"Loudness": On priors over preference relations (Brief technical note)

Benja Fallenstein and Nisan Stiennon

May 2014

Abstract

This is a quick writeup of a problem discussed at the May 2014 MIRI workshop: how to formally deal with uncertainty about preferences. We assume that the true preferences satisfy the von Neumann-Morgenstern (VNM) axioms, and can therefore be represented by a utility function. It may seem that we should then simply maximize the expectation of this function. However, in the absence of more information, this is not well-defined; in this setting, different choices of utility functions representing the same VNM preferences can lead the agent to make different choices. We give a formalization of this problem and show that the choice of a prior probability distribution over VNM preference relations together with the choice of a representative for each of these distributions is in a certain sense equivalent to the choice of a single number for every preference relation, which we call its "loudness". (Mathematically, a "loudness prior" can be seen as a probability distribution over preference relations, but this object does not have an epistemic interpretation.)

1 Introduction

If we had a powerful Artificial General Intelligence (AGI), smarter than any human being and even humanity as a whole, and if this AGI could be programmed to pursue any formally specified goal, what would we do with it? It seems that our best hope is to specify something along the lines of "do what we would have decided to do if we were smarter and had thought about the problem for a longer time", and then have the AGI use its superior intelligence to figure out what this would actually mean. Formally specifying a goal like this is an extremely difficult task. Here, however, we consider a problem that arises even when we have such a formal specification: How should the AGI act given that realistically, it will have uncertainty about what, exactly, humanity *would* want it to do if it were smarter and had thought about the problem longer? When different hypotheses disagree about which action is best, which action should it take? When should it decide to take an action that will help it reduce its uncertainty about its goals?

We assume that the formal specification describes preferences that obey the von Neumann-Morgenstern (VNM) axioms, although this description may make reference to facts about the world which the AGI may be uncertain about: an example would be a formalization of something like, "the VNM utility function we would choose if we were smarter and had thought about the problem longer". It may seem that if we define things in this way, the problem is trivial—our AGI should simply maximize *expected* utility. However, VNM utility functions are only defined up to positive affine transformations: a utility function scaled up by a factor of two still represents the same class of preferences, but scaling up the utility function representing one of the agent's hypotheses about the true preferences by a factor of two (without scaling any of the others) may affect the expected utility calculation. Thus, the expected utility calculation is sensitive to the way in which the representative utility functions are chosen for each set of preferences.

In order to make this problem clear, it will be helpful to introduce a simple formalism for modelling this problem. (This formalism is intended as a toy model, not as a formalism that would literally be used in the construction of an AGI as it stands.)

We first suppose that $\mathcal{Q}$ denotes a set of "outcomes" or "qualities", whose members are descriptions of all facts about the universe which the AGI may want to influence: for example, in the case of a "paperclip maximizer", which tries to maximize the expected number of paperclips in the universe, we could have $\mathcal{Q} = \mathbb{N}$, where elements $n \in \mathcal{Q}$ represent the actual number of paperclips anywhere in the universe; and in the case of an AGI that is uncertain whether it should maximize the number of paperclips or the number of staples, we could have $\mathcal{Q} = \mathbb{N} \times \mathbb{N}$, where $(n, n') \in \mathcal{Q}$ represents the actual number of paperclips (n) together with the actual number of staples (n'). (One could reasonably argue that the AGI might have sufficient uncertainty about what features are important that $\mathcal{Q}$ would have to become the set of *all* facts about the world, but we ignore this problem here.)

Next, we suppose that $\mathcal{M}$ denotes a set of "moralities", whose members describe all facts about the AGI's preferences that it is uncertain about. We assume that for every $m \in \mathcal{M}$, the preferences the AGI should follow if it knows that m is true are described by a preference relation on "lotteries" (meaning probability distributions) over outcomes[1], which satisfy the VNM axioms and can therefore be described by a utility function

$$u_m : \Delta(\mathcal{Q}) \to [0, 1]$$

(for every $m \in \mathcal{M}$), where we write $\Delta(\mathcal{Q})$ for the set of all probability distributions over $\mathcal{Q}$.

We further assume that Ω is a set of "possible worlds", that

$$M : \Omega \to \mathcal{M}$$

[1] We ignore questions of measurability, and in fact, we have only checked the assertions in this note for the case where $\mathcal{Q}$ is finite, but we hope that the results will extend or have analogs in settings where the space of outcomes is infinite and potentially uncountable.

is a function which specifies which morality is correct in each possible world, that $\mathcal{S}$ is a set describing all *strategies* the AGI could pursue, and that

$$Q : \Omega \times \mathcal{S} \to \mathcal{Q}$$

is a function which specifies the outcome that results if the AGI chooses a particular strategy in a particular possible world. Like pure strategies in an extensive-form game, a strategy $s \in \mathcal{S}$ is not just a single action, but a specification of which action the AGI would perform in any situation it might find itself in. In particular, there may be strategies in which the AGI will first perform an action that gives it more information about the true "morality" (for example, it might ask a human questions), and will then will take additional actions which depend on what information it has received in the first step. (Note that we allow the *outcome* to depend on the AGI's actions, but not the morality the AGI should follow.) Finally, we assume that we have a prior probability distribution[2]

$$\mathbb{P} \in \Delta(\Omega)$$

over the possible worlds.

It may now seem clear that the AGI should choose the strategy $s \in \mathcal{S}$ that maximizes the expectation $\mathbb{E}[u_{M(\boldsymbol{\omega})}(Q(\boldsymbol{\omega}, s))]$, where $\mathbb{E}$ is the expectation with respect to $\mathbb{P}$, and $\boldsymbol{\omega}$ is the random variable whose distribution is given by $\mathbb{P}$.

However, simply because the preferences of morality m are described by the utility function $u_m : \mathcal{Q} \to [0, 1]$, it does not follow that this is the *only* utility function describing these preferences; for example, $u'_m : \mathcal{Q} \to [0, 1]$ given by $u'_m(q) := u_m(q)/2$ also describes the same preferences, but replacing u_m by u'_m for one particular $m \in \mathcal{M}$ (while not changing the utility functions associated with other elements of $\mathcal{M}$) will in general change the way the AGI acts. (In particular, if it assigns m probability in $(0, 1)$, then after replacing u_m with u'_m, the AGI will give less weight to the preferences of morality m than it did before the change.)

How should we think about this problem? Should we choose to normalize the utility functions in some canonical way (e.g., such that $\inf_{q \in \mathcal{Q}} u_m(q) = 0$ and $\sup_{q \in \mathcal{Q}} u_m(q) = 1$ for every m such that u_m is not constant)? Or should we perhaps use something different from expected utility maximization?

In this note, we show that if we use the expected utility framework, then the choice of representatives u_m and the choice of the prior probabilities $\mathbb{P}[M(\boldsymbol{\omega}) = m]$ can be aggregated into a single choice of an artifically rescaled probability, which we call the morality's "loudness". In this framework, instead of choosing a prior probability for each morality and then separately needing to find a way to choose a representative u_m, we only choose a "loudness prior", which has the same number of degrees of freedom as a choice of a prior probability distribution on moralities alone. In a sense, this smaller amount of information is all the decision-relevant information we "really" choose in the expected utility framework.

[2] Again, we ignore issues of measurability, and say things that definitely make sense if Ω is finite and hopefully have analogs in infinite cases.

2 "Loudness"

Given a utility function $u : \mathcal{Q} \to \mathbb{R}$, define $\lfloor u \rfloor := \inf_{q \in \mathcal{Q}} u(q)$ and $\lceil u \rceil := \sup_{q \in \mathcal{Q}} u(q)$, and set

$$\mathrm{norm}(u) : \mathcal{Q} \to [0,1], \qquad \mathrm{norm}(u)(q) := \frac{u(q) - \lfloor u \rfloor}{\lceil u \rceil - \lfloor u \rfloor}$$

if the denominator is defined (i.e., if u is not constant), and $u(q) := 0$ otherwise. (Thus, if the denominator is $\neq 0$, we have $\lfloor \mathrm{norm}(u) \rfloor = 0$ and $\lceil \mathrm{norm}(u) \rceil = 1$.) Then $\mathrm{scale}(u) := \lceil u \rceil - \lfloor u \rfloor$ is the number by which $\mathrm{norm}(u)$ must be scaled in order for it to be weighted as much in expected utility calculations as u:

$$\begin{aligned}
& \mathbb{E}[u_{M(\boldsymbol{\omega})}(Q(\boldsymbol{\omega}, s))] \\
= & \sum_{\omega \in \Omega} \mathbb{P}(\omega) \cdot u_{M(\omega)}(Q(\omega, s)) \\
= & \sum_{\omega \in \Omega} \mathbb{P}(\omega) \cdot \Big(\mathrm{scale}(u_{M(\omega)}) \cdot \mathrm{norm}(u_{M(\omega)})(Q(\omega, s)) \; + \; \lfloor u_{M(\omega)} \rfloor \Big) \\
= & \sum_{\omega \in \Omega} \mathbb{P}(\omega) \cdot \mathrm{scale}(u_{M(\omega)}) \cdot \mathrm{norm}(u_{M(\omega)})(Q(\omega, s)) \; + \; \sum_{\omega \in \Omega} \mathbb{P}(\omega) \cdot \lfloor u_{M(\omega)} \rfloor .
\end{aligned}$$

Since the second sum is constant in s, it does not affect the preference ordering among strategies; thus, maximizing expected utility is equivalent to maximizing

$$\sum_{\omega \in \Omega} \mathbb{P}(\omega) \cdot \mathrm{scale}(u) \cdot \mathrm{norm}(u_{M(\omega)})(Q(\omega, s))$$

which in turn is equivalent to maximizing

$$\sum_{\omega \in \Omega} \left(\frac{\mathbb{P}(\omega) \cdot \mathrm{scale}(u_{M(\omega)})}{\sum_{\omega' \in \Omega} \mathbb{P}(\omega') \cdot \mathrm{scale}(u_{M(\omega')})} \right) \cdot \mathrm{norm}(u_{M(\omega)})(Q(\omega, s))$$

(where we assume that at least one $\omega' \in \Omega$ both has positive probability and corresponds to a non-constant utility function $u_{M(\omega')}$). Defining

$$\tilde{\mathbb{P}}(\omega) := \frac{\mathbb{P}(\omega) \cdot \mathrm{scale}(u_{M(\omega)})}{\sum_{\omega' \in \Omega} \mathbb{P}(\omega') \cdot \mathrm{scale}(u_{M(\omega')})},$$

we can rewrite this as

$$\sum_{\omega \in \Omega} \tilde{\mathbb{P}}(\omega) \cdot \mathrm{norm}(u_{M(\omega)})(Q(\omega, s)) \;\; = \;\; \tilde{\mathbb{E}}[\mathrm{norm}(u_{M(\boldsymbol{\omega})})(Q(\boldsymbol{\omega}, s))],$$

where $\tilde{\mathbb{E}}$ is the expectation with respect to the probability distribution $\tilde{\mathbb{P}}$.

In other words, expected utility maximization with respect to arbitrary representatives u_m of the VNM preferences associated with the different moralities $m \in \mathcal{M}$ is equivalent to expected utility maximization with respect to the normalized utility functions $\mathrm{norm}(u_m)$, together with rescaled "probabilities" $\tilde{\mathbb{P}}$.

It's clear that this fact would still hold if we used a different way of normalizing utility functions than we have used here, though the rescaled probabilities would then change as well.

Moreover, it is easy to see that

$$\mathbb{P}(\omega \mid M(\omega)) \;:=\; \mathbb{P}[\boldsymbol{\omega} = \omega \mid M(\boldsymbol{\omega}) = M(\omega)] \;=\; \tilde{\mathbb{P}}[\boldsymbol{\omega} = \omega \mid M(\boldsymbol{\omega}) = M(\omega)] :$$

i.e., our rescaling only changes the prior probability that a particular morality is the true morality, not the conditional probability that a particular possible world is the true world, given that its morality is the true morality. This is because the probabilities of all possible worlds with the same morality $m \in \mathcal{M}$ are scaled by the same factor, $\text{scale}(u_m)$.

One way of thinking about this is the following: When we set up an AGI using the above framework, it looks like we must choose prior probabilities over the different possible worlds, and a representative utility function u_m for each morality $m \in \mathcal{M}$. Choosing a prior over possible worlds is equivalent to choosing a prior over moralities, together with *conditional* prior probabilities of the possible worlds, given their moralities. But in a sense, it is redundant to specify both a prior probability $\mathbb{P}[M(\boldsymbol{\omega}) = m]$ for each morality $m \in \mathcal{M}$ and a representative u_m of the VNM preference relation corresponding to m: The same information about the AGI's decisions is conveyed by the scaled probability distribution $\tilde{\mathbb{P}}[M(\boldsymbol{\omega}) = m]$. We call $\tilde{\mathbb{P}}[M(\boldsymbol{\omega}) = m]$ the *loudness* of morality m and write it as $\mathbb{L}(m)$.

Thus, if we define $U : \mathcal{M} \times \mathcal{Q} \to [0, 1]$ by $U(m, q) := \text{norm}(u_m)(q)$, then a decision problem of the type we consider here is represented by a tuple $(\mathcal{Q}, \mathcal{M}, \Omega, \mathcal{S}, U, Q, M, \mathbb{L}(\cdot), \mathbb{P}(\cdot \mid \cdot))$, where $\mathbb{P}(\cdot \mid \cdot)$ denotes the function giving the conditional probabilities $\mathbb{P}(\omega \mid M(\omega))$. (Mathematically speaking, $\mathbb{L} \in \Delta(\mathcal{M})$ is a probability distribution, but the numeric values of $\mathbb{L}$ do not have an epistemic interpretation on their own, since they depend on the normalization function $\text{norm}(\cdot)$ being used.) In this formalization, the choice of a prior over $\mathcal{M}$ and of representatives u_m is replaced by the choice of the *loudness prior* $\mathbb{L}$.

3 Discussion

One might object that the prior probabilities have an intuitive epistemic meaning, which is lost if we aggregate them into the loudnesses $\mathbb{L}(m)$; even if these probabilities do not affect the agent's decisions except through their influence on loudness, if we have intuitions about how to choose prior probabilities and how to choose representative utility functions for each morality, then these intuitions give us a way to choose the loudness prior, which makes thinking in terms of these concepts useful.

However, although we clearly have *epistemic* intuitions, it is much less clear that we have intuitions about how to choose representatives. This requires us to be able to say "how strong" the preferences of different moralities are, compared to each other. MacAskill (2014, Chapter 4) argues that in many cases, our

intuitions suggest that there are grounds for making such comparisons, but it is far from clear that this is always the case.

This issue is sharpened if we change the problem under consideration to that of learning the utility function of a different VNM agent. In this case, our reason for using utility functions is simply that the other agent's behavior can be described by VNM preferences, and VNM preferences can be represented by utility functions; there is no reason to suppose that there is any well-defined sense in which under one hypothesis $m \in \mathcal{M}$ about what the other agent's preferences are, these preferences are e.g. "twice as strong" than under another hypothesis $m' \in \mathcal{M}$; there are no canonical grounds for comparing the strength of preferences between two different hypotheses.

In this situation, even if the choice of *probabilities* is non-arbitrary, as long as the choice of *representatives of utility functions* is somewhat arbitrary, so is the choice of loudnesses; given a prior probability distribution $\mathbb{P} \in \Delta(\mathcal{M})$, a loudness prior $\mathbb{L} \in \Delta(\mathcal{M})$ corresponds to a choice of scale factors $\mathbb{L}(m)/\mathbb{P}(m)$ for each morality that is assigned positive probability; we can then choose representatives $[\mathbb{L}(m)/\mathbb{P}(m)] \cdot u_m^{\text{norm}}$, where u_m^{norm} is the normalized utility function representing m. Thus, we might as well consider it our task to choose a good loudness prior, rather than to choose a good set of representatives u_m.

The problem of dealing with uncertainty about preferences, which we have discussed in this note, has a close relationship to the problem of social aggregation of the preferences of different members of society, and there is a large body of literature on the latter problem. However, there is a key difference between the two problems: When dealing with uncertainty about preferences, an agent has the opportunity to acquire *more information* (whereas in a society whose members have conflicting interests, it wouldn't be possible to acquire information about who is "right"). We will discuss the relationship between these problems further in a different technical note.

References

MacAskill, W. (2014). *Normative uncertainty*. PhD thesis, Oxford University.