On Robustness: An Undervalued Dimension of Human Rationality

Ardavan S. Nobandegani[†,1,4], **Kevin da Silva Castanheira**[†,4], **Timothy J. O'Donnell**[3], **& Thomas R. Shultz**[2,4]
{ardavan.salehinobandegani, kevin.dasilvacastanheira}@mail.mcgill.ca
{timothy.odonnell,thomas.shultz}@mcgill.ca
[1]Department of Electrical & Computer Engineering, [2]School of Computer Science, [3]Department of Linguistics, [4]Department of Psychology
McGill University
[†]**Co-primary authors**

Abstract

Human rationality is predominantly evaluated by the extent to which the mind respects the tenets of normative formalisms like logic and probability theory, and is often invoked by appealing to the notion of optimality. Drawing mainly on Simon's bounded rationality principle, there has been a surge in the understanding of human rationality with respect to the limited computational and cognitive resources the mind is faced with. In this work, we focus on another fairly underappreciated yet crucial facet of rationality, robustness: insensitivity of a model's performance to miscalculations of its parameters. We argue that an integrative pursuit of three facets (optimality, efficient use of limited resources, and robustness) would be a fruitful approach to understanding the extent of human rationality. We present several novel formalizations of robustness and discuss a recently proposed metacognitively-rational model of risky choice which is surprisingly robust to under- and over-estimation of its focal parameter, nicely accounting for well-known framing effects in human decision-making under risk. We close by highlighting the ubiquitous presence of robustness in natural as well as artificial realms, and the implications of our work for rationalistic approaches to understanding human cognition at the algorithmic level of analysis.

Keywords: bounded rationality; robustness; rational process models; heuristics; metacognition

1 Introduction

Practical applications of complex algorithms to solve problems may not always prove to be the ideal approach to real world problems. Indeed, there are circumstances in which simple heuristics outperform optimal process models (Gigerenzer, 2008, 2010). A good example is that of Harry Markowitz previously outlined by Gigerenzer (2010). Markowitz is best known for his optimal asset-allocation model known as mean-variance portfolio, for which he won a Nobel prize in economics. However, when it came to his investments for retirement, he relied on a simpler intuitive heuristic known as the $1/N$ heuristic: allocate your resources equally to each of N alternatives (Gigerenzer, 2007). In fact, it has been shown that $1/N$ heuristic outperforms mean-variance portfolio which is sensitive to sampling error unless there are sufficiently many samples. In contrast, except for N which can be trivially set based on the number of investment options available to the agent, $1/N$ does not have any free parameters to estimate (Gigerenzer, 2010). Even when N is relatively small ($N = 50$), one needs a large amount of data, approximately 500 years of stock data, in order to outperform the simple $1/N$ heuristic. This is far more data than are available to the average investment firm (DeMiguel, Garlappi, & Uppal, 2009). Surely, a model like this cannot reasonably be considered as truly rational as attempted implementations would prove to be impractical. Process models which require gargantuan amounts of data to provide accurate parameter estimates do not possess the robustness necessary to be considered as rational in uncertain environments (Gigerenzer, 2008). Considerations of robustness when evaluating rationality of process models are far too scarce in the psychological literature.

There have been many attempts to define human rationality with respect to normative formalisms like logic and probability theory. In doing so, the notion of optimality is often invoked. Anderson's (1991) rational analysis approach specifically characterizes rationality as the extent to which a model approximates or attains optimality with respect to some reasonable objective (see also Chater & Oaksford, 1999). However, recent work has drawn on Simon's (1957, 1972) principle of bounded rationality to temper rationality by placing limitations on this model (e.g., Icard, 2014; Griffiths, Lieder, & Goodman, 2015, Nobandegani, 2018).

In this paper, we focus on an often overlooked, yet, crucial, factor in understanding human rationality: robustness. To corroborate this view, we discuss a recent metacognitively-rational model of Availability bias which is surprisingly robust to under- and over-estimation of its focal parameter, and which accounts for well-known framing effects in human decision-making under risk: the fourfold pattern of risk preferences in outcome probability (Kahneman & Tversky, 1992) and magnitude (Markovitz, 1952; Scholten & Read, 2014). We further elaborate on the key role of robustness at the cognitive and meta-cognitive levels, and articulate how robustness, along with principles of optimality and efficient use of limited resources, naturally leads to a key, yet, often overlooked, cognitive level: *meta*-metacognition. We present several formalizations of the notion of robustness, and close by discussing how various recent rationalistic approaches to cognition at the algorithmic level (*rational process models*, Griffiths et al., 2009, 2012) could be integrated with robustness, simultaneously enabling the pursuit of optimality, efficient use of limited resources, and robustness.

2 Facets of Human Rationality

In what follows, we first overview the two main facets of rationality predominately discussed in the psychological literature, and then turn our attention to a key, yet often overlooked, dimension of human rationality: robustness.

Optimality Perhaps the best characterized and extensively

discussed facet of rationality is optimality. Optimality has been portrayed as the extent to which a model satisfies some objective (see Anderson, 1991, and Chater & Oaksford 1999). Models generally have as their objective the minimization or maximization of some objective function, or a combination thereof. For example, minimizing sum-of-squared error or cross entropy in training neural networks, minimizing probability of error in decision-making as in the Bayesian decision rule (Poor, 2013), maximizing expected utility as in expected utility theory (Von Neumann & Morgenstern, 1955), or minimizing the maximum probability of error as in the minimax decision rule (Poor, 2013). A model is considered optimal to the extent that it attains the set objectives. Thus, this facet of rationality depends on both the objective and the outcome, without regards to the context in which the cognitive system is operating. Surely, this cannot be taken as a comprehensive evaluation of rationality as it ignores many important factors affecting a cognitive system's performance, e.g., environmental uncertainty, lack of information, resource limitations, etc.

The importance of optimality—in evaluating what it means to be rational—is unquestionable. However we argue, like many others before us (Icard, 2014; Griffith et al., 2015; Nobandegani, 2018; Gigerenzer, 1998, Lewis, Howes, & Singh, 2014; Howes, Lewis, & Vera, 2009; Russell, Stuart & Subramanian, 1995; Russel 1997, *inter alia*), that there are other factors to take into consideration.

Economy In recent years, many have taken inspiration from Simon's (1957) bounded rationality to expand our understanding of human rationality. This concept is heavily based on the limitations of cognitive and computational resources imposed on the model when considering rationality. A boundedly rational agent need not fully optimize but find a solution which only *satisfices* certain criteria given the limitations at hand—both environmental and internal (Simon, 1957). The emphasis here is primarily on the circumstances and conditions under which the cognitive system operates,[1] highlighting the importance of the cognitive system's quest for *economy*: the economical use of limited computational and cognitive resources (e.g., time, memory, information). As opposed to optimality, which is predominately invoked with a disregard for such contextual limitations, the concept of economy is context-dependent. The concept of economy and its role in theorizing about human cognition are mainly pursued under titles like *ecological rationality* (Gigerenzer, 1998; Gigerenzer & Todd, 2012), *algorithmic rationality* (Halpern & Pass, 2011), *bounded-optimality* (Russell, Stuart & Subramanian, 1995; Russel 1997), *boundedly rational analysis* (Icard, 2014), *resource-rationality* (Griffiths et al., 2015), *computational rationality* (Lewis, Howes, & Singh, 2014), and *rational minimalist program* (Nobandegani, 2018).

Importantly, here there are broadly two approaches to economy. One assumes there is a necessary trade-off to be made between the two facets (e.g., Icard, 2014; Griffiths et al., 2015, Russell, Stuart & Subramanian, 1995; Russel 1997), while the other views the facets as largely independent (e.g., Gigerenzer, 2010; Nobandegani, 2018). For example, it has been surprisingly demonstrated that economical process models—often referred to as heuristics (*fast-and-frugal*, Gigerenzer, 2008)—can outperform optimal process models (Gigerenzer, 2010), thereby establishing that, at least in some settings, optimality and economy need not trade off.

Also interestingly, using limited knowledge, some algorithms can outperform or match algorithms which integrate all information available (i.e., multiple regression) (Gigerenzer, 2010). Drawing on the previously discussed example of investment, superior performance of the heuristic is chiefly due to its robustness with respect to uncertainty in parameter estimates (Gigerenzer, 2008). Only under extraordinary circumstances can the optimal, mean-variance portfolio model outperform the simple $1/N$ heuristic.

In the following sections, we shed light on another aspect of rationality which is not extensively discussed in the literature: robustness. Examples of robustness as an objective criterion are provided as well as several formalizations of it, providing formal and precise characterizations of this aspect and facilitating future evaluations of human rationality.

3 On Robustness

Although the concept of robustness is not new in the literature, it has been largely overlooked in discussions of rationality. Robustness has appeared previously in academic writing in a specialized and narrow sense (e.g., Gigerenzer, 2008; Lempert & Collins, 2007), largely without precise formal characterizations. In the field of decision-making, where attempts have been made to tackle the issue of uncertainty in model specifications (specifically the probability distributions of the parameters), robustness has been discussed (Lempert & Collins, 2007). There, importance is placed on not achieving the optimal solution, but dealing with uncertainty—trading off optimality for less sensitivity to violated assumptions (Lempert & Collins, 2007). We propose a similar view when evaluating process models of cognition in general. Robust models should be insensitive to inaccuracies of their parameters, with little or no decline in their performance. An agent should use models allowing them to perform optimally or near-optimally, regardless of the limitations imposed on them and possible miscalculations of model parameters.

At first, it may seem that robustness and economy are addressing the same concerns. However, further investigation of the implications of robustness as an independent facet of rationality reveals that these two facets are indeed distinct.

In fact, we can force a model to be economical (i.e., frugal) by restricting its use of resources (e.g., by limiting the amount of information the model is allowed to process). Nevertheless, this does not make the model robust with respect to miscalculations of its parameters. Let us elucidate this understanding in the context of a recent model by Piantadosi

[1] A reader familiar with Minimalist Program in linguistics (Chomsky, 1993), could see clear connections between the concept of *virtual conceptual necessity* and the topic under discussion here.

(2018). Surprisingly, Piantadosi (2018) presents a single-parameter model capable of fitting any scatter plot, on any number of points, to arbitrary precision. Despite having only a single parameter, this model is overly sensitive to parameter imprecision. We can have this model estimate its single parameter using only one randomly chosen point from the target scatter plot, thereby forcing the model to highly respect economy. Nonetheless, this does not alleviate the oversensitivity of the model with respect to parameter imprecision: Robustness is an intrinsic property of a model (either a model is sensitive to inaccuracies in parameter estimation or not), and it is independent of whether a model is economical.

Economy is primarily concerned with the strategic use of limited resources (e.g., computational, cognitive, etc.). In contrast, robustness is about insensitivity to inaccuracies in parameter estimation; the sources of these inaccuracies often boil down to the agent's incomplete knowledge of, and uncertainty about, its environment. However, incomplete knowledge and uncertainty are not the only factors responsible for an agent's inability to precisely estimate parameters.

There are several sources of uncertainty. First, uncertainty can come from changes in the environment. If one is attempting to estimate a value which changes over time, an estimate would be likely erroneous. Second, uncertainty can come from limited knowledge. An agent may not know all relevant information for the task at hand.[2] Third, even if an agent has all the relevant knowledge at their disposal, the computational power needed to accurately estimate parameters may be outside the agent's computational capacity.

Thus, miscalculations of parameters may be due to external (e.g., environmental changes) as well as internal (e.g., limited computational power) constraints. In that light, robustness can be characterized as preserved performance despite these constraints. Much like optimality and economy, robustness serves as a meta-level objective criterion for an intended cognitive level of analysis, reflecting on the quality of the model developed at that cognitive level.

4 Robustness as an Objective Criterion

In the following, we discuss in greater detail how robustness can serve as a meta-level objective criterion for human cognition at two distinct cognitive levels of analysis: the cognitive and meta-cognitive levels.

4.1 On the Cognitive Level

Reflecting on the robustness of cognitive models leads to a key level of analysis: metacognition. This level of analysis is analogous to the considerations afforded to the economy of cognitive models. To elaborate on the use of robustness as a meta-level objective criterion, we return to the investment example. The example of Harry Markowitz has been used to illustrate the success of heuristics over rational process models (Gigerenzer, 2010). Here, the optimal strategy is impractical to use as it requires a sizable amount of data (about 500 years of stock data) to accurately estimate parameter values. In other words, the optimal asset allocation strategy proposed by Markowitz (1952) would only result in the best outcome if the parameter values were known near-perfectly, as in a small world, but is inferior to heuristics in a large world, where parameter values need to be estimated from limited samples of information. The success of heuristics is largely due to the robustness of their performance afforded by insensitivity to imperfections in parameter estimates (see Gigerenzer, 2008).

Although the literature emphasizes limited number of samples as the main source of inaccuracies in parameter estimation, this account is incomplete. Gigerenzer (2010) argues that the optimality of mean-variance portfolio hinges on accurate parameter estimation using only a limited number of available samples. However, even if the samples would abound, one would need an extraordinary amount of computational power to estimate parameters adequately. Processing 500 years of stock data is no trivial task.

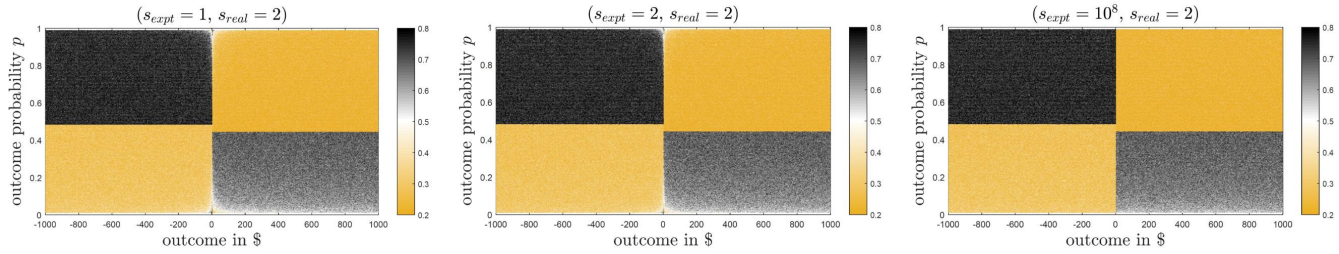
A noteworthy example of computational intractability being the primary source of miscalculations (as opposed to incomplete knowledge) eminently features in the problem of finding Nash equilibria in game theory. Even when everything about the game is known (aka complete-knowledge games), finding a (mixed) Nash equilibrium is computationally intractable (more precisely, it is **PPAD**-complete, Daskalakis et al., 2009), attesting to the claim that miscalculations may sometimes result from computational complexity barriers, not lack of information.

4.2 On Meta-Cognitive Level

Following the logic of the previous section, reflections on robustness can be applied to metacognitive models leading to another key level of analysis: meta-metacognition. "Meta-metacognition" is scarcely used in the literature. Previous uses have either been in a narrow sense (e.g., Arnold, 2013; Buratti & Allwood, 2012), or as a term whose content is not concretely specified, characterized broadly as "reflection" on the metacognitive level without articulating precisely what that reflection means (e.g., Renkl et al., 1996; Efklides & Misailidi, 2010). In what follows, we seek to clarify what meta-metacognitive considerations entail and provide concrete examples.

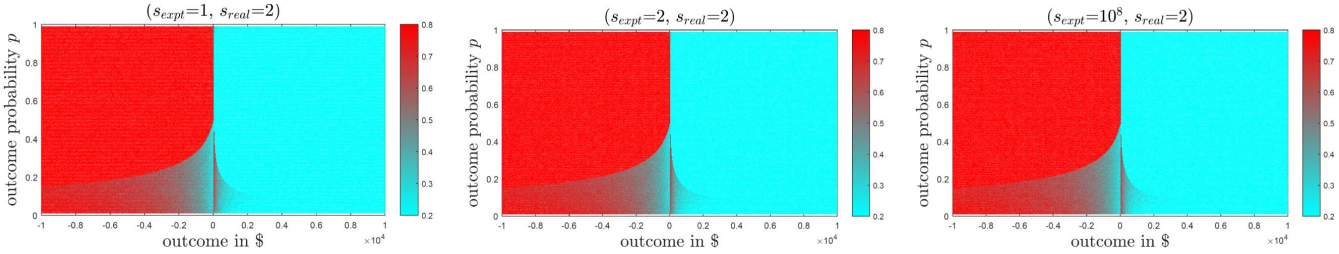
Research on metacognitively-rational models is still in its infancy, and little work is done on this exciting topic (e.g., Lieder & Griffiths, 2017; Nobandegani, da Silva Castanheira, Otto, & Shultz, 2018). A good example of such models is the recent work by Lieder and Griffiths (2017) on rational models of strategy selection. Despite its great performance, a pre-theoretic evaluation of this model suggests that it would not score high on robustness as its performance largely hinges on accurate parameter estimations. In this model, strategies (e.g., heuristics) are evaluated based on their previous performance on problems which share similar features (Lieder

[2] More precisely, uncertainty due to unanticipated environmental changes can be seen as an instance of incomplete knowledge with respect to future states of the environment. In that sense, the first source of uncertainty mentioned above is a special case of the second source.



(a) Under-estimation (b) Exact (c) Severe over-estimation

Figure 1: Accounting for the fourfold pattern of risk preferences in outcome probability (Kahnemnan & Tversky, 1992) using Nobandegani et al.'s (2018) metacognitively-rational model. Each figure plots the probability of choosing the risky choice, depending on the probability of outcome involved in the risky choice (p) and the amount of outcome in dollars; see Nobandegani et al. (2018) for details. A striking resemblance can be observed between **(a,b,c)**.



(a) Under-estimation (b) Exact (c) Severe over-estimation

Figure 2: Accounting for the fourfold pattern of risk preferences in outcome magnitude (Markowitz, 1952) using Nobandegani et al.'s (2018) metacognitively-rational model. Each figure plots the probability of choosing the risky choice, depending on the probability of outcome involved in the risky choice (p) and the amount of outcome in dollars; see Nobandegani et al. (2018) for details. A striking resemblance can be observed between **(a,b,c)**.

& Griffiths, 2017), and the perceived benefits from applying a given strategy (e.g., rewards) are contrasted with the costs (e.g., opportunity costs of the strategy's execution time) to supply a "value of computation" (Lieder & Griffiths, 2017; Horvitz, 1990). To evaluate the performance of a given strategy, accurate estimates of pivotal features of the problem are needed. However, in complex environments where these estimates are not readily available or easily computable, such a metacognitively-rational model for strategy selection would likely fail to satisfy robustness considerations. Therefore, one could say that such metacognitively-rational models are not *meta*-metacognitively rational.

Recent work on rational process models has proven to successfully capture all the three facets of optimality, economy, and robustness. In particular, drawing on the work by Lieder, Griffiths, & Hsu (2017) which applied importance sampling to estimate the expected utility of an action, recent work by Nobandegani et al. (2018) provides a metacognitively-rational process model of Availability Bias (Tversky & Kahneman, 1972), and the fourfold pattern of risk preferences in probability outcome (Tversky & Kahneman, 1992) and in outcome magnitude (Markowitz, 1952), by rationally taking into consideration the amount of time available for decision-making. Concretely, the model takes into account the number of samples the decision-maker draws before making a decision (Nobandegani et al., 2018). This model is in accord with a recent, empirically supported line of research suggesting that people draw very few samples (i.e., few mental simulations) in reasoning and decision-making (Vul et al., 2014; Battaglia et al. 2013; Lake et al., 2017; Gershman, Horvitz, & Tenenbaum, 2015; Hertwig & Pleskac, 2010; Griffiths et al., 2012; Gershman, Vul, & Tenenbaum, 2012; Bonawitz et al., 2014). Further investigation of this metacognitive model reveals that the performance of the model depends on the actual number of samples the model gets to draw (denoted by s_{real}), and not on the number of samples it anticipates drawing (denoted by s_{expt}, with s_{expt} serving as a priori estimate of s_{real}). In other words, the model is robust with respect to inaccurate estimation of the number of samples it gets to draw, both under- and over-estimations. Sensitivity analysis shows that the model is indeed surprisingly robust with respect to its focal parameter s_{expt}. Consistent with the past literature providing evidence for people drawing very few samples in reasoning and decision-making, when s_{real} assumes a value of 2, the fourfold pattern of risk preferences in outcome probability persistently emerges for exact, under- and severely over-underestimated values of s_{expt} (i.e., $s_{\text{expt}} = 2$, $s_{\text{expt}} = 1$, and $s_{\text{expt}} = 10^8$, respectively); see Fig. 1.

The model is also strikingly robust when producing more nuanced patterns of behavior like Markowitz's (1952) fourfold pattern of risk preference in outcome magnitude; see Fig. 2. In fact, it was this model which inspired our considerations of the importance of robustness in evaluating rational cognitive models. (Thus, process models satisfying near-

perfectly all three facets of rationality are indeed possible.) Nevertheless, Nobandegani et al. (2018) failed to provide a formal characterization of the robustness of their model.

In the following section, we formalize robustness to provide precise characterizations of this notion and facilitate future evaluations of rationality. As we show, our work additionally allows us to formally characterize the robustness of the Nobandegani et al.'s (2018) model and the $1/N$ heuristic.

5 Formalization of Robustness

We first consider robustness with respect to real-valued parameters, and finally show how these formalizations can be adapted to the discrete-valued parameters case.

Def. 1 (i^{th}-order locally-robustness) Model M_θ parameterized by θ is i^{th}-order locally-robust at $\theta = \theta_0$ iff M_θ's performance measure $T[M_\theta]$ is insensitive to infinitesimal deviations of θ from θ_0, all the way up to the i^{th}-order. That is, formally, $\forall j \leq i,\ \nabla^{(j)}_{\theta=\theta_0} T[M_\theta] = 0$, where $\nabla^{(j)}_{\theta=\theta_0}(\cdot)$ denotes the j^{th}-order gradient w.r.t. θ and evaluated at $\theta = \theta_0$.

Def. 2 (i^{th}-order ε-locally-robustness) Model M_θ parameterized by θ is i^{th}-order ε-locally-robust at $\theta = \theta_0$ iff M_θ's performance measure $T[M_\theta]$ satisfies: $\forall j \leq i,\ |\nabla^{(j)}_{\theta=\theta_0} T[M_\theta]| \leq \varepsilon$.

Definitions 1 and 2 are founded on an important understanding based on the concept of Taylor series in calculus: The more number of higher-order derivatives of function $f(x)$ are zero (or nearly-zero) at $x = x_0$, the wider and flatter $f(x)$ is in the neighborhood of $x = x_0$. Extending Definitions 1 and 2 to the case of multi-parameter models (as opposed to a single-parameter model M_θ), we arrive at the following:

Def. 3 (i^{th}-order singly-locally-robustness) Model $M_{\theta,\theta'}$ parameterized by $\{\theta, \theta'\}$ is i^{th}-order singly-locally-robust at $(\theta = \theta_0, \theta' = \theta'_0)$ iff $M_{\theta,\theta'}$'s performance measure $T[M_{\theta,\theta'}]$ is insensitive to infinitesimal deviations of θ from θ_0, all the way up to the i^{th}-order, when θ' is held fixed at θ'_0 (denoted by $\theta' := \theta'_0$). That is, formally, $\forall j \leq i,\ \nabla^{(j)}_{\theta=\theta_0|\theta':=\theta'_0} T[M_{\theta,\theta'}] = 0$, where $\nabla^{(j)}_{\theta=\theta_0|\theta':=\theta'_0}(\cdot)$ denotes the j^{th}-order gradient w.r.t. θ' and evaluated at $\theta' = \theta'_0$, when θ' is held fixed at θ'_0.

Def. 4 (i^{th}-order ε-singly-locally-robustness) Model $M_{\theta,\theta'}$ parameterized by $\{\theta, \theta'\}$ is i^{th}-order ε-singly-locally-robust at $(\theta = \theta_0, \theta' = \theta'_0)$ iff $M_{\theta,\theta'}$'s performance measure $T[M_{\theta,\theta'}]$ satisfies: $\forall j \leq i,\ |\nabla^{(j)}_{\theta=\theta_0|\theta':=\theta'_0} T[M_{\theta,\theta'}]| \leq \varepsilon$.

Definitions 1 to 4 can be straightforwardly adapted to to the case of discrete-valued parameters, with operations $\nabla^{(j)}_{\theta=\theta_0}(\cdot)$ and $\nabla^{(j)}_{\theta=\theta_0|\theta':=\theta'_0}(\cdot)$ being replaced, respectively, with the operations $D^{(j)}_{\theta=\theta_0}(\cdot)$ and $D^{(j)}_{\theta=\theta_0|\theta':=\theta'_0}(\cdot)$ defined as follows:

$$D^{(j)}_{\theta=\theta_0} f(\theta) \triangleq \big(f(\theta_0 + i) - f(\theta_0)\big)/i, \tag{1}$$

$$D^{(j)}_{\theta=\theta_0|\theta':=\theta'_0} g(\theta, \theta') \triangleq \big(g(\theta_0 + i, \theta'_0) - g(\theta_0, \theta'_0)\big)/i. \tag{2}$$

Eyeballing Figs. 1 and 2 reveals that Nobandegani et al.'s (2018) metacognitively-rational model is approximately[3] $(10^8 - 2)^{\text{th}}$-order singly-locally-robust at $(s_{\text{expc}} = 2, s_{\text{real}} = 2)$, with the performance measure being the most probable choice suggested by the model (i.e., the preference for the risky choice vs. the safe one). Note that, given that Definition 4 is a relaxation of Definition 3, Nobandegani et al.'s (2018) model is also approximately $(10^8 - 2)^{\text{th}}$-order ε-singly-locally-robust at $(s_{\text{expc}} = 2, s_{\text{real}} = 2)$, $\forall \varepsilon \in \mathbb{R}_+$.

Our formalism also allows us to provide a formal characterization of the robustness of $1/N$ heuristic. Using Def. 2, it is easy to mathematically show that, for any $N_0, i \in \mathbb{N}$, the $1/N$ heuristic is i^{th}-order 0.5-locally-robust at $N = N_0$, with the performance measure being the portion of resources to be allocated to each of N investment alternatives.

6 General Discussion

Examples of robustness in natural and man-made artifacts are abundant and often ensured by adding redundancy. In biological systems, robustness can be characterized as the maintenance of some functionality (e.g., phenotype) despite perturbations (e.g., genetic variation) and achieved through many means, one being redundancy (Kitano, 2004; Felix, 2015). At the genetic level, this can be seen as genes with overlapping products or at the network level with different mechanisms serving the same purpose: glycolysis and oxidative phosphorylation both produce ATP under different conditions (anaerobic and aerobic respectively) (Kitano, 2004). Modularity and decoupling of low-level variations from high-level functionality (e.g., genotype and phenotype) are also seen as sources of robustness in a biological system (Felix, 2015). Furthermore, a modular view of the brain fits nicely with the concept of robustness: Locally perturbations of a module should leave other modules unaffected. In fact, Fodor's (1983) view of low-level system modularity (e.g., perception and language) provides another example of modularity of biological systems. Similarly, decoupling of higher-level systems from lower-level systems is in accordance with the proposed view of robustness. The ubiquitous presence of robustness in biological systems suggests its importance in successful fulfillment of a system's goals (Felix, 2015).

Similarly, in artificial systems robustness is engineered into systems by particularly capitalizing on the benefits of adding redundancy to systems. For example, network architecture comprises several modules with overlapping functionalities, as opposed to a single integrative module (Kurose & Ross, 2009). In information theory, robustness is featured in error detection/correction codes for communicating informa-

[3] The rationale behind using the term "approximately" is that there could be some (x, y)-coordinates whose values are not perfectly invariant across Fig. 1(a-c) (and, likewise, across Fig. 2(a-c)). However, note that even if such (x, y)-coordinates do exist, they are very scarce, as evidenced by the striking resemblance of Fig. 1(a-c) (and, likewise, Fig. 1(a-c)). We could have provided a more rigorous characterization of this possibility using notions analogous to *almost-everywhere* in measure theory. However, for the sake of keeping the formalism simple, we refrained from that.

tion over a noisy medium, which by introducing redundancy into the transmitted code ensures that possible errors can be detected/corrected at the receiver (Cover & Thomas, 2012).

6.1 Toward Robust Models

Outperforming optimal models when accurate-enough parameter estimates cannot be obtained is evidence for their lack of robustness (Gigerenzer, 2008). The success of recent models at capturing and providing metacognitively-rational bases for intricate behavior patterns (Nobandegani et al., 2018) suggests that many of the findings in the psychology literature which are often considered "irrational" may be well-explained by appealing to metacognition or *meta*-metacognition. Indeed, the modeling work which inspired these reflections did not explicitly consider robustness. However, considerations of robustness should not be left up to serendipity. Rather, we believe that robustness of process models should be another factor in the modeler's objective set. Unlike considerations of sensitivity analysis, optimality and economy are not treated as an after-thought. Why should robustness be any different? Nobandegani et al.'s (2018) metacognitively-rational model achieves all three facets of rationality near-perfectly: optimality, economy and robustness.

Several frameworks for theorizing about cognitive process models have been proposed to simultaneously attain optimality and economy (e.g., Icard, 2014; Griffiths et al., 2015; Nobandegani, 2018). An important question is whether and how robustness can be integrated into these frameworks? Drawing on statistical learning theory and machine learning, we proposed a possible solution.

But first it is important to highlight a key connection between the concept of robustness discussed here and that of over-fitting in statistical learning theory and machine learning. If models (or theories) are selected from overly complex hypothesis sets, the learned model would likely overfit the observed data and would not generalize well. Importantly, an over-fitting model would also be fragile (as opposed to robust), since slight perturbations of the training patterns would lead to the selection of a radically different model. In that light, over-fitting models are fundamentally unrobust.

Inspired by these understandings, we suggest that current modeling frameworks should search for algorithms that satisfy some general characteristics ensuring robustness. A prominent such characteristic is for hypothesis sets to be not overly complicated, to avoid over-fitting. Importantly, several important theoretical measures of complexity of a hypothesis set have been already extensively studied in the statistical leaning theory, e.g., VC-dimentsion (Vapnik & Chervonenkis, 1971), Natarajan dimension (Natarajan, 1989), and Rademacher complexity (see Bartlett & Mendelson, 2002).

Although previous work has largely focused on the aspects of optimality and economy, underplaying the role of robustness in rationality, we hope to have given robustness the attention it truly deserves.

References

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences*, *14*(3), 471–485.

Bartlett, P. L., & Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, *3*(Nov), 463–482.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Bonawitz, E., Denison, S., Griffiths, T. L., & Gopnik, A. (2014). Probabilistic models, learning algorithms, and response variability: sampling in cognitive development. *Trends in cognitive sciences*, *18*(10), 497–500.

Buratti, S., & Allwood, C. M. (2012). The accuracy of meta-metacognitive judgments: Regulating the realism of confidence. *Cognitive processing*, *13*(3), 243–253.

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in cognitive sciences*, *3*(2), 57–65.

Chomsky, N. (1993). A minimalist program for linguistic theory. *The view from Building*, *20*, 1–52.

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Daskalakis, C., Goldberg, P. W., & Papadimitriou, C. H. (2009). The complexity of computing a nash equilibrium. *SIAM Journal on Computing*, *39*(1), 195–259.

DeMiguel, V., Garlappi, L., Nogales, F. J., & Uppal, R. (2009). A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science*, *55*(5), 798–812.

Efklides, A., & Misailidi, P. (2010). *Trends and prospects in metacognition research*. Springer Science & Business Media.

Félix, M.-A., & Barkoulas, M. (2015). Pervasive robustness in biological systems. *Nature Reviews Genetics*, *16*(8), 483.

Foder, J. A. (1983). *The modularity of mind: an essay on faculty psychology.* MIT Press.

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.

Gershman, S. J., Vul, E., & Tenenbaum, J. B. (2012). Multistability and perceptual inference. *Neural computation*, *24*(1), 1–24.

Gigerenzer, G. (1998). Ecological intelligence: An adaptation for frequencies. In *The evolution of mind* (pp. 9–29). Oxford University Press.

Gigerenzer, G. (2008). *Rationality for mortals: How people cope with uncertainty*. Oxford University Press.

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

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in cognitive science*, *1*(1), 107–143.

Gigerenzer, G., & Todd, P. M. (2012). Ecological rationality: the normative study of heuristics. In *Ecological rationality: Intelligence in the world* (pp. 487–497). Oxford University Press.

Griffiths, T., Levy, R., McKenzie, C. R., Steyvers, M., Tenenbaum, J., & Vul, E. (2009). Rational process models. In *Proceedings of the cognitive science society* (Vol. 31).

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*(4), 263–268.

Halpern, J. Y., & Pass, R. (2011). Algorithmic rationality: Adding cost of computation to game theory. *ACM SIGecom Exchanges*, *10*(2), 9–15.

Hertwig, R., & Pleskac, T. J. (2010). Decisions from experience: Why small samples? *Cognition*, *115*(2), 225–237.

Horvitz, E. J. (1990). *Computation and Action under Bounded Resources*. PhD Dissertation, Stanford University.

Howes, A., Lewis, R. L., & Vera, A. (2009). Rational adaptation under task and processing constraints: implications for testing theories of cognition and action. *Psychological review*, *116*(4), 717.

Icard, T. (2014). Toward boundedly rational analysis. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 36).

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive psychology*, *3*(3), 430–454.

Kitano, H. (2004). Biological robustness. *Nature Reviews Genetics*, *5*(11), 826.

Kurose, J. F., & Ross, K. W. (2009). *Computer networking: a top-down approach* (Vol. 4). Addison Wesley Boston, USA.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*.

Lempert, R. J., & Collins, M. T. (2007). Managing the risk of uncertain threshold responses: comparison of robust, optimum, and precautionary approaches. *Risk analysis*, *27*(4), 1009–1026.

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., & Hsu, M. (2017). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological review*.

Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological review*, *124*(6), 762.

Markowitz, H. (1952). The utility of wealth. *Journal of political Economy*, *60*(2), 151–158.

Natarajan, B. K. (1989). On learning sets and functions. *Machine Learning*, *4*(1), 67–97.

Nobandegani, A. S. (2018). The Minimalist Mind: On Mininality in Learning, Reasoning, Action, & Imagination. McGill University, PhD Dissertation.

Nobandegani, A. S., da Silva Castanheira, K., Otto, A. R., & Shultz, T. R. (2018). Over-representation of extreme events in decision-making: A rational metacognitive account. In: *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 2391-2396). Austin, TX: Cognitive Science Society.

Piantadosi, S. T. (2018). One parameter is always enough. *AIP Advances*, *8*(9), 095118.

Pollard, D. (2012). *Convergence of stochastic processes*. Springer Science & Business Media.

Poor, H. V. (2013). *An Introduction to Signal Detection and Estimation*. Springer Science & Business Media.

Renkl, A., Mandl, H., & Gruber, H. (1996). Inert knowledge: Analyses and remedies. *Educational Psychologist*, *31*(2), 115–121.

Russell, S. J. (1997). Rationality and intelligence. *Artificial Intelligence*, *94*(1-2), 57–77.

Russell, S. J., & Subramanian, D. (1995). Provably bounded-optimal agents. *Journal of Artificial Intelligence Research*, *2*, 575–609.

Scholten, M., & Read, D. (2014). Prospect theory and the forgotten fourfold pattern of risk preferences. *Journal of Risk and Uncertainty*, *48*(1), 67–83.

Simon, H. A. (1957). *Models of Man*. Wiley.

Simon, H. A. (1972). Theories of bounded rationality. *Decision and organization*, *1*(1), 161–176.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, *5*(4), 297–323.

Vapnik, V. N., & Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity* (pp. 11–30). Springer.

von Von Neumann, J., & Morgenstern, O. (1955). *The theory of games and economic behavior*. Princeton University Press.

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.