought to be an internalist conception, focused on optimal allocation of resources, and this allocation can produce the observed systematic deviations from classical rationality. Two major successes in psychological and neuroscientific decision theory can help to illustrate the nature and power of this focus on procedural rationality.

Impulsive Behavior and Reinforcement Learning

The highly influential framework of dual process models (Epstein, 1994; Sloman, 1996; Kahneman, 2003; Evans, 2008; Dolan & Dayan, 2013) purports to reconcile human rationality with human irrationality. One system is posited to be effortful and deliberative, and it serves as

[9]These results were not lost on ethicists. Baron (1994), Sunstein (2005), and Gigerenzer (2010) outlined sets of moral intuitions that seemed to correspond to decision-making biases. Horowitz (1998) took it further and aimed to undermine the validity of the 'doctrine of double effect' by arguing that the doctrine arises from standard decision-making biases which are not morally relevant. Greene (2008) leveraged neuroimaging work by himself and colleagues (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Greene, Nystrom, Engell, Darley, & Cohen, 2004) to relate humans' non-consequentialist tendencies to "less rational" areas of the human brain.

the rational ideal. The other is thought to be automatic and habitual, and is considered the paragon of human irrationality. The reinforcement learning (RL) community in computational cognitive neuroscience (Sutton, Barto, et al., 1998; Daw, Niv, & Dayan, 2005) formalized this distinction as model-free vs. model-based systems. These systems provide two different ways an agent can learn a *value function*, which helps them evaluate which action to take in a given state. In model-free (MF) learning, agents directly estimate the action value through trial-and-error experience and the subsequent updating of their stimulus-response mappings. In model-based (MB) learning, agents build an internal model of their environment and simulate potential trajectories from any queried state (Daw et al., 2005; Solway & Botvinick, 2012).[10]

For our purposes, the main difference between these approaches is the computations required by each. When confronted with a decision, the MF system uses fast retrieval mechanisms, whereas the MB system requires extensive, time-consuming deliberation. This extra computation provides the MB system with greater accuracy and flexibility since it enables the agent to directly model long-term dependencies. As a result, the agent is faced with a speed/accuracy tradeoff (Daw et al., 2005; Keramati, Dezfouli, & Piray, 2011): should she use the fast-but-perhaps-inaccurate MF system or the slow-but-more-accurate MB system?[11] This model of the human cognitive agent has been used to provide a rational account of habitual, compulsive, and impulsive decision-making as instances of this type of tradeoff (Daw et al., 2005; Keramati et al., 2011; Kool, Gershman, & Cushman, 2017).[12]

Probability Matching in Bayesian Cognitive Science

The Bayesian program in cognitive science (Tenenbaum & Griffiths, 2001; Chater & Oaksford, 2008) has been highly influential in its use of rational analysis (Marr, 1982; Anderson,

[10] Although initial work assumed that these two systems were separate, recent work has tried to integrate them, e.g. Keramati, Smittenaar, Dolan, & Dayan, 2016; Mattar & Daw, 2018.

[11] The RL logic has been adapted to moral decision-making, see Cushman, 2013; Crockett, 2013

[12] Of course, not all habits are necessarily value-based (despite the common assumption in MF decision-making), and thus this model does not claim that *all* habits are rational, see Miller, Shenhav, & Ludvig, 2019.

1990) as a fruitful method by which to explain human behavior. Bayesian accounts have been proposed for cognitive functions such as causal learning and inference (Schulz, Bonawitz, & Griffiths, 2007; Griffiths & Tenenbaum, 2009), motor control (Körding & Wolpert, 2004), word learning (Xu & Tenenbaum, 2007), and symbolic reasoning (Oaksford & Chater, 2001). Under this framework, agents are imbued with prior distributions, and their responses on different tasks are taken to reflect the integration of these priors with new evidence in a Bayes-optimal way. Part of the appeal of this approach is the underlying ethos of rationality: Bayesian updating is one optimal way of learning, and thus a human acting in a Bayes-consistent manner is presumptively rational.

One influential objection to the Bayesian program was that additional, untested assumptions are needed to actually claim that humans are acting in a rational manner (Mozer, Pashler, & Homaei, 2008; Eberhardt & Danks, 2011; Jones & Love, 2011). One notable concern arises because people often exhibit a phenomenon known as probability matching: if there are two possibilities A and B, then people often choose each of those options in proportion to the probabilities of that possibility. For example, if $P(A) = 0.1$ and $P(B) = 0.9$, then people will choose option A on 10% of the cases, even though the classically rational action is to always choose B.[13] Many experimental results provide evidence for Bayesian models only if we assume that people probability match (Griffiths & Tenenbaum, 2006; Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Eberhardt & Danks, 2011), but this assumption seems to contradict the assumption of (classical) rationality at the heart of much rational analysis.

Vul, Goodman, Griffiths, and Tenenbaum (2014) offered a compromise, proposing that the probability matching phenomenon could be rationalized by incorporating human constraints. They argued that the cognitive operation of computing the exact posterior probability of each possibility is costly. Instead, people should (on procedural rationality grounds) take limited samples from the complicated posterior distribution, where the exact number of samples depends

[13] In this case, probability matching leads to an 82% expected success rate, but always-B has a 90% expected success rate.

on the decreasing marginal value of each subsequent sample vs. the cost of time. If every individual were to sample from their posterior once and make a decision on the basis of that one sample, then people would essentially probability match. That is, if we think that people are procedurally rational rather than classically rational, then the empirical data do support the Bayesian models.[14]

In summary, a large line of decision-theoretic work in economics, psychology, and neuroscience has demonstrated that humans do not obey the classical *homo economicus* ideal. Humans are epistemically bounded, cognitively limited agents and thus a more reasonable standard of rationality is to optimize relative to these bounds. In particular, people should be understood as procedurally rational, even if they thereby exhibit (predictable) errors relative to classical standards. Precise computational and mathematical models (including these two examples, but not limited to them) have been developed to show that people respond appropriately when forced to tradeoff speed and accuracy in various ways. We now show how this idea can also illuminate issues about moral decision procedures.

IV. Maxim Consequentialism

The previous section's high-level description of the bounded/procedural rationality approach used largely qualitative terms since formal, quantitative derivations are available in other work. We now turn to the constructive ethical portion of this paper, using this framework to show how maxim consequentialism straightforwardly results from the combination of human computational limitations and an act-consequentialist criterion of rightness[15]. This section is, by necessity, more formal than the previous sections. One key point in favor of maxim consequentialism (as we derive it) is exactly its grounding in precise frameworks from

[14]The authors also explained why people give more classically rational responses in high-stakes situations when the (relative) cost of sampling is low (Vulkan, 2000; Shanks, Tunney, & McCarthy, 2002), as they should (on procedural rationality grounds) generate more samples.

[15]As mentioned before, we assume (but do not endorse) an act consequentialist criterion of rightness for convenience. We welcome efforts to apply similar techniques to other criteria of rightness.

(computational) cognitive science—with corresponding benefits of precision and predictions—rather than reliance on more qualitative arguments. As these analyses do not appear elsewhere, we make sure to "show our work" in this section.

Avoiding the Rule Consequentialism Objections

First, we formally demonstrate how maxim consequentialism can combat the traditional rule consequentialism objections stated in Section II. Consider a moral maxim M (*e.g.*, 'do no harm'). How ought a bounded agent choose whether to apply M in practice? The agent faces two decisions: (1) a meta-decision about whether to consider overriding M; and (2) if the decision in (1) is "yes, consider overriding," then a decision about whether to actually override M after deliberation.

Suppose that the agent finds herself in a situation where the default action is to simply apply maxim M, resulting in 0 utility.[16] Further suppose that the agent's prior belief is that overriding M has a 50% chance of net (positive) benefit and 50% chance of net (negative) loss, but that she could be 100% confident of which outcome if she spends t timesteps analyzing the dilemma. We can distinguish a LOW-STAKES case where the potential gain is 2ϵ and potential loss is -2ϵ, versus a HIGH-STAKES case in which the gain and loss are $2N$ and $-2N$, respectively. We use these names since we further suppose that $N \gg \epsilon > 0$.

The agent must first make the meta-decision to simply apply their default maxim M, or instead analyze the dilemma. The value of computation (VOC) is defined as the expected utility increase from analyzing the dilemma and acting accordingly (either overriding M if gain or following M if loss). Formally,

$$VOC = \mathbb{E}\left[\sum_{o \in O} p(o) U(DF(o))\right]$$

where $o \in O$ refers to the set of outcomes, $p(\cdot)$ is the agent's credence function of the outcomes,

[16]Setting the baseline to 0 is convenient, but not necessary. The baseline utility could be any arbitrary value, though the equations would be a bit more complicated.

$U(\cdot)$ refers to the utility function, and $DF(\cdot)$ refers to the decision function that outputs the agent's choice.

The decision function enables the agent to act rationally according to the results of the analysis. If the agent arrives at the belief that overriding M will have *positive* consequences (i.e., 2ϵ or $2N$), they ought to override M. But critically, if the agent conversely arrives at the belief that overriding M will have *negative* consequences (i.e., -2ϵ or $-2N$), they do not override M but rather apply M and receive 0 utiles. Formally speaking, the agent's decision function is

$$DF(o) = \underset{\{\text{APPLY M},\text{OVERRIDE M}\}}{\arg\max} \{0, U(o)\}$$

The $\arg\max$ in the decision function enables the agent to deliberate about a path with potential negative outcomes without committing to it, ensuring the value of computation is always non-negative. Applying these equations back to our example, we see that the VOC of LOW-STAKES is ϵ whereas the VOC of HIGH-STAKES is N.

The agent ought to take the time to analyze the dilemma—that is, she ought to consider overriding the maxim M—if the corresponding VOC outweighs the costs of deliberation, which we denote as $cost(t)$. (For the purposes of our example, we are agnostic as to the exact form of this function as long as it is monotonically increasing with t.[17]) For suitable values of N, the agent ought to consider overriding maxim M in HIGH-STAKES because $VOC = N > cost(t)$. Conversely, in LOW-STAKES, ϵ is small so almost certainly $VOC = \epsilon < cost(t)$, and the agent should simply apply M immediately, rather than deliberate about whether to apply M.

This example straightforwardly shows how maxim consequentialism overcomes RULE WORSHIP: the agent ought to override a maxim whenever (i) the expected gain of overriding is greater than the cost of deliberation; *and* (ii) deliberation dictates that overriding is the correct action.[18] We can overcome COLLAPSE by extending the above example to include a set of $\{t_i\}$ corresponding to different confidence levels (on the assumption that more

[17] In a consequentialist setting, this cost can be easily specified through factors such as opportunity costs (e.g. Agrawal, Mattar, Cohen, & Daw, 2021) and/or reward rate (e.g. Keramati et al., 2011).

[18] Both conditions are critical here. Condition (i) ensures that agents do not always perform the full (compu-

cognition/computation will lead to higher confidence in the resulting decision). If the agent has appropriate maxims available to her, then she only needs to engage in significant depth of computation (see Keramati et al., 2016; Sezener, Dezfouli, & Keramati, 2019; Agrawal et al., 2021) when very high confidence is required. As a result, the agent can readily work with coarse-grained maxims, as long as the decisions are relatively low stakes.

In summary, bounded agents can use maxims while rationally navigating RULE WORSHIP and COLLAPSE. Like our previous analogy to chess, maxims are helpful and rationally ought to be used in many cases, but they are not absolutes. In particular, an agent can rationally deviate from a maxim: if she is presented with a scenario in which the utility increase from violating the maxim is sufficiently high *and* it is rational to deliberate, then she ought to pursue the higher utility route.[19]

Consequentialist Maxims

The overall formal framework has the resources to avoid immediate objections, so we now apply it to specific moral dilemmas. While the overarching objective of our paper is to specify a normative theory of moral decision-making, we note that our formalization also connects maxim consequentialism with descriptive accounts of human behavior. In particular, when the

tationally intractable) act-consequentialist computations; condition (ii) ensures that they act rationally given their meta-decision.

[19]The argument in this section seems to 'beg the question': isn't the expected value formalization computationally intractable and thus isn't the meta-decision not boundedly rational? In other words, isn't there a fear of infinite regress of computational complexity? To block concerns of infinite regress, it is important that the meta-decision is computationally tractable. We are not proposing that agents calculate the stated expected value formalization, but instead propose that they are approximating it (see (Marr, 1982)). There is empirical evidence, as referenced in the reinforcement learning and Bayesian probability matching work (and, more anecdotally, how chess players operate; though see Russek et al., 2022 for modeling), that agents *are* adaptively making these meta-decisions. Understanding whether these agents are fully computing these meta-decisions or whether (and thus, how) they are using helpful heuristics is important in creating a complete theory of boundedly rational decision-making. If the latter, it is important to understand what the source of these heuristics are, e.g. development and/or evolution.

maxim-consequentialist decision procedure matches human behavior, then we have defeasible reasons to think that people are behaving (procedurally) rationally.

Lying.

> MISSING WEAPON (STANDARD): Your friend asks you where their weapon is. You know where their weapon is, but you would prefer them to not have it. Is it morally permissible to lie and say that you do not know?

Generally speaking, we will assume the morally permissible action is to tell the truth. But, given the potential downstream consequences of your friend having access to their weapon, it may be morally permissible to lie. From the maxim consequentialist perspective, the core question is whether to even engage in deliberation about whether to override the maxim (if we assume that the (local) act-consequentialist decision would be to override). The outcome of this meta-decision will depend on the expected gain from deliberation versus the cost of deliberation. In the STANDARD case, the balance is probably close, but in other cases, the meta-decision to deliberate might be much more obvious. Consider this higher-stakes situation:

> MISSING WEAPON (MENTAL HEALTH): Your friend asks you where their weapon is. You know where their weapon is, but you would prefer them to not have it as they have become ill and you believe they will use the weapon to inflict harm on someone. Is it morally permissible to lie and say that you do not know?

Here, there is a high expected value in lying (i.e., overriding the maxim), because you significantly decrease the (subjective) probability of someone being harmed. Moreover, the cost of deliberation is unlikely to be anything close to this high expected value. Maxim consequentialism thus implies that it is morally permissible to deliberate about whether to, and subsequently actually, override the maxim to lie in MISSING WEAPON (MENTAL HEALTH), in contrast to the more balanced case of MISSING WEAPON (STANDARD).[20]

[20] An 'avoid disaster' condition has been a consistent, but controversial, proposal to save rule consequentialism

Limits of Altruism. Many species, including humans, display altruistic behaviors. While these often reduce an individual's immediate utility, they are generally considered to contribute to a larger social utility function (which may increase the individual's long-term utility). The maxim consequentialist endorses altruistic behavior (to the extent it increases some social utility function), but the specifics are highly dependent on contextual factors.

One simple experimental paradigm that captures this idea is the dictator game (Forsythe, Horowitz, Savin, & Sefton, 1994; Engel, 2011), a game in which participant X is given a fixed amount of money and must choose how much to donate to participant Y, who must simply accept X's decision. In its simplest form, the "game" is a trivial one-shot decision, and the *homo economicus* prescription is for X to keep all the money.[21] Empirically, human participants systematically violate this prediction: people in the X position frequently give a portion of their money to the Y participant. Psychologists and economists often attribute this behavior to some kind of drive or impulse towards fairness (Forsythe et al., 1994; Bolton, Katok, & Zwick, 1998; Camerer, 2003).

This behavior is not observed in all situations; in some cases, this norm of fairness is eroded or violated. In particular, manipulations of the stakes are common in the literature (e.g. Forsythe et al., 1994; Carpenter, Verhoogen, & Burks, 2005; List & Cherry, 2008), and generally result in X allocating a smaller proportion to Y as the total stakes increase (Engel, 2011; Larney, Rotella, & Barclay, 2019). For example, a participant may allocate $\$5$ when given $\$10$ (50%) and $\$200$ when given $\$500$ (40%).[22] Maxim consequentialism predicts exactly this behavior: as the possible gain increases (due to the increasing stakes), the agent ought to be more likely to engage in

from its critics (Hooker, 1995; Arneson, 2005; Kahn, 2013). Here, we demonstrate that this condition naturally arises as part of the meta-decision procedure.

[21] An additional assumption needed here is that the game is one-shot; increasing the horizon of the game complicates the calculus, though the overall qualitative claims still hold.

[22] The effect of stake size is arguably smaller than one would expect on classical rationality grounds, suggesting the power of the norm is high and/or there may be other norms at play here. Questions about why this norm might focus on proportions rather than absolute numbers are interesting but out of scope for the present article.

deliberation about whether to override the norm of fairness. In at least some situations, the result of that deliberation may be a choice to (partially) override the fairness norm. That is, the maxim consequentialist generally abides by rules, but has the ability to restrict the scope of these norms when deliberation is warranted and results in a different decision given the contextual factors. We hasten to add that we use this example only to show that people's willingness to follow a maxim is a function of the stakes; we are not asserting that people are morally *right* to override the maxim in this case.

Incest as Overrepresentation of Extreme Events. We conclude with a final example that shows how one may start to derive substantive maxims themselves. Here, we consider the infamous example of an aversion to incest (Haidt, 2001):

> INCEST (HAIDT): Julie and Mark are brother and sister. They are traveling together in France on summer vacation from college. One night they are staying alone in a cabin near the beach. They decide that it would be interesting and fun if they tried making love. At very least, it would be a new experience for each of them. Julie was already taking birth control pills, but Mark uses a condom too, just to be safe. They both enjoy making love, but they decide not to do it again. They keep that night as a special secret, which makes them feel even closer to each other. What do you think about that, was it OK for them to make love?

Haidt and his colleagues (Haidt, Bjorklund, & Murphy, 2000) found that participants morally opposed the scenario but were 'dumbfounded' when pressed for a rationale. The scenario was constructed to be justified by the act-consequentialist calculus, and thus Haidt (2001) took participants' disapproval as evidence against the rationalist moral theories and towards his own social intuitionist theory.

We argue that the experimental participants are actually behaving in procedurally rational ways: a strong intuitive aversion to incest is justified on bounded rationality concerns, and so a norm against incest (that one is unlikely to deliberate about whether to overrule) is implied for maxim consequentialists. To illustrate, one can model incest as formally similar to a certain kind of RUSSIAN ROULETTE (Railton, 2014), where there is a large probability of a small gain versus a small probability of a large loss:

	Description	$p(\cdot)$	$U(\cdot)$
o_1	Thrill	$\frac{5}{6}$	1
o_2	Death	$\frac{1}{6}$	-10^9

How ought a bounded agent make decisions regarding RUSSIAN ROULETTE? Lieder, Griffiths, and Hsu (2018) proposed that, in these scenarios, agents ought to bias their deliberation process in order to maximize the expected utility of their outcome.

In their argument, the agent is assumed to simulate instances of the gamble,

$$X_1, \ldots, X_n \sim q$$

in which q is a distribution that can be specified as $q_i = w_i p_i$. After sampling, the agent computes the (estimated) expected value,

$$\widehat{U}_n = \frac{\frac{\sum_i U(X_i)}{w_i}}{n}$$

and then decides whether to take the gamble according to the valence of the estimate

$$DF = \underset{\{\text{REJECT},\text{ACCEPT}\}}{\arg\max} \{0, \widehat{U}_n\}$$

Lieder et al. (2018) ask what distribution q should the agent sample from in order to ensure they make the right decision regarding RUSSIAN ROULETTE?

The naive choice is to let $q = p$, the explicit distribution that specifies the gamble. But, because in the hypothesized cognitive process the agent only chooses REJECT if they sample the negative outcome, the agent has a higher-than-optimal probability of choosing ACCEPT (the optimal outcome). 10 samples ensure only an 83.85% chance of REJECT, and a total of 51 samples is needed for a 99.99% chance. This number of samples is perhaps too costly for a bounded agent, and thus a mechanism that ensures the agent chooses REJECT after only a few samples would be valuable.

Lieder et al. (2018) propose agents ought to approximate sampling from a biased distribution[23]

$$q(o) \propto p(o) \cdot \left| U(o) - \mathop{\mathbb{E}}_{p(o)}[U] \right|$$

When sampling from this distribution, the agent has a 99.99% chance of choosing REJECT after only one sample, and thus a boundedly rational agent ought to sample from this biased distribution in order to maximize decision-making utility.

We can extend a similar logic to explain strong aversions to incest. Consider specifying INCEST (CLASSIC) as

	Description	$p(\cdot)$	$U(\cdot)$
o_1	Thrill	p	1
o_2	Repercussion	$1-p$	-10^{-9}

Similar to RUSSIAN ROULETTE, INCEST (CLASSIC) has a low probability, extremely negative outcome and a high probability, slightly positive outcome. Under the Lieder et al. (2018) model, a bounded agent should have a strong, general aversion to incest.

Of course, the premise of INCEST (HAIDT) is that the downside is capped.[24] We formally specify INCEST (HAIDT) as

	Description	$p(\cdot)$	$U(\cdot)$
o_1	Thrill	p	1
o_2	Repercussion	$1-p$	0

In this setting, in which there is no negative outcome, why ought there still remain a (maxim) consequentialist aversion to incest?

[23]The details of this derivation can be found in the original paper.

[24]The Haidt (2001) example only eliminated the biological repercussions; other psychological repercussions could have factored into participants' responses (Royzman, Kim, & Leeman, 2015). Our formal characterization is generous in that we have completely eliminated the downside, and we aim to show that the aversion is nonetheless rational even in this setting.

To understand this aversion, recall that maxim consequentialism involves first making a meta-decision on whether to deliberate at all. Actual deliberation would have costs, particularly since a new distribution q used for sampling would need to be constructed.[25] Thus, when presented with INCEST (HAIDT), the participant first should decide whether to deliberate at all, or simply follow the "no incest" maxim. The small expected gain of deliberation in INCEST (HAIDT) is admittedly positive, but highly unlikely to be greater than the cost of deliberation. Hence, people ought not even consider whether to override the maxim, and should simply say "do not engage in incest." That is, people ought to act exactly as they do in these experiments.

We note two experimental predictions about these cases, given the assumption that people are procedurally rational (and maxim consequentialists). First, if the expected gain of deliberation was sufficiently high (e.g., the act of incest is the only way to save the world), then maxim consequentialism prescribes that people ought to deliberate about whether to override the norm. Second, if people had more experience and exposure to cases like these, then they should develop finer-grained maxims to use. In general, people ought to use maxims that mostly work in most situations, but identification of such maxims may require experience, either by the individual or a teacher (in the case of a social norm). Moral decisions about incest arguably do not arise in the daily lives of Haidt's participants, and so they have no (procedurally) rational reason to learn more fine-grained maxims. Additional experiences could change the maxims that one ought to use.[26]

[25] We assume that the default sampling distribution is the one used for INCEST (CLASSIC). When and how this default distribution is constructed has been explored elsewhere, see Bear, Bensinger, Jara-Ettinger, Knobe, and Cushman (2020) and Griffiths (2020).

[26] We conjecture that something like this phenomenon might explain changing moral behavior in "Trolley Problems" from 2000 to 2020. As those cases became more widely-known, people had increasing experiences with them, and so plausibly (and rationally) developed more fine-grained maxims.

V. Conclusion

There is a long history of attempts to show that computationally and epistemically bounded[27] agents, including us humans, rationally ought to employ some kind of rule-based moral decision procedure. These attempts have been largely unsuccessful, as they have failed to show (in a precise, non-question-begging way) when people ought to use those rules as opposed to overriding them in some particular context. We have proposed that advances in computational cognitive science over the past forty years provide the necessary conceptual, formal, and quantitative tools. The maxim consequentialism that we proposed and developed here implies that people ought to use maxims in much of their moral decision-making, while retaining the flexibility to override a maxim when (a) it is rational to meta-decide in favor of deliberation about whether to override; and (b) deliberation rationally implies that one should override. We have shown how this approach can address various concerns about rule consequentialism, and even provide rational justification for (some of) the substantive content of a moral maxim.

We acknowledge that this paper only scratches the surface of maxim consequentialism. We suggest that there are two key directions that should be explored in the future. First, this paper has considered only a few examples, and so cannot reveal the full scope and complexity of maxim consequentialism. The present paper shows how to answer many different questions about maxim consequentialism, but the actual effort remains to be done. Second, and more importantly, the framework of bounded/procedural rationality does not provide a criterion of goodness, but rather presupposes one. In this paper, we have focused on an act-consequentialist criterion that can be captured in a utility function. However, it will be critical to consider alternative criteria of goodness. For example, a standard concern about deontological theories is that they often cannot explain why different rules are preferred in different contexts. Can this framework help to

[27] Our focus was primarily on computational, as opposed to epistemic, bounds. The influence on rational moral decision-making of bounded rationality based in epistemic bounds (e.g. Icard, 2021) is an intriguing direction for future research.

represent and resolve that concern?[28] Regardless, we propose that "maxim X" accounts of normative moral decision-making, grounded in precise computational models of our bounded cognition, provide an intriguing way to integrate psychology and morality.

[28]For example, perhaps people rationally construct a default ordering of deontological rules, and then rationally make a meta-decision about whether to revise that ordering in a specific situation by considering whether deliberation has a positive expected value.

References

Adams, R. M. (1976). Motive utilitarianism. *The Journal of Philosophy*, *73*(14), 467–481.

Agrawal, M., Mattar, M. G., Cohen, J. D., & Daw, N. D. (2021). The temporal dynamics of opportunity costs: A normative account of cognitive fatigue and boredom. *Psychological review*.

Anderson, J. R. (1990). *The adaptive character of thought*. Psychology Press.

Ariely, D., & Jones, S. (2008). *Predictably irrational*. Harper Audio New York, NY.

Arneson, R. (2005). Sophisticated rule consequentialism: some simple objections. *Philosophical issues*, *15*, 235–251.

Awad, E., Levine, S., Loreggia, A., Mattei, N., Rahwan, I., Rossi, F., ... Kleiman-Weiner, M. (2022). When is it acceptable to break the rules? knowledge representation of moral judgement based on empirical data. *arXiv preprint arXiv:2201.07763*.

Bales, R. E. (1971). Act-utilitarianism: account of right-making characteristics or decision-making procedure? *American Philosophical Quarterly*, *8*(3), 257–265.

Baron, J. (1994). Nonconsequentialist decisions. *Behavioral and Brain Sciences*, *17*(1), 1–10.

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

Bolton, G. E., Katok, E., & Zwick, R. (1998). Dictator game giving: Rules of fairness versus acts of kindness. *International journal of game theory*, *27*(2), 269–299.

Brandt, R. B. (1984). *A theory of the good and the right*. Clarendon Press.

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

Campbell, M., Hoane Jr, A. J., & Hsu, F.-h. (2002). Deep blue. *Artificial intelligence*, *134*(1-2), 57–83.

Carpenter, J., Verhoogen, E., & Burks, S. (2005). The effect of stakes in distribution experiments. *Economics Letters*, *86*(3), 393–398.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive psychology*, *4*(1), 55–81.

Chater, N., & Oaksford, M. (2008). *The probabilistic mind: Prospects for bayesian cognitive science*. Oxford University Press, USA.

Crockett, M. J. (2013). Models of morality. *Trends in cognitive sciences*, *17*(8), 363–366.

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and social psychology review*, *17*(3), 273–292.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, *8*(12), 1704–1711.

de Groot, A. D. (1978). Thought and choice in chess.

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312–325.

Eberhardt, F., & Danks, D. (2011). Confirmation in the cognitive sciences: The problematic case of bayesian models. *Minds and Machines*, *21*(3), 389–410.

Edwards, W. (1954). The theory of decision making. *Psychological bulletin*, *51*(4), 380.

Engel, C. (2011). Dictator games: A meta study. *Experimental economics*, *14*(4), 583–610.

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American psychologist*, *49*(8), 709.

Evans, J. S. B. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.*, *59*, 255–278.

Forsythe, R., Horowitz, J. L., Savin, N. E., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic behavior*, *6*(3), 347–369.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278.

Gigerenzer, G. (2010). Moral satisficing: Rethinking moral behavior as bounded rationality. *Topics in cognitive science*, *2*(3), 528–554.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive science*, *32*(1), 108–154.

Greene, J. D. (2008). The secret joke of kant's soul. *Moral psychology*, *3*, 35–79.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*(2), 389–400.

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

Griffiths, T. L. (2020). Understanding human intelligence through human limitations. *Trends in Cognitive Sciences*, *24*(11), 873–883.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological science*, *17*(9), 767–773.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological review*, *116*(4), 661.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological review*, *108*(4), 814.

Haidt, J., Bjorklund, F., & Murphy, S. (2000). Moral dumbfounding: When intuition finds no reason. *Unpublished manuscript, University of Virginia*, 191–221.

Hare, R. M. (1981). *Moral thinking: Its levels, method, and point*. Oxford: Clarendon Press; New York: Oxford University Press.

Harrod, R. F. (1936). Utilitarianism revised. *Mind*, *45*(178), 137–156.

Harsanyi, J. C. (1977). Morality and the theory of rational behavior. *Social research*, 623–656.

Hooker, B. (1990). Rule-consequentialism. *Mind*, *99*(393), 67–77.

Hooker, B. (1995). Rule-consequentialism, incoherence, fairness. In *Proceedings of the aristotelian society* (Vol. 95, pp. 19–35).

Hooker, B. (2002). *Ideal code, real world: A rule-consequentialist theory of morality*. Oxford University Press.

Horowitz, T. (1998). Philosophical intuitions and psychological theory. *Ethics*, *108*(2), 367–385.

Icard, T. (2021). Why be random? *Mind*, *130*(517), 111–139.

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? on the explanatory status and theoretical contributions of bayesian models of cognition. *Behavioral and brain sciences*, *34*(4), 169.

Kahn, L. (2013). Rule consequentialism and disasters. *Philosophical studies*, *162*(2), 219–236.

Kahneman, D. (2003). A perspective on judgment and choice: mapping bounded rationality. *American psychologist*, *58*(9), 697.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological review*, *80*(4), 237.

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

Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS computational biology*, *7*(5), e1002055.

Keramati, M., Smittenaar, P., Dolan, R. J., & Dayan, P. (2016). Adaptive integration of habits into depth-limited planning defines a habitual-goal–directed spectrum. *Proceedings of the National Academy of Sciences*, *113*(45), 12868–12873.

Kool, W., Gershman, S. J., & Cushman, F. A. (2017). Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychological science*, *28*(9), 1321–1333.

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244–247.

Larney, A., Rotella, A., & Barclay, P. (2019). Stake size effects in ultimatum game and dictator game offers: A meta-analysis. *Organizational Behavior and Human Decision Processes*, *151*, 61–72.

Lewis, R. L., Howes, A., & Singh, S. (2014). Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in cognitive science*, *6*(2), 279–311.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*.

Lieder, F., Griffiths, T. L., & Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological review*, *125*(1), 1.

List, J. A., & Cherry, T. L. (2008). Examining the role of fairness in high stakes allocation decisions. *Journal of Economic Behavior & Organization*, *65*(1), 1–8.

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information.

Mattar, M. G., & Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nature neuroscience*, *21*(11), 1609–1617.

Mill, J. S. (1861). Utilitarianism. *Collected Works of John Stuart Mill*, *10*.

Miller, K. J., Shenhav, A., & Ludvig, E. A. (2019). Habits without values. *Psychological review*, *126*(2), 292.

Morgenstern, O., & Von Neumann, J. (1953). *Theory of games and economic behavior*. Princeton university press.

Mozer, M. C., Pashler, H., & Homaei, H. (2008). Optimal predictions in everyday cognition: The wisdom of individuals or crowds? *Cognitive science*, *32*(7), 1133–1147.

Oaksford, M., & Chater, N. (2001). The probabilistic approach to human reasoning. *Trends in cognitive sciences*, *5*(8), 349–357.

Parfit, D. (2011). *On what matters* (Vol. 1). Oxford University Press.

Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, *372*(6547), 1209–1214.

Portmore, D. W. (2007). Consequentializing moral theories. *Pacific Philosophical Quarterly*, *88*(1), 39–73.

Portmore, D. W. (2009). Consequentializing. *Philosophy Compass*, *4*(2), 329–347.

Railton, P. (1984). Alienation, consequentialism, and the demands of morality. *Philosophy & Public Affairs*, 134–171.

Railton, P. (2014). The affective dog and its rational tale: Intuition and attunement. *Ethics*, *124*(4), 813–859.

Rawls, J. (1955). Two concepts of rules. *The philosophical review*, *64*(1), 3–32.

Ross, W. D. (1930). *The right and the good.some problems in ethics*. Clarendon Press.

Royzman, E. B., Kim, K., & Leeman, R. F. (2015). The curious tale of julie and mark: unraveling the moral dumbfounding effect. *Judgment & Decision Making*, *10*(4).

Russek, E., Acosta-Kane, D., van Opheusden, B., Mattar, M. G., & Griffiths, T. (2022). Time spent thinking in online chess reflects the value of computation.

Scanlon, T. M. (1982). Contractualism and utilitarianism.

Schulz, L. E., Bonawitz, E. B., & Griffiths, T. L. (2007). Can being scared cause tummy aches? naive theories, ambiguous evidence, and preschoolers' causal inferences. *Developmental psychology*, *43*(5), 1124.

Sezener, C. E., Dezfouli, A., & Keramati, M. (2019). Optimizing the depth and the direction of prospective planning using information values. *PLoS computational biology*, *15*(3), e1006827.

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making*, *15*(3), 233–250.

Sidgwick, H. (1913). *The methods of ethics.* Macmillan and Co.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., ... others (2018). A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, *362*(6419), 1140–1144.

Simon, H. A. (1955). A behavioral model of rational choice. *The quarterly journal of economics*, *69*(1), 99–118.

Sims, C. A. (2003). Implications of rational inattention. *Journal of monetary Economics*, *50*(3), 665–690.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological bulletin*, *119*(1), 3.

Smart, J. J. C. (1956). Extreme and restricted utilitarianism. *The Philosophical Quarterly*

(1950-), *6*(25), 344–354.

Solway, A., & Botvinick, M. M. (2012). Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates. *Psychological review*, *119*(1), 120.

Sunstein, C. R. (2005). Moral heuristics. *Behavioral and brain sciences*, *28*(4), 531–541.

Sutton, R. S., Barto, A. G., et al. (1998). Introduction to reinforcement learning.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and bayesian inference. *Behavioral and brain sciences*, *24*(4), 629–640.

Thaler, R. H., & Ganser, L. (2015). Misbehaving: The making of behavioral economics.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, *185*(4157), 1124–1131.

Urmson, J. O. (1953). The interpretation of the moral philosophy of js mill. *The Philosophical Quarterly (1950-)*, *3*(10), 33–39.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive science*, *38*(4), 599–637.

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of economic surveys*, *14*(1), 101–118.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological review*, *114*(2), 245.