# Implementation of Moral Uncertainty in Intelligent Machines

Kyle Bogosian
kbogosia@tulane.edu

**Abstract:** The development of artificial intelligence will require systems of ethical decisionmaking to be adapted for automatic computation. However, projects to implement moral reasoning in artificial moral agents so far have failed to satisfactorily address the widespread disagreement between competing approaches to moral philosophy. In this paper I argue that the proper response to this situation is to design machines to be fundamentally uncertain about morality. I describe a computational framework for doing so and show that it efficiently resolves common obstacles to the implementation of moral philosophy in intelligent machines.

## Introduction

Advances in artificial intelligence have led to research into methods by which sufficiently intelligent systems, generally referred to as artificial moral agents (AMAs), can be guaranteed to follow ethically defensible behavior. Successful implementation of moral reasoning may be critical for managing the proliferation of autonomous vehicles, workers, weapons, and other systems as they increase in intelligence and complexity.

Approaches towards moral decisionmaking generally fall into two camps, "top-down" and "bottom-up" approaches (Allen et al 2005). Top-down morality is the explicit implementation of decision rules into artificial agents. Schemes for top-down decisionmaking that have been proposed for intelligent machines include Kantian deontology (Arkoudas et al 2005) and preference utilitarianism (Oesterheld 2016). Bottom-up morality avoids reference to specific moral theories by developing systems that can implicitly learn to distinguish between moral and immoral behaviors, such as cognitive architectures designed to mimic human intuitions (Bello and Bringsjord 2013). There are also hybrid approaches which merge insights from the two frameworks, such as one given by Wiltshire (2015).

## The Problem of Moral Disagreement

A problem which has been cited as an obstacle to the development of top-down approaches (Bello and Bringsjord 2013) as well as to machine ethics more generally (Shulman et al 2009, Bostrom 2014) is the lack of agreement among moral philosophers on which theory of ethics should be followed. The 2009 PhilPapers survey of philosophy faculty revealed that 26% accepted or leaned towards deontology, 24% accepted or leaned towards consequentialism, 18% accepted or leaned towards virtue ethics, and the remaining 32% favored other approaches entirely. For someone who believes in a particular approach to ethics, the correct system for implementation in artificial moral agents may be obvious, but the path forward for society as a whole remains unclear. Companies, governments, and researchers will have to decide which system to use for AMAs and will be faced with a difficult choice between competing moral paradigms.

This paper will describe a computational framework that will determine how we should design moral machines in the light of this disagreement, but first it is necessary to determine exactly why moral disagreement is a problem. Just because there exists disagreement does not imply that research and development programs cannot succeed in their objectives, nor does it imply that we need to worry about differences in opinion. However, there are two reasons for why moral disagreement can pose a serious problem for the project of machine ethics.

First, it could be a pragmatic problem as value differences interfere with projects that require cooperation. In a worst-case scenario, decisions over which research programs to fund will turn into bitter political battles, research agencies will be bogged down in disputes, and developers will split up their resources and devolve into a competitive mindset which reduces information sharing and slows research progress, thus making it difficult if not impossible to construct moral machines.

Second, it could be a moral problem. If we are so uncertain about morality and are split across many different moral principles, then the likelihood that anyone's particular moral system is entirely correct is statistically extremely low (Shulman et al 2009). Therefore, if we do build AIs grounded in any particular moral system, then they will probably be making many poor moral decisions.

Bello and Bringsjord (2013) argue that disagreement provides a reason to avoid top-down approaches to machine ethics in favor of bottom-up or hybrid approaches that copy or take inspiration from human moral thinking. However, it is not clear how this would solve the problem of disagreement. The authors claim that people can agree on examples of good moral behavior despite disagreeing over specific theories, but people's personal moral judgements can differ as widely as moral theories do when people are faced with moral dilemmas (Greene et al 2001) and when they are considering politicized issues such as racial fairness, animal farming, and economic inequality. Therefore, it is perfectly plausible that researchers and government regulators with strong value disagreements will come into conflicts over the kinds of moral intuitions and data which should or shouldn't be included in automated reasoning.

Moreover, in the cases where people do agree on moral choices despite disagreeing over moral theories, there is neither a pragmatic nor a philosophical problem with the top-down approach: a utilitarian engineer has no moral reason to care whether a robot's computational pattern is internally utilitarian or not, as long as it is still maximizing utility; more generally, if we know that machines will act in the morally right way, then we have no reason to consider the specific programming of machine code to be morally significant. So to the extent that moral disagreements between ethical theories might pose pragmatic and moral problems for the development of top-down computational ethics, they would pose equal pragmatic and moral problems for the development of bottom-up computational ethics.

We can illustrate this with a case of self-driving cars which must be programmed to adjudicate between preserving the lives of many pedestrians in the street ahead of the vehicle and saving a single pedestrian on a nearby sidewalk. The act-utilitarian approach to ethics would say that each pedestrian's life is equal, at least in the absence of specific information about the people, and would therefore order a swerve into the one person on the sidewalk in cases where at least two

people were in the street. A hardline deontological approach would say that no matter how many people are in the street, it is wrong to swerve and kill the one person. Modeling the car upon human intuitions and cognition would not give a clear solution, as people have widely differing thoughts about such cases. Training the driving system to determine the correct action through supervised learning would only rephrase the issue by forcing us decide which training examples ought to be classified as right or wrong.

Instead, the engineers and regulators might compromise: based on the number and earnestness of stakeholders on both sides, or the strength of various arguments in moral philosophy, they could agree to order a swerve only in the case where a relatively high threshold of pedestrians was at risk, such as four or five people. This method alleviates the pragmatic problem, as it satisfies everyone with a partial compromise and provides an incentive for people to contribute to machine ethics projects rather than stonewalling them. It ensures that different groups will have to make approximately equal sacrifices to project goals, ameliorating perceptions of unfairness. Whether it solves the moral problem is another issue which we will address in more detail.

**Dealing with Moral Disagreement**

Moral disagreement has also been recognized as a problem for normative ethics proper, independent of any concerns regarding machine ethics. A direct approach to overcoming it is to assume that there is a correct moral theory which we are searching for, acknowledge that we are fundamentally uncertain about which moral theory is correct, and then act in such a way as to give some weight to the judgements of different theories. To continue the self-driving car example, a human driver who is uncertain about whether hardline deontology or act-utilitarianism is right might make up her mind to only swerve in cases where at least four or five pedestrians were in the way, as erring too far on either side could be especially morally problematic.

More generally, given the failure of moral philosophy to reach satisfactory conclusions, it can be argued that we should adopt a framework of reasoning which takes multiple moral views into account (Lockhart 2000). However, moral uncertainty is controversial, and many philosophers have attempted to address the various mathematical and philosophical problems involved with the concept (Żuradzki 2016). In particular, it has been argued that making comparisons between the values and judgements of different moral theories is impossible (Nissan-Rozen 2015).

A recent proposal aimed at answering these concerns was developed by William MacAskill in an extensive thesis (2014). It avoids objections based on incomparability and incommensurability by developing moral uncertainty as a voting problem among moral theories (MacAskill 2016). Voting provides a close analogy with the pragmatic problems faced by machine ethics: as voting is the general process by which decisions are made from the preferences of a population, computations of moral uncertainty represent a process by which agents can act in accordance with the diverse values of humanity. Since it works as a voting system where theories have equal say adjusted by their probability of being correct, it approximates the framework which has been suggested for moral reasoning by Nick Bostrom (2009). The rest of this paper will develop and defend the use of this model for machine ethics.

**Normative Uncertainty**

The scheme presented in "Normative Uncertainty" (MacAskill 2014) is to make action-guiding judgements based on all moral theories in which the agent has some level of credence. MacAskill defines his approach as a *metanormative theory*, characterized as follows:

The agent investigates a *decision-situation* comprising a quintuple $\langle S, t, \mathcal{A}, T, C\rangle$ where $S$ is the decisionmaker, $t$ is the time, and $\mathcal{A}$ is the set of possible actions to take. $T$ is the set of normative theories being considered, where a theory $T_i$ is a function of decision-situations that produces a cardinal or ordinal choiceworthiness ranking of actions $CW_i(A)$ for all actions $a \in \mathcal{A}$. $C(T_i)$ is a credence function assigning values in [0, 1] to every $T_i \in T$. A metanormative theory is a function of decision-situations that produces an ordering of the actions in $\mathcal{A}$ in terms of their appropriateness.

MacAskill distinguishes between theories which assign cardinal rankings to options, as utilitarianism would, and theories which only assign ordinal rankings. When it comes to cardinal theories, MacAskill also distinguishes between sets of theories with moral values which are directly comparable and sets of theories which are incomparable. (The existence of comparable theories is not necessary for the system to work.) He proposes the metanormative theory of maximizing expected choiceworthiness (MEC), the steps of which are as follows:

1. Each set $\mathcal{K}$ of $k$ moral theories in which the choiceworthiness rankings of options are cardinal and intertheoretically comparable are aggregated into single theories, where

$$C(T_{\mathcal{K}}) = \sum_{i=1}^{k} C(T_i)$$

$$CW_{\mathcal{K}}(A) = \frac{\sum_{i=1}^{k} CW_i(A)C(T_i)}{\sum_{i=1}^{k} C(T_i)}$$

In other words, the credence in the new theory equals the sum of the agent's credences in each of the individual theories in the set, and the choiceworthiness of an option according to the new theory is the credence-weighted average of the choiceworthiness of the option in all of the individual theories.

2. The rankings of options according to each ordinal theory are used to provide choiceworthiness scores of options using a modified Borda scoring rule that is designed to properly account for ties.

$$CW_o'(A) = |a \in \mathcal{A} : CW_o(a) < CW_o(A)| - |a \in \mathcal{A} : CW_o(a) > CW_o(A)|$$

This violates the independence of irrelevant alternatives, but MacAskill provides reasons to allow the violation. First, independence of irrelevant alternatives is the least essential of all the axioms present in Arrow's Impossibility Theorem, and if a different axiom were to be violated then there would be no prospects for a satisfactory metanormative account involving ordinal moral theories. Second, the primary motivation for independence of irrelevant alternatives is that

it combats tactical voting, but tactical voting is not a problem with moral theories: theories aren't agents, and a moral agent can't conceal information from itself. Third, there are some moral cases where we would expect the independence of irrelevant alternatives to be violated.

3. The options' scores provided by each of the aggregated sets of cardinal comparable theories, the scores provided by each of the ordinal theories, and the scores provided by each of the cardinal incomparable theories *p* are all divided by the respective standard deviations of the moral rankings which the moral theories (or sets of intertheoretically comparable cardinal theories) provide over a general set $\mathcal{G}$ of moral options. This provides normalized choiceworthiness rankings and is the only possible way of equalizing the value of voting for each value system (Cotton-Barratt 2013), which (as MacAskill argues) performs the role of giving each theory equal say.

$$CW_{\mathcal{K}}^{N}(A) = \frac{CW_{\mathcal{K}}(A)}{\sigma(CW_{\mathcal{K}}(\mathcal{G}))}$$

$$CW_{o}^{N}(A) = \frac{CW_{o}'(A)}{\sigma(CW_{o}'(\mathcal{G}))}$$

$$CW_{p}^{N}(A) = \frac{CW_{p}(A)}{\sigma(CW_{p}(\mathcal{G}))}$$

A note on $\mathcal{G}$: it is necessary for variance normalization that a representative set of actions be defined for computing the variance. This provides the background against which theories can tell if a decision is comparatively important or comparatively unimportant. MacAskill thinks that a broad account featuring many actions in this set, rather than just the ones under consideration in the present decision-situation, is theoretically desirable. Here I model the process as if we are providing moral machines with large arrays of data representing moral decisions across the spectrum of human and machine experience, so that calculating the variance only needs to be done once before the resulting number is used and shared by many agents.

Variance normalization also violates the independence of irrelevant alternatives, but it's a necessary violation for this application. In order to perform comparisons across moral theories, we need to determine the amount of weight which a value system places on a particular issue compared to other issues. Since different computational approaches to ethics can have wildly differing numerical outputs, it would be impossible to do this fairly without normalizing.

4. Each option is scored with the credence-weighted average of the variance-normalized scores provided by the normative theories. This provides a ranking of expected choiceworthiness:

$$E(A) = \frac{\sum_{i=1}^{n} CW_{i}^{N}(A)\, C(T_i)}{\sum_{i=1}^{n} C(T_i)}$$

Then the agent chooses the action *A* which maximizes *E*(*A*). An agent could be given a broad directive to maximize or otherwise respond to the expected choiceworthiness of all its actions.

The algorithm is linear with respect to the number of actions and linear with respect to the number of moral theories, as adding another action requires a fixed series of operations to be done with each moral theory and adding another moral theory requires a fixed series of operations to be performed on each action. We might expect the variety of actions available to agents to increase as they become more complex and more intelligent, though it's not clear if the number of credible moral theories will significantly increase in the future.

The effect of this system is that an agent will make prudent decisions that avoid the most severe infractions in accordance with various moral theories. For instance, when presented with the opportunity to commit a great deception in order to gain a small increment of happiness, an agent will refrain even if they view consequentialism as more likely than nonconsequentialist theories, because the relative wrongfulness of the act according to nonconsequentialist theories is significant whereas the relative benefit under consequentialism is minor. Likewise, even if an agent believes that it is probably not morally obligated to refrain from actions which cause significant indirect harms to others, it will nevertheless refrain from being careless in such a way, because the scale of the moral harms it would be committing if it were wrong would be significant.

Such a scheme would provide balanced judgements about optimal moral behavior. Since an agent under this directive is most likely to follow a given moral theory in cases where the decision at hand is particularly important to that moral theory, there is also a sound game-theoretic reason to promote this model in a world of heterogeneous moral attitudes: it would provide compromises that seek to satisfy the most critical interests of any common value set. The full description and defense of each step in the theory is given by MacAskill (2014).

## On How to Build an Uncertain Machine

### *Ordinal and cardinal scoring*

One of the requirements of this framework is that moral theories actually provide rankings over actions. Perhaps the majority of moral theories provide neither cardinal nor ordinal scores at first glance. However, this is a surmountable issue. First, it simply seems obvious that just about any moral theory should be able to rank certain impermissible actions, such as torturing a large number of people, as worse than other impermissible actions, like stealing someone's coffee. So any theory should at least accommodate ordinal rankings across moral actions in order to be satisfactory. But even if a full ranking is not possible, we could still assign a rudimentary score system rooted in deontic logic, such as 0 for all impermissible actions and 1 for all permissible actions.

Defining the exact numbers would require varying degrees of input from human operators depending on the complexity of the moral theory. However, computational approaches towards morality may provide numerical values and rankings which are otherwise absent from moral theory. Computational reasoning differs from human thinking and the procedures required for providing moral judgements in artificial intelligence systems may involve functions from which meaningful ordinal, integer or real-valued scores for actions can be extracted.

However, a top-down approach to machine ethics is not required with this proposal. The 'theories' implemented in the system need not all be explicit moral theories in the philosophical sense. One or more bottom-up decisionmaking systems might be included; for instance, the output of a complicated learned function can be treated as a choiceworthiness ranking. As long as it ranks different actions by normative criteria, it satisfies the criteria proposed by MacAskill for being a moral theory.

*Credences*

The outcomes of comparisons will always crucially depend upon the credence function *C*, and there are two possible ways to create this function, mirroring the two manifestations of the problem of moral disagreement.

The first approach would be to envision the machine as an ideal reasoning agent which aims towards philosophical correctness as much as possible. The machine could start by referencing the positions of moral philosophers, which would make the credence in consequentialism 0.24 and the credence in deontological ethics 0.26. Refining these probabilities could be achieved with a Bayesian approach combining a prior belief for a moral theory with new evidence. Possible sources of evidence include insights regarding people's moral beliefs from experimental philosophy and cognitive neuroscience and deduction or induction conducted by a moral machine. Lengthy and difficult approaches which rely on extensive human input are acceptable, as information modifying the credence function could be generated just once and then distributed to many agents via an update. However, a machine designed to be purely ethical could fail to act in ways that are profitable or otherwise satisfy the interests of its creators, so this model may be unfeasible to use.

The other approach is to envision the machine's reasoning as a voting system taking the interests of its stakeholders into account, so that they represent the values of the corporate shareholders, research faculty or government agencies which are responsible for the project. There could be compromises between legal requirements, owner interests, lobbyists, and other affected parties. However, this opens the door for ethical values to be determined by corporate and political interests, which may be morally unacceptable. Therefore, the right approach would be a synthesis of the philosophical and pragmatic frameworks, where both philosophical theory and sociopolitical systems are given weight.

*Limited approaches*

Implementing this system of normative uncertainty would require multiple moral decisionmaking systems to function in a moral agent. Some systems could be quite simple, such as a self-driving car programmed with traffic laws as a deontological module and a basic utilitarian calculator for death and injuries. Other systems could be more complex. The complexity of deriving appropriateness scores is linear with respect to the number of actions under consideration and linear with respect to the number of moral theories; implementing a large number of moral theories may be too computationally expensive for the framework to be practical in all applications.

More basic systems of value comparisons involving computational shortcuts and heuristics would be easier to implement, and may also seem to be more philosophically defensible by dint of being simpler. In reality, they will have to sacrifice some theoretical properties in order to achieve this simplicity, such as equal say among theories, fine-grained representation of moral values, or additional axioms of Arrow's Theorem.

The pragmatic and moral problems of disagreement must generally be addressed for machine ethics to be successful, so comparisons and decisions between value systems will still have to be made in some fashion even if some machines use simplified approaches. MEC represents one of the most recent and philosophically rigorous methods for adjudicating between different value systems, and has desirable theoretical properties including equal say and effective comparisons across all theories. Therefore, the system here can serve as the standard against which simplified schemes are judged.

## Long Term Progress in Ethics and Artificial Intelligence

Sufficiently competent machines which adhere to a particular value function are likely to pursue it endlessly with pathological consequences (Bostrom 2012). While I do not aim to provide a long term solution for value selection and alignment for arbitrarily intelligent agents, it is nevertheless valuable to start by designing systems that cooperate among different value systems, as that will improve the fail-safety of agents which become very intelligent but have imperfect goal functions (Gloor 2016). The framework proposed here clearly fulfills that criterion.

However, as machines become more advanced, they might perform more moral functions autonomously, such as updating credences in moral theories and generating new ones, depending on the degree to which progress in artificial intelligence enables them to investigate questions in moral philosophy. MacAskill points out that MEC provides an intrinsic motivation for agents to actively pursue moral truth, so not only would a morally uncertain system be cooperative, but it would also be likely to update and improve its moral beliefs. This is desirable insofar as we think there might be cases of widespread moral wrongdoing which humans have consistently failed to identify (Williams 2015), which machines should be incentivized to avoid.

## Objections

### *Skepticism about Moral Uncertainty*

Numerous objections can be leveled against MEC. One could argue that moral theories still cannot be compared in a voting process as it assumes they can, that assigning numerical rankings to actions is a post-hoc process not supported by all moral theories, that infectious nihilism prevents metanormative theories from having any grounding (MacAskill 2013), or that there is simply no compelling reason to act with regard to uncertainty. These claims are contentious and are yet to be fully explored in the literature. However, they all point towards the same position: that we are not obligated to give consideration to moral uncertainty, or at least not in the way that MEC implies.

Answering these arguments is beyond the scope of this paper, but even if we accept them, they don't provide us with a reason to abandon the proposal described here. The question of how an ideal agent ought to act does not alter the question of how we can most efficiently and fairly overcome the issues surrounding moral disagreement in machine ethics. Again the analogy with voting is illuminating; very few people believe that the winner of an election is always the one who would be the best leader, but most people agree with the system, as it is a necessary and effective compromise. As long as we disagree with each other about ethics, we should still build moral machines based on uncertainty even if we reject the idea as a guide for human behavior.

The system in this paper is useful for addressing the problem of disagreement even if one happens to be completely confident in a single moral theory and completely dismissive of moral uncertainty. Suppose that someone is a fanatic with the choice between pushing for her chosen value system in artificial moral agents and pushing for the compromise described here. If she pushes for her own values, there is no incentive for other parties to compromise with her, and they will push for their own values as well. In the long run, this results in a patchwork of agents with competing values; some fraction of agents, *x*, will consistently adhere to her values based on her share of agent production, implying that a randomly distributed proportion *x* of all agents' actions will be judged according to her values. This is equivalent to a metanormative theory which randomly does what her particular theory demands *x* proportion of the time, while ignoring her values otherwise.

On the other hand, suppose she agrees to compromise through the framework described in this paper, and wins a credence that is equal to her share of agent production. Then a portion *x* of the decisionmaking weight in each agent will be in accordance with her values. The resulting system will be more ethically desirable according to her own criteria because the voting weight will be focused on the decisions which are most important according to her own values, as opposed to being channeled into a randomly selected fraction of all agents' actions. So no matter how little consideration one gives towards moral uncertainty, the framework of moral uncertainty is still more desirable than accepting competition among value systems.

*Infinite Regress*

A further objection would not necessarily claim that MacAskill's approach is incorrect, but just that it is contentious, as evidenced by the aforementioned arguments. Therefore, the objection goes, we have not actually eliminated disagreement, but we have merely removed it from the moral domain to the meta-moral domain—where pragmatic and moral (or meta-moral) problems still arise. People may disagree over what framework of moral uncertainty to use, how to assign credences, how to score options, how to construct the general set $\mathcal{G}$, or other issues. Any framework for compromising over these disputes will also be subject to disagreement, prompting further disputes, and so on.

While there is certainly potential for the pragmatic problem to arise, I claim that it is much less severe in the meta-moral case of disagreement than it is in the moral case. This is evidenced by the relative lack of controversy over voting methods and procedures in organizations and political institutions. While there are efforts to reform voting procedures in Western democracies

such as the United States, these campaigns are much less popular and much less violent than object-level campaigns over contentious ideological and policy disputes.

In machine ethics, while different people may advocate for different systems of moral uncertainty, it won't be clear which one would best achieve anyone's particular value system, and disputes over core moral principles will be overshadowed by more mundane issues like fairness, computational complexity and the axioms of voting theory. So the pragmatic problem is mostly, though not completely, solved. Certainly there are large challenges ahead, but the point is that they are more technical than they are partisan.

The moral problem is not completely solved either – we don't know if the system of moral uncertainty presented by MacAskill would lead to the morally best outcome, and the answer may be different depending on whose value system is actually correct. In fact, since it makes compromises among moral claims, it is almost guaranteed to commit moral errors at some point. However, given our lack of knowledge about morality, it represents a solid best guess that maximizes the expected choice-worthiness of our agents' actions. The only alternatives are to unilaterally assume a single value system (whether top-down or bottom-up) with large risks of controversy and moral failure, or to choose a method of meta-moral compromise which sacrifices some of the desirable attributes of MEC while failing to avoid the pitfalls mentioned here.

**Summary**

I have investigated the problem of moral disagreement and explored why and how it poses an obstacle to our efforts to design artificial moral agents. I have shown that simple approaches like relying on human intuition do not remove the problem, and I have described an alternative strategy of maximizing expected choiceworthiness derived from literature on social choice and moral uncertainty which has been extensively defended by William MacAskill. I have argued that it is a feasible and desirable system for providing guidance in moral machines, with various pragmatic and moral advantages in the short term as well as the long term. Finally, I have defended it against objections of skepticism and regress regarding moral uncertainty.

**References**

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, 149-155. http://doi.org/10.1007/s10676-006-0004-4

Arkoudas, K., Bringsjord, S., & Bello, P. (2005). Toward ethical robots via mechanized deontic logic. *Machine Ethics: Papers from the 2005 AAAI Fall Symposium*, 17–23. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Toward+ethical+robots+via+mechanized+deontic+logic#0

Bello, P., & Bringsjord, S. (2013). On How to Build a Moral Machine. *Topoi*, *32*(2), 251–266. http://doi.org/10.1007/s11245-012-9129-8

Bostrom, N. (2009). Moral uncertainty – towards a solution? http://www.overcomingbias.com/2009/01/moral-uncertainty-towards-a-solution.html

Bostrom, N. (2012). The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. *Minds and Machines*,*22*(2), 71-85. http://doi.org/10.1007/s11023-012-9281-3

Cotton-Barratt, O. (2013). Geometric reasons for normalising variance to aggregate preferences. http://users.ox.ac.uk/~ball1714/Variance%20normalisation.pdf

Greene, J. D., Sommerville, R. B., Nystrom, L., Darley, J., & Cohen, J. (2001). An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science*, *293*(5537), 2105–2108. http://doi.org/10.1126

Gloor, L. (2016). Suffering-focused AI safety: Why "fail-safe"measures might be our top intervention. *Foundational Research Institute, Report FRI-16-1*. https://foundational-research.org/files/suffering-focused-ai-safety.pdf

Lockhart, T. (2000). *Moral Uncertainty and Its Consequences*. Oxford University Press. ISBN 978-0195126105

MacAskill, W. (2013). The Infectiousness of Nihilism. *Ethics*, *123*(3), 508-520.

MacAskill, W. (2014). Normative Uncertainty. http://commonsenseatheism.com/wp-content/uploads/2014/03/MacAskill-Normative-Uncertainty.pdf

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

Nissan-Rozen, . (2015). Against Moral Hedging. *Economics and Philosophy*, *3*, 1-21. http://doi.org/10.1017/S0266267115000206

Oesterheld, C. (2015). Formalizing preference utilitarianism in physical world models. *Synthese*. http://doi.org/10.1007/s11229-015-0883-1

PhilPapers Foundation (2009). Preliminary Survey Results. http://philpapers.org/surveys/results.pl

Shulman, C., Tarleton, N., & Jonsson, H. (2009). "Which Consequentialism? Machine Ethics and Moral Divergence." *AP-CAP 2009: The Fifth Asia-Pacific Computing and Philosophy Conference*. https://intelligence.org/files/WhichConsequentialism.pdf

Williams, E. (2015). The Possibility of an Ongoing Moral Catastrophe. *Ethical Theory and Moral Practice, 18*(5). 971-982. https://doi.org/10.1007/s10677-015-9567-7

Wiltshire, T. J. (2015). A Prospective Framework for the Design of Ideal Artificial Moral Agents: Insights from the Science of Heroism in Humans. https://doi.org/10.1007/s11023-015-9361-2

Żuradzki, T. (2016). Meta-Reasoning in Making Moral Decisions under Normative Uncertainty. *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon, 2015, 2*. 1093-1104.