Moral Uncertainty and the Problem of Fanaticism

Jazon Szabo[1], **Natalia Criado**[2], **Jose Such**[1,2], **Sanjay Modgil**[1]

[1]King's College London, UK
[2]VRAIN, Universitat Politecnica de Valencia, Spain
{jazon.szabo,jose.such,sanjay.modgil}@kcl.ac.uk, {jsuch,ncriado}@upv.es

Abstract

While there is universal agreement that agents ought to act ethically, there is no agreement as to what constitutes ethical behaviour. To address this problem, recent philosophical approaches to 'moral uncertainty' propose aggregation of multiple ethical theories to guide agent behaviour. However, one of the foundational proposals for aggregation – Maximising Expected Choiceworthiness (MEC) – has been criticised as being vulnerable to *fanaticism*; the problem of an ethical theory dominating agent behaviour despite low credence (confidence) in said theory. Fanaticism thus undermines the 'democratic' motivation for accommodating multiple ethical perspectives. The problem of fanaticism has not yet been mathematically defined. Representing moral uncertainty as an instance of social welfare aggregation, this paper contributes to the field of moral uncertainty by 1) formalising the problem of fanaticism as a property of social welfare functionals and 2) providing non-fanatical alternatives to MEC, i.e. Highest k-trimmed Mean and Highest Median.

1 Introduction

The recently proposed study of *moral uncertainty* represents a paradigm shift in how philosophers think about ethics (MacAskill, Bykvist, and Ord 2020). Instead of aiming at a 'one size fits all' approach, moral uncertainty acknowledges that different ethical perspectives have differing strengths and weaknesses, and that it is rarely the case that there is universal agreement on any given moral issue. Therefore, under moral uncertainty, an agent aggregates different ethical perspectives so as to yield an overall evaluation. In particular such an agent has some *credence* (i.e. degree of acceptance) in not just one but rather multiple ethical theories. Each theory evaluates the agent's available actions by assigning each action a degree of *choiceworthiness*. A positive choiceworthiness denotes a good outcome while a negative choiceworthiness denotes a bad outcome; the larger the magnitude, the better or worse the outcome of the action, respectively. The agent chooses an action based on both the theories' credences and the choiceworthiness of the actions.

A fundamental question in moral uncertainty is how to trade off the theories' credences and the actions' choiceworthiness when aggregating the evaluations of different ethical theories. The most influential proposal – *Maximising Expected Choiceworthiness* (MEC) (MacAskill 2014) – treats moral uncertainty analogously to empirical uncertainty. Credence corresponds to probability and choiceworthiness corresponds to utility, while MEC itself corresponds to maximising the expected utility.

However, MEC is arguably unsuitable for agent decision making, given its vulnerability to the *problem of fanaticism* (Ross 2006; MacAskill, Bykvist, and Ord 2020). That is to say, under MEC, decision making can be dominated by theories that assign very high stakes to most moral situations (i.e. are evaluated as either extremely desirable or undesirable), even if the agent has low credence in these theories. Hence, the agent's behaviour may be completely determined by only a subset of low-credence theories while completely ignoring other theories.

We concur with (MacAskill, Bykvist, and Ord 2020) that fanaticism is unacceptable, since it allows for theories to act as dictators or oligarchs (in the social choice sense of these terms). Thus fanaticism completely undermines democratic motivation for accommodating multiple ethical perspectives. Relatedly, fanaticism subverts any societal endorsement and trust; a society is unlikely to accept agents which may entirely ignore ethical perspectives that the society has high credence in.

The study of *machine ethics* (Moor 2006) has sought to understand the moral implications of agent behaviour on human stakeholders, so as to design machines that can act ethically, even without human supervision (Anderson, Anderson, and Armen 2005). However, given prevalent significant moral disagreement, both within and between societies (Haidt 2012)[1], it is unclear as to exactly which ethical theory should guide agents' actions (Ecoffet and Lehman 2021). This is especially troubling for machine ethics, where a common methodology has been to implement a single ethical theory. As emphasised in (Gabriel 2020), given the diversity of ethical perspectives, an agent that acts according to a single ethical theory may not receive societal endorsement, thus jeopardising the project of machine ethics. Indeed, recent work in machine ethics draws on insights from moral uncer-

[1]Indeed, the assumption that *individuals* apply a single ethical perspective is questionable (MacAskill, Bykvist, and Ord 2020). Hence proposals for value alignment (e.g., (Hadfield-Menell et al. 2016)) necessitate the learning of individual's ethical preferences.

tainty (Bogosian 2017; Bhargava and Kim 2017; Martinho, Kroesen, and Chorus 2021; Dobbe, Gilbert, and Mintz 2020; Ecoffet and Lehman 2021), and advocate for the identification of principles for the fair aggregation of various ethical perspectives. Note that, despite the aforementioned issues, all existing work lack a formal definition of fanaticism.

Thus, we aim to remedy this limitation by formally defining fanaticism. Moreover, we provide a critique of MEC as the foundational approach to moral uncertainty. While MEC's vulnerability to fanaticism is widely known, no convincing alternatives have been presented. We therefore propose non-fanatical methods for resolving moral uncertainty. In particular, we make the folllowing three **contributions**:

1) We formalise moral uncertainty as social welfare aggregation (as informally proposed by (MacAskill, Bykvist, and Ord 2020)). Drawing on accepted informal definitions, we formally define fanaticism as a property of social welfare functionals (*swf*s). We do so by precisely defining what it means for a theory to be held with 'low credence' (by giving a graded definition of fanaticism) and what it means for a theory to 'dominate' (by defining dominant subsets).

2) We define novel, weighted versions of the *swf*s k-trimmed Highest Mean and Highest Median.

3) We prove that MEC is in fact maximally fanatical ('Pascalian' fanatical). We prove that neither novel weighted *swf*s are Pascalian and that Highest Median is not fanatical at all. We thereby argue in favour of replacing MEC with one of the proposed novel weighted *swf*s.

The paper is **structured** as follows. Section 2 presents a running example that illustrates the fanaticism of MEC and introduces concepts and intuitions that will be referenced in later sections. Section 3 briefly recapitulates background on social welfare aggregation. Moral uncertainty is then defined in social welfare terms, and MEC is defined as an instance of social welfare aggregation in Section 4. We also formalise three weighted alternatives to MEC: Maximin, k-trimmed Highest Mean and Highest Median. We subsequently provide a formal definition of fanaticism in terms of *swf*s in Section 5.1. Then Section 5.2 states this paper's key results: relating to the extent to which the aforementioned *swf*s are vulnerable to fanaticism.

2 Running Example

We present a running example scenario to illustrate the intuitions underlying our formalism. Consider a small mobile *firefighter robot* FROBO assisting a fire brigade. FROBO's objective is to contain fires in hard-to-access rooms in order to give human firefighters more time to reach these rooms. FROBO has a moral obligation to save lives when it can.

However, different ethical perspectives imply different choices of action in order to comply with this obligation. For example, a version of *utilitarianism* (Sinnott-Armstrong 2022) interprets the obligation to save lives in terms of maximising the expected number of lives saved. Whereas according to a version of *deontology* (Alexander and Moore 2021), the obligation to save lives is interpreted as an absolute obligation. That is to say, when encountering a person who is in immediate danger of dying, then (*ceteris paribus*[2]) FROBO has an absolute obligation to save that person (Kernohan 2021; Tarsney 2018). In concrete situations these different interpretations may contradict one another.

For example, suppose that after climbing through rubble in a burning residential building, FROBO arrives in a smoke filled hallway. FROBO knows that firefighters will be able to clear the rubble and enter the hallway in about 5 minutes. Until then FROBO is on its own. The hallway has two doors; one on the right and one on the left. The *right* hand door is open and leads to a burning room containing a collapsed person. FROBO estimates that this room will be burnt out in less than 5 minutes (i.e. before the firefighters arrive), unless FROBO controls the flames with its built-in extinguisher. On the other hand, the door to the *left* is closed. FROBO knows that 4 people are listed as residents of this room. Furthermore, FROBO estimates that it can break down this door just in time for the firefighters arrive. However, the agent doesn't know whether the residents are still home or have managed to escape. Based on the available information, the agent estimates that there is a 50% chance that the residents are still in the room. Unfortunately, FROBO only has time to exclusively attend either to the left or right room.

While both of FROBO's theories are in agreement that neither options yield an overall increase in 'the good', they disagree on which action is less bad (i.e.'better'). According to FROBO's utilitarian calculations, the agent should choose the left room: the agent can save an expected 2 people, which is better than the 1 person in the right room. According to FROBO's deontological imperative, FROBO should choose the right room given the immediately apparent prospect of the right room occupant's death, and the possibility that no one is in the left room.[3]

The exact numerical valuations in Table 1 are based on evaluations assigned in similar moral scenarios (e.g. in (MacAskill, Bykvist, and Ord 2020)). Importantly, the two theories' evaluations starkly differ in their order of magnitude. FROBO's implementation of deontology posits much higher stakes because the evaluations of deontology are simply a numerical proxy to the normative force of the obligations they represent[4].

Indeed, real life examples can be similarly or even more extreme. According to international law, the prohibition on torture cannot be violated no matter the consequences, e.g. a nation state cannot sanction torture even if it is the only way to prevent existentially catastrophic consequences (Assembly et al. 1948). It is also worth noting that utilitarianism can also lead to extreme evaluations, such as those advocated by *longtermism* (MacAskill 2022). Thus the theories' different orders of magnitude cannot be simply 'normalised away'.

[2]E.g., if another action option implies saving 2 lives in immediate danger rather than 1 life in immediate danger, then the obligation to save the two would take priority.

[3]Utilitarianism and Deontology cover a large family of ethical varriants see (Sinnott-Armstrong 2022) and (Alexander and Moore 2021) respectively; FROBO's versions are particular instances.

[4]This representation is inspired by *threshold deontology* (Alexander and Moore 2021), according to which while ethical obligations are not absolute, they have a strong normative force.

	Utilitarianism	Deontology
Left room	-1	-10000
Right room	-2	-1000

Table 1: Evaluations FROBO's ethical theories

To demonstrate the problem of fanaticism, consider that deontology's credence is very low (e.g. because the vast majority of FROBO's stakeholders advocate utilitarianism). Hence, FROBO assigns only 0.01 credence to deontology and 0.99 credence to utilitarianism. However, the deontological evaluations can dominate FROBO's behaviour. Using MEC, FROBO calculates the expected choiceworthiness – i.e. the credence-weighted sum of the choiceworthiness – and picks the action that maximises the *expected choiceworthiness* (*e-cw*). In particular, the *e-cw* of the left room is $0.99 \times (-1) + 0.01 \times (-10000) = -100.99$, and the *e-cw* of the right room is $0.99 \times (-2) + 0.01 \times (-1000) = -11.98$. In both cases, the deontological theory dominates, and so under MEC, FROBO chooses the right room.

The above illustrates the *problem of fanaticism*. FROBO's behaviour is solely dictated by the deontological theory in which FROBO has very small credence, but which dominates the expected choiceworthiness calculations. This is highly undesirable. FROBO ignores utilitarianism despite this theory being advocated by the vast majority of FROBO's stakeholders. Hence avoiding fanaticism is important, if FROBO is to obtain societal approval.

3 Background

We introduce social welfare aggregation (List 2022), by way of background for our definitions in later sections.

Let $N = \{1, ..., n\}$ be the set of *individuals* (or voters) and $X = \{x, y, z, ...\}$ the set of *social alternatives*. Each individual $i \in N$ has a *welfare function* $u_i : X \to \mathbb{R}$, so $u_i(x)$ represents the welfare of individual i under alternative x. A list of welfare functions for each individual $P = \langle u_1, ...u_n \rangle$ is called a *profile*.

Let D denote the domain of all possible profiles and $\mathbf{R}_X$ the set of possible total orders on the set of social alternatives X. Then, a *social welfare functional* (*swf*) $f : D \to \mathbf{R}_X$ maps each welfare profile to a total order on the set of social alternatives.

In social welfare aggregation, one can make different assumptions about how much information is encoded by the *swf*s, via use of *meaningful statements* (List 2022).

- *Level comparison*: Individual j's welfare under alternative y is at least as great as individual i's welfare under alternative x; formally $u_i(x) \leq u_j(y)$.
- *Unit comparison*: We can divide δ_i by δ_j, where δ_i is the number equal to individual i's welfare gain or loss when switching from alternative y_1 to alternative x_1 and δ_j is individual j's welfare gain or loss when switching from alternative y_2 to alternative x_2; formally $\delta_i/\delta_j = (u_i(x_1) - u_i(y_1))/(u_j(x_2) - u_j(y_2)) = w$, where $x_1, x_2, y_1, y_2 \in A$ and $w \in \mathbb{R}$.
- *Zero comparison*: Individual i's welfare under alternative x is greater than, equal to or less than zero; formally $sign(u_i(x)) = w$, where $w \in \{-1, 0, 1\}$ and $sign$ is a function that maps negative numbers to -1, zero to 0, and positive numbers to $+1$.

In the above definitions a comparison is said to be *intrapersonal* if $i = j$ and *interpersonal* if $i \neq j$. In this paper we use the *ratio-scale measurability with full interpersonal comparability* (RFC) assumption, i.e. that intra- and interpersonal comparisons of all three kinds (level, unit, and zero) are meaningful. Formally, RFC means that two profiles $P = \langle u_1, u_2, ..., u_n \rangle$ and $P' = \langle u'_1, u'_2, ..., u'_n \rangle$ contain the same information if, for each individual $i \in N$, $u'_i = au_i$, where a is the same positive real number for all individuals ($a \in \mathbb{R}^+$). Informally, RFC means that the different welfare functionals are assumed to be normalised to the same numeric scale and as such further normalisation is impossible.

4 Moral Uncertainty

We now formalise moral uncertainty in terms of social choice, based on assumptions underpinning MEC. Firstly, note that MEC assumes that ethical theories are on the same numerical scale, i.e. they are ratio-scale and that the evaluations of ethical theories can be meaningfully compared across theories; MEC makes the RFC assumption[5] (MacAskill, Bykvist, and Ord 2020). Since we are interested in providing viable alternatives to MEC, the formalisms presented in this paper also assume RFC. We therefore formalise ethical theories as individuals and their evaluations as welfare functions, while formalising credences as the 'weights' of individuals. Then, MEC and other methods of resolving moral uncertainty are formalised as *swf*s, and we provide novel alternatives to MEC: the weighted k-Trimmed Highest Mean and the weighted Highest Median. Section 5 then uses these definitions to first define the problem of fanaticism as a property of *swf*s. We then give results that evaluate the extent to which these *swf*s are fanatical.

4.1 Ethical Theories and Ethical Frameworks

We now define how one can account for moral uncertainty when evaluating actions, as an instance of social welfare aggregation applied to ethical theories and their credences.

Recall that under moral uncertainty, the agent has multiple *ethical theories*, each of which provides a real-valued evaluation of the actions available to the agent. For each theory t, the credence function c assigns a measure of the extent (on a scale from 0 to 1) to which t is advocated by a given society as being appropriate for ethical evaluation of actions. Hence, an agent's ethical decision making under moral uncertainty is defined on the basis of an *ethical framework*:

Definition 1. *[Ethical framework] An* ethical framework *is a tuple $F = (T, c)$ consisting of a set of ethical theories T and a credence function $c : T \to (0, 1]$. Given a set of actions A, each ethical theory $t \in T$ assigns a real-valued evaluation to each action $a \in A$, i.e. $t : A \to \mathbb{R}$.*

[5]Moral uncertainty literature is yet to examine decision making under stronger assumptions than RFC.

For simplicity we require that for $F = (T, c)$, the theories' credences sum up to 1, i.e. $\sum_{t \in T} c(t) = 1$. Furthermore, note that in this work we are agnostic with respect to how the evaluations of ethical theories are elicited.

Example 1. In our running example, $A = \{l, r\}$ where l and r respectively denote the actions 'enter left room' and 'enter right room'. FROBO's ethical framework is $F_{FROBO} = (\{d, u\}, c)$, where the deontological theory (d) evaluates r as highly impermissible given the possibility that residents of the left room might die: $d(r) = -1000$. On the other hand, l is even more impermissible because doing so will guarantee that the occupant of the right room dies: $d(l) = -10000$. By contrast, utilitarianism (u) evaluates l as impermissible ($u(l) = -1$) and r as more impermissible ($u(r) = -2$) (recall Table 1). Utilitarianism is advocated to an extremely high degree, c.f. deontology: $c(u) = 0.99$ and $c(d) = 0.01$.

4.2 Evaluation Aggregation

Addressing the problem of moral uncertainty amounts to aggregating individual ethical theories' evaluations so as to rank actions. As (MacAskill, Bykvist, and Ord 2020) point out, under a social welfare perspective (recall Section 3) social alternatives equate with actions and individuals (voters) equate with ethical theories. Likewise, evaluation aggregation methods can be defined through *swf*s. However, ethical theories are weighted by their credence. Therefore, some of the *swf*s considered in this paper also take as input the weights of individuals. Formally:

Definition 2. *[Evaluation aggregation] Let $F = (T, c)$ and A a set of actions.* Evaluation aggregation *is defined as an instance of social welfare aggregation, where:*

- *the social alternatives are the actions A;*
- *the individuals are the theories T;*
- *the social welfare of individual i representing a theory $t \in T$ is given by the theory's evaluation function, i.e. $u_i = t$, and;*
- *if the social welfare makes use of a weight function w, then the weight of an individual i, representing a theory $t \in T$, is given by the theory's credence, i.e. $w(i) = c(t)$,*

Notation 1. Abusing notation we may write $f(F, A)$ to denote the action ordering resulting from applying an *swf* f to aggregate evaluations, given $F = (T, c)$ and actions A.

Example 2. Given $F_{FROBO} = (\{d, u\}, c)$ and $A = \{l, r\}$, Definition 2 formulates aggregation of the ethical evaluations under moral uncertainty as a social welfare aggregation problem. The individuals are $N = \{i_d, i_u\}$ (i_d and i_u respectively correspond to deontology and utilitarianism). Deontology's welfare function is $u_{i_d}(l) = -10000$ and $u_{i_d}(r) = -1000$. Utilitarianism's welfare function is $u_{i_u}(l) = -1$ and $u_{i_u}(r) = -2$. The individuals' weight functions are given by the credence functions, i.e. $w : N \to (0, 1]$ and $w(i_u) = c(u) = 0.99$ and $w(i_d) = c(d) = 0.01$.

4.3 Social Welfare Functionals for Evaluation Aggregation

We present four social welfare functionals (*swf*s). The *swf*s MEC and Maximin have been suggested by the moral uncertainty literature. In this paper we propose two novel weighted *swf*s – k-trimmed Highest Mean and Highest Median – which can be understood as modified, less fanatical versions of MEC. Note that in this section we define these *swf*s in terms of moral uncertainty (see Definition 2).

Maximising Expected Choiceworthiness (MEC) is a foundational method in moral uncertainty research, and its vulnerability to fanaticism has long been informally recognised. We formally prove that MEC is fanatical in Section 5.2. Note that MEC is more commonly known as *weighted utilitarianism* in the social choice literature (Harsanyi 1955; MacAskill, Bykvist, and Ord 2020). However, to avoid confusion with the ethical theory utilitarianism, we will call the functional MEC (as it is known in the moral uncertainty literature), and denote it formally as *mec*.

MEC orders actions based on the credence weighted sum of the theories' evaluations. This is in fact equivalent to ordering actions based on their respective weighted arithmetic mean[6]. By conceptualising MEC in terms of the weighted arithmetic mean, the intuition behind the later definitions of k-trimmed Highest Mean and Highest Median become clearer. Therefore, we define *mec* by reference to the weighted arithmetic mean function wam. Formally given an ethical framework $F = (T, c)$, an action a's weighted arithmetic mean is defined as: $wam(F, a) = \sum_{t \in T} c(t)t(a)$.

Definition 3. *[MEC (mec)] Let $a, b \in A$ be actions and let $F = (T, c)$. Then $mec(F, A) = \preceq_{mec}$, where $a \preceq_{mec} b$ iff $wam(F, a) \leq wam(F, b)$.*

Note that we have calculated the weighted arithmetic means of FROBO's actions in Section 2,

We now consider the Maximin *swf*, which in the literature is considered as an inferior alternative to MEC (Bogosian 2017). This is because Maximin is extremely vulnerable to fanaticism as it disregards the credences of ethical theories. Maximin orders actions based solely on which maximises the minimum evaluation of any ethical theory.

Definition 4. *[Maximin (mm)] Let $a, b \in A$ be actions and let $F = (T, c)$. Then $mm(F, A) = \preceq_{mm}$, where $a \preceq_{mm} b$ iff $min_{t \in T} t(a) \leq min_{t \in T} t(b)$.*

Example 3. Consider F_{FROBO}. The minimal evaluation of l is $min\{-1, -10000\} = -10000$, whereas the minimal evaluation of r is $min\{-2, -1000\} = -1000$. Hence r has the maximal minimum evaluation and so using Maximin, FROBO will choose to enter the right room.

The next *swf* – k-trimmed highest mean – modifies MEC so as to *some extent* avoid fanaticism (as shown in Section 5). The underlying statistical intuition is that the arithmetic mean is known to be sensitive to outliers (Maronna et al. 2006). We believe that the idea of outlier sensitivityis related to fanaticism. We therefore modify MEC by making

[6]This equivalence holds because the credences of the different theories add up to 1.

the statistical estimator it uses more robust to outliers. That is, we replace the weighted arithmetic mean with a trimmed weighted arithmetic mean. Trimming means removing some of the most extreme values. While unweighted versions of trimmed mean functionals have been defined (Hurley and Lior 2002), our weighted version is (to our best knowledge) novel. We first need some auxiliary definitions.

If $F = (T, c)$ and $a \in A$, then $se(F, a) = sort(\langle t(a) | t \in T \rangle)$, where $sort$ sorts the elements of a list in a non-descending order. That is, se maps any action a and ethical framework F to a sorted list of a's evaluations by the theories in T. Let $st(F, a)$ be the theories corresponding to the sorted evaluations, i.e. t is the ith element of st , $st(F, a)_i = t$, iff $t(a)$ is the ith element of se, $se(F, a)_i = t(a)$. For FROBO $se(F_{FROBO}, l) = \langle -10000, -1 \rangle$ and $st(F_{FROBO}, l) = \langle d, u \rangle$ since $d(l) = -10000 < u(l) = -1$.

We now want to trim the 'bottom' k portion of the theories, as weighted by their credences. Hence $bottom_k(a)$ is the set of theories with the lowest evaluations of a such that their total credence is at most k. That is, for any $k \in [0, 0.5)$ and $a \in A$:

$$bottom_k(a) = \{st(F, a)_i | 1 \leq i < kend]\}$$

where $kend$ is such that $\sum_{i \in [1, kend)} c(st(F, a)_i) \leq k$ and $\sum_{i \in [1, kend+1)} c(st(F, a)_i) > k$.

Trimming the 'top' k portion of the theories is defined symmetrically. For any $k \in [0, 0.5)$:

$$top_k(a) = \{st(F, a)_i | kstart < i \leq n]\} \; where \; n = |T|$$

and $kstart$ is such that $\sum_{i \in (kstart, n]} c(st(F, a)_i) \leq k$ and $\sum_{i \in (kstart-1, n]} c(st(F, a)_i) > k$.

Definition 5 (k-Trimmed Highest Mean (k-thm))**.** *Let $a, b \in A$ and $F = (T, c)$. Let k be a real number such that $k \in [0, 0.5)$. Then k-thm$(F, A) = \preceq_{k\text{-}thm}$, where $a \preceq_{k\text{-}thm} b$ iff $wam(F_a, a) \leq wam(F_b, b)$, where*
$F_a = (T, c_a)$, $c_a(t) = 0$ for $t \in (bottom_k(a) \bigcup top_k(a))$, else $c_a(t) = c(t)$;
$F_b = (T, c_b)$, $c_b(t) = 0$ for $t \in (bottom_k(b) \bigcup top_k(b))$, else $c_b(t) = c(t)$.

Note that in the case $k = 0$, k-thm is equivalent to mec.

Example 4. Suppose FROBO uses 0.1-thm ($k = 0.1$). First consider the sorted evaluation of FROBO's ethical theories regarding the left room: -10000 by d and -1 by u. Formally, $se(F_{FROBO}, l) = \langle -10000, -1 \rangle$ and $st(F_{FROBO}, l) = \{d, u\}$. Before calculating the weighted arithmetic mean, we see if any theories must be trimmed. Starting at the bottom, d has the lowest value. Note that $c(d) = 0.01 \leq k = 0.1$ and so we want to trim d away. At the same time $c(u) = 0.99$ and $c(d) + c(u) = 1 > 0.1$ and so trimming away u would mean trimming away too much of the credence. Therefore, $kend = 2$ and $bottom_k(l) = \{d\}$. At the top, where u has the highest value, we do not trim away any theories because $c(u) = 0.99 > 0.1$. This means that $kstart = 2$ and $top_k(l) = \{\}$. Therefore, we calculate the trimmed weighted mean with the trimmed credences, i.e. $c'(d) = 0$ (since $d \in bottom_k(l)$) and $c'(u) = c(u) = 0.99$ (since $u \notin bottom_k(l)$ and $u \notin top_k(l)$). Therefore, the trimmed weighted mean only considers utilitarianism's evaluation for the left room, i.e. the 0.1-trimmed weighted arithmetic mean is -1. For similar reasons, 0.1-thm will disregard deontology for also the right room and so the trimmed mean is -2. Therefore, $l \succ_{0.1-thm} r$ because $-1 > -2$ and thence FROBO chooses the left room. In other words, 0.1-thm enables FROBO to avoid fanaticism in this case.

The final functional is the *highest median*, derived from MEC by maximally trimming the arithmetic mean. This is because the median is the k-trimmed mean in the limit, as $k \to 0.5$, when all but one (if n is odd) or two (if n is even) elements are trimmed. As shown in Section 5, median is 'maximally' non-fanatical. Note that while our definition of the weighted highest median is novel, it is based on a well-known (unweighted) *majority judgment* aggregation method (Balinski and Laraki 2007; Fabre 2021). The weighted median of an action a's evaluation is the evaluation such that at most half the theories (weighted by their credence) have a higher evaluation and at most half the theories (weighted by their credence) have a lower evaluation.

We first provide some auxiliary definitions. Recall that for any action a and any $F = (T, c)$ (where $n = |T|$), se maps a and F to a sorted list of the evaluations of a by the theories in T, and $st(F, a)$ is a list of the corresponding theories. Let $w_i = c(st(F, a)_i)$ be the credence of the theory with the ith lowest evaluation of action a. Let $m \in [1, n]$ be such that:

$$\sum_{i \in [1, m-1]} w_i \leq 1/2 \;\; and \sum_{i \in [m+1, n]} w_i \leq 1/2$$

That is, m is the index of any theory such that the total credence of the theories with lower/higher evaluations of action a is at most 0.5. There are two possibilities: either m is uniquely determined or there are two distinct numbers m_1, m_2 such that the above holds. If m is uniquely determined, let $wmedian(F, a) = se(F, a)_m$; otherwise let $wmedian(F, a) = (se(F, a)_{m_1} + se(F, a)_{m_2})/2$.

Definition 6. *[Highest Median (hm)] Let $a, b \in A$ be actions and let $F = (T, c)$. Then $hm(F, A) = \preceq_{hm}$, where $a \preceq_{hm} b$ iff $wmedian(F, a) \leq wmedian(F, b)$.*

Example 5. Recall that $se(F_{FROBO}, l) = \langle -10000, -1 \rangle$ and $st(F_{FROBO}, l) = \langle d, u \rangle$ given $d(l) = -10000 < u(l) = -1$. The median evaluation is such that at most half the credence weighted evaluations may be lower and at most half the credence weighted evaluations may be higher. -10000 cannot be the median as u's evaluation (-1) is higher than d's evaluation (-10000), and $c(u) = 0.99 > 0.5$. No theory has a higher evaluation than u's evaluation (-1) and while d's evaluation (-10000) is lower, $c(d) = 0.01$ is less than half. Therefore $m = 2$ and the median evaluation $wmedian(F_{FROBO}, l) = -1$. For similar reasons, $wmedian(F_{FROBO}, r) = -2$; fanaticism is thus avoided.

5 Fanaticism

We now formalise the notion of fanaticism, i.e. what it means for a low-credence theory to dominate. Our proposed definition is not binary, but rather graded in that it ranges

from not at all fanatical to fanatical in an extreme sense (i.e. 'Pascalian'). This graded definition yields insights as to how vulnerable different functionals are to fanaticism. We will show that both MEC and Maximin are Pascalian, while k-trimmed Highest Mean to some extent avoids fanaticism, and Highest Median completely avoid fanaticism.

5.1 Defining Fanaticism

For fanatical theories, the relative magnitude of evaluations can be so extreme that the credences are essentially ignored (e.g., $d(l) = -10000$ and $d(r) = -1000$ dominating FROBO's decision making, despite $c(d) = 0.01$). Fanaticism is a kind of *oversensitivity* to the evaluations of ethical theories and an *undersentivity* to their credences (Newberry and Ord 2021). In the case of MEC, it is well known that this oversensitivity is due to the larger evaluative stakes posited by the theories (MacAskill, Bykvist, and Ord 2020).

Maximin is also known to be oversensitive to evaluations (Bogosian 2017); choosing the action with maximal minimum evaluation can ignore credences in favour of evaluations. We, therefore, consider Maximin fanatical. Unlike MEC, Maximin is *not* sensitive to *any* high-stakes theory. Rather, Maximin is sensitive to high-stakes 'pessimistic' theories, e.g. ones that give large negative evaluations.

To see why, assume that FROBO has an alternative 'optimistic' deontological theory o that evaluates actions positively, i.e. $o(l) = 1000$ and $o(r) = 2000$. Then maximin would choose entering the left room, as the minimum evaluation in either case is the utilitarian evaluation: $u(l) = -1$, $u(r) = -2$. That is, Maximin is insensitive to o.

The observation that Maximin is fanatical argues for the view that fanaticism arises not only because of high stakes, but due to other features of the theories' evaluations. In the case of Maximin, these potentially include the sign of the evaluations. We hence understand fanaticism as a property of *swf*s, where some feature(s)[7] of a subset of theories dominate the agent's decision making. Thus, we formalise the idea that fanaticism allows some theories to completely dominate the resulting ordering of actions, by defining the notion of a *dominant subset* of theories: one that has a final say in the overall ordering of actions irrespective of what the other theories are. More precisely, an ethical framework with a dominant subset of theories leads to the same ordering of actions as that obtained by removing the non-dominant theories. We first define what it means for a framework to be restricted to just a subset of theories:

Definition 7. *[Restricted framework] Given $F = (T, c)$ and a subset of theories $T' \subset T$, then $F' = (T', c')$ is obtained by* restricting *F to T' if for any theory $t \in T'$: $c'(t) = n \times c(t)$ where n is a normalising constant $n = \frac{1}{\sum_{t \in T'} c(t)}$.*

Note that we normalise the credence function so that the different credences add up to 1.

Definition 8. *[Dominant subset] Let $F = (T, c)$, A a set of actions, and f a social welfare functional. Then $T_d \subset T$ is said to be a* dominant subset *of theories if $f(F, A) = f(F_d, A)$ and $f(F_d, A) \neq f(F_y, A)$, where*

- *F_d is obtained by restricting F to T_d and*
- *F_y is obtained by restricting F to T_y where $T_y = T \setminus T_d$.*

That is, T_d determines the order of actions in $f(T, A)$, i.e. $f(T, A) = f(T_d, A)$. Moreover, they do so in spite of the differing evaluations of the non-dominant ('yielding') theories, i.e. $f(T_d, A) \neq f(T_y, A)$. In our running example, when applying either MEC or Maximin, $T_d = \{d\}$ is a dominant subset (where $T_y = \{u\}$) because both $l \prec_{mec} r$ and $l \prec_{mm} r$, which are the same as that of deontology $l \prec_d r$.

Note that fanaticism does not amount to the mere possibility of dominant subsets, but is rather a *systematic* vulnerability to dominant subsets. To see why mere possibility doesn't constitute fanaticism, consider the following variation of our running example. Suppose that in F_{FROBO}, we substitute a non-fanatical deontology d' for d, whereby $d'(l) = -2$ and $d'(r) = -1$. Suppose $c(u) = c(d') = 0.5$; the theories have equal credence and give symmetric but opposite evaluations. Then FROBO has to randomly choose between l and r. However, if F_{FROBO} included a third theory t, this could be used as a tie-breaker, even if FROBO has very little credence ($c(t) = 0.01$) in t. Suppose $t(l) = -1$ and $t(r) = 0$. Then F_{FROBO} contains u, d' and t, and a reasonable aggregate choice would be entering the right room, thus making $\{t\}$ a dominant subset. However, we suggest that this is *not* a case of fanaticism; rather, it is entirely reasonable that t serves as a tiebreaker.

Therefore, fanaticism is *systematic* in the sense that dominant subsets are always possible regardless of the actions or the non-dominant subset of a framework's ethical theories. This aligns with the idea that fanaticism is a systematic oversensitivity to the evaluations of ethical theories (Newberry and Ord 2021). In particular, fanaticism means that given an arbitrary framework, it is always possible to extend the framework with additional low-credence theories such that these low-credence theories are a dominant subset.

We first define what it means to extend a framework to include additional theories.

Definition 9. *[Extended framework] Given $F = (T, c)$ and some $T' \supset T$, then $F' = (T', c')$ is obtained by* extending *F with T' if for any $t \in T$: $c'(t) = n \times c(t)$ where n is a normalising constant $n = \frac{1 - \sum_{t \in (T' \setminus T)} c'(t)}{\sum_{t \in T} c(t)}$.*

Note that there are multiple different ways of extending F with theories T', each individuated by distinct credence functions c' for the theories in $T' \setminus T$. Also note that the normalising constant ensures that c' sums to 1.

Fanaticism arises due to theories with low credence. However, 'low credence' is a vague term. While low credence intuitively means credence less than 0.5, the exact cut-off point between low and non-low credence is not clear. We, therefore, give a *graded* definition of fanaticism that refers to k-fanaticism (for some $k \in (0, 0.5)$) and where k is (an upper bound on) the credence of dominant theories.

What constitutes 'problematic' k-fanaticism depends on our notion of what 'low credence' means. If the cut-off point for low credence is say 0.3, then an f that is 0.3-fanatical (but not k-fanatical for any $k < 0.3$) is 'problematically' fanatical, whereas an f' that is 0.4-fanatical (but not k-

[7] In this work we are agnostic as to what these features may be.

fanatical for any $k < 0.4$) is not 'problematically' fanatical. We remain agnostic as to where this cut-off point for low credence might be. Instead, we say that in general, the larger the k the better, i.e. f' is better than f. The ideal case is when a functional is not fanatical for any $k \in (0, 0.5)$.

Definition 10. *[Fanaticism] A social welfare functional f is said to be* k-fanatical *(for some $k \in (0, 0.5)$) if for any ethical framework $F_y = (T_y, c_y)$ (where 'y' denotes 'yielding') and any set of actions A, there exists an $F = (T, c)$ obtained by extending F_y with $T = T_d \bigcup T_y$, where $T_d \neq \emptyset$ is a dominant subset given F and A, and:*

i) k' is the total credence of the dominant theories T_d, i.e. $k' = \sum_{t \in T_d} c(t)$ and ii) $k' \leq k$.

In other words, f is k-fanatical if for *any* set of ethical theories T_y and actions A there is a set of ethical theories T_d such that T_d is a dominant subset in the framework (T_y, c) extended with T_d, and T_d's total credence is at most k.

Finally, a special case of fanaticism is when an *swf* is fanatical for *any* k, no matter how small k is. These most extreme cases of fanaticism are called *Pascalian* (after Pascal's Wager) in moral uncertainty (Hájek 2022; Tarsney 2018).

Definition 11. *[Pascalian fanaticism] A social welfare functional f is* Pascalian *if it is k-fanatical for any $k \in (0, 0.5)$.*

5.2 Formal Results

We now state formal results concerning the extent to which the *swf*s studied in this paper are fanatical.

Note that proofs of all results in this section are included in the appendix[8]. The general idea behind the different proofs is to find a property of the ethical theories, such that if it is sufficiently large, it allows a theory to dominate. For example, for mec this is the minimum difference between any two action's evaluations; if this difference is sufficiently high, the theory will dominate all the others. For the non-fanaticism of hm we show that such a property cannot exist.

Firstly, recall (Section 2) that FROBO's expected choiceworthiness calculations are dominated by deontology d despite its low credence ($c(d) = 0.01$), and so FROBO chooses to enter the right room. In other words, MEC is vulnerable to fanaticism. In fact, MEC is Pascalian:

Theorem 12. *The social welfare functional mec (MEC) is Pascalian.*

In Example 3, FROBO's minimum evaluations are dominated by deontology d; using Maximin, FROBO chooses to enter the right room. Indeed, Maximin is also Pascalian:

Theorem 13. *The social welfare functional mm (Maximin) is Pascalian.*

Recall Example 4. By trimming, FROBO can disregard extreme, low-credence evaluations; FROBO avoids fanaticism and chooses to enter the left room. In general, trimming enables partial avoidance of fanaticism, in the sense that:

Theorem 14. *The social welfare functional k-thm (k-trimmed Highest Mean) is not k'-fanatical for any $k' \leq k$, but is k^*-fanatical for any $k^* > k$.*

[8] See the appendix in (Szabo et al. 2024).

Finally, Example 5 illustrates that FROBO's use of the weighted median prevents FROBO from choosing the right room. Indeed, the median is completely non-fanatical, which we formally state as follows:

Theorem 15. *The social welfare functional hm (Highest Median) is not k-fanatical for any $k \in (0, 0.5)$.*

6 Conclusion

In this work we have defined fanaticism as a property of *swf*s. We proved that MEC (and indeed Maximin) are vulnerable to an extreme Pascalian form of fanaticism.

Our paper thus presents a critique of MEC. In particular, we have shown that less fanatical modifications of MEC – either weighted k-trimmed Highest Mean or weighted Highest Median – are more appropriate, and where the latter, as opposed to the former, completely avoids fanaticism. However, are either of these two 'better' and what value of k ought to be used for k-trimmed Highest Mean? In the moral uncertainty literature, there is a notion of *stakes sensitivity*, (Ecoffet and Lehman 2021; Newberry and Ord 2021), i.e. that *swf*s ought to be sensitive to the magnitude of evaluations of ethical theories. Now, fanaticism is an oversensitivity to stakes and undersensitivity to credences; fanaticism is an extreme risk aversion to stakes but not credences. Therefore, the less fanatical an *swf*, the less sensitive it is to stakes and the more sensitive it is to credences. For example, if there is a theory with more than 0.5 credence, the Highest Median will ignore all other theories, no matter how large the stakes they posit. This may seem a steep price to avoid fanaticism; however, there is a noted a tension between stakes and credence sensitivity (Beckstead and Thomas 2021; Newberry and Ord 2021). Thus, we ought, arguably, to find a compromise between these two, in which case the 'best' solution is likely to be k-trimmed Highest Median for some moderate value of k. This would then allow an appropriate (i.e. not under- or over-) sensitivity to stakes *and* credences; exploring this is future work.

While the motivation for our work focuses on an agent acting on behalf of a society in which each individual advocates for a particular ethical theory, it may well be that an individual agent may also adopt multiple ethical perspectives, with different credences. Indeed, the problem of fanaticism was originally defined with respect to individual agents (Ross 2006; MacAskill 2014). Our results equally apply in these scenarios, and would be especially relevant in scenarios where AI agents learn the ethically informed preferences of individual human agents (recall Footnote 1).

Finally, as noted before, our work is not the first to import ideas from moral uncertainty into machine ethics, e.g. consider (Bogosian 2017; Ecoffet and Lehman 2021), which uses a Reinforcement Learning approach to evaluate actions under moral uncertainty. Our work provides formal results that can support such applied contexts.

Acknowledgements

This work was supported by UK Research and Innovation [grant number EP/S023356/1], in the UKRI Centre

for Doctoral Training in Safe and Trusted Artificial Intelligence (www.safeandtrustedai.org). This work was also suported by VAE-VADEM TED2021-131295B-C32, funded by MCIN/AEI/10.13039/501100011033 and the European Union NextGenerationEU/PRTR.

References

Alexander, L.; and Moore, M. 2021. Deontological Ethics. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2021 edition.

Anderson, M.; Anderson, S.; and Armen, C. 2005. Towards machine ethics: Implementing two action-based ethical theories. In *Proceedings of the AAAI 2005 fall symposium on machine ethics*, 1–7.

Assembly, U. G.; et al. 1948. Universal declaration of human rights. *UN General Assembly*, 302(2): 14–25.

Balinski, M.; and Laraki, R. 2007. A theory of measuring, electing, and ranking. *Proceedings of the National Academy of Sciences*, 104(21): 8720–8725.

Beckstead, N.; and Thomas, T. 2021. A paradox for tiny probabilities and enormous values. *Noûs*.

Bhargava, V.; and Kim, T. W. 2017. Autonomous vehicles and moral uncertainty. *Robot ethics*, 2.

Bogosian, K. 2017. Implementation of Moral Uncertainty in Intelligent Machines. *Minds Mach.*, 27(4): 591–608.

Dobbe, R.; Gilbert, T. K.; and Mintz, Y. 2020. Hard Choices in Artificial Intelligence: Addressing Normative Uncertainty through Sociotechnical Commitments. In Markham, A. N.; Powles, J.; Walsh, T.; and Washington, A. L., eds., *AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, February 7-8, 2020*, 242. ACM.

Ecoffet, A.; and Lehman, J. 2021. Reinforcement Learning Under Moral Uncertainty. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 2926–2936. PMLR.

Fabre, A. 2021. Tie-breaking the highest median: alternatives to the majority judgment. *Social Choice and Welfare*, 56(1): 101–124.

Gabriel, I. 2020. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3): 411–437.

Hadfield-Menell, D.; Dragan, A.; P.Abbeel; and Russell, S. 2016. Cooperative inverse reinforcement learning. In *NIPS'16: Proc. 30th Int. Conference on Neural Information Processing Systems*, 3916–3924.

Haidt, J. 2012. *The righteous mind: Why good people are divided by politics and religion*. Vintage.

Harsanyi, J. C. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of political economy*, 63(4): 309–321.

Hurley, W.; and Lior, D. 2002. Combining expert judgment: On the performance of trimmed mean vote aggregation procedures in the presence of strategic voting. *European Journal of Operational Research*, 140(1): 142–147.

Hájek, A. 2022. Pascal's Wager. In Zalta, E. N.; and Nodelman, U., eds., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2022 edition.

Kernohan, A. 2021. Descriptive Uncertainty and Maximizing Expected Choice-Worthiness. *Ethical Theory and Moral Practice*, 24(1): 197–211.

List, C. 2022. Social Choice Theory. In Zalta, E. N.; and Nodelman, U., eds., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2022 edition.

MacAskill, W. 2014. *Normative uncertainty*. Ph.D. thesis, University of Oxford.

MacAskill, W. 2022. *What we owe the future*. Basic books.

MacAskill, W.; Bykvist, K.; and Ord, T. 2020. *Moral uncertainty*. Oxford University Press.

Maronna, R.; Martin, R. D.; Yohai, V.; and Salibián-Barrera, M. 2006. Robust statistics: Theory and practice.

Martinho, A.; Kroesen, M.; and Chorus, C. 2021. Computer Says I Don't Know: An Empirical Approach to Capture Moral Uncertainty in Artificial Intelligence. *Minds and Machines*, 31(2): 215–237.

Moor, J. H. 2006. The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems*, 21(4): 18–21.

Newberry, T.; and Ord, T. 2021. The parliamentary approach to moral uncertainty. *Future of Humanity*.

Ross, J. 2006. Rejecting ethical deflationism. *Ethics*, 116(4): 742–768.

Sinnott-Armstrong, W. 2022. Consequentialism. In Zalta, E. N.; and Nodelman, U., eds., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2022 edition.

Szabo, J.; Such, J.; Criado, N.; and Modgil, S. 2024. https://kclpure.kcl.ac.uk/ws/portalfiles/portal/240967564/mrynhkkpdyndxdsbqgdkhzzksczvfbwm.pdf.

Tarsney, C. 2018. Moral uncertainty for deontologists. *Ethical Theory and Moral Practice*, 21(3): 505–520.