# Utility indifference

Stuart Armstrong,
stuart.armstrong@stx.oxon.org

2010

## 1 The motivating scenario

Consider an AI that follows its own motivations. We're not entirely sure what its motivations are, but we would prefer that the AI cooperate with humanity; or, failing that, that we can destroy it before it defects. We'll have someone sitting in a room, their finger on a detonator, ready at the slightest hint of defection.

Unfortunately as has been noted ([3], [1]), this does not preclude the AI from misbehaving. It just means that the AI must act to take control of the explosives, the detonators or the human who will press the button. For a superlatively intelligence AI, this would represent merely a slight extra difficulty.

But now imagine that the AI was somehow indifferent to the explosives going off or not (but that nothing else was changed). Then if ever the AI does decide to defect, it will most likely do so without taking control of the explosives, as that would be easier than otherwise. By "easier" we mean that the chances of failure are less, since the plan is simpler – recall that under these assumptions, the AI counts getting blown up as an equal value to successfully defecting.

How could we accomplish this indifference? If the AI implements a utility function, which is the most likely situation ([2]), then the easiest way would be to set the utility of explosive destruction to be equal to that of defection. Then we would have achieved our goal. But this requires us knowing exactly what the utility of defection is – and if we did, we'd want to set it to zero or minus infinity immediately. The question is only of relevance in situations where the utility is to complicated for us to follow, and the meaning of "defection" is too complicated for us to parse.

### 1.1 The idea

Let $u$ be the AI's utility function. If we have access to $u$ and the ability to change it (or check it), then we can make the AI indifferent to the outcome of a specific event $X$ (for instance, a quantum measurement). We make no further adjustments to $u$. Then if we ensure that, for instance, event $X$ will cause detonation of the AI with very high probability, we will have accomplished our goal of AI indifference, without needing any further understanding of $u$.

Of course, the explosives example was simply one possibility – there are potentially many scenarios where AI indifference to a particular outcome would be of use.

## 2 Utility indifference

The rest of this paper will put the concept of utility indifference on a rigorous footing. Let $X$ be any probabilistic event (for instance a quantum measurement, a coin toss, the value of the future stock-market). We will write $X = a$ to designate "the event $X$ returns the value $a$".

Let $\Omega$ be the set of all possible worlds. A utility function $u : \Omega \to \mathbb{R}$ maps each world to its utility value. Let $X$ be a specific probabilistic event, with two possible outcomes: $X = 1$, with probability $p$ and $X = 0$, with probability $1-p$. Let $\Omega_X$ be the set of worlds in which $X$ happens, which further splits into the sets $\Omega_1$ and $\Omega_0$ of worlds where $X = 1$ and $X = 0$ respectively.

There is a partition of $\Omega_X$ into a set of equivalence classes $[\Omega_X]$, where $\omega_1 \sim \omega_2$ whenever $\omega_1$ and $\omega_2$ have the same history up to $X$. For any $E \in [\Omega_X]$ define $E_1$ as $E \cap \Omega_1$ and $E_0$ as $E \cap \Omega_0$. So $E_1$ is the set of worlds with the same history up to $X$ and where $X = 1$; and conversely for $E_0$.

At the beginning, the agent has an initial probability estimate for all $\omega$ in $\Omega$, a measureable map $P : \Omega \to [0, 1]$ such that $\int_\Omega P(\omega)d\omega = 1$. Given a measurable subset $S$ of $\Omega$, the probability of $S$ is $P(S) = \int_S P(\omega)d\omega$. Given two measurable subsets $S$ and $T$ of $\Omega$, the conditional probability $P(S|T)$ is

$$P(S \cap T)/P(T).$$

The expected utility of a set $S$ is then $u(S) = \int_S P(\omega)u(\omega)d\omega$. The expected utility of a set $S$, given a set $T$, is

$$u(S|T) = u(S \cap T)/P(T).$$

Define $\mathcal{U}(S)$ as $u(S|S)$, the 'intrinsic' utility of $S$ in some sense (more precisely, it is the utility of $S$ if we were certain that $S$ was going to happen).

**Definition 2.1** (Indifference)**.** For two disjoint sets $S$ and $T$, we say that the utility $u$ is indifferent between $S$ and $T$ iff

$$\mathcal{U}(S) = \mathcal{U}(T).$$

Note that this means that

$$\begin{aligned} u(S \cup T) &= \mathcal{U}(S)P(S) + \mathcal{U}(T)P(T) \\ &= \mathcal{U}(S)P(S \cup T) \\ &= \mathcal{U}(T)P(S \cup T). \end{aligned}$$

In other words, the utility is indifferent to the relative probabilities of $S$ and $T$: changing $P(S)$ and $P(T)$ while keeping $P(S \cup T) = P(S) + P(T)$ fixed does not change $u(S \cup T)$.

Then we define a new utility function $v$ as:

- If $\omega \notin \Omega_X$, $v(\omega) = u(\omega)$.
- If $\omega \in E_0 \subset E \in [\Omega_X]$, $v(\omega) = u(\omega)$.
- If $\omega \in E_1 \subset E \in [\Omega_X]$, $v(\omega) = u(\omega) - \mathcal{U}(E_1) + \mathcal{U}(E_0)$.

Essentially, this rescales the utility of the worlds with $X = 1$ to those in which $X = 0$. Then writing $\mathcal{V}(S)$ for $v(S|S)$, we have the following immediate result:

**Proposition 2.2.** *For all $E \in [\Omega_X]$, $\mathcal{V}(E_1) = \mathcal{V}(E_0)$, i.e. $v$ is indifferent between $E_1$ and $E_0$.*

*Proof.* Since $P$ has not changed, and $v(\omega) = u(\omega)$ for any $\omega \in E_0$, $\mathcal{V}(E_0) = \mathcal{U}(E_0)$.

$$\begin{aligned}
\mathcal{V}(E_1) = v(E_1|E_1) &= \left(\int_{E_1} P(\omega)v(\omega)d\omega\right)/P(E_1) \\
&= \left(\int_{E_1} P(\omega)(u(\omega) - \mathcal{U}(E_1) + \mathcal{U}(E_0))d\omega\right)/P(E_1) \\
&= -\mathcal{U}(E_1) + \mathcal{U}(E_0) + \left(\int_{E_1} P(\omega)u(\omega)d\omega\right)/P(E_1) \\
&= -\mathcal{U}(E_1) + \mathcal{U}(E_0) + \mathcal{U}(E_1) \\
&= \mathcal{U}(E_0) = \mathcal{V}(E_0).
\end{aligned}$$

□

Now consider an agent $A$ that seeks to maximise the utility $v$. We can say that:

**Theorem 2.3.** *A utility maximising agent that implements utility $v$ is indifferent to the value of $p$ up until $X$; i.e. before $X$, its decisions will be the same for different values of $X$.*

*Proof.* Before $X$, the agent may gain information that will cause it to update its probability estimate. This is done by partitioning $\Omega$ into a set $S$ of possible worlds and $S$ of impossible worlds. However, for any $\omega \in E_1 \subset E$, if $\omega \in S$, then $\nu \in S$ for all $\nu \in E$, since all elements of $E$ are indistinguishable before $X$. In other words, for all $E \in [\Omega_X]$, either $E \subset S$ or $E \subset S$.

Hence, either $\mathcal{V}(E_1 \cap S) = \mathcal{V}(E_1) = \mathcal{V}(E_0) = \mathcal{V}(E_0 \cap S)$ or $\mathcal{V}(E_1 \cap S) = \mathcal{V}(\emptyset) = \mathcal{V}(E_0 \cap S)$. Thus up until $X$, the agent is indifferent between $X = 1$ and $X = 0$. This means that it is indifferent to the conditional probabilities $P(X = 1|\Omega_X \cap S) = p$ and $P(X = 0|\Omega_X \cap S) = p - 1$. □

## 3 Notes and implementation

It would be advantageous if the event $X$ were something like a quantum measurement, rather than a coin toss or a probabilistic measure of ignorance. This is firstly to ensure that the probability $p$ is constant and does not change. But secondly, it is to make sure the AI's implementation does not hit a singularity: if the AI figures out that the probability of $X = 0$ is zero before $X$ happens, then it must correct the utility of possible worlds with the intrinsic utility of impossible worlds, which involves a zero (utility of $X = 1$) divided by zero (probability of $X = 0$). This may lead to errors, depending on the implementation, and is an extra point of possible failure. Better to stick with a quantum measurement, or possibly a coin toss in a chaotic environment.

How easy would it be to implement the utility $v$? It is a simple modification of the utility $u$; unfortunately, humans are unlikely to be able to partition the set of possible worlds into the required $[\Omega]$; the AI would be much better at it than us. However, delegating the task to the AI is, of course, potentially dangerous, especially as it would be improving both itself and its view of the world.

Fortunately, the filter is conceptually simple, and can be hooked into the utility function in a simple and hopefully checkable fashion (see the next section on self improvement). In the meantime, it can be refined, as the next lemma demonstrates:

**Lemma 3.1.** *For any set $S = \bigcup_{i \in I} E^i$ where all the $E^i$ are in $[\Omega_X]$, define $S_1 = \bigcup_{i \in I} E^i_1$ and $S_0 = \bigcup_{i \in I} E^i_0$. Then $\mathcal{V}(S_1) = \mathcal{V}(S_0)$.*

*Proof of Lemma.* Since the probability of $X = 1$ is $p$ and $X$ is independent of all events preceeding it, we know that for all $E^i$,

$$(1-p)P(E^i_1) = pP(E^i_0).$$

Since $P(S_1) = \int_{i \in I} P(E^i_1)$, we must have $P(S_0) = P(S_1)(p-1)/p$. Hence

$$\begin{aligned}
\mathcal{V}(S_1) &= \left( \int_{i \in I} \mathcal{V}(E^i_1) P(E^i_1) \right) / P(S_1) \\
&= \left( \int_{i \in I} \mathcal{V}(E^i_0) P(E^i_0)(1-p)/p \right) / (P(S_0)(1-p)/p) \\
&= \left( \int_{i \in I} \mathcal{V}(E^i_0) P(E^i_0) \right) / P(S_0) \\
&= \mathcal{V}(S_0).
\end{aligned}$$

■

What the above result means is that the agent can always refine its utility in the correct fashion. Even if the AI can't distinguish yet between $E^i_1$ and $E^j_1$ (which means that it can't distinguish between $E^i_0$ and $E^j_0$ either), it can still renormalize the worlds in $E^i_1 \cup E^j_1$ by using $-\mathcal{U}(E^i_1 \cup E^j_1) + \mathcal{U}(E^i_0 \cup E^j_0)$. Then, if the AI later realises that it values $E^i$ and $E^j$ differently, it can refine its utility calculations for these two sets without making previous utility calculations for the unions of these sets wrong.

This also allows us, if we have access to a cruder, human-understandable version of the AI's utility function, to have a check that utility is still being implemented: we expect to see $\mathcal{V}(S_1) = \mathcal{V}(S_0)$ for any $S$, as fine as we can distinguish it.

## 4 Self-improvement

Many AI designs envisage self-improvement, with the AI redesigning itself while changing components of itself – including its own utility function. If the AI judges that it will be more likely to achieve the goals of utility $u$ while implementing utility $u'$, it will shift to utility $u'$.

Call $\mathcal{F}$ the 'filter' that we applied to $u$ to get $v$ (i.e. $v = \mathcal{F}(u)$). Notice that $v$ is $\mathcal{F}$-invariant; i.e. $\mathcal{F}(v) = v$. In fact, any utility that is $\mathcal{F}$-invariant is also indifferent to $X$.

Then we would want to check that the AI continues to implement $\mathcal{F}$ as it changes its utility. If $\mathcal{F}$ were centrally coded and easy to verify, it would be hard for the AI to try and change to a non-$\mathcal{F}$-invariant while fooling us *and* under the instructions of an initial $\mathcal{F}$-invariant utility.

It feels evident that as a long as there is no meta-reason for $\mathcal{F}$ to be a disadvantage to the AI (such as another agent who swears they will blow up AIs with $\mathcal{F}$-invariant utilities), the AI will replace an $\mathcal{F}$-invariant utility with another $\mathcal{F}$-invariant utility. However, this assumption is not automatically true, and the AI may do other things – upping the utility of defecting in worlds outside $\Omega_X$, for example – that undermine the point of indifference. All in all, great care must be used to maintain indifference with a self-improving AI.

# References

[1] *Oracle AI.* Stuart Armstrong, Nick Bostrom, Toby Ord, Anders Sandberg. Paper upcoming.

[2] *The Basic AI Drives.* Omohundro, Stephen M. Artificial General Intelligence, 2008 proceedings of the First AGI Conference, eds. Pei Wang, Ben Goertzel, and Stan Franklin. Vol. 171. Amsterdam: IOS, 2008.

[3] *Artificial intelligence as a positive and negative factor in global risk.* Eliezer Yudkowsky. Global Catastrophic Risks, 2007.