The Bounded Rational Frontier: Rational Uncertainty Aversion and Implications for AI, Economics and Philosophy

Jonathan Harris
September 28, 2024

jonathan@total-portfolio.org

TL;DR: This note's toy model shows how uncertainty aversion can be rational and maximize expected utility for bounded rational agents. This underappreciated possibility has significant implications for fields including AI, economics and philosophy.

Summary: Bounded rationality - the idea that our decision-making is constrained by limited information and processing power - is often overlooked in the idealized models that dominate various fields. This research note introduces an accessible toy model of bounded rationality that is designed to be widely applicable. It highlights several underexplored insights:

- Rational agents should seek out different, independent sources of information to reduce their uncertainty, and bounded rationality amplifies the benefits of doing this.
- Under certain forms of bounded rationality, risk and ambiguity aversion can actually maximize expected value.
- These insights can all be elegantly visualized using 'frontier curves'.

These insights have important implications for various topics, including: capabilities versus safety trade-offs in AI, moral uncertainty in philosophy, diversification in philanthropy, and robustness in economics. By bridging the gap between idealized models and practical realities, this note aims to encourage more nuanced approaches to complex problems in these fields and beyond.

I welcome feedback and am open to collaborations to further develop these ideas.

Introduction

Consider the following decision problem: You have $1,000 to give to charity. Conventional thinking would recommend giving the whole amount to the 'best' charity. This simplifies the decision. It is optimal when the best charity can be clearly identified. Yet, you find yourself hesitating. You feel almost clueless about how to compare different impacts – from improving global health to reducing animal suffering to mitigating catastrophic risks from AI. You encounter a wide range of conflicting views while doing your research. You also observe that some charities are high-risk bets that could turn out to be hugely important while others seem like much safer bets. This brings up many questions: should you be pessimistic and lean towards low-risk opportunities? Or be optimistic about upside of the high-risk ones? Should you spread your donations across two or more charities?

This scenario illustrates a broader challenge: how can decision-makers, whether human or artificial, make optimal choices when their resources—time, information and computing power—are constrained? While classical models often assume perfect rationality, real-world decision-makers always operate under significant constraints. It can't be rational in an idealistic sense, but it can be 'bounded rational' (e.g., Genewein et al., 2015) or 'resource rational' (e.g., Bhui, Lai, and Gershman (2021); Lieder and Griffiths (2019)) - that is, following the optimal policy given the practical constraints that they face. For simplicity, I will use the term 'bounded rational' to refer to all related frameworks.

Bounded rationality and related frameworks cover a huge space that crosses many fields and has been around for a long time. Simon (1955) is often cited as one of the first relevant writings. Bounded-rational models are generally harder to solve than normal models. Yet, exciting developments in cognitive science and AI are starting to offer tractable models for optimal agent behavior under realistic constraints. This note offers a taste for those unfamiliar with the concept and, hopefully, a novel perspective for experts.

This note presents a simple toy model of bounded rational decision-making to highlight three key concepts that are often overlooked:

1. **Decision-making as inference**. The decision-maker in the model combines multiple imperfect signals to *infer* the true state of their environment, producing better results than if they relied on a single signal.
2. **Rational Uncertainty Aversion**. The more volatile the potential outcomes, the more a bounded-rational decision-maker will tend to underperform compared to ideal conditions. This can make it rational for them to prefer less risky, less ambiguous options and to diversify. Even if it would be rational for them to ignore uncertainty under ideal rationality (e.g., because their utility function is linear so 'risk-neutral').
3. **Frontier Curves**. The results are visualized as curves that show how different strategies trade off expected outcomes against risk and other factors—similar to the 'efficient frontier' used in finance. As the saying goes, 'a picture is worth a thousand words', and these frontier curves are a simple, yet powerful tool to make the results of complex models more transparent.

As AI rapidly advances and global challenges become increasingly complex, understanding bounded rationality and its implications is more crucial than ever. These concepts are important because:

- The mental models we use influence everything from policy discussions to AI system design. 'Inference' models, where it is best to combine multiple points of view, can serve as powerful reminders of the importance of open-mindedness in these times of increasing

polarization and conflict. Yet, in many areas such models aren’t the norm. The dominant paradigm is more for different models to compete with each other to be the one that makes the decision, without any real mixing of models occurring.
- The potential for uncertainty aversion to be rational affects everything from philathropy to AI alignment. Yet, the main versions of it in this note have been rarely, if ever, explored.
    - Rational uncertainty aversion is nuanced. It pushes against both extremes of fanatical bets and vague 'do no harm'/precautionary principle dogmas that themselves cause harm by killing innovation.
- Much of the related literature lacks clear visualizations (like the frontier curves), despite their importance for effective communication.

In AI, philosophy, economics, and other fields, there's a persistent gap between practical results and theoretical understanding of why those practices work. This gap is particularly evident with uncertainty aversion, which is often either presumed or ignored entirely. It is important to investigate when and why uncertainty aversion is a reasonable assumption.

This note is not a final paper but an invitation to researchers from various disciplines to collaborate in exploring these important concepts. The goal of the note is to demonstrate the importance of these concepts in a simplified setting that can be adapted to a range of fields in the future, including philosophy and economics. By doing so, this note aims to spark further exploration and collaboration among researchers in AI, economics, philosophy, and related fields. I am actively seeking collaborators who are interested in building on these ideas to formalize, extend, or adapt them to their areas of expertise.

Structure of this note and of uncertainty aversion

We will see that the toy model can illustrate four different types of uncertainty aversion:

- **Noise aversion**. Where the agent prefers options with less noisy information.
- **Opportunity-level risk aversion**. Where the agent prefers opportunities with less volatile outcomes when choosing between two opportunities.
- **Ambiguity aversion** (or aversion to Knightian uncertainty, epistemic uncertainty, model uncertainty). Where the agent prefers options for which they have greater confidence in their probability model, all else equal.
- **Policy-level risk aversion**. Where the agent prefers policies that produce less volatile results given the available mix of opportunities.

The structure of this note is as follows:

1. **Inference setup**. Introduces the basic setup of the toy model.
    - 1.1 presents the frontier curves for this setup, to show that even with ideal rationality noise aversion is rational.
2. **Bounded rational inference I**. Introduces the popular form of 'information-theoretic' bounded rationality. This increases noise aversion but doesn't lead to the other types of

uncertainty aversion.

3. **Bounded rational inference II**. Introduce another bounded rationality constraint.
    - 3.1-3.3. Show how this generates rational opportunity-level risk aversion, ambiguity aversion, and policy-level risk aversion.
4. **Connections to and implications for different fields**. Readers are encouraged to focus on the subsection most relevant to their field.
5. **Research ideas and questions**.
6. **Conclusion**.

The code, a brief technical explanation, and references are at the end of the note.

1. Inference setup

The setup is designed to align with philosophical and economic thought experiments - for example it is similar to Buchak (2023). A 'donor' agent has decided to donate 1,000 dollars to charity. Their default option is Opportunity A - a benchmark charity that produces 1 unit of impact per dollar.

They have also asked their friend to recommend a risky Opportunity B that produces either $s = 0$ or $s = v > 1$ units of impact per dollar, where v is a 'risk level' that the agent specifies to help narrow down the recommendation. For example, asking for a low v opportunity might mean seeking charities backed by extensive empirical evidence, while asking for a high v opportunity corresponds to wanting to see a new, 'startup' charity.

They choose the fraction a to donate to Opportunity B in increments of 0.1, with $1000 * (1 - a)$ going to Opportunity A. For example, if they choose $a = 0.7$ and B is in state $s = 2$, then the impact per dollar of their donation will be $0.3 * 1 + 0.7 * 2 = 1.7$. Figure 1 shows the impact per dollar for the different state and action pairs for $v = 2$.

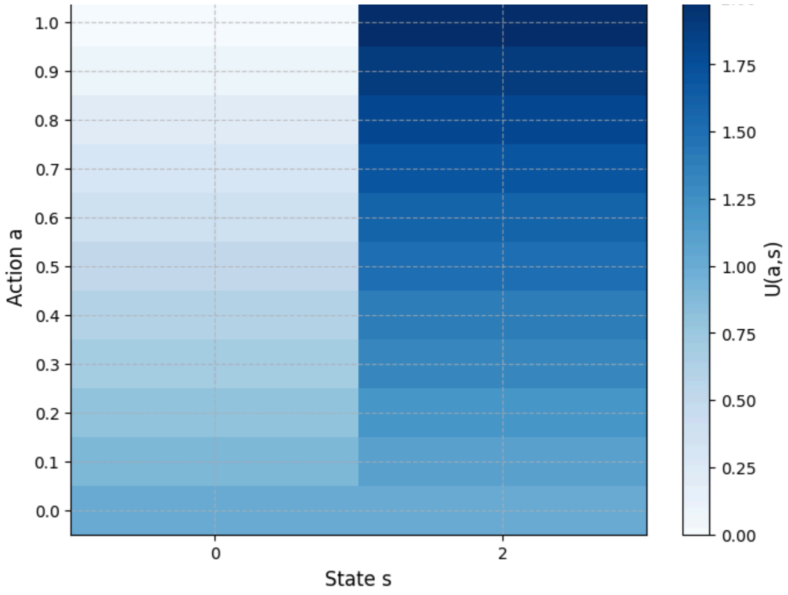

Figure 1: The impact per dollar for each action and state with v=2.

The setup is 'fair' in that, without additional information, the probability of each state is such that both charities have an expected impact equal to 1. This means that higher values of v correspond to lower upside probabilities for Opportunity B.

The goal of the agent is to maximize their impact per dollar. The agent needs to:

1. Choose v.
2. Receive and review analyses of Opportunity B to 'infer' the probability B is in either state.
3. Choose a.

To inform their decision, the agent has access to two analysts who provide different observations about Opportunity B's state:

- One provides a quantitative score (0, 1, or 10)
- The other provides a qualitative rating (x, y, or z)

The agent has a probabilistic model for the relationship between observations and states. They have a prior belief about the likelihood of each state and about the likelihood of each observation given each state. The latter beliefs can be combined into conditional probabilities for the observation pairs given each state as in Figure 1 (left). They can invert this, using Bayes' theorem, to get the probability of each state given each observation pair as in Figure 1 (right). These probabilities can then be used to calculate the expected impact of each action given an observation pair. This approach allows them to combine observations from the analysts even if they aren't easily comparable (i.e., quantitative and qualitative, non-commensurable).

In this note we consider several observation scenarios:

ID	Scenario Name	Description
0	Perfect information	At least one analyst identifies the state with 99+% accuracy
1	Good, independent observations	Both analysts have a 'good' ability (65+%) to identify the state and are otherwise uncorrelated
2	Correlated observations	Same as Scenario 1 but the analysts views are partially correlated (e.g., using similar methods and evidence)
3	Only one good observation	One analyst is good, the other is clueless; or equivalently, both analysts are 100% correlated
4	Relatively clueless	One analyst is clueless, the other offers only a slight edge in identifying the true state

Figure 1 illustrates the agent's model for the observations given the states for the 'Good, independent observations' scenario. Similar charts for other scenarios can be generated using the code provided at the end of this notebook.

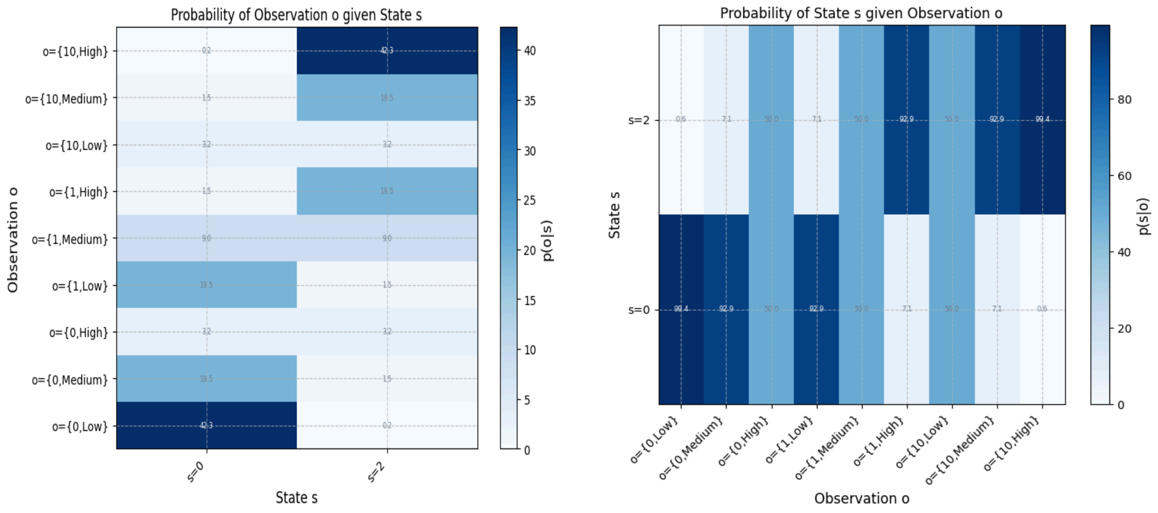

Figure 2: The left chart shows the agent's beliefs about the probabilities of the observations given the states for the 'Good, independent observations' scenario (with v=2). That is, their prior beliefs about these conditional probabilities. Combined with their beliefs about the probability of each state, using Bayes' theorem they can calculate the probability of each state given each observation, as shown in the right chart. The probabilities for each cell are indicated in percentage points.

1.1 'Noise' aversion

A natural way to present the results of bounded rationality models is with frontier curves that show the 'frontier' of the highest possible expected value (for utility, impact or whatever fits the context) given a fixed value for another model statistics or parameter. This note focuses on expected value-volatility frontier curves that compare the expected impact and volatility (square-root of the variance) of the impact a policy is expected to produce.

This is similar to the 'Efficient Frontier' return-volatility curve in finance and the 'Efficient Impact Frontier' return-impact curve in impact investing. These visualizations are a powerful way to understand the implications of different forms of bounded rationality.

Figure 3 presents these frontier curves for the toy model just based on the inference setup above (with no bounded rationality). The curves are produced by allowing the agent's chosen v to vary from 1 to 100. Increasing v naturally increases the volatility of the agent's policy.

The figure illustrates that:

- Better observation scenarios produce better results.
- Having two good analysts is better than one as long as they have independent opinions.
- Expected value is increasing in volatility, so the agent should not be risk averse and should choose v as high as possible.

So, the inference setup means that opportunities with more certain observations will be preferred. That is, the agent should rationally exhibit 'noise' aversion. But, this is just a natural result of being Bayesian (as, for example, in the 'optimizer's curse'). It is just a natural preference for better observations.

In the next sections we will see if adding bounded rationality can cause the agent to become uncertainty averse in their preferences regarding v and a.

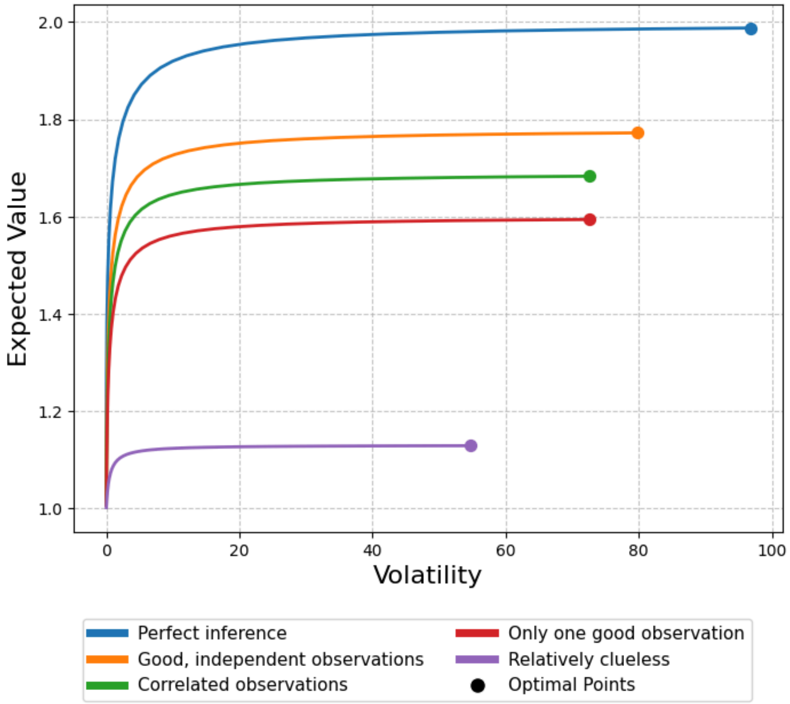

Figure 3: Frontier curves for each observation scenario. Expected Value and Volatility are, respectively, the expected value and the square root of the variance of the impact per dollar donated. This demonstrates 'Observation-uncertainty aversion' as for a given Volatility the agent will always prefer the higher quality observation scenario. However, this does not change that the underlying setup means that for a given observation scenario an expected-value maximizing agent will always prefer higher Volatility.

2. Bounded rational inference I

The agent in our setup needs to choose their action a given the observations o. Generalizing, we can think of their policy as the probability $p(a|o)$ that they take action a given observation o. In our simple, linear toy model, for a given observation o, an ideal rational agent will always have $p(a^*|o) = 1$ for one best action a^* and the rest of the probabilities as zero. A key insight of bounded rationality is that such policies are more 'complex' in the sense of require more computations to implement than randomly choosing an action.

Implementing a policy for the toy model may seem trivial, but in general an ideal policy requires more computational 'bandwidth'. It requires storing the policy in memory, capturing the observations, looking up the corresponding action and executing it. Whereas a random strategy just requires randomly choosing an action and executing it. This has deep connections to concepts like entropy in information theory and physics that we won't get into here.

So, one of the most popular and promising class of models for bounded rationality are 'information theoretic' models that subtract a complexity cost from the agent's utility (see, for example, Genewein et al. (2015) and Lai and Gershman (2024)). Figure 4 shows the optimal policy for different 'prices of complexity'. It is calculated numerically using an iterative algorithm to find the policy that maximizes the expected impact minus the expected complexity cost (see the 'Technical notes' section at the bottom for more). The figure illustrates that when the price of complexity is high, the optimal policy is more random. There is still some randomness even with a low price of complexity because for some observations the expected values of the different actions are very similar.

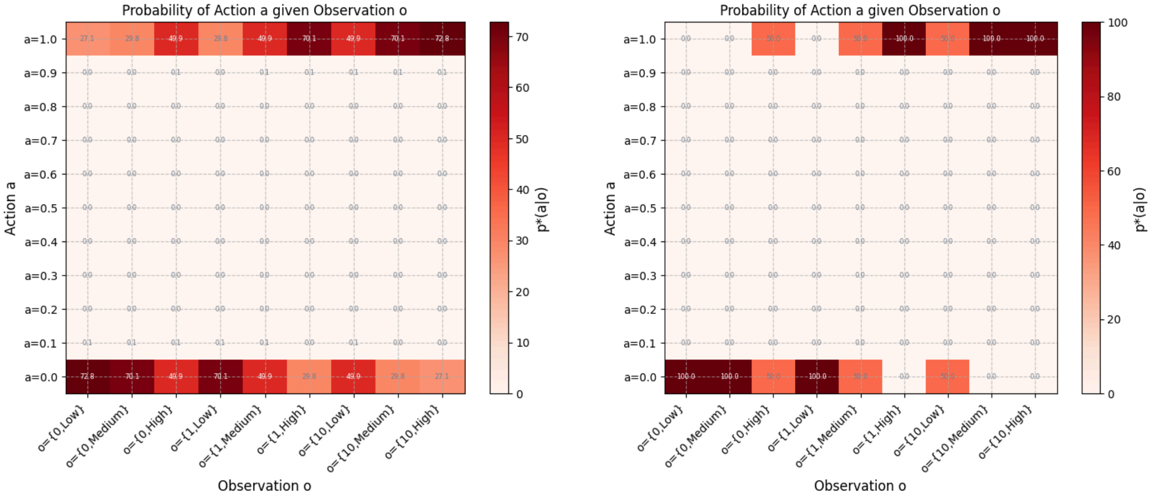

Figure 4: Optimal action probabilities given each observation for the "Good, independent observations" scenario with v=2 with a high price of complexity (left) and low price of complexity (right). The policy given a high price of complexity costs is much more random (i.e., less complex). The policy given a low price of complexity still includes some randomness because for some observations the expected values of the actions are very similar.

2.1 Strengthening 'observation-noise' aversion

Figure 5 shows that introducing information-theoretic bounded rationally in this way just shifts down the frontier curves and that the shift is greater for the weaker observation scenarios. This means this form of bounded rationality amplifies the agent's uncertainty aversion: an agent with a high price of complexity should be willing to pay more to improve their observation scenario (if possible). But, again, to be precise this is still 'noise' aversion not risk or ambiguity aversion. So, in the next section we turn the price of complexity all but off and explore an alternative form of bounded rationality that generates other forms of uncertainty aversion.

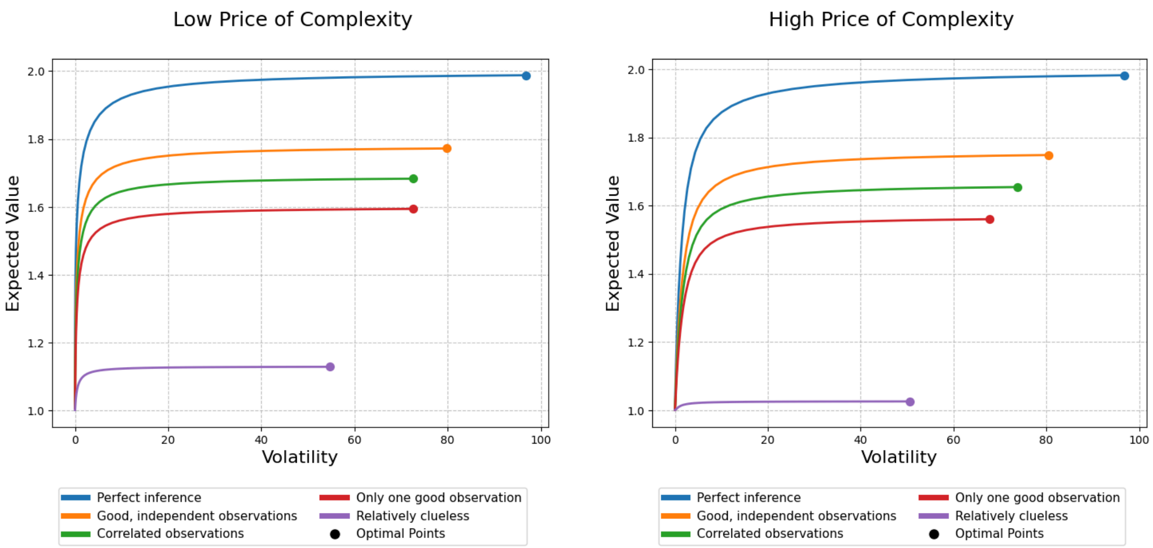

Figure 5: This figure compares the frontier curves from Figure 2 for a low price of complexity cost (left) and the frontier curves with a high price of complexity (right). The shape of the curves is the same. But, the curves for the high-price case are all below the respective ones for the low-price case and much more so for the weaker observation scenarios. This indicates that more bounded rational agents (with a high price of complexity) should be willing to pay more to upgrade their observations.

3. Bounded rational inference II

Modelling bounded rationality can be a bit like playing 'whack-a-mole' because when you once try to constrain one implicit assumption in conventional models you realize there is yet another one to pin down (or they your constraint comes with its own implicit assumptions). An implicit assumption we've overlooked so far is that the agent can compute the expected value of each action without error. Error in such estimates is a practical issue in machine learning, exemplified by systems like DeepMind's Go-playing AI which can only sample from a fraction of the possible games. But this practical point has received surprisingly little attention in the broader bounded rationality literature. By incorporating estimation error into bounded rationality models, we can better reflect the realities faced by both artificial and human decision-makers.

We can add this issue to the toy model by supposing that the agent can't directly compute the expected impact of each action. Instead, they must estimate it by randomly sampling the possible outcomes and taking an average across the samples as in Lieder, Hsu and Griffiths (2014). This may not seem to be an issue in the context of our toy model, but it would be reality in more complex scenarios with many possible states and actions.

Policies that are optimized based on noisy estimates will be strictly worse than the true optimal policy (i.e., optimal policies like in Figure 4). Thus, even symmetric errors in the estimation don't cancel out when it comes to the expected value of the policies derived given those errors. This results in a performance gap between the expected value of the true optimal policy and the average policy with estimation error. If the performance gap increases significantly in riskier scenarios, then low-risk scenarios could offer higher expected value in practice even if higher-risk scenarios would offer much higher expected value under ideal conditions.

To test the results we restrict the agent to 2, 8, 32 or 128 samples for their estimates. The number of samples can be thought of as a proxy for how well the agent is able to approximate the optimal policy (i.e., as in Figure 4). Or, in other words, as a proxy for how much time you have to spend on something and how useful spending more time on that problem is likely to be.

The following subsections illustrate how these sampling constraints can generate different forms of uncertainty aversion. For more detail, see the brief technical explanation at the bottom of this note. Note that in line with this all being a game of 'whack-a-mole' these results are just a claim about this specific toy model and what happens with the assumptions we have made. It could be that there is a smart way for the agent to get around the constraints and assumptions we've put in place here. And then the assumptions about how much of that smartness makes sense to allow would have to be examined. And so on.

3.1 'Opportunity-level' risk aversion

Figure 4 presents frontier curves generated by varying v from 1 to 100. With perfect information or a large number of samples the agent should prefer higher volatility (i.e., $v = 100$) as they can achieve expected impact of almost 2. But, if restricted to only 2 samples the performance in the non-perfect scenarios (1 to 4) degrades significantly for higher volatilites - so much so that it is optimal for the agent to prefer the lower risk levels generated by v around 5 to 15. For most of the scenarios the optimal risk level increases significantly as the number of samples increases. However, it remains quite low in all cases for the 'Relatively clueless' scenario.

In other words, Figure 4 illustrates a whole range of behaviors. If you are in a situation where you are confident in your model and have decent information (red or better) then you should be looking for more risky opportunities. But, if you have a lot of model uncertainty or you feel 'clueless' then you should stick to relatively safe opportunities.

These results illustrate that it is possible for this sampling constraint to make it optimal for the agent to prefer lower values of v, despite higher values of v offering more upside under ideal conditions. Even though the estimation errors are symmetric, they result in asymmetric risk preferences, favoring lower-risk options. This is a particular form of risk aversion - we might call it 'opportunity-level' risk aversion. The next two subsections explore other features that result from the sampling restriction.

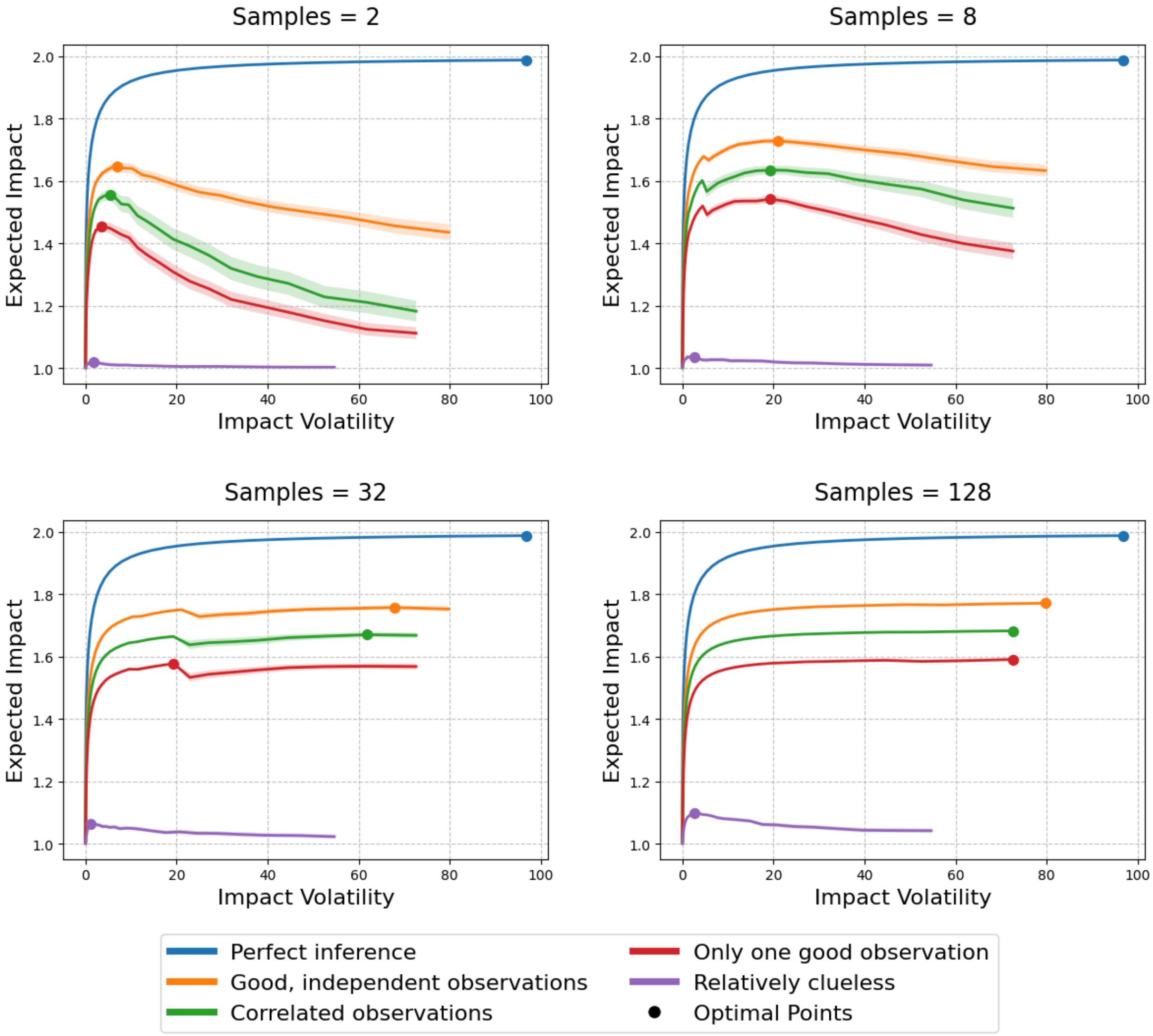

Figure 6: Expected impact-risk level frontier curves for each observation scenario and number of samples. The Expected Impact values are averages across many simulations with two standard deviations of sampling error represented by the shaded areas.

3.2 'Ambiguity' aversion

Ambiguity (or model uncertainty) is often presented as cases where the probabilities themselves are uncertain. This isn't explicitly a part of the toy model. However, the sample restrictions can be viewed as producing a similar, arguably equivalent, effect.

Figure 5 reorganizes the results of Figure 4 to make it easy to confirm that as the number of samples increases the Expected Value increases, for all values of v. So, it is rational for the agent to prefer lower 'ambiguity' (higher sample) situations. Thus, the toy model generates both rational opportunity-level risk aversion (preference for smaller v) and rational ambiguity aversion (preference for situations that are equivalent to having more samples).

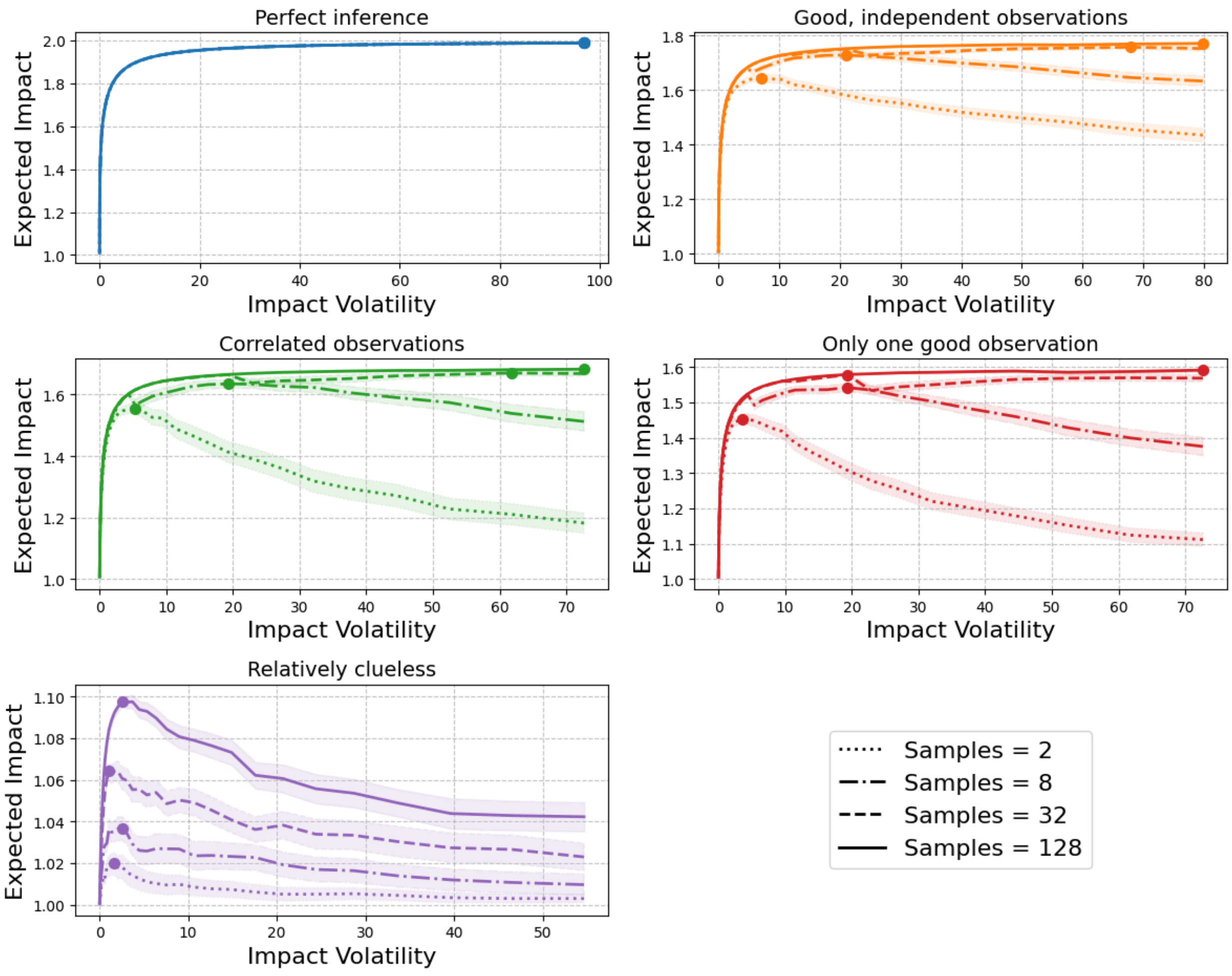

Figure 7: The frontier curves from Figure 4 grouped by observation scenario. This demonstrates 'Ambiguity aversion' as for a given observation scenario and Volatility, the agent always prefers the higher sample case. This is different from risk aversion as moving to the higher sample case can increase the optimal Volatility level.

3.3 'Policy-level' risk aversion and diversification

The model above generates uncertainty aversion but not diversification. Generally, the agent's samples will either confirm to them that giving everything to A or B is the best policy. Diversification in this toy model will take the form of 'fractional actions', $0 < a < 0.1$. In general, these actions may be underexplored in other models because they aren't a thing in many prominent domains (e.g. you have to label a picture with a single label, not some mix).

Note that the strategy we've allowed the agent is admittedly a bit naive. With a bit more thought they might choose to preemptively include risk aversion in their model so they underweight actions with more volatile impacts. This should steer them towards determining their policy based on actions that they have better estimates and mitigate some of the effect of their sampling constraint.

For risk aversion like this to have a chance of being optimal, we need to consider situations where the choice between A and B isn't so obvious given good observations. When B is either $s = v = 100$ or $s = 0$, it's almost always going to be clear that is it better or worse than B. We need to consider situations that are both close and uncertain, so we'll consider $v = 1.1$ and allow the agent 2 samples.

The top chart in Figure 6 shows the frontier curves when the agent's policy is determined by a biased model that subtracts, for each action given each observation, the condtional variance of the impact multiplied by a 'price of risk'. The results demonstrate that in this case using a non-zero price of risk is rational in that it improves the expected impact. This relates to the classic idea that biased solutions can have lower MSE. But the framing of this toy model offers a fundamental reason for the standard assumption that lower MSE is better: with bounded rationality the policies with lower MSE can also be the policies with higher expected value.

The bottom chart in Figure 6 shows for each of the expected impact maximizing points the percentage of time a 'mixed' action of 0.1 to 0.9 is taken. This shows that rational risk aversion results in diversification even in this binary situation with no inherent risk aversion in the agent's utility function.

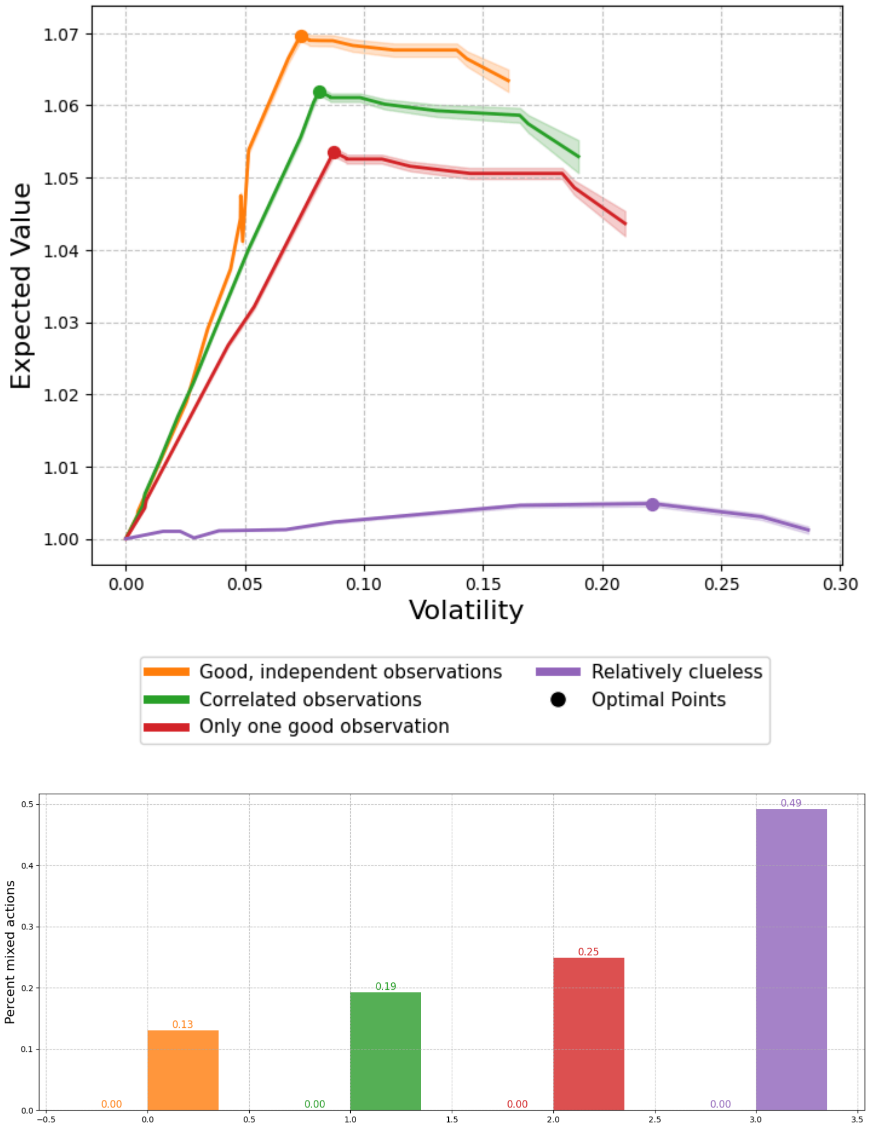

Figure 8: The top chart shows the frontier curves for each observation scenario with v=1.1 and Samples=2. The curves are generated by increasing the agent's 'price of risk'. The Expected Impact values are averages across many simulations with two standard deviations of sampling error represented by the shaded areas. The bottom chart compares the percentage of time a 'mixed' action, 0 < a < 1, is taken with zero price of risk (left bars) and with the optimal level of the price of risk (right bars).

4. Connections to and implications for different fields

This section offers a selection of implications the toy model, and bounded rationality in general, have for different fields. Readers are encouraged to focus on the subsection most relevant to their field.

General

- **Bottom-up paradigm**: Research across related fields is dominated by a top-down paradigm of adding features to models based on heuristic reasoning and seeing what works to improve performance. This toy model, and 'bounded rationality' in general, suggests an alternative 'bottom-up' paradigm of first focusing on better defining the optimization problem including all its constraints. Solving the properly defined problem then naturally leads to optimal performance. In other words, better theory can lead to better models, and the better theory that may be needed is a better description of the practical constraints faced by decision-making agents.

- **Uniting risk and ambiguity aversion**: In the model the different forms of uncertainty aversion all come from the same source of estimation error. This highlights that the distinction between risk and ambiguity aversion may not be as clear as is often assumed in economics and philosophy.

- **Frontier curves**: The note highlights the value of visual tools like frontier curves for analyzing optimal policies.

- **Two types of problems**: This model highlights a fundamental distinction between problems where uncertainty can be reduced through more data and compute, and those with irreducible uncertainty. In data-driven contexts, the appropriate level of 'uncertainty aversion' can be empirically calibrated (e.g., via cross-validation). However, for problems with irreducible uncertainty - perhaps like many in moral philosophy or long-term policy - such data-driven automation is not possible, and a careful, uncertainty-averse approach may be the only viable strategy.

- **Bridging theory and practice**: This model highlights a significant gap between theoretical frameworks and practical decision-making across various fields. It demonstrates how theory can be improved by incorporating realistic constraints like bounded rationality, potentially leading to more nuanced and applicable theoretical models. Improved theories, in turn, can enhance practice by providing more accurate guidance for real-world decision-making under uncertainty.

Philanthropy and impact investing

- **Open-mindedness**: The inference framing emphasizes the potential benefits of open-minded consideration of different perspectives, especially in challenging, complex decisions like those to do with impact on society.

- **Preference for safer bets**: The toy model doesn't rule out a 'hits-based' approach of making multiple high-risk, high-reward bets. But the results do point out that in situations of greater uncertainty it can be rational to stick to safer bets.

- **A fundamental reason for diversification**: The model highlights estimation error as a fundamental reason for diversification, whether across charities, investments, strategies, causes, or time. This aligns with the fact that many philanthropic organizations diversify and goes beyond traditional explanations like diminishing-returns to scale, reputation management, and combinatorial effects. The issue of estimation error applies to donors and investors of all sizes and it applies most strongly to more complex and controversial topics like how much to allocate to philanthropy versus impact investing.

AI alignment

- **Nuanced AGI behavior and alignment strategies**: While not diminishing the potential for catastrophic risks from AI, this model suggests that advanced AI systems might develop uncertainty aversion and open-mindedness as rational responses to complex environments. This challenges simplistic views of AI as pure expected utility maximizers and could inform more nuanced AI alignment strategies.

- **Effect of search limitations**: The sampling constraint form of bounded rationality relates to 'search'/'inference compute'/'test-time compute' in AI. For example, ChatGPT o1 uses additional compute at inference time to explore different possible responses, analogous to game-playing AIs exploring moves, or the agent in the toy model sampling actions to estimate expected values. This highlights a critical question: are there fundamental problems that increased inference compute cannot solve due to some kind of irreducible uncertainty? This could have significant implications, suggesting some challenges may remain intractable even with vast computational resources (at inference time).

- **Capability-safety alignment**: If it's optimal for AIs facing complex, highly uncertain situations to consider multiple perspectives and exhibit uncertainty aversion, this could mean that training AIs to acknowledge their own bounded rationality could be a win-win for capabilities and safety. This insight could be particularly relevant for AI systems designed to operate in real-world environments with irreducible complexity, such as self-driving cars navigating through unpredictable, heavy-tailed environments.

- **Generalized alignment**: In situations so complex that both humans and AIs are more or less 'clueless', it may be more productive to focus on improving global coordination and robustness in general, rather than solely on AI alignment. This could involve developing coordination mechanisms (including AI training algorithms) that are explicitly aligned with bounded rational inference. Such an approach acknowledges that unlimited computation doesn't guarantee unbounded rationality, as logical inconsistencies often arise from the nature of problems themselves, not just from a lack of resources.

- **Implications for AI safety research**: This model highlights the importance of considering bounded rationality in AI safety research. It suggests that safe and capable AI systems might need to be designed with the ability to handle uncertainty, combine multiple perspectives, and make robust decisions under constraints. This could lead to new research directions in AI safety, focusing on how to implement these bounded rationality principles in AI architectures and training processes.

Moral philosophy

- **Practical ethics**: Bounded rationality challenges moral theories that assume unlimited reasoning capabilities, particularly pure utilitarianism. It emphasizes the importance of practical ethics - balancing ideal outcomes with realistic human behavior. It suggests that simple, clear recommendations (e.g., donating a fixed percentage of income) may be optimal, in alignment with, for example, Peter Singer's work on making ethics accessible and actionable as in Singer (2009).

Furthermore, practical ethics may be more aligned with uncertainty aversion than with pure utility calculations. For instance, a pure utilitarian might argue: "Eating meat is wrong because it net destroys utils." Whereas an uncertainty-averse framing might be: "What if eating meat is really bad? Maybe you can reduce your moral risk by being reducetarian." This uncertainty-averse framing may align more closely with how people actually make moral decisions and could lead to more widespread adoption of more ethical practices.

- **Bridging normative and descriptive ethics**: Bounded rationality models offer a way to bridge the gap between normative ethical theories (what we ought to do) and descriptive accounts of moral decision-making (what people actually do). By showing how uncertainty aversion can be rational under realistic constraints, it provides a framework for understanding why people's moral intuitions often diverge from pure utilitarian calculations, without necessarily invalidating the underlying utilitarian concerns.
- **Morality as inference**: The inference setup suggests framing existing moral theories as imperfect inferences of an ideal theory, assuming a common ground that different theories attempt to capture. This perspective allows for combining theories to better approximate the ideal, even when they seem incommensurable. This approach offers a new angle on moral uncertainty, potentially informing frameworks like 'variance voting' and the 'moral parliament' (MacAskill, Bykvist and Ord, 2020). For example, it suggests that the credences assigned to different moral theories in these existing approaches could be informed by Bayesian calculations based on rough models for how each theory approximates the ideal theory.
- **Inseparability of empirical and moral uncertainty**: The model highlights that adjustments for empirical uncertainty cannot necessarily be treated separately from moral uncertainty, as the optimal strategy depends on both the agent's observation scenario and their model uncertainty (as proxied by the number of samples) and this dependence can be complex.
- **Fanaticism vs. caution**: The model endogenously generates that in situations of cluelessness safe bets are preferred over high-risk, 'fanatical' bets, while the opposite can be true in situations that are closer to ideal.
- **Rational uncertainty aversion**: The model suggests that some degree of risk aversion in moral decision-making can be rational, providing a new perspective on issues like cause-area diversification in philanthropy. This could complement existing arguments for moral

hedging and provide a formal basis for intuitions about moral uncertainty (MacAskill, Bykvist and Ord, 2020).

- **Computational moral philosophy**: This model demonstrates the potential for integrating insights from neuroscience and machine learning into philosophically relevant thought experiments. It offers a mathematical framework for moral uncertainty that complements existing abstract work (e.g., on 'global consequentialism') and challenges idealized moral theories by incorporating realistic cognitive constraints.

Techno-philosophies

- **Alignment with d/acc (defensive accelerationism)**: The results of the model are perhaps aligned with Vitalik Buterin's d/acc over the extreme optimism of 'e/acc' and the extreme pessimism of 'doomers'. This is because it highlights that humility and risk aversion can be optimal in the most complex, uncertain situations (like advancing global progress), though it fully supports aggressive, high-risk bets in certain situations.

Economics, statistics and machine learning

- **Endogenous robustness**: This model demonstrates how uncertainty aversion can arise endogenously and rationally, contrasting with approaches that treat robustness as an intrinsic good.
- **Explanation for evil-agent framing**: A popular way to frame robustness is as a two-player game between the agent and the environment, where the environment has the ability to distort the relevant probability distribution and it has the 'evil' goal of minimizing the agent's utility subject to constraints on how much distortion it can cause. The present toy model highlights that asymmetry of this evil-agent framing (why should the environment be anything but neutral?) makes sense because any deviation in the agent's information will cause sub-optimal performance. This connection is not apparent in most of the literature on this topic as conventionally model estimation is separated from the agent choosing their action policy.
- **Optimal robustness**: By potentially inferring the estimation abilities of different actors, this approach might help define optimal levels of robustness rather than leaving it as a subjective hyperparameter.
- **Adversarial framing**: The model challenges the asymmetric presumption of adversariality in some robustness models, instead suggesting a mechanism to generate this asymmetry from symmetric estimation errors.
- **Exploration and uncertainty aversion**: This framework shows how agents can be both uncertainty averse and explorative, reconciling seemingly contradictory aspects of decision-making under uncertainty.
- **Non-maximum likelihood estimation**: If estimating a model (i.e. for $p(s|o)$) to support utility maximization under bounded rationality it is possible, likely, that the best possible model will depend on both the data and U - not just the data. That is, the model should be fit to produce the best results in practice, not to fit the data the best.

5. Research ideas and questions

This model highlights several promising directions for future research, including:

- Extending the model to address its limitations and to capture more complex, multi-period decision scenarios, and applying it to classic problems in decision theory and economics.
- Investigating how different forms of bounded rationality interact, particularly in multi-agent competitive or cooperative settings.
- Exploring the moral philosophical implications regarding topics like cluelessness, open-mindedness and fanaticism.
- Developing a unified framework that reconciles different types of uncertainty (e.g., risk, ambiguity, Knightian uncertainty) within the context of bounded rationality and computational constraints.
- Developing practical tools and heuristics based on these insights for real-world decision-making in fields such as finance, policy-making, and AI development.

6. Conclusion

This research note introduces a simple toy model that highlights the important implications of bounded rational inference, including that:

- Combining multiple, independent models in an inference setting leads to superior outcomes.
- Both risk and ambiguity aversion can be expected-value maximizing under some forms of bounded rationality.
- Frontier curves provide a powerful way to visualize complex decision-making trade-offs.

These insights challenge conventional thinking across multiple disciplines. This includes offering a new perspective on topics like diversification in philanthropy and impact investing, capabilities versus safety trade-offs for AGI, moral uncertainty in philosophy, and robustness in economic models.

I encourage researchers from AI, economics, philosophy, and beyond to explore these ideas further. If you find models like this intriguing and see potential relevance to your field, I'd love to hear from you. Please reach out, whether you're interested in integrating these ideas into your work or developing a paper or applied project together.

As we face increasingly complex global challenges, further study of models like this can help develop a more nuanced understanding of rational decision-making – one that embraces our limitations while striving for optimal outcomes. Future research should focus on refining these models and developing practical applications to enhance decision-making strategies in an uncertain world.

Code

This section presents the code for this example. It was developed by adapting code from Genewein et al. (2015). The full codebase, including this notebook and all supporting files, is available on GitHub here.

In [ ]:

```
#Import necessary functions
import Modules.Rational_Frontiers_imports
import importlib
importlib.reload(Modules.Rational_Frontiers_imports)
from Modules.Rational_Frontiers_imports import *
```

In [ ]:

```
# Generate results for different complexity costs
gamma_ao_values = np.logspace(np.log10(.0001), np.log10(100), num=50)  # Define th
results_b, scenarios_b = generate_results(da=0.1,v_values=[2],gamma_ao_values=gamm
```

In [ ]:

```
# Plot probability models for a specific scenario
plot_scenario_prob_dist(scenarios_b,scenario_id=1,gamma_ao=1)
```

In [ ]:

```
# Plot Utility (impact) matrix for a specific scenario
scenario = scenarios_b[1]
visualize_matrix(scenario['U_mat'], scenario['s_values'], scenario['a_values'], xl
```

In [ ]:

```
# Generate results for different risk levels and sample sizes
v_values = np.logspace(np.log10(1.01), np.log10(100), num=30) # Define the range f
results_v0, scenarios_v0 = generate_results(da=0.1,v_values=v_values,gamma_ao_valu
results_v0b, scenarios_v0b = generate_results(da=0.1,v_values=v_values,gamma_ao_va
```

In [ ]:

```
# Generate results for different risk levels and sample sizes
v_values = np.logspace(np.log10(1.01), np.log10(100), num=30) # Define the range f
results_v, scenarios_v = generate_results(da=0.1,v_values=v_values,n_samp_values=[
```

In [ ]:

```
# Generate results for different levels of explicit risk aversion
g_values = np.concatenate(([0], np.logspace(np.log10(0.0001), np.log10(1), num=30)
results_g, scenarios_g=generate_results(da=0.1,v_values=[1.1],n_samp_values=[2],ga
```

In [ ]:

```
# # Save res_v and res_g
# with open('Results/results_v.pkl', 'wb') as f:
#     pickle.dump({'results_v': results_v, 'scenarios_v': scenarios_v}, f)
# with open('Results/results_g.pkl', 'wb') as f:
#     pickle.dump({'results_g': results_g, 'scenarios_g': scenarios_g}, f)

# Load res_v from the pickle file
with open('Results/results_v.pkl', 'rb') as f:
    loaded_data = pickle.load(f)
results_v = loaded_data['results_v']; scenarios_v = loaded_data['scenarios_v']
# Load res_g from the pickle file
with open('Results/results_g.pkl', 'rb') as f:
    loaded_data = pickle.load(f)
results_g = loaded_data['results_g']; scenarios_g = loaded_data['scenarios_g']
```

In [ ]:

```
# Plot frontier curves for risk level vs expected utility
plot_frontiers(results_v0,scenarios_v0,x_field='Var_U',y_field='E_U',x_label='Vola
plot_frontiers(results_v0b,scenarios_v0b,x_field='Var_U',y_field='E_U',x_label='Vo
```

In [ ]:

```
# Plot results for different risk levels and sample sizes
plot_results_v(results_v,scenarios_v)
plot_results_v_byscenario(results_v,scenarios_v)
```

In [ ]:

```
# Plot frontier curves for impact volatility vs expected utility with explicit ris
plot_frontiers(results_g,scenarios_g,x_field='Vol_U_sample',y_field='E_U_sample',x
bar_by_scenario(results_g)
```

Technical Notes

Consider a decision-making as inference context where an agent uses observations o to infer the world state s and decide a policy $p(a|o)$ for actions a that will maximize their expected utility:

$$\underset{p(a|o)}{\arg\max}\ \mathbf{E}_{p(s,o,a)}[U(s,a)]$$

where in the ideal case the only constraint is that $p(a|o)$ be a valid probability distribution.

Information-theoretic frameworks like 'bounded rationality' (Genewein et al., 2015) and 'policy compression' (Lai, Gershman, 2024) add a term to account for the computational cost of converting observations into actions:

$$\underset{p(a|o)}{\arg\max}\ \mathbf{E}_{p(s,o,a)}[U(s,a)] - \gamma_{ao} I(O;A),$$

where $I(O;A)$ is the mutual information between the distribution of observations and actions and γ_{ao} is the 'price of complexity' (at least in this note). In general they may also add other terms, including $I(O;S)$.

The solution is given by:

$$\begin{aligned} p^*(a|o) &= \frac{1}{Z(o)} p(a) \exp\left(\frac{1}{\gamma_{ao}} E[U(s,a)|a,o]\right) && (1) \\ E[U(s,a)|a,o] &= \sum_s p(s|o) U(s,a) && (2) \\ p(a) &= \sum_{s,o} p(s) p(o|s) p^*(a|o), && (3) \end{aligned}$$

where $Z(o)$ denotes the corresponding normalization constant. In most cases it is not possible to solve these equations analytically. But, generally the solution can be found numerically via the

Blahut-Arimoto algorithm.

However, the above solution assumes that the agent can accurately calculate $E[U(s,a)|a,o]$. If this expectation is complex and costly to calculate then the agent will only have a noisy estimate. Even unbiased noise inside the exponential term in the solution can result in suboptimal policies. The simple model in this note assumes the agent makes a noisy estimate of $\hat{E}[U(s,a)|a,o]$ and then repeatedly uses this during the Blahut-Arimoto algorithm. This introduces a fundamental asymmetry: even when the noise in the estimates is symmetric, it results in asymmetric risk preferences. This is because any deviation from the true optimal policy, whether positive or negative, leads to suboptimal performance.

The model with explicit risk aversion is a modification to:

$$\arg\max_{p(a|o)} \mathbf{E}_{p(s,o,a)}[U(s,a)] - \gamma_{ao} I(O;A) - \gamma \mathbf{E}_{p(s,o,a)}[Var[U(s,a)|a,o]],$$

with solution,

$$p^*(a|o) = \frac{1}{Z(o)} p(a) \exp\left(\frac{1}{\gamma_{ao}}(\hat{E}[U(s,a)|a,o] - \gamma Var[U(s,a)|a,o])\right) \tag{4}$$

where γ is the risk aversion parameter and $Var[U(s,a)|a,o]$ is the variance of U conditional on a and o. This assumes that the agent is endowed with estimates of $Var[U(s,a)|a,o]$. This may seem unlikely as the key idea here is that the agent struggles to estimate $E[U(s,a)|a,o]$. But, it is plausible the agent has some intuitive idea of the level of $Var[U(s,a)|a,o]$. Future research could make this model more realistic.

References

Bhui R., Lai L., and Gershman S. J. (2021) Resource-rational decision making. Current Opinion in Behavioral Sciences, Volume 41.

Buchak, L. (2023) How Should Risk and Ambiguity Affect Our Charitable Giving? Utilitas 35(3):175-197.

Genewein, T., Leibfried, F., Grau-Moya, J. and Braun, D. A. (2015) Bounded Rationality, Abstraction, and Hierarchical Decision-Making: An Information-Theoretic Optimality Principle. Front. Robot. AI 2:27.

Lai, L., Gershman, S. J. (2024) Human decision making balances reward maximization and policy compression. PLOS Computational Biology 20(4).

Lieder, F., Griffiths, T. L. (2019) Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. Behav Brain Sci.

Lieder, F., Hsu, M., and Griffiths, T. L. (2014). The high availability of extreme events serves resource-rational decision-making. Proceedings of the Annual Meeting of the Cognitive Science Society, 36.

MacAskill, W., Bykvist, K., and Ord, T. (2020). Moral uncertainty. Oxford University Press.

Simon, H. A. (1955) A Behavioral Model of Rational Choice. Quarterly Journal of Economics, 69(1): 99–118.

Singer, P. (2009) The Life You Can Save: Acting Now to End World Poverty. Random House.