Meta normativity: Questions of moral and empirical uncertainty

Nicholas Kluge Corrêa[1] e Nythamar de Oliveira[2]

[1]Master of Electrical Engineering and Ph.D. student in Philosophy - PUCRS
nicholas.correa@acad.pucrs.br, ORCID: 0000-0002-5633-6094

[2] Ph.D. in Philosophy (State University of New York, 1994), Full Professor - PUCRS
nythamar@yahoo.com, ORCID: 0000-0001-9241-1031

Graduate Program in Philosophy of the Pontifical Catholic University of Rio Grande do Sul –
Av. Ipiranga, 6681 - Partenon, Porto Alegre - RS, 90619-900.

Abstract

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? An emerging field of study provides an interdisciplinary disposition for questions like these. The area of research in artificial intelligence safety, which within philosophy is one of the contemporary aspects of machine ethics and AI ethics, has been gaining rapid expansion in recent years. One of the main objectives of this area is the alignment of values between artificial autonomous agents and humans, that is, how to imbue our moral preferences robustly and clearly to autonomous processes. At the heart of the alignment problem are several important philosophical issues, issues that go beyond any technical limitation we currently face, one of them being the problem of decision making under situations of moral uncertainty. After all, 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. Given its similarity to the theory of expected utility, we will show that it is possible to integrate both models to address decision-making problems in situations of empirical and moral uncertainty.

Keywords: Moral uncertainty, Normative uncertainty, Meta normativity, Moral pluralism, Maximization of expected choice-worthiness**.**

I. Introduction

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

— Jeremy Bentham

Humanity is vast and multifaceted, our species is currently divided into approximately 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, in a statement[2] said: "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...". Even so, our world is not just conflict, even if this is our status quo, this doesn't mean that we are destined to an existence of eternal conflict and violence. We can learn, cooperation is possible, parties A and B can compromise divergent goals in favor of their common goals, but the possibility of something like this should not overlap the practical complexity of such a task.

One of the most significant differences between societies, and individuals, are their moral values, their preferences, their normative principles, their ethics. The problem of 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?". Another area is also interested in the problem of preference aggregation and moral decision making in the field of Artificial Intelligence (AI) Safety, an emerging research area that has gained popularity in recent years (JOBIN et al, 2019; HAGENDORFF, 2020). The subfields of AI safety research have short and long term interests and applications, ranging from "how to

[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.

[3] A search o "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.

make existing AI techniques safer and more robust?" (AMODEI et al, 2016), to "how to ensure that human values are preserved and understood by super-intelligent artificial agents?" (BOSTRON, 2014). Even so, the common motivation for short- and long-term strategy is the same: "how to make the interaction between humans and AI safe? (JURIĆ et al, 2020). As AI becomes more and more autonomous and proficient, the task of imbuing artificial agents with ethical principles becomes more and more important.

Alignment and value learning is one of the most important long-term research subfields in AI safety. Relevant philosophical issues arise in the context of alignment between AI and humans, promoting a rich area of interdisciplinary study. The real challenge of value alignment is not to identify the "one and true ultimate moral theory", but to understand which principles define a fair and egalitarian form of alignment. One question that arises from this challenge is: "Which, or from whom, values should we align our IAs, and how can we aggregate different moral theories?" Thus, at the heart of the alignment problem, we find not technical problems, but fundamentally philosophical questions, such as:

- What are the characteristics that define a virtuous and moral individual?
- Should the preferences of individuals who act unethically be disregard?
- Which method should be used to rank preferences?
- Should we calculate the preferences of the destitute with more weight?

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 Problem. 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,

[4] Available at: https://www.moralmachine.net/ Accessed on: August 25, 2020.

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) a decision in situations of moral uncertainty, that is: *how to act when we don't know how to act*?

II. 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, meta-ethical statements, or meta-normative statements, deal with questions like "What is wrong? What is goodness? What does it mean for something to be wrong or right?".

Meta normativity can be defined as the study of norm structures in general, when we refer to meta normative we are referring to *norms about norms*, i.e., sets of normas that can help in situations where there is uncertainty about 1st order norms. (ŻURADZKI, 2016). Meta normative 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 meta normative strategy to accomplish such a process. However, it's important to make a distinction between moral pluralism (pluralism of

values) and moral uncertainty. Moral pluralism is the metaethical vision, popularized by the philosopher Isaiah Berlin (1997), which proposes the existence of several different moral values or moral theories. In situations of moral uncertainty adopting a pluralistic meta normative 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 (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 uncertain when several moral theories point to different or conflicting decisions, this agent then can use a 2nd order norm 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-meta normativity" (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 meta-normative reasoning? There are 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...*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 meta-normative hierarchy and reaches the n_{max}-order, its "maximum normative epistemic reach", there being no convergence, then this agent is irreparable uncertain.

III. Uncertainty

There is a dichotomy in the economic literature about the difference between uncertainty and risk, popularized by economist Frank Knight (1885-1972) by the name of *Knightian uncertainty*. Knight (1921), differentiates uncertainty and risk in the following way: risk means an uncertain event, whose chances of occurrence can be predicted and measured, while uncertainty can also be considered an uncertain event, whose chances of occurrence cannot be predicted and measured. That is, risks can be assigned probabilities and uncertainties are impossible to assign such measures (e. g, black swans). Knight's proposal, in a simplistic way, that there are events to which no form of probability can be attributed. We argue against this distinction in the context of moral uncertainty, and similar arguments can be found in the literature (ROSER, 2017).

First of all, we would like to ask if such a distinction makes any sense? Could this distinction be applied in normative decision-making? In the case of the infinite regress problem cited above, "true uncertainty", the epistemic inability to attribute subjective or objective probability to normative statements, would certainly lead us to an impasse (specifically in situations of unlimited rationality). We believe that such distinction of words, like "uncertainty" and "risk", "known unknown" and "unknown unknown" if accepted it can lead to several problems and dubious definitions. For example, can we make a precise distinction between "known" and "unknown" probabilities? What does it mean to assign 100% probability to a 50% probability, and assign 50% probability to a 100% probability? Why does there need to be a dichotomy between risk and uncertainty? Why can't there be a continuum instead of a binary relationship? A less dualistic proposal would be: agents may have more or less information to estimate probabilities, but such epistemic states are not opposites.

We argue that the idea that probabilities cannot be attributed to certain types of uncertainty, especially in the case of phenomena that have a physical effect on the environment, is mistaken. And if there are propositions to which no type of probability

can be attributed, we believe that these should be the minority (and perhaps a minority without any moral relevance). It is understandable that through a frequentist interpretation we cannot attribute probabilities to events that never occurred, "what is the probability of an extraterrestrial invasion occurring by the end of the year?", even so, if there is no objective data we can still use the Bayesian method and infer subjective probabilities that represent our beliefs. In fact, some events are so complex, and dependent on so many factors, that it is difficult to estimate precisely some kind of probability. Let's use as an example the Milky Way space colonization project: *will humanity become an interstellar civilization*? We have no precedent on this (except films and science fiction books), and experts in the field may have extremely divergent opinions, however, it is not as if we knew nothing (LANDIS, 1998).

- How many planets are in the habitable zone in our galaxy?
- How far are the nearest ones?
- What is the atmospheric composition of these planets?
- How close can we travel to the speed of light?
- Can we design ships that transport colonies out of the solar system?
- Can humans survive extremely long space travel?
- What would be the cost of this type of project?
- How likely is it that the human species will end before this happens?

A quick survey on the topics shows that scientists have some knowledge of such issues, some better informed than others. Even for distant, and possibly unlikely events like this, we can assign subjective probabilities. Perhaps after collecting a lot of information we can answer (will humanity become an interstellar civilization?): yes with a 1% probability[5]. Now, what is the probability that humanity will become an interstellar species tomorrow (definitely less than 1%)? Just because we don't have enough information about the concreteness or veracity of a hypothesis doesn't mean that we can't assign probabilities to it. Or else the probability of "humanity will become an interstellar civilization" tomorrow, or 1 million years from now would be the same! And we believe that if we define uncertainty that way, the whole concept of subjective probability and Bayesianism will collapse as a result. We argue that even on under

[5] The authors' subjective estimation does not represent their true beliefs.

Knightian uncertainty the epistemic agent must be able to assign different probabilities to hypotheses such as:

- H_1) God Exists;
- H_2) God exists, and, he is a ball of flying spaghetti[6].

When we are dealing with propositions that have observable consequences, surely there must be a form of epistemic updating so that we can adjust our beliefs about a hypothesis, that is, a way to acquire more knowledge. For example, in the case "will humanity will become an interstellar civilization?" we can use what we already know about space travel, the data related to the Apollo Program, we can evaluate all the questions raised in the list above and what the state of the art about them is. And based on the data collected (our reference model) create a prediction for, say, the next one thousand years. When dealing with subjective (and objective) probabilities we suggest a rule of non-dogmatism, that is, "1 and 0 are not probability measures", nothing is impossible or certain, a principle reminiscent of the philosophy of Radical Probabilism (SKYRMS, 1996). For that the Laplace Rule of Succession seems a likely alternative, for example: if of the 50 conditions stipulated for a "safe" interstellar journey the 50 conditions prove to be intractable with current technology, we should not assign a 0% probability of success in this goal, instead: If X1, ..., Xn+1 are conditionally independent random variables that can assume the value 0 or 1 if we know nothing else about them, $P(X_{n+1} = 1|X_1 + \cdots + X_n = s) = \frac{s+1}{n+2}$, that is, $\frac{51}{52} = 0.98$. So we still have a 2% chance of being successful in the Milky Way colonization project. Non-dogmatism seems to be an important principle for a meta- normative strategy to solve uncertainty situations, by attributing probabilities (our belief) to different moral theories, we should never estimate 0 or 1.

If we think of cases where the influence in the world is zero given the truth or falsity of a proposition, that is, the influence of such a fact cannot be observed, then we would have cases where the accumulation of information would not be useful, and no form of objective probability could be attributed. For example: "How many angels can dance on the head of a pin?", what is the meaning of this question? Would there be any answer more correct than another in an objective sense? Would there be any observable

[6] The laws of probability ensure that the probability of the conjunction of two events is always less than or equal to the probability of only one occurring, i. e., $P(H_1) \geq P(H_2)$.

influence in the world that could help an agent to update his belief regarding the answer to this question? Questions like this are something we cannot evaluate objectively, and any subjective evaluation (even if it is possible in principle) is lacking in practical meaning. Our point is that questions (or propositions) like these, totally unverifiable and not falsifiable, whose truth or falsity has no impact on the real world are vague and meaningless. As Carl Sagan would say:

> [...]Now, what's the difference between an invisible, incorporeal, floating dragon who spits heatless fire and no dragon at all? If there's no way to disprove my contention, no conceivable experiment that would count against it, what does it mean to say that my dragon exists? Your inability to invalidate my hypothesis is not at all the same thing as proving it true. Claims that cannot be tested, assertions immune to disproof are veridically worthless, whatever value they may have in inspiring us or in exciting our sense of wonder (SAGAN, 2011, p. 171).

We believe that the same is applied to normative reasoning, that theories or moral hypotheses whose influence can be observed, that is, that can influence the decision of an agent, can also be subject to subjective probabilistic measures. While questions like "does morality exist?", we suggest that the normative uncertainty generated by questions that have no influence should not be reasoned during the entire meta normative hierarchy, because what matters if morality exists or not (objectively) if the agent continues to act as if he prefers certain things to happen whether they are right or wrong? In this case, perhaps the best we can achieve is to accept a form of projectivisim or moral pragmatism.

IV. Moral and Empirical Uncertainty

What is the difference between normative and epistemic reasoning? To illustrate, we can put ourselves in the following situation of moral uncertainty: what is the moral value, or the importance of the welfare, of animals compared to the moral value of human beings? Questions of moral uncertainty often occur because of our empirical uncertainty about certain issues relevant to the subject, such as: do animals feel pain? Do animals have a conscience? Do animals have emotions? Do animals feel grief when they lose a member of their offspring or group? These questions are not trivial from a scientific point of view, and at the same time, when we are completely clear about these

facts we may still be in a state of moral uncertainty, i. e., divided between moral theories that assign (or not) moral value to non-human beings.

Something difficult to delimit is what differentiates moral uncertainty from empirical uncertainty? After all, as we saw in the example above, one seems to be part of the other, which raises the question of whether there is a difference between moral acting and rational acting, that is, what would be the difference between rationality and morality? Questions that ask the moral value of animals, and other questions of this sort, can be deconstructed into different questions, empirical and moral, e. g., what is the scientific evidence that animals have sentience? Should we assign moral value to sentient beings? We could define the two forms of action as follows:

- *Rational Agent:* the rational agent seeks to choose his actions based on his preferences and beliefs to maximize (choose the alternative less desirable than all the others) it's utility/pleasure/well-being (subjective value assigned to certain results);
- *Moral Agent:* the moral agent seeks to choose actions based on his credited moral theories, to maximize (choose the alternative less wrong than all the others) their moral value.

Certainly, this definition cannot bridge Hume's Is-Ought gap, but from a pragmatic point of view, it can be useful for the normative reasoner. According to the definition suggested above, we can see that there is a similarity between rational and moral action: both employ the concept of preference, i. e., some comparative relationship that differentiates what is better from what is worse, and choice, i. e., intention/action. In situations of empirical uncertainty, we have a robust picture on how to update our hypotheses (Bayesian inference) and make decisions (expected utility maximization), so what would be the analog for cases of moral uncertainty? MacAskill and Ord (2018), and Greaves and Cotton-Barratt (2019) suggest that possibly the distinction between concepts such as “rationality” and “morality” is just a semantic dispute of contextual analysis, where both interpretations, of what “should” be done, end up reaching the same forms of conclusion, being no practical difference between rationality and morality. Reasonably in different contexts, the meaning of the terms is indeed different, but in a pragmatic approach to decision-making under moral uncertainty, we believe that an equivalence between the two concepts can be achieved. Besides, if we see the

question of normative rationality as a question of rationality, we avoid having to define what in fact would be a moral action, and consequently, import all tools from rational choice theory.

The theory of Expected Utility of von Neumann-Morgenstern (1944), first introduced by Daniel Bernoulli in 1738 with the St. Petersburg Paradox, defines as rational, the behavior of an agent who "follows what must be done" according to his beliefs and preferences, trying to maximize his gain, given the constraints of the environment and his constitution. Thus the theory of expected utility, no matter how many interpretations there are of its meaning, seeks to define how a rational agent must act (in a coherent and transitive way) to fulfill his preferences given his beliefs. We then have implicitly a form of normativity in rational theory, in a situation of moral uncertainty, to apply the concept of rationality, we do not need to define any position as correct or wrong, we only need to assume that the moral agent (rational) when in a state of moral uncertainty tries to perform the right action and avoid the wrong act. At the same time, the moral agent is indifferent to correct (or incorrect) actions that have the same moral (or immoral) value. If such assumptions can be accepted, then the theoretical principles of rational choice theory can be applied to meta normative 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.

There is the argument that morality is something transcendent to rational thinking, generally expressed by Hume's guillotine (2009 [1739]). However, if the question of "what should one do when in doubt about what to do?" be considered transcendental and unattainable, then why should we care about morality in the first place (in the practical sense)? We would like to propose an analogy in defense of the applicability of meta normative strategies for solving moral conflicts: π is a transcendental number, it is not algebraic and it is not the solution of any polynomial equation. If a carpenter, when trying to build a wheel for a carriage asks "what numerical value do I need to determine the length of the circumference of my wheel's ?" is answered with something like "this is transcendental knowledge" that answer will not help the carpenter to build his wheel. But a mathematician or an engineer could answer "I can give you a good enough

approach to your problem", which in our view is a much more empathetic and reasonable answer. Any approximation, 3.1415, is better than no answer. In this sense, meta normative strategies are practical ways of getting approximate solutions to (possibly) moral (intractable) problems. If such a decision is "even the right answer" (something reminiscent of Moore's open-ended question argument), maybe in any practical sense something meaningless. Thus, for critics who would argue something like "*Ignoramus et ignorabimus*" to the concept of normative decision making, we counter-argue paraphrasing David Hilbert's motto: "*Wir müssen wissen. Wir werden wissen*". We need to know, and we will know.

V. Meta normative Strategies

As we argued in the last session, the difference between empirical and moral uncertainty can be confusing. However, it cannot be said that no progress or strategy has been proposed to deal with this problem, one of them being "My Favorite Theory" (GUSTAFSSON, TORPMAN, 2014). 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 10U$ and 5U$ respectively, for both moral theories. According to T_1 the death of a cow is evaluated as -100U$, making a meatloaf worth -90U$. According to the T_2, Ana only needs to choose between a cheese loaf (5U$) and a meatloaf ($10U), because the value of cattle life is not considered. *What should Ana do*?

	$\mathbf{T_1 - 30\%}$	$\mathbf{T_2 - 70\%}$
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[7]. 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 (MEW), Variance Voting (VV), and Borda Rule (BR). Unlike MTF, MEW, 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 valuation-choice that the theories attribute to different actions. MTF is a noncomparative meta normative 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:

- *Cardinal Moral 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 moral theories.
- *Ordinal Moral 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, na ordinal moral theory, as some deontological version of Kantianism may dictate that "lying is wrong". However, such a theory does not

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

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 moral theories.

We can say that cardinal moral 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 moral theories are the less 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. For example, how can we compare the utility received by a cleaning agent/robot, trained through reinforcement learning (RL) to clean an office ($U_{\text{Cleaningbot}} = \text{collect dirt}$), and an RL (Reinforcement Learning) agent who plays Mario Kart ($U_{\text{AgentMK}} = \text{get first to the finish line}$), how can we compare "collect dirt" and arrive at "first place"? Or, how can we exchange the concept of "pleasure" from hedonic utilitarianism and the concept of "utility" from other forms of utilitarianism? Similarly, how can we compare cardinal theories with ordinal theories, such as consequentialist moral theories and deontological moral theories?

VI. 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*: (MEW): the MEW 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 $(\mathrm{EW(A)})$ is given as follows:

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

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 MEW we arrive at the following result:

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

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

So, according to MEW, Ana should buy a cheese loaf. MEW seems to have a better result than theories like MTF, being a way to aggregate several normative theories even when there is a uniform distribution of belief. Another advantage of the MEW strategy is that it is possible to use it as a kind of "standard rule" or heuristic. For example, let's

imagine that Bob is in a moral stalemate between cheating in his final exam or not. Bob is divided between two moral theories, he has a lot of credibility in a form of utilitarianism that evaluates the action of "cheating" in such a circumstance as acceptable, although, Bob also haves little credibility in another consequentialist theory that evaluates "cheating" as definitely wrong. Bruno may consider that cheating would increase his well-being, however, while this action would only bring a small EW value (small positive value multiplied by a high probability), the action of cheating for the other theory has a high moral "magnitude" (low probability multiplied by a high negative value). Perhaps the best alternative according to MEW in Bob's case is simply to study.

In case the moral theories evaluated do not have a consistent exchange rate between units of EW, 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 MEW. 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 (CW_M), then we calculate the variance (σ) by adding the quadratic differences from the mean, divided by the number of possible actions (n):

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

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

Let's imagine the example used above again, but now the T_1 and T_2 theories are not inter-theoretically comparable. Ana has the same belief distribution between the moral theories as to the previous example ($T_1 = 30\%$ e $T_2 = 70\%$), the utility of a meatloaf

and cheese loaf for Ana are 10U$ and 5U$ respectively. However, moral theory T_2 assigns 100 times more value to meat (Meatloaf = 1000U$ and Cheese loaf = 5U$) than T_1. *What should Ana do?* According to VV we first need to normalize the values by variance and then apply the MEW*:*

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

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

$$VV_{T_{1(\text{Meatloaf})}} = \frac{-90 - (-42.5)}{2{,}256} = -0.02$$

$$VV_{T_{1(\text{Cheese loaf})}} = \frac{5 - (-42.5)}{2{,}256} = 0.02$$

$$VV_{T_{2(\text{Meatloaf})}} = \frac{1000 - 502.5}{247{,}506} = 0.002$$

$$VV_{T_{2(\text{Cheese loaf})}} = \frac{5 - 502.5}{247{,}506} = -0.002$$

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

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

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

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

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 MEW 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 vegetarian dilemmas: 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?

- [C]: An goes to the Steakhouse;
- [N]: An goes to the vegetarian restaurant;
- [I]: An goes to 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 ($A \succeq B, A \succeq C, A \succeq D \ldots A \succeq Z$) 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 ($A \succeq B, B \succeq C, A \succeq C$) the final result can still be non-transitive ($A \succeq B, B \succeq C, C \succeq A$), 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 normative meta-decision strategies. That is because elections rarely have to deal with a group of voters whose numbers vary, 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 normative meta decision strategies. Therefore, we use the Borda Rule. 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, for the Borda Rule method, the winner would be C_2. The argument in favor of the Borda Rule is that the magnitude of 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, $T_{DE} = [F \succeq V \succeq S]$.

	$T_{Utilitarianism}$	$T_{Common-sense}$	$T_{Deontological}$	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.

VII. Integrating MEW with Expected Utility Theory

The formalism created by MacAskill (2014) is similar to the formalism of rational choice theory, 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. As we discussed in the session Moral Uncertainty and Empirical Uncertainty the difference between both concepts can be vague and difficult to clarify, especially if we try to define another way beyond rationality for decisions on moral uncertainty. 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 MEW model with the expected utility theory as follows[8]:

[8] 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.

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

Where, $\mathrm{EW(A)}$ is the value of the expected choice-worthiness of an action A, $\mathrm{C(T_i)}$ is the credibility of the decision-maker in moral theory $\mathrm{T_i}$. And now instead of valuing the action, we value the $\mathrm{CW(O_j)}$ observation, i. e., the $\mathrm{P(O_j \mid A)}$ consequence of action A according to $\mathrm{T_i}$. Now we have a model that unifies empirical and moral uncertainty, where to find the 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 $\mathrm{T_i}$), multiplied by the credibility of the decision-maker in moral theory $\mathrm{T_i}$. 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 ($\mathrm{T_1} = 30\%$ e $\mathrm{T_2} = 70\%$). In $\mathrm{T_1}$, the choice to eat meatloaf causes $-100\mathrm{U\$}$ in Ana, and both theories guarantee $10\mathrm{U\$}$ for eating the meatloaf and $5\mathrm{U\$}$ 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 MEW, integrated with the Expected Utility Theory:

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

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

Again the model recommends Ana choose the cheese loaf. Just like we did with the MEW 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 MEW model with the expected utility theory can also be applied to Variance Voting and the Borda Rule. In the case of the Variance Voting, remembering that the final 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, al least in a pragmatic sense, or not dichotomic concepts.

VIII. Conclusion

Human beings are always subject to uncertainty, whether empirical or moral. What this study has tried to show is that a sharp distinction between concepts such as rationality and morality may not be necessary, in the same way, that differentiating "uncertain probabilities" from "known probabilities" is something, from a pragmatic point of view, without meaning. We argue that the agent is capable, at least by a Bayesian subjective sense, of inferring probabilities, well, or poorly informed. These are the basic assumptions needed to defend a meta normative strategy to solve problems of moral uncertainty. The global landscape of human morality is something pluralistic, however, perhaps a pluralistic approach is not possible when trying to build global ethical guidelines, as in the case of AI ethics. As the voting paradox shows, even if all the preferences of all cultures could be expressed in a coherent and transitive way, we have no guarantee that the final set of preferences will be transitive.

The methods proposed by William MacAskill, Maximization of Expected-Choice-Whortines, Variance Voting, and Borda Rule, are promising meta normative 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 MEW with Expected Utility Theory allows simple and intuitive integration of both methodologies. That's because MEW 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 problem 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*"[9]. However, the meta normative project it's

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

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 MEW model through the Borda Rule. From a meta normative point of view, moral deontological theories (Ordinal), are a specific case of a more general class of normative structures (Cardinal). The meta normative analysis 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.

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.

Acknowledgments

We would like to thank the Academic Excellence Program (PROEX) of CAPES Foundation (Coordination for the Improvement of Higher Education Personnel), and the Graduate Program in Philosophy of the Pontifical Catholic University of Rio Grande do Sul, Brazil.

Funding

Research supported by the Academic Excellence Program (PROEX) of CAPES Foundation (Coordination for the Improvement of Higher Education Personnel), Brazil.

References

AMODEI, D.; OLAH, C.; STEINHARDT, J.; CHRISTIANO, P.; SCHULMAN, J.; MANÉ, D. Concrete problems in AI safety. ArXiv preprint, 2016. Available at: https://arxiv.org/abs/1606.06565 Accessed in: August 30, 2020.

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.

BOSTROM, N. Superintelligence: Paths, dangers, strategies. OUP Oxford, 2014. ISBN: 9781306964739

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.

HAGENDORFF, T. The Ethics of AI Ethics: An Evaluation of Guidelines. Minds and Machines, 30, 99–120, 2020. Available at: https://doi.org/10.1007/s11023-020-09526-7 Accessed in: August 30, 2020.

HARMAN, E. The Irrelevance of Moral Uncertainty. In 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.

HUME, D. Tratado da Natureza Humana. Tradução de Débora Danowiski. 2ª ed. São Paulo: Editora da UNESP, 2009 (1739).

JOBIN, A.; IENCA, M.; VAYENA, E. The global landscape of AI ethics guidelines. Nat Mach Intell, 1, 389–399, 2019. Available at: https://doi.org/10.1038/s42256-019-0088-2 Accessed in: August 30, 2020.

JURIĆ, M.; ŠANDIĆ, A.; BRCIC, M. AI safety: state of the field through quantitative lens. 43rd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2020. Disponível em: https://arxiv.org/ftp/arxiv/papers/2002/2002.05671.pdf 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

KNIGHT, F. H. Risk, Uncertainty, and Profit. Boston, MA: Hart, Schaffner & Marx; Houghton Mifflin Company, 1921.

LANDIS, G. A. The Fermi Paradox: An Approach Based on Percolation Theory. Journal of the British Interplanetary Society, London, 51, pp. 163-166, 1998.

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.

MACASKILL, W. ORD, T. Why Maximize Expected Choice-Worthiness?. Noûs. 10.1111/nous.12264, pp. 1–27, 2018.

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

ROSER D. The Irrelevance of the Risk-Uncertainty Distinction. Sci Eng Ethics. 23(5), pp. 1387-1407, 2017. DOI:10.1007/s11948-017-9919-x

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.

SAGAN, C. The Demon-Haunted World: Science As a Candle in the Dark. Editora Random House Publishing Group, 2011. ISBN: 0307801047

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.

SKYRMS, B. The structure of radical probabilism. Erkenntnis, 35, pp. 439–60, 1996.

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.

About the authors

Corrêa, N. K is a Ph.D. student in philosophy, and a Master in electrical engineering, currently working in the area of machine ethics at PUCRS University in Brazil. He has an academic interest in the area of machine ethics and the problem of moral alignment between AI and humans.

De Oliveira, N.F. Ph.D. in Philosophy from the State University of New York at Stony Brook. He is currently a full professor at the Pontifical Catholic University of Rio Grande do Sul (PUCRS), Coordinator of Philosophy at CAPES (2018-22), Coordinator of the Neurophilosophy Research Group (Brain Institute, InsCer), Editor of Veritas magazine, member of the Clinical Bioethics Committee and member of the coordinating committee of the Brazilian Center for Research in Democracy, created in 2009.