THE HARSANYI-RAWLS DEBATE: POLITICAL PHILOSOPHY AS DECISION THEORY UNDER UNCERTAINTY

RAMIRO ÁVILA PERES
https://orcid.org/0000-0002-9376-2105
Banco Central do Brasil
Departamento de Supervisão de Cooperativas e Instituições não Bancárias
Porto Alegre, R.S.
Brazil
ramiro.peres@bcb.gov.br

Article info
CDD: Aguardando
Received: 30.11.2020; Accepted: 26.05.2021
https://doi.org/10.1590/0100-6045.2021.V44N2.RP

Keywords
Uncertainty
Political Philosophy
Difference Principle
Utilitarianism

Abstract: Social decisions are often made under great uncertainty – in situations where political principles, and even standard subjective expected utility, do not apply smoothly. In the first section, we argue that the core of this problem lies in decision theory itself – it is about how to act when we do not have an adequate representation of the context of the action and of its possible consequences. Thus, we distinguish two criteria to complement decision theory under ignorance – Laplace's principle of insufficient reason and Wald's maximin criterion. After that, we apply this analysis to political philosophy, by contrasting Harsanyi's and Rawls's theories of justice, respectively based on Laplace's principle of insufficient reason and Wald's *maximin* rule – and we end up highlighting the virtues of Rawls's principle on practical grounds (it is intuitively attractive because of its computational simplicity, so providing a salient point for convergence) – and connect this argument to our moral intuitions

and social norms requiring prudence in the case of decisions made for the sake of others.

Introduction[1]

How should we act in social contexts of great uncertainty – when we find it hard to apply our standard political principles and face some sort of decision paralysis? Since an action aims to an end, the decision to act is irrational if we cannot justifiably believe that we can achieve the end – or it is self-defeating, if acting in accordance with the decision prevents us from reaching that end.

The decision theory called *subjective expected utility* solves this problem by assigning subjective probabilities to possible scenarios, measuring the agent's uncertainty about them. However, one can now reply with a new problem: how to assign these probabilities and how to define such an evaluation, especially when it includes different individuals' preferences and judgments – computing all possible outcomes and all possible values for all possible actions is at least intractable. After all, a rational agent would need to have an idea of the *risks* and the *opportunity costs* of each alternative in comparison to others. If this is a common problem for individual decision-making, it is pervasive for social decisions. Too often, we risk falling into "decision

[1] I would like to express my gratitude towards people who read previous versions of the manuscript, particularly: (a) my colleagues in the Research Network of the Central Bank of Brazil; (b) former colleagues from UFRGS' Graduate Program in Philosophy, especially M. K. Oliveira and Paulo MacDonald, who generously shared with me their knowledge on contractualist theories; (c) finally, the anonymous reviewers, whose critiques and suggestions greatly improved the text.

paralysis" due to *cluelessness* – situations where we cannot assess the value, nor the long-term consequences of our actions (Greaves, 2016).

The subfield of decision theory that deals with scenarios where there is no probability distribution over possible outcomes is called *decision under ignorance*; there are four different criteria to complement decision theory under ignorance: Laplace's principle of insufficient reason (a.k.a. "principle of indifference"), Wald's maximin criterion, Savage's minimax regret and the Hurwicz's criterion. In the first half of this paper, we will on decision theory and: a) provide an introduction to a canonical model of decision theory, the subjective expected utility theory, and also to the common obstacles this model faces concerning Knightian uncertainty; b) though we do not equate decision under ignorance to Knightian uncertainty (actually, our intent is to highlight their differences), we explain how Laplace's principle and Wald's maximin aim to overcome them; c) we argue that detaching the notion of *risks* from subjective probabilities is not a solution to the problem posed by uncertainty – we criticize Pritchard (2015) as an example of this failed proposal. In the last half of the text, we extrapolate this discussion to political philosophy: first, we aim to make clear that this problem is not restricted to consequentialist theories; second, we present and contrast two competing conceptions of theories of justice with contractualist grounds – i.e., Harsanyi's utilitarianism and Rawls's difference principle[2]. We show that the former uses

[2] It must be noticed that this is not an exegetical exercise: we don't claim that Rawls should be interpreted as endorsing our position; he was more concerned with the basic structure of an ordered society than with applying decision theory to real-world political problems. However, we do assume the reasoning that parties perform in the original position reflect the decision procedure and

Laplace's principle of indifference to cope with the uncertainty of the contratualist thought experiment, while the latter uses a version of Wald's *maximin* rule – leading to the much-debated difference principle. We highlight how framing the original position as a social contract favors the difference principle on the grounds that it better incentivizes *ex post* stable cooperation, and that, thanks to its simplicity, it works as a salient point[3]; this is consistent with common social practices regarding decisions made for the sake of others, such as norms requiring decision-makers to be prudent, which display high uncertainty-aversion.

1. Decision theory, risk and uncertainty

Our classic notion of risk is derived from *lotteries*, *chances* or frequentist probabilities: we say a fair coin flip has a 50% risk of resulting in tails. Since in real life we usually do not have immediate access to frequencies, the standard conception used by risk analysts is based on *bets*, *odds* and subjective probabilities – according to Savage's (1972) subjective expected utility theory. This theory aims to explain

---

the arguments here exposed; our focus is on the decision-making procedure under uncertainty that leads to principles of justice.

[3] We thank an anonymous reviewer for highlighting that this could be read as a Humean argument, which could be accepted by utilitarians (as Harsanyi himself briefly touches on something like it), in favor of Rawls's conception of justice as fairness. On the other hand, as we sketch in our last section and in our conclusion, this is a very limited defense of this account, as our aim is to provide an explanation of why our intuitions as rational agents, in some situations with potential for conflict and cooperation (including the Original Position), bend towards maximin-like principles.

how agents make predictions and forecasts (or how they *would consistently* make them, if not for mistakes) in relation to how they value states of affairs – so that the relative probabilities assigned to states s_1 and s_2 are connected to an agent's willingness to accept a wager on them. This notion applies to *decision problems* characterized by:

> A) an agent (the "basic decision unit" within a model representing the decision problem);
> B) a set of mutually exclusive actions that one can take;
> C) a set of possible outcomes (i.e., states of affairs, or possible worlds, following the action).

The decision involves a procedure that defines how the agent maps a probability distribution over possible outcomes, so representing their uncertainty: a) begins with a *prior* probability distribution that assigns a corresponding value to each possible world; b) as they receive new information, the agent updates their probability distribution by *conditionalization* (e.g., by using Bayes's theorem) - resulting in a *posterior* probability distribution (Bostrom, 2014, p. 10). Furthermore, the agent needs a preference ordering, which defines how they evaluate states of affairs; thus, a utility function is attributed to them - a *complete*, *continuous* and *transitive order* of preferences that associates a number called *utility* with each state of affairs, representing its corresponding 'desirability' (as proposed by the theory of *revealed preference*). When making a decision, the agent selects the action with the greatest expected utility; for this, they must compute: a) for each possible world M and each action A, the corresponding *posterior* probability p to reach the possible world M conditioned to the choice of action A; b) for each possible world M, the product (p x u) of that

probability *p* and the corresponding utility *u*; c) for action, the *sum* of each of these products (Bostrom, 2011, p. 11).

However, the "rational agent model" underlying subjective expected utility theory demands a probability distribution across all possible states – which (due to combinatorial explosion) is computationally unfeasible in real life, and would also imply that an agent cannot have "unknown unknowns[4]." Even so, we must highlight that "putting a number" in a judgement allows one to strive for calibrated *accuracy* in representing one's own uncertainty. Besides, probability calculus allows us aggregate information from different agents – as defenders of prediction markets (such as Sunstein, 2006) argue.

1.1 Decision theory under ignorance and Knightian uncertainty

Uncertainty is a vague and ambiguous term; we can see it as opposed to the concept of information itself. In general, it is used in a broad sense applied to decision situations where there is no complete and perfect knowledge. It is common to adopt Knight's distinction (1921, p. 46) between *risk* (negative events with accurately measurable probability - i.e., the "known risk") and *uncertainty*[5] ("unknown or

[4] In its underlying epistemic logical calculus, this would equate to something like $\sim K(\sim Kp)$ ("the agent does not know that she does not know that p"), which (assuming knowledge is *reflexive*: $Kp \rightarrow KKp$) would imply Kp – i.e., "the agent knows p" (Binmore, 2009).

[5] The term "risk-aversion" is also ambiguous:

- in game theory, it is equivalent to the *concavity* of a utility-function (Binmore, 2009, p. 50); i.e., if I am indifferent between 1 chocolate and 50% chance of winning two chocolates, then my risk-aversion

unmeasurable risk"); it assumes a stricter sense where one does not have an adequate definition of the probabilities of possible states of affairs – the sense used by decision theory under complete ignorance. However, even in this case, uncertainty can be increased if the agent is unaware of relevant possible states of affairs (or of relevant long-term consequences of those states of affairs) or even of his own utility/value function (a type of normative uncertainty).

Decision under complete ignorance has often been represented as *a game against Nature* (Binmore, 2009, p. 154), where *Nature* is a fictional "unreasoning entity whose choice affects its pay-off, but which has no awareness of, or interest in, the outcome of the game." (Straffin, 2010, 56) One of the favorite candidate criteria to fill this lack of a probability

---

is zero. Henceforth we shall use the expression *in this* sense, unless we state it otherwise.

- in finance, it refers to volatility - i.e., the (square root of) *variance* relative to the expected return of a portfolio (Mandelbrot & Hudson, 2008). If an investment of $ 100 has an expected return of 1%, with a volatility of 3%, then, roughly, one can expect to receive between $ 98 and $ 104. This is an important factor, since the investor has a lower bound for the losses she can bear (i.e., her capital) - which explains the asymmetry between risk-aversion as loss (limited to capital) and as gains.

This second notion offers a conservative way of modeling "grave uncertainty," though still applying Laplace's principle - as an unknown (and potentially unlimited) *variance*. Buchak (2013, p. 44) approaches this possibility in a particularly interesting way, arguing that risk-aversion should be a *third* subjective dimension of the decision, not reducible to the other two (subjective probability and preference functions) - where the most rational decision is the one with the greatest subjective *risk-weighted* expected utility; this approach might be interpreted as consistent with financial practice.

distribution is Laplace's *principle of insufficient reason* – whereby, in the absence of other information, for reasons of *symmetry*, we assign to possible states of affairs the same probability; it is a *judgement* rule (it tells us how to form our *credences*), leading to a natural extension of the theory of subjective expected utility (Binmore, 2009, p. 154-155).

Another criterion is the more conservative *maximin*, by which agents minimize their exposure to risk, as if *Nature* were an adversary playing a zero-sum game against them (Binmore, 2005, p. 33) – so selecting the action that maximizes the outcome in the worst-case scenario. This is not a procedure for defining the probability of states of affairs (i.e., measuring our corresponding uncertainty), but a rule for selecting actions. These two criteria are at the basis of two competing theories of justice: Harsanyi's utilitarianism and Rawls's liberalism, respectively – as we shall see in section 2.

The literature discusses two additional criteria that extend decision theory to situations of complete ignorance and satisfy the axiomatic *desiderata* of *symmetry, strong dominance* and *linearity*[6] (Milnor, 1954; Binmore, 2009, p. 158); these are the Hurwicz's criterion[7] (Arrow & Hurwicz, 1977) and Savage's

[6] However, decision theory under bounded rationality also provides attractive criteria that do not fully satisfy these axioms.

[7] It provides a function combining both the maximin criterion with a *maximax* criterion – i.e., actions are evaluated according to both their worst and their best possible results, pondered by a new parameter, the "Hurwicz constant" h (with a positive value smaller than 1) that reflects one's optimism / pessimism concerning the pay-offs (measured in utilities). Suppose *a* is the worst pay-off of an action, and *b* its best pay-off; then the agent should choose the action that maximizes the value of sum of: a) the product of (1 - *h*) times *a*, and b) the product of *h* times *b*. When *h* = 0, this criterion

(1972) *minimax regret*[8]. The canonical analysis of these four criteria, with a formal comparison between them, is found in Straffin (2010) and in Luce & Raiffa (2012). However, we notice Hurwicz's and Savage's criteria are mostly absent from debates in political philosophy[9]; a possible explanation would be that they are too complex for the thought

---

is reduced to the maximin criterion; when $h = 1$, it's reduced to the *maxmax* criterion.

[8] Savage clearly express that his subjected expected utility theory should be almost entirely confined to what he calls "small worlds" (Savage, 1972, p. 82) – i.e., decision problems with well-defined boundaries and measurable subjective probabilities. In "large worlds," he recommends us to adopt a special type of minimax algorithm to select actions: starting with the pay-off matrix of a decision problem, we should then compute the *regret matrix* by subtracting the pay-off of each possible outcome from the highest pay-off we would receive if we had taken *another action*; i.e., this provides us with how much we could "regret" if we had taken such an action. Then, we should choose the action with the *lowest maximum* regret. One major problem is that, since the regret matrix takes into consideration the pay-off of *all the other options*, this criterion ends up violating *transitivity* (in Von Neumann-Morgenstern axioms) and the sure-thing principle (in Savage's axioms). Savage "bites this bullet," saying his minimax rule is a "practical rule of thumb in contexts where the concept of 'best' is impractical. […] it would not be strange, for example, if a banquet committee about to agree to buy chicken should, on being informed that goose is available, finally compromise on duck" (Savage, 1972, p. 206-207).

[9] Particularly surprising in the case of Savage's minimax regret, since he is explicitly concerned with group decisions, pointing out the similarity between an *ambiguous* decision problem (i.e., where many different probabilities distributions fit the data) and a decision where different agents are in disagreement over the best fitting probability distribution.

experiments philosophers use to expose their theories – but that is not quite accurate, because, particularly after Rawls (1971), moral and political philosophers have developed a taste for complexity and economic concepts. Another possibility is that, though these criteria are adequate for decisions under *ignorance*, they would not extrapolate well for decisions under *deep uncertainty*, where the parties cannot agree upon "(1) the appropriate models to describe the interactions among a system's variables, (2) the probability distributions to represent uncertainty about key variables and parameters in the models, and/or (3) how to value the desirability of alternative outcomes" (Lempert, Popper & Bankes, 2003, p. 3). Our conjecture is that, unlike Laplace's and Wald's criteria, they are not *intuitive*; in the last sections, by comparing Harsanyi's and Rawls's theories of justice, we hope to throw some light on why some principles might be considered intuitive – they have salient properties that allow for bounded rational agents, in situations of scarce information, to converge.

1.2 Uncertainty in real life: ambiguity aversion

One way to illustrate the divergence between these criteria is the Ellsberg's (1961) paradox, in this simple formulation exposed by Gintis (2009):

> Consider two urns. Urn A has 51 red balls and 49 white balls. Urn B also has 100 red and white balls, but the fraction of red balls is unknown. One ball is chosen from each urn but remains hidden from sight. Subjects are asked to choose in two situations. First, a subject can choose the ball from urn A or urn B, and if the ball is red, the subject wins $10. In the second situation,

> the subject can choose the ball from urn A or urn B, and if the ball is white, the subject wins $10. Many subjects choose the ball from urn A in both cases. This violates the expected utility principle no matter what probability the subject places on the probability p that the ball from urn B is white. For in the first situation, the payoff from choosing urn A is [0.51u(10) + 0.49u(0)] and the payoff from choosing urn B is [(1-p)u(10) + pu(0)], so strictly preferring urn A means p > 0.49. In the second situation, the payoff from choosing urn A is [0.49u(10) + 0.51u(0)] and the payoff from choosing urn B is [pu(10) + (1-p)u(0)], so strictly preferring urn A means p < 0.49. This shows that the expected utility principle does not hold.
> […]
> The usual explanation of this behavior is that the subject knows the probabilities associated with the first urn, while the probabilities associated with the second urn are unknown, and hence there appears to be an added degree of risk associated with choosing from the second urn rather than the first. If decision makers are risk-averse and if they perceive that the second urn is considerably riskier than the first, they will prefer the first urn. (Gintis, 2009, p. 17-18)

What is striking about Ellsberg's paradox is that, even after noticing the inconsistency, subjects often remain attracted to picking the "safe" urn A (with known probabilities). First, they arguably do not regard it as a contradiction: they do not consider the information that, in urn B, the probability of withdrawing a red ball is *negatively*

correlated with that of removing a white ball; i.e., they choose in both situations as if they had completely different urns. After all, if, in the second situation, instead of the urn B we had another urn C (identically described: with an unknown proportion *q* of red balls, but with no correlation to B's distribution), there would be no *contradiction* – but, instead, an arbitrary treatment between the subjective probability distribution regarding two identical information sets (B and C). It is arguable that, if we framed both choices as a single bet, subjects could realize (after reflexion) that picking both balls from urn A or urn B results in the same expected *pay-off*.

However, subjects usually *do not* make this decision by determining a probability distribution over the outcomes, but by just preferring the urn they have more information about. Binmore (2009, p. 89) equates this preference to "preferring to settle issues by tossing a coin known to be fair, rather than tossing a coin about which nothing is known at all," which usually regarded as a sound policy. We can contrast this preference with experiments made by Binmore, Stewart and Voorhoeve (2012), showing ambiguity-aversion decreases as subjects are assured that outcomes will be determined by a random process with no human interference - in which case it would make sense to adopt Laplace's principle.

1.3 Deep uncertainty: trying to evade probability assessments

If our problem (uncertainty) is that we cannot estimate adequately the probability of a hazard, why not to adopt a distinct conception of *risk* that does not appeal to probabilities? This is the aim of Pritchard's (2015) conception, which advocates for a "modal account" of risk. He distinguishes it from a "probabilistic conception" by

comparing two equiprobable events: suppose that a bomb will explode if (i) in a lottery, a specific number out of 14 million is withdrawn, or if ((ii) a conjunction of bizarre events (e.g., the spontaneous pronouncement of a Polish phrase in the Queen's next speech, plus the victory of an underdog at the *Grand National*...) occurs, with an assigned probability of 1 in 14 million. According to the argument, the second event would be a smaller risk, since it is "modally farther away;" i.e., it is in a less "plausible" possible world, demanding a greater number of coincidences. A supporter of this argument, however, should exchange a loss of $ 100.01 under condition (a) to *get a 6 in a dice roll*, for a loss of $100 under condition (b) a *head-head-tails sequence in three unbiased coin flips* (*under the condition that*, *if all three flips turn out identical, the coin is flipped thrice again*)[10]. Although the second condition is more complex - and, in the sense defined by Pritchard, "modally more distant" - both are equally likely, and the exchange implies an expected loss of $ 0.01.

There is no simple escape from the probabilistic conception: once we begin to assign probabilities to a class of events, updating them by Bayes's rule is the only way for the agent to obtain new information without dynamically breaking consistency and transitivity requirements – and so hedging oneself against sure losses such as *dutchbooks* (Skyrms, 1987). So why is Pritchard's conception appealing? One possible answer is that it uses an *intuition pump* (we are borrowing the expression from Dennett, 2013) enhanced by a *representation bias*: even though we know that both options are equiprobable, we prefer the one that is simpler to conceive. Something analogous happens when people judge,

[10] The probability of any specific outcome would usually be 2^{-3}; but in the case of (b) it is 1/6, since the results *HHH* and *TTT* are excluded. Any finite (additive) probability can be simulated by means of a similar process.

e. g., that the frequency of a particular letter is higher in the beginning of words than in the middle, since it is easier to remember words beginning with a certain letter than words containing it in the middle; or when tossing a fair coin, "people regard the sequence HTHTTH to be more likely than the sequence HHHTTT, which does not appear to be random," although by definition both have the same probability $2^{(-6)}$ of occurring (Tversky & Kahneman, 1977).

These biases arise because we extrapolate from cognitive *heuristics* – i.e., judgment and decision procedures that usually lead to satisfactory results in situations of bounded rationality for which our intelligence has evolved. And simplicity is often a good proxy for probability, indeed – some conceptions of induction depend on that[11]. Besides, simplicity implies lower computational costs, and is usually more robust to *random noise*; in our example, before assigning probabilities to the different scenarios, it would be harder to compute the more complex one, and it would be more prone to errors. In fact, if we added to Pritchard's thought experiment the real and deep uncertainty of the real world, the first hard problem would be to decide how to assign a credence to the second event; and, even after assigning it, we would still distinguish both events according to the amount of available information.

[11] Besides, simplicity implies smaller computational costs, and is usually more robust to *random noise*. In our example, before assigning probabilities to the different scenarios with a probability of 10^{-14}, it would be harder to compute the odds of the more complex one, and doing so would be more prone to errors (it would require working with many probability distributions with high variance). That's not the case of the coin toss, which is demonstrably equiprobable with the dice roll, defined by similar stochastic processes.

Thus, in scenario (i) we have *complete information* (we know, as well as any other agent, what there is to know about the relevant variables - the lottery), and our estimate is very close to the objective probability of the event[12] - it is identical to the estimate of any other agent. On the other hand, in scenario (ii), we have less information, so that estimates tend to be *ambiguous* (instead of a well-defined probability assignment, the agents would estimate it within a relatively wide range – their probability distribution would show more variance). So it would be tempting to perceive the latter as less risky thanks to our aversion to scenarios with greater information in loss situations.

Also, this points to a feature associated with the "decision discomfort" we often associate with uncertain situations – that they are uncertain because of *our* limits, because of our lack of information; maybe we could be less uncertain if we had more data, or time, or intelligence, or minuteness, even if we do not know how much more – so that the parameters of our decision problem (actions, world states, credence and utilities) are drastically affected by our previous choice on when to stop reflecting about the problem at hand to finally make our decision. In everyday deliberations, it is obvious that the information gain from spending more time

[12] It is unnecessary to start an ontological discussion on the meaning of objective vs, subjective probabilities here (and even more on modal realism); this remark works as well with a *relative* notion of subjective probability (in the style of Gilboa & Schmeidler, 2003): e.g., I believe that a coin *n* has an *objective* probability of .5 to result in heads because I have analyzed its previous history (i.e., the *heads* frequency approaches the limit 0.5), whereas a coin *m*, over which I have no information, is deemed to have a *subjective* probability of 0.5 of resulting in heads, because in this case I apply the principle of Laplace's insufficient reason. Thus, we usually call a probability assignment *more* subjective insofar as it is performed with *less information*.

reflecting on, e.g., the choice between two ice cream flavors would be offset by the increased costs of doing so; on the other hand, that's not so clearly the case for life-or-death decisions – so we can be expect to deploy different decision-making processes in these examples.

We would better interpret Pritchard's theory as a descriptive model of how people *actually* assess risks (though this would require empirical tests and a psychology inquiry) – competing with, e. g., Tversky & Kahneman's (1977) *prospect theory*. It would represent a set of heuristics adopted because, in ordinary cases, they approach the reasoning of an unbounded rational agent, but at much lower processing costs. Indeed, the idea of modal distance from Pritchard's epistemology seems compatible with a *case-based decision theory* (Gilboa & Schmeidler, 2003), where, instead of well-defined judgments about probabilities, an agent reasons based on the *similarity* between observed events – a kind of instance-based learning. Unlike Pritchard, however, Gilboa & Schmeidler do acknowledge that:

a) their theory is indeed a theory of bounded rationality (Gilboa & Schmeidler, 2003, p. 41), combining statistical, deductive, and analogical reasoning[13] about how people actually decide in very narrow information contexts - such as when they cannot map

[13] Case-based theory gives prominence to the analogical reasoning associated with legal practice - precisely the type of scenario focused by Pritchard (2015). The evaluation of evidence in a judicial process is hard to capture in an expected utility model because the decision-making process is marked by a division of the burden of proof between parties and by shared rules and precedents. Such rules aim not only to save time and other scarce resources, but at values such as publicity and fairness, too, thus restricting the quest for accuracy and efficiency (Sunstein, 1994).

all possible scenarios, or adequately estimate their likelihood or utility (Gilboa & Schmeidler, 2003, p. 1-2);

b) in the limit, in a scenario with complete and perfect information, both cased-based and expected utility theory would make equivalent predictions, so that one can be described as embedded in the other (Matsui, 2000). In fact, Gilboa & Schmeidler (2003, p. 93) argue that both are different *conceptual frameworks*, in which specific theories focused on certain types of decision problems are formulated, rather than competing theories.

So, it is plausible that our problem with uncertainty will not be solved by just adopting an alternative conception of risk – i.e., by using a different framework to measure uncertainty, unrelated to subjective probability functions (even if it might provide a good heuristic for some decision problems)[14].

2 Uncertainty and normative political philosophy

Normative political philosophers may be tempted to deflect the problem of uncertainty by claiming that theories of justice say nothing about how agents make judgments about what will happen, but focus on how they *should* behave – i.e., it provides prescriptions. However, judgments of fact about how people will behave and what may occur are relevant insofar as they define what can possibly occur

[14] An exhaustive analysis of other decision theories would have to include many competing proposals such as Spohn's (2017) ranking decision theory; however, besides our lack of space for this, Spohn himself acknowledges that we still lack an account of how alternative uncertainty measures can be applied to decision problems.

(Sorensen, 1995, p. 267). Epistemic limitations may make it impossible to identify the act recommended by a particular theory of justice: for these cases, the corresponding theorist usually recommends special procedures, so that the right act becomes the one determined by these procedures; but if the method and the corresponding principle diverge, the theorist's recommendations assume a paradoxical, if not contradictory, character (Kagan, 2018). We can then extend the problem of uncertainty from decision theory to the theory of justice: one theory is *inapplicable* if you cannot justifiably believe you can reach the goals it prescribes, and it is self-defeating if, by acting in accordance with the principles it prescribes, you cannot reach those goals.

Such an argument is often used against utilitarianism – i.e., that acting as a utilitarian often ends up in an agent taking sub-optimal actions, so concluding with the recommendation that agents should not profess utilitarianism (if they wish to achieve the best possible states of affairs). But utilitarians actually "bite the bullet" and see no problem in this objection (cf. Sorensen, 1995, p. 250): they often accept the immediate inapplicability of act-utilitarianism on a case-by-case basis, and then argue in favor of something like general policies or even rule-utilitarianism – on the principle that adopting general rules and policies aiming at general utility would imply *ex ante* more general expected utility than individually and constantly deciding which action has the most general expected utility (see, e.g., Harsanyi, 1985).

It is a mistake to think this objection is confined to consequentialist or teleological theories. After all, the consequences of an action may also be relevant to deontological theories[15] - otherwise, these theories would be,

[15] Especially when dealing with distributive justice; in the preface to *Sovereign Virtue*, Dworkin (2000) argues that every action or

according to the words of a deontologist himself, simply "insane" (Rawls, 1971, p. 26)[16]. For instance, deontological conceptions of individual rights must include considerations of consequences when dealing with risks, harms and compensation – as when determining if one's action creates an acceptable risk of harm for others (e.g.: Nozick, 1990, p. 73; Dworkin, 2011, p. 290). Even a Kantian would not be completely immune to the objection: one may have to worry about the consequences of an action when one defines a maxim to apply the categorical imperative on (Gibbard, 2007, p. 207) – most obviously in the case of so-called *imperfect duties*[17].

omission of the state benefits someone, and that we can estimate who is this person; so we should be able to justify the *status quo* to the worst-off - since we could have made different decisions. However, if no one is able to predict or adequately estimate the relevant consequences of a public policy, then one cannot reliably provide such justification.

[16] One can argue whether uncertainty could threaten even Rawls's difference principle: "(...) that social and economic inequalities must be arranged so that they are of the greatest benefit to the least advantaged. Insurmountable problems about measuring social and economic status might make a mystery of what counts as satisfying this obligation" (Sorensen, 1995, p. 248).

[17] In Kantian philosophy, "perfect duties" are inviolable limitations that moral law imposes on agents (e.g., never make a promise without intending to fulfill it); unlike "imperfect duties", which refer to cases where moral law requires one to adopt an *end*, and so allow for the possibility that the action might be objected on other grounds – for instance, assisting others is an imperfect duty, which might be trumped by perfect duties (if the project is morally objectionable) or additional imperfect duties – e.g., if the cost of doing so is too high (Johnson & Cureton, 2016).

2.1 Uncertainty and impartiality in a social contract

One can see the problem of deciding under uncertainty as an embarrassment of riches: the absence of information prevents us from constraining our "search space" in a way that would allow for a straight application of expected utility maximization. So, instead of deflecting the objection, a theorist may claim that political philosophy can provide *additional* criteria to complement decision theory – such as impartiality or fairness.

The most famous thought experiment to test a principle's impartiality is the Rawlsian *original position*, in which, under a *veil of ignorance*, subjects are unaware of their particular features and interests (Freeman, 2016). However, Harsanyi (1953; 1975) presents an analogous thought experiment (the *impartial observer*) with a utilitarian conclusion: by applying the principle of insufficient reason, an individual in such a situation would assign the same *ex ante* probability to occupying anyone's position in a society. Therefore, without appealing to one's own utility function (i.e., one's personal preferences) and position in society, the rational agent would choose principles leading to the greatest general expected utility. For Rawls (1971), though, in face of a similar uncertain scenario, an agent would want some hedge against risky alternatives – so including, among the principles of justice, the *difference principle*, whereby wealth and income inequalities shall be arranged as if to benefit the worst-off (Rawls, 1971, p. 72).

The two authors get to different conclusions mainly because, despite these similarities, they in fact use *qualitatively* different experiments with different objectives (Moehler, 2015). First, there's a distinction on what is being distributed in those scenarios: in Rawls (1971, p. 65), there are agents who are in the process of defining the basic structure of the distribution of the "benefits and burdens of cooperation" of

a stable society – i.e., resources. In Harsanyi, they must choose which distribution of *utility* they would prefer – i.e., what possible world they would prefer to live in, when they only know such distribution (thus not including information on how cooperation will occur). Thus, this seems like asking someone about who one would like to be, or in which world one would like to live; assuming a constant population and an equal probability of occupying the place of any individual, a rational agent would prefer to live in the world with the highest average expected utility (Harsanyi, 1975, p. 598) – i.e., the situation where an agent is more likely to satisfy a greater number of their own preferences. But in this situation, it makes little sense to talk about *a social contract*: it is like asking whether one would rather be born in a paleolithic egalitarian community, with a 30-year life expectancy and an income Gini-index close to zero, or in the 21st century[18].

On the other hand, despite defending the utilitarian principle as the right foundation of morality, Harsanyi himself considers that a *maximin* principle (used to justify the difference principle) can be a good proxy for the utility principle in real-world applications, such as the problem of "optimal income distribution or of optimal taxation," so

[18] One problem with this distinction is that it allows us to ask whether, in fact, the original position is the best context from which to determine the principles of justice that *our* society must adopt. This is a point often made by Sen (1999; 2009, 93): the question about which principles to adopt depends on the type of information available - even in the original position. Also, this explains why the "equality of what?" debate (resources vs. utility) is *irrelevant* to our present analysis of the original position (though crucial to post-Ralws political philosophy). After all, if agents have *only* information about utility in the original position, then their decision would reflect *that*; but if all they know is the distribution of something else, such as resources, they will take *another* decision.

making use of a distinction between "basic philosophical principle" and principles applicable to decision-making (Harsanyi, 1975, p. 606). In other words, the difference principle might be a good heuristic for these pragmatic problems[19]; it is attractive particularly because it tends to be less informationally demanding – one needs only to identify the alternative with the best lowest pay-off.

We can extend this argument to the reasoning about the interaction with other agents (modeled as a game of partial conflict): even if one regards Harsanyi's argument as a solid ground for utilitarianism as a *comprehensive philosophical doctrine* – at least in the situation where we have no other information

[19] There is an alternative heuristic, often neglected by philosophers and economists: while the average expected utility principle makes us choose the action associated with the highest **average** utility, the *median heuristic* has us choosing the action with the highest *median* in the same distribution: "Statistically, expected value is the central tendency of the distribution embodied in a risky gamble. For highly non-Gaussian distributions, the mean is not considered a valid estimator (…). The median, an alternative estimator of central tendency, is robust to noise and is often favored for highly skewed distributions." (Hayden & Platt, 2009)

Statisticians prefer to use the median as an estimator because, in the absence of additional information, the variable (in our case, the hypothetical agent) is equally likely to result placed *above* or *below* the median position; it works even for *pathological* probability distributions, such as the Cauchy distribution, that have no definite mean. This is arguably the case of inequality: the distribution of income in complex societies is incompatible with a normal distribution (Mandelbrot & Hudson, 2008). Thus, the median heuristic in the original position (i.e., choosing a society that maximizes the utility of the individual in the median position) would tend to converge more towards the difference principle than towards utilitarianism - though it does not coincide with none of them.

on the decision problem – it does not provide a stable reason for anyone already embedded in a society to support corresponding utilitarian institutions; on the contrary, it may seem revolting for the worst-off to hear from the best-off: "I regret your misfortune, but we had the same priors for success, and my happiness offsets your misery." Harsanyi's solution, therefore, underestimates the need for acceptance, coordination and cooperation (as argued by Regan, 1985; Gauthier, 1984, p. 71). In a utilitarian society, those favored by the benefits of cooperation would not be the neediest, but those with the greatest potential for increasing overall marginal utility; this may coincide with the worst-off – given the law of diminishing returns[20] – but not necessarily. As in the management of a productive activity, an efficiency-oriented society would be tempted to transfer resources to agents capable of increasing the "production" of utility (Sen, 1979, p. 9)[21].

---

[20] "It is a perennial idea of the utilitarian school that if utilities are concave, egalitarian consequences will follow from the sum of average rules." (Mongin & Pivato, 2016, p. 734) However, neither utilitarianism nor *maximin* give us reasons to prefer, in the original position:

1. distribution of 1 *utility* to A (with probability of 50%), or to B (with probability of 50%), or

2. distribution of 1 *utility* to A.

A Rawlsian agent would choose option (i), not for the difference principle itself, but because it matches the spirit of the equal opportunity principle.

[21] The difference principle also allows redistributing resources 'upwards', to more productive people, as an incentive - but only in order to latter improve the situation of the worst-off. This is the 'argument from incentives': "Supposedly, [...] the greater expectations allowed to entrepreneurs encourages them to raise the prospects of laboring class. Their better prospects act as incentives

Thus, in Rawls, the support for the difference principle involves more than the simple fact of uncertainty. The original position is not only a decision under complete ignorance against *Nature*; it anticipates interactions and bargains, either conflict or cooperation, that might occur after the withdrawal of the veil of ignorance[22]. So the principle aims to provide a basis for a social contract, assuring the worst-off citizens that the only way to improve their *ex post* relative condition would have been by putting others in a similar or worse position. This justification appeals to demands of stability, publicity and reciprocity (Freeman, 2016; Rawls, 2001, p. xvii): if a society must organize itself according to principles that everyone can identify and accept, then it has to be able to sacrifice potential efficiency for the sake of increased fairness and stability[23]. Hence, Binmore (2005, p. 170) argues that liberal egalitarianism would be the option for a social contract in the absence of an external entity to define and enforce social

---

so that the economic process is more efficient, innovation proceeds at a faster pace, and so on." (Rawls, 1971, p. 68)

[22] This characterization may be controversial - but it seems to follow from the idea that there must be a link between the original position and society after the veil's withdrawal. Gauthier (1984) and Binmore (2005) rather conceive the original position as a *game* (a non-parametric decision situation); however, they model it as a bilateral *bargaining problem*, which might not adequately represent various aspects of a social contract with a plurality of parties, such as the emergence of coalitions, reputations, and conventions (Sugden, 2001, p. 235).

[23] For Binmore (2005, p. 5-9), this is the most important *desideratum* of social norms and the greatest appeal of political liberalism, for an evolutionary reasoning - i.e., considering the simple fact that unstable norms by definition do not persist for a long time (either they are changed, or society dissolves).

norms – i.e., where social norms would have to be enforced by all individuals.

This highlights that, for Rawls, the decision in the original position is seen as a reason for any individual to comply with the social contract even *after* lifting the veil of ignorance - for, as we have pointed out, it is a matter of defining the principles on which we would cooperate. Thus, one can argue that the difference principle provides better incentives for this goal and enhances social trust: in a Rawlsian well-ordered society, the best-off could say to the worst-off that improving the prospects of the latter would imply an even worse condition for someone; but, in a utilitarian society, all they can say is "my joy compensates your pain," which, in real life, may fuel class struggle.

Even without the possibility of being abused, utilitarianism provides a reason only from a more *detached* point of view, an *ex ante* perspective, and it would be tempting for citizens to complain about it *ex post*:

> (…) current individuals would not see the reasons for the chooser in Harsanyi's original position as reflecting their reasons. In this way, the justificatory link between the model of rational choice in the original position and the reasons of actual individuals in society would be severed. Individuals will not see Harsanyi's justification of average utility as justification for *them*. (Gaus & Thrasher, 2015, p. 57)

Why is the difference principle supposedly more *acceptable*, more susceptible to foster stable cooperation? Since it can be justified by reasons compatible with different reasonable philosophical frameworks, appealing to individuals in different situations, and since, as Harsanyi admits, it is computationally simpler, Ralws's difference

principle may work as a *Schelling* (2015) point for people with distinct comprehensive doctrines and cognitive resources – a salient equilibrium to which all can converge.

3. Coping with uncertainty: moral intuitions and responsibility in groups

The reasoning above contrasts with other non-consequentialist arguments, which philosophers often claim to reflect primitive and intuitive moral principles. Instead, we argue that the intuitive appeal of the difference principle actually flows from its simplicity (i.e., it is cognitively less demanding), making it a more acceptable principle for coordination (particularly under the threat of disagreements that may affect social cooperation).

A similar argument might explain the origins of the distinction between actions and omissions (e.g., the difference between "killing and letting die"), underlying the intuitive appeal of a *status quo* bias: we usually do not hold people accountable for "doing nothing", particularly under uncertainty, because that's what is expected from them as a *default* action. However, "doing nothing" is not the most accurate description of someone's behavior, since we can always be described as subjects of some trivial action (e.g.: breathing); in fact, "doing nothing" usually means "doing whatever one is already doing by *default*, making no additional conscious decision…"

From an individual point of view, this is often a sensible principle, guiding how we distribute burdens and responsibilities; think about a simple cooperative interaction, like playing volleyball: when a ball falls between many players, whoever takes the initiative must communicate this to others, thus assuming the responsibility for catching it. The others must "do nothing" (i.e., wait and stay alert), so

minimizing the risk of interference and harmful collisions. On the other hand, if the ball falls closer to one specific player, we assume it is her responsibility – it is the alternative with the lowest marginal cost. Too often, though, amateurs let the ball drop among them because it is not clear who should move to catch it – they face a (micro) social dilemma; "doing nothing" is a bad decision procedure in this case. This does not happen among professionals, who have a strong bias for claiming responsibility for themselves when they are uncertain, so minimizing the collective risk of an adverse score. We can wonder if this decision procedure is optimal: should not those players reason about who is the best catcher for each ball?

Extrapolating from this rough analogy to social decision-making, "doing nothing" would be like this individualistic heuristics: in the absence of social norms and collective concerns, "do nothing," lest you often assume responsibility for harming others unnecessarily. Harsanyi's utilitarianism, if taken to its extreme, would recommend us to find out the optimal solution (in the case of volleyball, the best move to maximize the chances of scoring a point for each ball); but utilitarians are often more pragmatic than that, and would recommend a *maximin* principle that all defenders could follow: first, minimize our risk and catch the ball, and so define a strategy to score a point later on.

However, here is the difference between this "volleyball" metaphor and the theory of justice (and the debate over its underlying decision theory): it is very clear what are our goals in playing volleyball, and also most of our constrains, so the only remaining uncertainty is in the facts. Notice we are not so arguing for the difference principle as a proxy that would help us satisfy the average utilitarianism, as Harsanyi argued – because, just like the original position (and our metaphor of the volleyball game), Harsanyi's *impartial observer* is just *another* thought experiment; Instead, in explaining the *intuitive*

appeal of the difference principle in a social contract situation like the original position, we highlighted features that might be identified in other social decisions.

3.1 Deciding for others: prudence and precaution

We often demand that authorities decide in ways that deviate from utilitarianism. For instance, consider the claim, from authors such as Buchak (2017) and Otsuka & Voorhoeve (2009), that the shift from the perspective of an individual decision to that of *interpersonal* decisions changes the principles we must use to decide. This results in the so-called *risk-principle:* "When making a decision for an individual, choose under the assumption that he has the most risk-avoidant attitude within reason unless we know that he has a different risk attitude." (Buchak, 2017, p. 632)

In environmental law and sanitary policies, companies are often required to comply with a *precautionary principle* requiring them to demonstrate that an activity doesn't offer risks for others; a bad interpretation for this principle is that it complies with some sort of *status quo* bias argument. However, it has the advantage of balancing the information asymmetry between the public and the innovator, so imposing the burden of researching and publishing effects on the latter (Picavet & Lafaye, 2012) – it is a practical solution for the problem of cooperation in an uncertain context. Analogously, the *prudence* principle in accountancy determines that an asset of an uncertain value must be written according to its lowest assessment, in order to avoid overestimation (because assessing the value of an asset is usually a task for the security's owner, not for their creditors). The principle was so incorporated into professionals' practice (because of the advantage of a reputation of prudence) that, even after recent legal changes, firms keep

adopting it – it is often less costly than developing a precise and debatable model to classify different risks (Kronbauer et al., 2017).

Moreover, this also echoes a concern about the relationship between managers and investors; in financial markets, institutions are required to disclose information about their risk *appetite* - the higher it is, the more information to disclose. Also, risks are to be managed aiming at institutional robustness, in order to optimize resistance to *failure* – I.e., losses an institution may not recover from. A system is robust if it can keep functioning, despite internal and external risks, without drastic changes in its structure (Lempert & Collins, 2007); by *modus tollens*, if a system is not robust, in an environment with risks arising from random processes, either it will break, or its structure will change. Thus, in some activities involving externalities, managers are required to adopt measures to resist to uncertain events, minimizing the risk of failures that could threaten economic activity, even if their probability is small or unknown (where they would be underestimated in a day-by-day cost-benefit analysis).

Similarly, that's how we often reason in preventing social harms: Shrader-Frechette (2014, p. 194) argues that we should favor Rawls's theory over Harsanyi's in cases of decision-making with high uncertainty having: a) a society as a scope (and not just an individual), b) unequal distribution of adverse outcomes and c) potential for catastrophe[24]. Even

[24] It's hard to define catastrophe. Cooke (1985) uses as an example of a catastrophic event disasters that extinguish life - implying a state of things with no measurable value; this is different from a state of affairs with 'zero utility'. On the other hand, Yudkowsky (2008) uses the term 'catastrophe' to refer events of global impact (such as the death of millions). This does not correspond to the *relative* sense we are use using, which includes impact events within an activity. However, it also does not correspond to the more usual

if a society committed to preventing such scenarios compromises part of the well-being of some citizens, that is considered an acceptable price – otherwise it would risk compromising the well-being of the worst-off, who have fewer chances of protecting themselves[25].

Conclusion

First, we have seen that the problem of uncertainty is *pervasive*: we cannot escape the problem of ignoring the consequences of a decision – otherwise such a theory might be inapplicable. We argued that, in situations of social risks, the adoption of the difference principle, according to the *maximin* criteria, justifies the action to all parties involved; moreover, as noted by Harsanyi himself, this principle is easier to apply than the utilitarian principle because it has lower informational requirements - it is epistemically simpler to identify and avoid worst outcomes. We showed how this

---

sense of catastrophe used by Shrader-Frechette (2014), which includes incidents involving the death or disability of a large group of people.

This seems to include both the idea of a state of things difficult to assess (of very high impact - in which the prospect of recovery is nil, or improbable) or to estimate (great uncertainty); this is a practical problem for the subjective expected utility theory, since it often associates the subjective probability assigned to an event with an agent's propensity to bet on it (as in Ramsey, 1931).

[25] It may be unnecessary, but we must remark that this reasoning cannot be extrapolated to justify authoritarian control, incompatible with a liberal society – at least for Rawls; after all, by the same argument of the original position, the parties would choose the principle of equal freedom and the principle of equal opportunities - which take precedence over the difference principle. An autocratic society can be seen as the first risk the parties in the original position would seek to secure against.

reasoning may explain our moral intuitions and how it is consistent with social norms concerning risk allocation.

We must remark, though, the limitations of our argument: it doesn't mean that the *maximin* is a good criterion for decision theory in general; it does not extrapolate to individual decision-making, nor even to cases where the boundaries of a decision problem can be well-defined. We only argued that, in uncertain social contexts, it provides a more *acceptable* justification for policies than utilitarianism. It is a *decision* rule, for coping with uncertainty, not a *judgement* procedure, for *reducing it*; i.e., it is a policy to select actions in the face of uncertainty, not a procedure to precisify our credences when we lack information – so it doesn't solve the problem of assigning probabilities to different possible states. Finally, we highlight that we ignored population ethics and intergenerational conflict – i.e., our argument explicitly appeals to the need for stable cooperation between present agents, not future ones[26].

[26] Harsanyi points out, e.g., that Rawl's theory doesn't pay enough attention to intergenerational justice, and possibly implies zero net savings (since future generations tend to be better-off thanks to technological progress).That's true: Rawls proposes a *just savings principle* according to which we only have a duty to legate to next generations wealth enough for them to sustain a well-ordered society (it does not mean there are no other reasons to increase this endowment, though). But Harsanyi (1975, p. 602) himself complains that utilitarianism requires too much savings for the next generations (unless we make use of a social discount rate, which defeats the egalitarian value of utilitarianism, by privileging the present). We avoid this discussion here because, despite its huge importance, population ethics is hard, no matter what principle one chooses, so both the difference principle and the average-utilitarianism do not provide satisfying answers for long-term matters.

Rawlsians may dislike this conventional, even naturalistic, account of a theory of justice; it seems to lack the normative 'flavor' we usually expect from arguments of principle. However, instead of thinking of this as a reduction of a normative theory of justice to a non-normative theory of conventions, we suggest one should see it as an argument over the conditions under which principles of justice can be applied: even in the absence of a common agreement on what precise norms should be chosen and followed, or on what is the best conception of good, bounded rational agents can converge in a meta-level, particularly if they know they need to cooperate with each other. Actually, we dare to conclude by suggesting that this might be the main function of a normative theory – a theory about how agents should proceed: to allow for some guidance for the cooperation of bounded rational agents under uncertainty. If we could determine a cardinal utility function for each agent, and a corresponding precise probability distribution over outcomes, we would have no need for a normative theory of any kind; game theory would be enough to provide us with an answer about what decisions would be observed.

References

Adamou, A.; Peters, O. "Dynamics of inequality". *Significance,* 13(3): 32-37, 2016. doi:10.1111/j.1740-9713.2016.00918.x

Aristotle. *The Nicomachean Ethics of Aristotle*. Translation: W. D. Ross. London: Oxford University Press, 1966.

Arrow, K.; Hurwicz, L. "Appendix: An optimality criterion for decision-making under ignorance". In K. Arrow & L. Hurwicz (Eds.), *Studies in Resource Allocation Processes*, pp. 461-472. Cambridge: Cambridge

University Press, 1977. doi:10.1017/CBO9780511752940.015

Barros, G. "Herbert A. Simon and the concept of rationality: boundaries and procedures". *Brazilian Journal of Political Economy*, 30(3), 455-472, 2010. doi:10.1590/S0101-31572010000300006

Binmore, K. G. *Game theory: A very short introduction*. New York: Oxford University Press, 2007.

Binmore, K. G. *Natural Justice*. Princeton: Princeton University Press, 2005.

Binmore, K. G. *Rational Decisions*. Princeton: Princeton University Press, 2009, 2009.

Binmore, K.; Stewart, L.; Voorhoeve, A. "How much ambiguity aversion? Finding indifferences between Ellsberg's risky and ambiguous bets". *Journal of Risk and Uncertainty*, 45:215–238, 2012. doi:10.1007/s11166-012-9155-3

Bostrom, N. *Superintelligence: Paths, dangers, strategies*. London: Oxford University Press, 2014.

Buchak, L. "Taking Risks behind the Veil of Ignorance". *Ethics*, *127*(3), 610-644, 2017. doi:10.1086/690070

Bun, M.; Sarafidis, V.; Kelaher, R. "Crime, Deterrence and Punishment Revisited". UvA-*Econometrics Working Papers* N. 16-02, Universiteit van Amsterdam, Dept. of Econometrics, 2016. Available at: <https://ideas.repec.org/p/ame/wpaper/1602.html >

Cooke, R. "Conceptual fallacies in subjective probability". *Topoi* 5 (1):21-27, 1986. doi: 10.1007/BF00137826

Dennett, D. C. *Intuition Pumps and Other Tools for Thinking*. New York: W. W. Norton & Company, 2013.

Dworkin, R. *Sovereign Virtue: The Theory and Practice of Equality*. Cambridge, Mass: Harvard University Press, 2000.

Dworkin, R. *Justice for Hedgehogs*. Cambridge, Massachussets: Belknap Press (Harvard University Press, 2011.

Ellsberg, D. "Risk, Ambiguity, and the Savage Axioms". *Quarterly Journal of Economics*, 75 (4): 643–669, 1961. doi:10.2307/1884324

Elster, J. *Sour grapes: Studies in the subversion of rationality*. Cambridge, UK: Cambridge University Press, 2016.

Freeman, S. "Original Position" in Zalta, E. N. *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/win2016/entries/original-position/>, 2016.

Gaus, G, & Thrasher, J. "Rational choice and the original position: The (many) models of Rawls and Harsanyi". In Hinton, T. (ed.) *The Original Position* (Classic Philosophical Arguments), pp. 39-58. Cambridge: Cambridge University Press, 2015. doi:10.1017/CBO9781107375321.003

Gauthier, D. *Morals by Agreement*. Oxford: Oxford University Press, 1987, 1987.

Gibbard, A. "Thinking How to Live with Each Other". *The Tanner Lectures on Human Values,* na Universidade da California, Berkeley, de 28 de fevereiro a 2 de março de 2006, 2007.

Gilboa, I, & Schmeidler, D. *A Theory of Case-Based Decisions*. Cambridge, England: Cambridge University Press, 2003.

Gintis, H. *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*. Princeton, N.J.: Princeton University Press, 2009.

Giocoli, N. "From Wald to Savage: Homo Economicus Becomes a Bayesian Statistician". SSRN Electronic Journal, 2011.doi:10.2139/ssrn.1944218

Gray, J. "The Friedrich Hayek I knew, and what he got right - and wrong". *Newstatesman* online, July 31, 2015. Available at: <https://www.newstatesman.com/politics/2015/07/john-gray-friedrich-hayek-i-knew-and-what-he-got-right-and-wrong>

Greaves, H. "Cluelessness". *Proc Aristot Soc*, 116(3): 311–339, 2016. doi:10.1093/arisoc/aow018

Harsanyi, J. C. "Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking". *Journal of Political Economy* 61(5): 434–35, 1953.

Harsanyi, J. C. "Can the Maximin Principle Serve as the Basis for Morality? A Critique of John Rawls's Theory". *American Political Science Review* 69: 594–606, 1975. Available at: <https://www.jstor.org/stable/i306770>

Harsanyi, J. C. "Does Reason Tell Us What Moral Code to Follow and, Indeed, to Follow Any Moral Code at All?".*Ethics*, 96(1), 42-55, 1985. Available at: <http://www.jstor.org/stable/2381321>

Hayden, B. Y.; Platt, M. L. "The mean, the median, and the St. Petersburg paradox". *Judgment and Decision-Making,* 4(4): 256–272, 2009.

Hirschman, A. O. *The Rhetoric of Reaction*. Cambridge: Harvard University Press, 1991.

Holton, R. *Willing, wanting, waiting*. Oxford: Clarendon Press, 2011.

Johnson, R.; Cureton, A.. "Kant's Moral Philosophy". In Zalta , Edward N. (ed.), *The Stanford Encyclopedia of Philosophy*, 2016. Available at: <plato.stanford.edu/archives/spr2019/entries/kant-moral/>

Kagan, S. "The paradox of methods". *Politics, Philosophy and Economics,* 17 (2):148-168, 2018. doi: 10.1177/1470594X17717737

Knight, F. *Risk, Uncertainty, and Profit*. Boston, MA: Houghton-Mifflin, 1921.

Korsgaard, C. "The right to lie: Kant on dealing with evil". *Philosophy and Public Affairs,* 15(4): 325-349, 1986. Available at <http://nrs.harvard.edu/urn-3:HUL.InstRepos:3200670>

Kronbauer, C. A.; Marquezan, L. H.; Barbosa, M. A.; Diehl, C. A. "Analysis of the effects of conservatism in accounting information after the 2011 change in the basic conceptual pronouncement." *Review of Business Management*, 19(65), 453-468, 2017. doi:10.7819/rbgn.v19i65.2742

Lempert, Robert J.; Collins, Myles T. "Managing the risk of uncertain threshold responses: comparison of robust, optimum, and precautionary approaches". *Risk Analysis*. 27 (4): 1009–1026, August 2007. doi:10.1111/j.1539-6924.2007.00940

Lempert R. J.; Popper, S. W.; Bankes, S. C. *Shaping the Next One Hundred Years*: New

Methods for Quantitative, Long-Term Policy Analysis. Santa Monica, CA: RAND Corp, 2003.

Lo, A. W.; Mueller, M. T. "Warning: Physics Envy May be Hazardous to Your Wealth!" *SSRN Electronic Journal*, 2010. doi:10.2139/ssrn.1563882

Luce, R. D, & Raiffa, H. *Games and Decisions: Introduction and Critical Survey*. North Chelmsford, MA: Courier Corporation, 2012.

MacAskill, W. "Normative Uncertainty as a Voting Problem." *Mind,* 125 (500):967-1004, 2016.doi; 10.1093/mind/fzv169

Mandelbrot, B.; Hudson, R. L. *The Misbehavior of Markets: A Fractal View of Financial Turbulence*. New York: Basic Books, 2008.

Matsui, A. "Expected utility and case-based reasoning." *Mathematical Social Sciences,* 39(1): 1-12, 2000. doi:10.1016/s0165-4896(99)00008-6.

Milnor, J.W. "Games against nature". In: Thrall R.M.; Coombs C.H.; Davis R.L. (eds) . *Decision Processes*, pp. 49-60. New York: Wiley, 1954.

Moehler, M. "The Rawls-Harsanyi Dispute: A Moral Point of View". *Pacific Philosophical Quarterly*, 99(1): 82-99, 2015.doi:10.1111/papq.12140

Mongin, P, & Pivato, M. "Social Evaluation under Risk and Uncertainty." In Adler, A. D.; Fleurbaey, (eds) M. *The Oxford Handbook of Well-Being and Public Policy*, 2016.doi: 10.1093/oxfordhb/9780199325818.013.23

Nozick, R. *Anarchy, state, and utopia*. Oxford: Basil Blackwell, 1990.

Otsuka, M, & Voorhoeve, A. "Why It Matters That Some Are Worse Off Than Others: An Argument against the Priority View." *Philosophy & Public Affairs*, 37(2):

171-199, 2009. doi:10.1111/j.1088-4963.2009.01154.x

Picavet, E.; Lafaye, C. G. "La précaution, l'éthique et la structure de l'action." *Revue de métaphysique et de morale*, 4 (76): 593-609, 2012. doi : 10.3917/rmm.124.0593

Pritchard, D. "Risk" *Metaphilosophy*, 46(3): 436-461, 2015.doi:10.1111/meta.12142

Ramsey, F. P. "Truth and Probability" in Braithwaite, R.B. (eds) *The Foundations of Mathematics and Other Logical Essays*, p. 156-198. London: Routledge and Kegan Paul, 1931.

Rawls, J. *A Theory of Justice*. Cambridge, MA: Belknap Press of Harvard University Press, 1971.

Rawls, J, 2005) *Political Liberalism*. 3ª Ed. New York: Columbia University Press.

Regan, D. "On Preferences and Promises: A Response to Harsanyi". Ethics, 96(1): 56-67, 1985.doi:10.1086/292718

Savage, L. *The Foundations of Statistics*. 2ª Revised Edition. New York: Dover, 1972.

Schelling, T. C. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press, 2015.

Sen, A. K. "Equality of What?" *Tanner Lecture on Human Values*, in the Stanford University, on May 22, 1979. Available at: <http://www.ophi.org.uk/wp-content/uploads/Sen-1979_Equality-of-What.pdf>

Sen, A. K, 2009) *The Idea of Justice*. Cambridge, MA: Belknap Press, 2009.

Shrader-Frechette, K. *Tainted: How philosophy of science can expose bad science*. New York: Oxford University Press, 2014.

Spohn, W. "Knightian Uncertainty Meets Ranking Theory." *Homo Oecon* 34, 293–311, 2017. doi:10.1007/s41412-017-0060-5

Sorensen, R. A. "Unknowable Obligations". *Utilitas*, *7*(2): 247-271, 1995.doi:10.1017/s0953820800002077

Straffin, P. D. *Game Theory and Strategy*. Washington, DC: Mathematical Association of America, 2010.

Sugden, R. "Ken Binmore's Evolutionary Social Theory." *The Economic Journal*, 111(469), 213-243, 2001.doi:10.1111/1468-0297.00604

Sunstein, C. R. "Political Conflict and Legal Agreement". *The Tanner Lectures on Human Values*, in Harvard University, from November 29 to December 1st, 1994. doi:10.2139/ssrn.2544359

Sunstein, C. R. *Infotopia: how many minds produce knowledge*. New York: Oxford University Press, 2006.

Tversky, A., & Kahneman, D. "Judgment under uncertainty: Heuristics and biases." *Science*, 185 (4157), 1124-1131, 1973. doi: 10.1126/science.185.4157.1124

Yudkowsky, E. "Cognitive Biases Potentially Affecting Judgment of Global Risks." In Bostrom, N, & Ćirković, M. M. *Global catastrophic risks*. New York: Oxford University Press, 91–119, 2008. Available at: <intelligence.org/files/CognitiveBiases.pdf>