Metanormativity: Questions of moral and empirical uncertainty[a]

Metanormatividade: Questões de incerteza moral e empírica

NICHOLAS KLUGE CORRÊA [b]

NYTHAMAR DE OLIVEIRA[c]

[a] Research supported by the Academic Excellence Program (PROEX) of CAPES Foundation (Coordination for the Improvement of Higher Education Personnel), the Brazilian National Research Council (CNPq), and the Graduate Program in Philosophy of the Pontifical Catholic University of Rio Grande do Sul, Brazil.

[b] Pontifical Catholic University of Rio Grande do Sul (PUCRS), Porto Alegre, RS, Brazil. Master of Electrical Engineering and Ph.D. student in Philosophy. E-mail: nicholas.correa@acad.pucrs.br

[c] Pontifical Catholic University of Rio Grande do Sul (PUCRS), Porto Alegre, RS, Brazil. Ph.D. in Philosophy (State University of New York, 1994), Full Professor – PUCRS. E-mail: nythamar@yahoo.com

Abstract

How can someone reconcile the desire to eat meat, and a tendency toward vegetarian ideals? How should we reconcile contradictory moral values? How can we aggregate different moral theories? How individual preferences can be fairly aggregated to represent a will, norm, or social decision? Conflict resolution and preference aggregation are tasks that intrigue philosophers, economists, sociologists, decision theorists, and many other scholars, being a rich interdisciplinary area for research. When trying to solve questions about moral uncertainty a meta understanding of the concept of normativity can help us to develop strategies to deal with norms themselves. 2nd-order normativity, or norms about norms, is a hierarchical way to think about how to combine many different normative structures and preferences into a single coherent decision. That is what metanormativity is all about, a way to answer: *what should we do when we don't know what to do?* In this study, we will review a decision-making strategy dealing with moral uncertainty, Maximization of Expected Choice-Worthiness. This strategy, proposed by William MacAskill, allows for the aggregation and inter-theoretical comparison of different normative structures, cardinal theories, and ordinal theories. In this study, we will exemplify the metanormative methods proposed by MacAskill, using has an example, a series of vegetarian dilemmas. Given the similarity to this metanormative strategy to expected utility theory, we will also show that it is possible to integrate both models to address decision-making problems in situations of empirical and moral uncertainty. We believe that this kind of ethical-mathematical formalism can be useful to help develop strategies to better aggregate moral preferences and solve conflicts.

Keywords: Moral uncertainty. Normative uncertainty. Metanormativity. Maximization of expected choice-worthiness.

Resumo

Como é que alguém pode conciliar o desejo de comer carne, e uma tendência para com ideais vegetarianos? Como devemos reconciliar valores morais contraditórios? Como podemos agregar diferentes teorias morais? Como é que preferências individuais podem ser agregadas de forma justa para representar uma vontade, norma, ou decisão social? A resolução de conflitos e a agregação de preferências são tarefas que intrigam filósofos, economistas, sociólogos, teóricos da decisão, e muitos outros estudiosos, sendo uma área interdisciplinar rica para a investigação. Ao tentar resolver questões sobre incerteza moral, uma meta compreensão do conceito de normatividade pode ajudar-nos a desenvolver estratégias para lidar com as próprias normas. A normatividade de segunda ordem, ou normas sobre normas, é uma forma hierárquica de pensar sobre como combinar muitas estruturas normativas e preferências diferentes em uma única decisão coerente. É disso que se trata a metanormatividade, uma forma de responder: o que devemos fazer quando não sabemos o que fazer? Neste estudo, iremos rever uma estratégia de tomada de decisão que trata da incerteza moral, Maximização da Escolha Valiosa-Esperada. Uma estratégia proposta por William MacAskill que permite a agregação e comparação interteórica de diferentes estruturas

normativas, teorias cardeais e teorias ordinais. Neste estudo, vamos exemplificar os métodos metanormativos propostos por MacAskill, usando como exemplo uma série de dilemas vegetarianos. Dada a semelhança desta estratégia metanormativa com a teoria da utilidade esperada, mostraremos também que é possível integrar ambos os modelos para abordar problemas de tomada de decisão em situações de incerteza empírica e moral. Acreditamos que este tipo de formalismo ético-matemático pode ser útil para ajudar a desenvolver estratégias para melhor agregar preferências morais e resolver conflitos.

Palavras-chave: *Incerteza moral. Incerteza Normativa. Metanormatividade. Maximização da Escolha Valiosa-Esperada.*

Introduction

"*...the rarest of all human qualities is consistency*".

— Jeremy Bentham

Humanity is vast and multifaceted, as our species is currently spread out throughout over 195 countries. And in this landscape, humanity has not yet achieved the status of a single cosmopolitan society. However, our shared environment forces us to have to deal with each other, something that is often a reason for conflict, given our differences in their most diverse forms. Situations where our differences are aggravated often involve some sort of moral disagreement, and this is one of the most persistent sources of conflict in human life. ACLED[1] (Armed Conflict Location & Event Data Project) is an interactive online infographic that shows in which countries occur armed confrontations between state forces and civil, rebel groups. It informs us how the occurrence of conflicts in our world is something sadly common. In March 2020, UN Secretary-General António Guterres, given the current pandemic caused by the new coronavirus COVID-19, said in a statement[2]: "The fury of the virus illustrates the madness of war [...] For the warring parties, I say: withdraw from hostilities. Silence the weapons; stop the artillery; stop the air raids. This is crucial...".

One of the most significant differences between societies, and individuals, are their moral values, their preferences, their normative principles, their ethics. The problem of

[1] Armed Conflict Location & Event Data Project (ACLED). Available at: https://acleddata.com/2020/08/18/mid-year-update-10-conflicts-to-worry-about-in-2020/ Accessed on: August 25, 2020.

[2] Transcript of the Secretary-General's virtual press encounter on the appeal for global ceasefire. United Nations Secretary-General, Statements/Reports. 23 March 2020. Available at: https://www.un.org/sg/en/content/sg/press-encounter/2020-03-23/transcript-of-the-secretary-generals-virtual-press-encounter-the-appeal-for-global-ceasefire Accessed on August 25, 2020.

aggregating conflicting preferences and solving moral dilemmas is something that intrigues philosophers from ancient times till the present day (LOCKHART, 2000; ŻURADZKI, 2016; TARSNEY, 2018; HICK, 2018; BARRY & TOMLIN, 2016; 2019), but surprisingly, when compared to the study of empirical uncertainty we see that the study of moral uncertainty is a much less explored field[3] . Moral uncertainty research has applications from the most micro level, "how can an individual reconcile contradictory preferences?", to the macro, "how can societies (and the world) aggregate their preferences into a single coherent ordered structure?".

An interesting study, with empirical findings supporting moral pluralism, showing the difference between moral principles among different cultures is the Moral Machine experiment conducted by Awad et al. (2018; 2020). The Moral Machine is an experiment implemented on an online platform[4], to explore moral dilemmas faced by autonomous vehicles, using the formal framework of the well-known Trolley Problems. The platform achieved a very large reach, gathering 40 million decisions, in ten languages, from 10 million people in several different countries. In the experiment global moral preferences were summarized in nine different groups, which characterize certain decision-making patterns, like a preference for saving pedestrians, preferring to spare the young, and others. Using the individual variations in preferences based on the demographic data of the participants, transcultural ethical variations were observed, which were grouped into three major groups of countries: *Eastern* (mainly formed by Islamic and Confucian countries and cultures), *Western* (formed by Protestant, Catholic, and Orthodox countries in Europe), and *Southern* (formed by Latin American countries in Central and South America and also several African countries). The results revealed marked differences between the preferences among the three groups.

Findings like those of Awad et al. only reinforce the idea that we live in a morally pluralistic world. Given this reality, how can we reach a consensus between different cultures and individuals? To preserve our cultural and moral pluralism it is important to develop techniques to aggregate moral preferences and solve moral conflicts. Some tools to help us deal with this problem can be obtained from areas such as Social Choice Theory, Expected Utility Theory, and Voting Theory. In this study, we will investigate heuristics to reach a (social) consensus or (individual) decision in situations of moral uncertainty, namely: *how to act when we don't know how to act*?

---

[3] A search on "Google Scholar" can show that the results for "empirical uncertainty" (4,150,000 results) double the ones related to "moral uncertainty" (2,180,000 results) in August 2020.
[4] Available at: https://www.moralmachine.net/ Accessed on: August 25, 2020.

2nd-Order normativity

First of all, it is important to define some terms. Normativity is something that implies a need for action, something that should be pursued, such as "what should be" or "what should be done". When we talk about normative reasoning we are talking about a form of decision-making that is based on some kind of normative principle, such as deontological rules, a utility function, or common sense itself. A more systematic approach to the study of normativity would be metaethics, which also tries to define what the nature of "good" is, or rather the nature of normative statements. While first-order normative statements like "killing's are wrong" clearly imply a form of action or behavior, defining "wrong" as something that should not be done, metaethical statements, or metanormative statements, deal with questions like "What is wrong? What is goodness? What does it mean for something to be wrong or right?".

Metanormativity can be defined as the study of norm structures in general, when we refer to the metanormative we are referring to *norms about norms*, i.e., sets of norms that can help in situations where there is uncertainty about 1st-order norms (ŻURADZKI, 2016). Metanormative strategies are heuristics to evaluate between different first-order normative structures. If we think of ways to aggregate preferences, values, different moral theories, we need a metanormative strategy to accomplish such a process. However, it's important to make a distinction between moral pluralism (BERLIN, 1997) and moral uncertainty. In situations of moral uncertainty adopting a pluralistic metanormative strategy is a possibility. Instead of distributing our belief between two (or more) different moral theories (e. g., Kantianism and Utilitarianism), the agent can unite both theories into a new first-order normative theory (KING, 2008). An individual can be equally convinced given the merit of two different moral theories, having no uncertainty about which is the more valid, and importing principles from both theories into his new theory (e. e., quasi-utilitarianism). For example, the subject believes that the maximization of "well-being" should be pursued given certain deontological restrictions, such as "lying is wrong", or "don't murder children". However, to aggregate two or more moral theories is not the same as being uncertain. In the above case, the subject did not doubt the merit or value of Kantianism or Utilitarianism, so there is no moral uncertainty. Only if the individual possesses uncertainty about the validity of some Y principle in comparison with X principle, then we can assign moral uncertainty to this agent.

An agent is under normatively uncertainty when several moral theories point to different or conflicting decisions, so that this agent can then use a 2nd-order norms to solve a moral dilemma. A plausible conclusion from this argument is: "an agent may also be uncertain about which 2nd-order norm to apply", in which case we would need "meta-metanormativity" (3rd-order norms). Thus, we can see that the concept of normativity, ethics, rules in general, implies an infinite hierarchy of norms that a "strongly uncertain" agent may have to recursively explore. Should the possibility of an n-order hierarchy discourage metanormative reasoning? Are there normative questions that can lead to infinite recursions, such as "Is God good"? Perhaps during the process of evaluating different moral theories, and becoming uncertain about the validity of one theory versus another, we become obliged to use a 2nd-order normative rules, and by becoming uncertain about which 2nd-order normative rule to use we have to resort to a higher metanormative level *ad infinitum*. However, we argue that such cases are the exception contrary to the norm. Hardly moral questions have no form of influence in the physical world so that no subjective or objective attribution of probability can be made. Another argument against the problem of infinite recursion is that the agent only needs to regress until a decision can be made. This is imperative in any scenario where the agent is rationally limited since only agents with infinite rational capabilities could perform such recursive processes. If the agent follows the recursive path through the metanormative hierarchy and reaches the n_{max}-order, its "maximum normative epistemic reach", there being no convergence, then this agent is irreparably uncertain.

Metanormative Strategies

One of the most know metanormative strategies to deal with the problem of moral uncertainty is "My Favorite Theory" (GUSTAFSSON & TORPMAN, 2014). To exemplify it, let us imagine the following problem, involving a dilemma that many people dabbling with vegetarian ideas have to face:

Ana finds herself in a moral dilemma. Ana is undecided about whether to buy a meatloaf or a cheese loaf, and Ana has beliefs in different moral theories. Ana has 30% belief in a moral theory (T_1) that assigns moral value to cattle life. Meanwhile, Ana has 70% belief in another moral theory (T_2) that does not assign any moral value to the life of cattle. The utility of a meatloaf and cheese loaf for Ana are $\$10$ and $\$5$ respectively, for both moral theories. According to T_1 the death of a cow is evaluated as $-\$100$, making a meatloaf worth $-\$90$. According to the T_2, Ana only needs to choose between a cheese

loaf ($5) and a meatloaf ($10), because the value of cattle life is not considered. *What should Ana do*?

	T_1 – 30%	T_2 – 70%
Meatloaf	−90	10
Cheese loaf	5	5

My Favorite Theory (MTF) proposes that we make our choice based on the moral theory that we have the greatest belief, thus, in the dilemma above Ana using the MTF strategy would choose the moral theory T_2, and would buy the meatloaf. An obvious question for the reader might be: "can't we do better than that?", and how should Ana act if her beliefs in T_1 and T_2 are the same? In situations where beliefs are equally distributed among the moral theories under consideration. MTF does not provide us with a satisfactory solution. We assume that the option of "throwing a fair coin" would not be a moral or rational attitude[5]. MTF also recommends the individual to make "morally risky" decisions when her belief is divided almost indifferently. For example, if Ana has 49% belief in $T_1$and 51% belief in T_2, MTF still recommends meatloaf, even though more than 49% of Ana's moral beliefs are committed to a penalty almost 10 times greater than the gain of a meatloaf.

In our view, better solutions than MTF were proposed, such as the theoretical negotiation approach of Greaves and Cotton-Barratt (2019), and "Moral Hedging" (HICKS, 2019). However, in this study, we will explore the propositions made by William MacAskill (2014), known as the Maximization of Expected Choice-Worthiness (MEC), Variance Voting (VV), and Borda Rule (BR). Unlike MTF, MEC, Variance Voting, and Borda Rule are comparative approaches. Comparative approaches suppose that the normative agent decision-making should not be based only on the credence to different moral theories, but also on the degree of choice-worthiness that the theories attribute to different actions. MTF is a noncomparative metanormative strategy. In order that an inter-theoretical comparison between different moral theories can be made, MacAskill first defines different types of moral theories as follows:

[5] Imagine Anna attending animal rights protests on Monday and gutting a cow on Tuesday.

- *Cardinal Theories*: Moral theories are cardinally measurable if beyond an order of preference, "what is better than what" the theory can say how much something is better than the other. That is, besides saying that $A \succeq B$, the theory says how much A is better, through a quantifier (γ):

$$T_i = \{\gamma A \succeq B, \gamma B \succeq C, \gamma C \succeq D, \dots\}$$

For example, $\gamma(100)A \succeq B$, where $\gamma(100)$ means that A is 100 units of value better than B. Consequential moral theories, such as utilitarianism, are examples of Cardinal theories.

- *Ordinal Theories*: moral theories are ordinal if they only present an ordinal preference relationship, i. e:

$$T_i = \{A \succeq B \succeq C \succeq D, \dots\}$$

For example, an ordinal theory, as some deontological version of Kantianism may dictate that "lying is wrong". However, such a theory does not tell us how much worse lying is than another action, it just provides us with an order of preferences;

$$\text{Deontological System}_i = \{\text{lying} \succeq \text{steal} \succeq \text{assault} \succeq \text{kill}\}$$

Moral deontological theories are usually ordinal theories.

We can say that cardinal theories are the ones that provide us with the most information, since besides a preference ranking, they make available to us a comparative magnitude between preferences, while ordinal theories are the least informative normative structures. The best kind of situation in a decision-making under moral uncertainty is when we have to compare different moral theories that are cardinal and are inter-theoretically comparable (it is when we have more information). Inter-theoretical comparability refers to the fact that not always cardinal theories are inter-theoretically comparable, that is, there is not always a nonarbitrary exchange rate between the units of "Choice-Worthiness" between theories.

Maximization of Expected Choice-Worthiness, Variance Voting, and Borda Rule

To solve the decision problem of moral uncertainty, and the problem of inter-theoretical comparability, MacAskill (2014) recommends the following methods:

1) *Maximization of Expected Choice-Worthiness* (MEC): the MEC is used if all the moral theories considered by the agent are cardinal and inter-theoretical comparable theories;
2) *Variance Voting* (VV): used when the moral theories under consideration are cardinal but not inter-theoretically comparable;
3) *Borda Rule* (BR): if all the theories under consideration are ordinal. It is important to realize that it is possible to reduce a cardinal theory to an ordinal. However, much information (the magnitude of preferences) is lost in the process.

All of the strategies cited aim at maximizing the "expected choice-worthiness" of the decision-maker under of moral uncertainty. This value is the decision-makers' belief in a particular moral theory, multiplied by the moral value of a certain action. In the ideal case, where the theories evaluated are all cardinal and inter-theoretically comparable, the value of the expected choice-worthiness of an action, $\mathrm{EC(A)}$, is given as follows:

$$\mathrm{EC(A)} = \sum_{i=1}^{n} \mathrm{C(T_i)CW_i(A)}$$

Where $\mathrm{C(T_i)}$ represents the credibility (belief) of the decision-maker in $\mathrm{T_i}$ (some particular moral theory), while $\mathrm{CW_i(A)}$ represents the "choice-worthiness", according to $\mathrm{T_i}$, of A (an action that the decision-maker can choose). Let's use Ana's example again in her choice between buying a meatloaf or a cheese loaf, divided between two different moral theories, $\mathrm{T_1}$ (30% belief) and $\mathrm{T_2}$ (70% belief), which have different opinions about the moral value of animal life. The moral values attributed to each action are described again below:

	$\mathbf{T_1 - 30\%}$	$\mathbf{T_2 - 70\%}$
Meatloaf	-90	10
Cheese loaf	5	5

Using the MEC we arrive at the following result:

$$\mathrm{EC(Meatloaf)} = (0.3 \times -90) + (0.7 \times 10) = -20$$

$$\mathrm{EC}(\text{Cheese loaf}) = (0.3 \times 5) + (0.7 \times 5) = 5$$

In case the moral theories evaluated do not have a consistent exchange rate between units of choice-worthiness, MacAskill proposes to first normalize in some way the choice-worthiness values. In areas like statistics, normalization is common practice when we are evaluating values measured at different scales, so normalization brings the values to a new common "fictitious" value, so we normalize the moral theories evaluated before calculating the traditional MEC. Normalization by Variance Voting is done by the variance of the choice-worthiness values in each moral theory. However, other forms of normalization are possible, such as standardization or Z-score. Variance is a measure that tells us about the scattering of data distribution, that is, how far the scored values of each moral theory tend to be from the mean. Normalizing by variance intuitively means letting each moral theory individually choose its exchange rate, using the dispersion of its CW values as a scale. Thus, the VV of an action is calculated as follows: first, we obtain the "mean value" of the CW values of a given moral theory ($\mathrm{CW_M}$), then we calculate the variance (σ) by adding the quadratic differences from the mean, divided by the number of possible actions (n):

$$\sigma_{\mathrm{T_i}} = \frac{\sum(\mathrm{CW_i} - \mathrm{CW_M})^2}{\mathrm{n}},$$

$$\mathrm{VV_{T_i}} = \frac{\mathrm{CW_i} - \mathrm{CW_M}}{\sigma_{\mathrm{T_i}}}$$

Let's imagine the example used above again, but now the $\mathrm{T_1}$ and $\mathrm{T_2}$ theories are not inter-theoretically comparable. Ana has the same belief distribution between the moral theories as to the previous example ($\mathrm{T_1} = 30\%$ and $\mathrm{T_2} = 70\%$), the utility of a meatloaf and cheese loaf for Ana are $\$10$ and $\$5$ respectively. However, moral theory $\mathrm{T_2}$ assigns 100 times more value to meat ($\text{Meatloaf} = \$1000$ and $\text{Cheese loaf} = \$5$) than $\mathrm{T_1}$. *What should Ana do*? According to VV we first need to normalize the values by variance and then apply the MEC:

$$\sigma_{\mathrm{T_1}} = \frac{(-90 - (-42.5))^2 + (5 - (-42.5))^2}{2} \approx 2{,}256$$

$$\sigma_{\mathrm{T_2}} = \frac{(5 - (502.5))^2 + (100 - (502.5))^2}{2} \approx 247{,}506$$

$$\mathrm{VV}_{\mathrm{T}_{1(\mathrm{Meatloaf})}} = \frac{-90 - (-42.5)}{2256.25} = -0.02$$

$$\mathrm{VV}_{\mathrm{T}_{1(\mathrm{Cheese\ loaf})}} = \frac{5 - (-42.5)}{2256.25} = 0.02$$

$$\mathrm{VV}_{\mathrm{T}_{2(\mathrm{Meatloaf})}} = \frac{1000 - 502.5}{247{,}506} = 0.002$$

$$\mathrm{VV}_{\mathrm{T}_{2(\mathrm{Cheese\ loaf})}} = \frac{5 - 502.5}{247{,}506} = -0.002$$

	$\mathbf{T_1 - 30\%}$	$\mathbf{T_2 - 70\%}$
Meatloaf	-0.02	0.002
Cheese loaf	0.02	-0.002

Using the MEC in the normalized choice-worthiness values we have the following result:

$$\mathrm{EC}(\mathrm{Meatloaf}) = (0.3 \times -0.02) + (0.7 \times 0.002) = -0.0046$$

$$\mathrm{EC}(\mathrm{Cheese\ loaf}) = (0.3 \times 0.02) + (0.7 \times -0.002) = 0.0046$$

Normalization by variance allows an inter-theoretical comparison between moral theories with completely different scales of value, allowing moral theories themselves to define their exchange rate based on how much the choice-worthiness values are distributed. In the above case, Ana again should choose the cheese loaf, because of the high variance in T_2 that decreases the normalized values of choice-worthiness for each available action. That is, a high variance in the choice-worthiness values causes a penalization in the MEC evaluation in theories with very sparse distributions in choice-worthiness. Now, for the case where we have to compare moral theories, some cardinal and others ordinal, the method that MacAskill recommends is the Borda Rule (BR). In BR the information of consequentialist theories, which give the magnitude of a preference, is lost, and we can only count on the ordering of preferences of each moral theory.

We will use as an example the following case, adapted from MacAskill (2014, p. 63), but still involving a vegetarian dilemma: Ana is going to dinner, and she has a considerable belief that animals are worthy of moral value. However, Ana is also divided into going to a

steakhouse (she has not yet fully transitioned to vegetarianism but sympathizes with the cause). The steakhouse is closer than the vegetarian restaurant. Besides, in the middle of the way between the steakhouse and the vegetarian restaurant, there is a fast-food franchise where maybe there are vegetarian options, but not as healthy as the vegetarian restaurant options. Ana is hungry and has just left home, what restaurant should Ana go?

[S]: the Steakhouse;
[V]: the vegetarian restaurant;
[F]: the fast-food franchise.

Ana has greater credibility in that eating meat, given her current state of hunger, is morally justifiable according to a variant of Utilitarianism. At the same time, Ana has a strong belief in a Common sense moral theory that prefers the vegetarian restaurant, or secondly, the fast-food which may have some vegetarian option. And finally, Ana has less belief in a moral Deontological theory which dictates that as long as she doesn't eat meat, the sooner she can satisfy her hunger the better. Ana's distribution of credibility among the moral theories she credits are:

35% of credibility in a variant of utilitarianism, $\mathrm{T_{UT}} = [\mathrm{S} \succeq \mathrm{V} \succeq \mathrm{F}]$;
34% of credibility in a variant of common sense, $\mathrm{T_{CS}} = [\mathrm{V} \succeq \mathrm{F} \succeq \mathrm{S}]$;
31% of credibility in a deontological theory, $\mathrm{T_{DE}} = [\mathrm{F} \succeq \mathrm{V} \succeq \mathrm{S}]$.

According to MTF, the right choice is to eat at the steakhouse ($\mathrm{T_{UT}} = 35\%$) tries to find the best decision using Voting Theory tools, and within this theory, the "gold standard" is the Condorcet method. A voting system uses a Condorcet method when: if most voters prefer A to B, then A is the Condorcet winner (the elected one), if there are multiple candidates, then the candidate who is preferred in all possible pairs of comparisons $(\mathrm{A} \succeq \mathrm{B}, \mathrm{A} \succeq \mathrm{C}, \mathrm{A} \succeq \mathrm{D} \ldots \mathrm{A} \succeq \mathrm{Z})$ will be the Condorcet winner. This method compares all possible pairs of preferences (candidates) and declares the winner the preference that outperforms all others in a head-to-head tournament. However, it is not always possible for a Condorcet winner to emerge, so extensions to this method are necessary. The Condorcet method is also susceptible to the voting paradox, which occurs when the aggregation of social preferences becomes non-transitive, even if the preferences of all individuals is transitive. That is, even if all voters have transitive preferences in an election $(\mathrm{A} \succeq \mathrm{B}, \mathrm{B} \succeq \mathrm{C}, \mathrm{A} \succeq \mathrm{C})$ the final result can still be non-transitive $(\mathrm{A} \succeq \mathrm{B}, \mathrm{B} \succeq \mathrm{C}, \mathrm{C} \succeq \mathrm{A})$, the voting paradox is a classic example of the composition fallacy (just because

all voters have transitively ordered preferences that don't mean that the social choice will be transitively ordered) (GEHRLEIN & VALOGNES, 2001).

Condorcet extensions, such as the Condorcet Minimax method and the Schulze method (LEVIN & NALEBUFF, 1995), exist to solve the voting paradox. However, as much as these methods are the best alternatives in electoral systems, they are not appropriate for metanormative decision strategies. That is because elections rarely have to deal with a group of voters whose numbers fluctuate, and on the contrary, our beliefs (the electorate) constantly vary among different moral theories. MacAskill (2016) argues that, at the very least, increasing our belief in a particular moral theory should not warm the ordering of preferences of that theory, something that Condorcet extensions do not accomplish very well. Condorcet extensions like Condorcet Minimax can make the "best preference" of moral theories with greater credence as sub-optimal options (MACASKILL 2014, pp. 68-71). Given this limitation of Condorcet extensions, MacAskill proposes the following fitness condition:

- *Consistency Update*: For all possible preferences and actions A, and for all moral theories T_i, if A is maximally appropriate according to T_i, and the decision-maker increases his credibility in T_i, keeping the relationships of his beliefs proportional among all other moral theories, A should still be maximally appropriate.

As Condorcet extensions fail to preserve this condition of consistency update, this shows that Condorcet extensions are not appropriate for metanormative decision strategies. Therefore, we use the Borda Rule. This forme of score counting is generally the way championships of sporting events, like football, are judged, where we assign points in ascending order to favorite (best placed) competitors. To visualize the Borda Rule, and how it evaluates votes differently than Condorcet extensions, we can imagine a competition with n candidates, where each pair of possible candidates will face one against the other. In this situation, the Condorcet Minimax selects as the winner the candidate whose greatest pairwise defeat is smaller than the greatest pairwise defeat of any other candidate, using the magnitude of the biggest defeat as a tie-breaker criterion. The Borda Rule simply adds up the number of points of all the candidate's victories in all possible contests/pairs (in the case of football, the winner is the one with the highest number of goals scored in victories). Thus, in a tournament where competitors C_1 and C_2 are the two best, and C_1 won all

matches (including C_2 by 1×0), but with a very small margin (scored few goals), this tournament evaluated by the Condorcet Minimax method would attribute the victory to competitor C_1. However, C_2 was a much better competitor, lost only to C_1, but won all the other games with many points (goals) of difference. Condorcet Minimax gives much more weight to the victory and not the magnitude of the victories, while for the Borda Rule method the winner would be C_2. The argument in favor of the Borda Rule is that the magnitude of a victory should matter in preference elections when only ordinal theories are being evaluated, the moral theories being like the voters, ordering their preferences with a variable electorate (the credibility they have) and the actions being the candidates.

The definitions of the Borda Rule for decisions in a situation of moral uncertainty are as follows (MACASKILL, 2016):

- *The Borda Score* of option A, for any T_i theory, is equal to the number of possible options worse than A according to T_i, less the number of possible options better than A according to T_i;
- *The credence-weighted Borda Score* of an option A is the sum, by all moral theories T_i that the decision-maker has credence, of the Borda Score of A according to the T_i theory multiplied by the decision maker's belief in T_i;
- *Borda Rule*: An option A is more appropriate than an option B if, and only if, A has a higher credence-weighted border score than B. If A and B have the same credence-weighted border score, A and B are equally appropriate.

Let's go back to Ana's situation, with a dilemma between three possible restaurants (steakhouse, vegetarian, fast-food), evaluated according to three different moral theories as follows:

35% of credibility in a variant of utilitarianism, $T_{UT} = [S \succeq V \succeq F]$;

34% of credibility in a variant of common sense, $T_{CS} = [V \succeq F \succeq S]$;

31% of credibility in a deontological theory, $T_{DE} = [F \succeq V \succeq S]$.

	$\mathbf{T_{Utilitarianism}}$	$\mathbf{T_{Common-sense}}$	$\mathbf{T_{Deontological}}$	**Credence Weighted Borda Score**
Steakhouse	$2-0=2$	$0-2=-2$	$0-2=-2$	-0.6
Vegetarian	$1-1=0$	$2-0=2$	$1-1=0$	0.68
Fast − food	$0-2=-2$	$1-1=0$	$2-0=2$	-0.08

The result of the table above can be explained as follows: in the case of $T_{UT} = [S \succeq V \succeq F]$, and the "Steakhouse" decision, the utilitarian theory has two options lower than the "Steakhouse" option and none higher, so $2-0=2$, a value which we multiply by Ana's credibility in $T_{UT} = 35\%$. To know the credence-weighted Borda score of each action, we add the contributions of each moral theory to each possible action:

$$BR(\text{Steakhouse}) = (0.35 \times 2) + (0.34 \times -2) + (0.31 \times -2) = -0.6$$

$$BR(\text{Vegetarian}) = (0.35 \times 0) + (0.34 \times 2) + (0.31 \times 0) = 0.68$$

$$BR(\text{Fast} - \text{food}) = (0.35 \times -2) + (0.34 \times 0) + (0.31 \times 2) = -0.08$$

The Borda Rule recommends Ana to walk a little more and satisfy her hunger in the vegetarian restaurant. That's because both the "Steakhouse" and "Fast-Food" options were the least preferred actions by at least one of the three moral theories that Ana has considerable credence, while the choice to go to the vegetarian restaurant was not the least preferred of any theory.

Integrating MEC with Expected Utility Theory

The formalism created by MacAskill (2014) is similar to the formalism of rational choice theory and Expected Utility Theory (von NEUMANN & MORGENSTERN, 1944), which allows us to integrate the two models into a single model, capable of assessing both the empirical uncertainty and the moral uncertainty of the decision-maker. If such assumptions can be accepted, then the theoretical principles of rational choice theory can be applied to metanormative strategies and reasoning under moral uncertainty (LOCKHART 2000; ROSS, 2006; SEPIELLI, 2009; BYKVIST, 2017). Critics of this view consider "duty" as something purely moral (HARMAN, 2015; WEATHERSON, 2002).

However, these authors do not provide a way in which decisions under moral uncertainty can be made by only applying moral principles. After all, what would be a moral agent, if not a rational agent that makes its decisions according to its normative beliefs and preferences? What a rational and moral decision-maker can do in situations of uncertainty is to dissolve the problem into its empirical and moral components. For example, in the case where Ana needed to choose between a meatloaf and a cheese loaf, several forms of empirical uncertainty can be added to the problem, such as:

- What is the probability that if Ana stops buying the meatloaf, this will have a positive effect (less animal suffering) on the environment? Perhaps the lack of consumption will cause even worse situations for animals in captivity.
- How sure is Ana about the sentience of large mammals like cows and bulls?
- Could low meat consumption have even worse negative influences on human lives?
- How much does the consumption of cheese, which is a dairy product (probably cow milk), harms animal life?

All of these questions can help Ana's choice because some of the conclusions of these facts can help to change Ana's beliefs about a moral theory that places more value on animal life, or not. That is, the ability to acquire more information helps the agent to restrict the space of moral theories to those that best represent the real world. Thus, we integrate the MEC model with the expected utility theory as follows[6]:

$$\mathrm{EC}(\mathrm{A}) = \sum_{\mathrm{i}=1}^{\mathrm{n}} \sum_{\mathrm{j}=1}^{\mathrm{n}} \mathrm{P}(\mathrm{O_j} \mid \mathrm{A}) \mathrm{CW_i}(\mathrm{O_j}) \mathrm{C}(\mathrm{T_i})$$

Where, $\mathrm{EC}(\mathrm{A})$ is the value of the expected choice-worthiness of an action A, $\mathrm{C}(\mathrm{T_i})$ is the credibility of the decision-maker in moral theory $\mathrm{T_i}$. And now instead of valuing the action, we value the $\mathrm{CW_i}(\mathrm{O_j})$ observation, i. e., the $\mathrm{P}(\mathrm{O_j} \mid \mathrm{A})$ consequence of action A according to $\mathrm{T_i}$. Now we have a model that unifies empirical and moral uncertainty, where to find the

[6] We would like to point out that the first one to suggest this integration, to our knowledge, was Michael Aird, a Research Fellow at the Center on Long-Term Risk, in his LessWorng post "Making decisions when both morally and empirically uncertain". Available at: https://www.lesswrong.com/s/4NFwxwzLzpiikfkk3/p/eYiDjCNJrR3w3WcMM Accessed on: August 25, 2020.

action with the highest choice-worthiness. The agent now also evaluates each possible result for the action taken (i. e., the purchase of meatloaf increasing rather than decreasing the suffering of animals), multiplying by the value that the result would bring (according to moral theory T_i), multiplied by the credibility of the decision-maker in moral theory T_i. Let us now return to the example of Ana and her vegetarian dilemma to exemplify this approach:

Ana has the same belief distribution between the moral theories of the first example ($T_1 = 30\%$ and $T_2 = 70\%$). In T_1, the choice to eat meatloaf causes $-\$100$ in Ana, and both theories guarantee $\$10$ for eating the meatloaf and $\$5$ for the cheese loaf. Now about the empirical uncertainty, let's imagine that Ana believes with 80% credibility that buying meat increases animal suffering, and 20% chance that buying cheese leads to the same result. *What should Ana do*? According to MEC, integrated with the Expected Utility Theory:

$$\mathrm{EC(Meatloaf)} = (0.8 \times -90 \times 0.3) + (0.8 \times 10 \times 0.7) = -16$$

$$\mathrm{EC(Cheese\ loaf)} = (0.2 \times 5 \times 0.3) + (0.2 \times 5 \times 0.7) = 1$$

Again the model recommends that Ana should choose the cheese loaf. Just like we did with the MEC model, this extension can be spread to calculate several empirical uncertainties, and can also be applied in a heuristic way. If the agent is able to assign any naive notion of value and probability, we can come to the conclusion that a small risk in making a big "evil" does not justify the small gain of a meatloaf. We believe that such a form of reasoning seems a promising application of rational choice theory to decision problems involving moral uncertainty. The integration of the MEC model with expected utility theory can also be applied to Variance Voting and the Borda Rule. In the case of the Variance Voting, we only need to remember that the final choice-worthiness values must first be normalized by the variance. In the same way, the Borda Rule can also be extended, we just need to take into account the empirical uncertainty of the facts, for example, what credibility does Ana have that the vegetarian restaurant is open? What is the probability that the fast-food has a vegetarian menu? We just need to multiply this new probability values by each relevant consequence and multiply the result by each moral theory that Ana credits. The similarities between Maximization of Expected Choice-Worthiness and Maximization of Expected Utility may serve as an example that morality and rationality, at least in a pragmatic sense, are not dichotomic concepts.

Conclusion

The methods proposed by William MacAskill, Maximization of Expected-Choice-Whortines, Variance Voting, and Borda Rule, are promising metanormative strategies, and they are not the only ones. Perhaps these are the kind of tools we need to define a way to carry out reflective normative reasoning in states of moral uncertainty. The similarity of MEC with Expected Utility Theory allows simple and intuitive integration of both methodologies. That's because MEC was most likely inspired by rational choice theory. Perhaps this is even what defines a good process of rational and moral decision, the careful analysis of the parts that involve the problema. Marcus Aurelius Antoninus, in his Meditations, has a passage that reflects on this idea: "*nothing is as productive for the elevation of the mind as being able to examine methodically and truly every object that presents itself to you in life*"[7]. However, the metanormative project it's not finished, many rules still seem arbitrary, and perhaps they can be improved, such as normalization by variance, and the choice between methods that are Condorcet extensions or not. Depending on the chosen method, different moral theories and preferences will be better ranked.

Perhaps the reader believes that this method expresses a certain bias towards consequentialism. However, we argue that what is truly stated is that cardinal theories possess a greater amount of information about the moral value of possible actions and outcomes. Even so, an inter-theoretical comparison is still possible, and deontological models can be worked within the MEC model through the Borda Rule. From a metanormative point of view, moral deontological theories (Ordinal), are a specific case of a more general class of normative structures (Cardinal). The metanormative analysis also allows us a redefinition of concepts such as morality, so we suggest the following:

- Agents can distribute their beliefs among moral theories as they wish, and moral theories can order choice-worthiness values of actions/observations in any way (moral pluralism). However, if the decision-maker is under moral uncertainty, and after updating its choice-worthiness values, chooses an action less valuable than another available action, according to its limitations and the moral theories it has credence on, the agent is non-normative. Or at least we cannot assign choice-worthiness values and moral beliefs to its choices consistently.

[7] Meditations, by Marcus Aurelius Antoninus. Book 3, 11.

One last point we would like to mention is the need for the moral agent to deal with bounded rationality and the lack of logical omniscience. We defend the idea that subjective probabilities can be attributed, but as it is known, perfect Bayesian inference is something intractable for rationally bounded agents. We believe that a better understanding of concepts such as counterfactuality and uncertain probabilities can help us to develop better normative reasoning heuristics. After all, what hyperpriors can we use to estimate probabilities about unknown events? How do we assign probabilities to *probabilities*? We intend to try to answer these questions in further studies, with the help of concepts like complexity, similarity, and simplicity.

References

AWAD, E.; DSOUZA, S.; KIM, R.; SCHULZ, J.; HENRICH, J.; SHARIFF, A.; BONNEFON, J. F.; RAHWAN, I. The Moral Machine Experiment. *Nature*. 563, 2018. DOI: 10.1038/s41586-018-0637-6

AWAD, E.; DSOUZA, S.; SHARIFF, A.; RAHWAN, I.; BONNEFON, J. F. Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proc. Natl. Acad. Sci.*, 2020 DOI: 10.1073/pnas.1911517117.

BARRY, C. & TOMLIN, P. Moral uncertainty and permissibility: Evaluating Option Sets. *Canadian Journal of Philosophy*, 46 (6), 1-26, 2016. DOI: 10.1080/00455091.2016.1198198

BARRY, C. & TOMLIN, P. *Moral Uncertainty and the Criminal Law*. In K. Ferzan & L. Alexander (Eds.), *Handbook of Applied Ethics and the Criminal Law*. New York, USA, Palgrave, 2019.

BERLIN, I. The *Proper Study of Mankind: An Anthology of Essays*. HARDY, H. HAUSHEER, R. (eds.) Chatto and Windus. pp. 238, 1997. ISBN 0701165278.

BYKVIST, K. Moral uncertainty. *Philosophy Compass*, 12(3), 2017. Disponível em: https://doi.org/10.1111/phc3.12408

GEHRLEIN, W. & VALOGNES, F. Condorcet efficiency: A preference for indifference. *Soc Choice Welfare,* 18, pp. 193–205, 2001. Available at: https://doi.org/10.1007/s003550000071 Accessed in: August 30, 2020.

GREAVES, H. & COTTON-BARRATT, O. A bargaining-theoretic approach to moral uncertainty. *Global Priorities Institute*, Oxford, UK, 2019. Available at: https://globalprioritiesinstitute.org/wp-content/uploads/2020/Cotton-Barratt_%20Greaves_bargaining_theoretic.pdf Accessed in: August 30, 2020.

GUSTAFSSON, J. E. & TORPMAN, O. In Defence of My Favourite Theory. *Pacific Philosophical Quarterly*, 95(2), pp. 159-174, 2014. Available at: https://doi.org/10.1111/papq.12022 Accessed in: August 30, 2020.

HARMAN, E. The Irrelevance of Moral Uncertainty. *Oxford Studies in Metaethics*. 10, pp. 53-79, 2015. Available at: http://www.princeton.edu/~eharman/documents/UncorrectedProofsIrrelevanceUncertainty.pdf Accessed in: August 30, 2020.

HICK, A. Moral Uncertainty and Value Comparison. *Oxford Studies in Metaethics*, 13, 2018. DOI: 10.1093/oso/9780198823841.003.0008

HICKS, A. Moral Hedging and Responding to Reasons. *Pacific Philosophical Quarterly,* 100 (3), pp.765-789, 2019. Available at: https://philarchive.org/archive/HICMHA Accessed in: August 30, 2020.

KING, I. *How to Make Good Decisions and Be Right All the Time: Solving the Riddle of Right and Wrong*. Bloomsbury Publishing, 2008. ISBN: 9781441149862

LEVIN, J. & NALEBUFF, B. An Introduction to Vote-Counting Schemes. *Journal of Economic Perspectives*, 9(1), pp. 3–26, 1995.

LOCKHART, T. *Moral Uncertainty and its Consequences*. Oxford, UK, Oxford University Press, 2000. DOI: 10.1093/mind/111.443.693

MACASKILL, W. Normative Uncertainty as a Voting Problem. *Mind*, 125(500), 2016. DOI:10.1093/mind/fzv169

MACASKILL, W. *Normative Uncertainty*. Thesis for the degree of Doctor of Philosophy. St Anne's College, University of Oxford, February 2014. Available at: http://commonsenseatheism.com/wp-content/uploads/2014/03/MacAskill-Normative-Uncertainty.pdf Accessed in: August 30, 2020.

MACKIE, J. L. *Ethics: Inventing right and wrong*. UK, Penguin, 1990. ISBN: 0141960094

PARFIT, D. *On what matters: volume one* (Vol. 1). Oxford University Press, 2011. DOI:10.1093/acprof:osobl/9780199572809.001.0001

ROSS, J. Rejecting Ethical Deflationism. *Ethics*, 116, pp. 742–768, 2006. DOI: 10.1086/505234. Available at: https://www.jstor.org/stable/10.1086/505234 Accessed in: August 30, 2020.

SEPIELLI, A. *What to Do When You Don't Know What To Do*. IN Oxford Studies in Metaethics, R. Shafer-Landau (Ed.), Oxford University Press, 2009.

TARSNEY, C. Moral Uncertainty for Deontologists. *Ethical Theory and Moral Practice*, 21 (3), pp. 505-520, 2018. DOI: 10.1007/s10677-018-9924-4

von NEUMANN, J. & Morgenstern, O. *Theory of Games and Economic Behavior*. 1st ed. Princeton, NJ, Princeton University Press, 1944.

WEATHERSON, B. Review of Ted Lockhart's "Moral Uncertainty and Its Consequences". *Mind*, 111, pp. 693–696, 2002.

ŻURADZKI, T. *Meta-Reasoning in Making Moral Decisions Under Normative Uncertainty*. In D. Mohammed & M. Lewiński (Eds.), *Argumentation and Reasoned Action*. College Publications. pp. 1093-1104, 2016.