Impossibility and Uncertainty Theorems in AI Value Alignment

or why your AGI should not have a utility function

Peter Eckersley
Partnership on AI & EFF
pde@partnershiponai.org

Abstract

Utility functions or their equivalents (value functions, objective functions, loss functions, reward functions, preference orderings) are a central tool in most current machine learning systems. These mechanisms for defining goals and guiding optimization run into practical and conceptual difficulty when there are independent, multi-dimensional objectives that need to be pursued simultaneously and cannot be reduced to each other. Ethicists have proved several impossibility theorems that stem from this origin; those results appear to show that there is no way of formally specifying what it means for an outcome to be good for a population without violating strong human ethical intuitions (in such cases, the objective function is a social welfare function). We argue that this is a practical problem for any machine learning system (such as medical decision support systems or autonomous weapons) or rigidly rule-based bureaucracy that will make high stakes decisions about human lives: such systems should not use objective functions in the strict mathematical sense.

We explore the alternative of using uncertain objectives, represented for instance as partially ordered preferences, or as probability distributions over total orders. We show that previously known impossibility theorems can be transformed into uncertainty theorems in both of those settings, and prove lower bounds on how much uncertainty is implied by the impossibility results. We close by proposing two conjectures about the relationship between uncertainty in objectives and severe unintended consequences from AI systems.

1 Introduction

Conversations about the safety of self-driving cars often turn into discussion of "trolley problems", where the vehicle must make a decision between several differently disastrous outcomes (Goodall 2014; Riek and Howard 2014; Belay 2015; Nyholm and Smids 2016; Bonnefon, Shariff, and Rahwan 2015; Lin 2016). For the engineers building robots and machine learning systems, however, these questions seem somewhat absurd; the relevant challenges are correctly identifying obstacles, pedestrians, cyclists and other vehicles from noisy and unreliable data sources, and reasonably predicting the consequences of the vehicle's control mechanisms in various conditions. The conceivable circumstances under which a self-driving car's systems might accurately foresee and deliberately act on a genuine "trolley problem" without having been able to avoid that problem in the first place are a tiny portion of possibility space, and one that has arguably received a disproportionate amount of attention.[1]

There are, however, other kinds of AI systems for which ethical discussions about the value of lives, about who should exist or keep existing, are more genuinely relevant. The clearest category is systems that are *designed* to make decisions affecting the welfare or existence of people in the future, rather than systems that only have to make such decisions in highly atypical scenarios. Examples of such systems include:

- autonomous or semi-autonomous weapons systems (which may have to make trade-offs between threats to and the lives of friendly and enemy combatants, friendly and enemy civilians, and combatants who have surrendered).[2]
- risk assessment algorithms in criminal justice contexts (which at least notionally are making trade-offs between risk to society and the harm that life-shortening incarceration does to individual defendants in the criminal justice system and their families).
- decision support systems in high-stakes bureucratic settings. For instance a system that helps to allocate scarce resources like doctors' time, laboratory test access and money in a medical system would implicitly or explicitly have to make trade-offs between the life span, quality of life, and fertility of numerous people.
- autonomous types of open-ended AI taking important actions without their designers' input.

Governmental organizations already have to make decisions that involve weighing benefits to present or future people against risks to present or future people, or trading off benefits (such as wealth) to one group against benefits of another sort (such as freedom) to another. Therefore, we might

[1] There are related and more fruitful questions like, "how should self-driving cars make trade-offs between time saved by driving more aggressively and accident risk?" See for instance (Gerdes and Thornton 2016).

[2] The creation of autonomous weapons systems may well be an extremely bad idea (see, eg (Roff 2014), but one of the main arguments made by their proponents is that such systems will be able to make ethical decisions in real world equivalents of trolley situations at least as well as if not better than human pilots, and some military-funded work in that direction is already underway (see, for instance, (Arkin 2009; Govindarajulu and Bringsjord 2017)).

think that the decision-making tools that they employ could be implemented by any narrow AI system that has to make similar choices. A variety of metrics have been proposed and used to perform cost-benefit analyses in such situations. For example, the US Environmental Protection Agency uses the 'value of a statistical life' (VOSL) metric, which identifies the cost of an intervention that puts people at risk with people's willingness to pay to avoid similar risks (for a critique, see (Broome 1985)) or the QALY metric used by the UK's National Health Service, which assigns a given value to each life-year saved or lost in expectation, weighted by the quality of that life. The problem is that existing impossibility theorems apply equally to each of these metrics and therefore to an AI system that implements them.

2 Impossibility Theorems in Ethics

2.1 Example: An Impossibility Theorem in Utilitarian Population Ethics

Economists and ethicists study individual utility functions and their aggregations, such as social welfare functions, social choice functions and formal axiologies, and have discovered various impossibility theorems about these functions. Perhaps the most famous of these is Arrow's Impossibility Theorem (Arrow 1950), which applies to social choice or voting. It shows there is no satisfactory way to compute society's preference ordering via an election in which members of society vote with their individual preference orderings. Fortunately, Arrow's theorem results from some constraints (incomparability of individual's preferences, and the need for incentive compatibility) which may not apply to AI systems.

Unfortunately, ethicists have discovered other situations in which the problem isn't learning and computing the tradeoff between agents' objectives, but that there simply *may not be* such a satisfactory tradeoff at all. The "mere addition paradox" (Parfit 1984) was the first result of this sort, but the literature now has many of these impossibility results. For example, Arrhenius (Arrhenius 2000) shows that all total orderings of populations must entail one of the following six problematic conclusions, stated informally:

- **The Repugnant Conclusion** For any population of very happy people, there exists a much larger population with lives barely worth living that is better than this very happy population (this affects the "maximise total wellbeing" objective).
- **The Sadistic Conclusion** Suppose we start with a population of very happy people. For any proposed addition of a sufficiently large number of almost-as-happy people, there is a small number of horribly tortured people that is a preferable addition (this affects the "maximise average wellbeing" objective).
- **The Very Anti-Egalitarian Conclusion** For any population of two or more people which has uniform happiness, there exists another population of the same size which has lower total and average happiness, and is less equal, but is better.
- **Anti-Dominance** Population B can be better than population A even if A is the same size as population B, and every person in A is happier than their equivalent in B.
- **Anti-Addition** It is sometimes bad to add a group of people B to a population A (where the people in group B are worse of than those in A), but *better* to add a group C that is larger than B, and worse off than B.
- **Extreme Priority** There is no n such that create of n lives of very high positive welfare is sufficient benefit to compensate for the reduction from very low positive welfare to slightly negative welfare for a single person (informally, "the needs of the few outweigh the needs of the many").

The structure of the impossibility theorem is to show that no objective function or social welfare function can simultaneously satisfy these principles, because they imply a cycle of world states, each of which in turn is required (by one of these principles) to be better than the next. The structure of that proof is shown in Figure 1.

2.2 Another Impossibility Theorem

Arrhenius's unpublished book contains a collection of many more uncertainty theorems. Here is the simplest, which is a more compelling version of Parfitt's Mere Addition Paradox. It shows the impossibility of an objective function satisfying the following requirements simultaneously:

- **The Quality Condition:** There is at least one perfectly equal population with very high welfare which is at least as good as any population with very low positive welfare, other things being equal.
- **The Inequality Aversion Condition:** For any triplet of welfare levels A, B, and C, A higher than B, and B higher than C, and for any population A with welfare A, there is a larger population C with welfare C such that a perfectly equal population B of the same size as A∪C and with welfare B is at least as good as A∪C, other things being equal.
- **The Egalitarian Dominance Condition:** If population A is a perfectly equal population of the same size as population B, and every person in A has higher welfare than every person in B, then A is better than B, other things being equal.
- **The Dominance Addition Condition:** An addition of lives with positive welfare and an increase in the welfare in the rest of the population doesn't make a population worse, other things being equal.

The cyclic structure of the proof is shown in Figure 2.

2.3 Ethical Impossibility Derives From Competing Objectives

The particular impossibility theorems summarized above, and those in the related literature, result from the incompatibility of several different utilitarian objectives: maximizing total wellbeing, maximizing average wellbeing, and avoiding suffering. Similar problems are also likely to arise when considering almost any kinds of competing objectives, such as attempting to simultaneously maximize different notions of wellbeing, freedom, knowledge, fairness, or other widely accepted or domain-specific ethical goods (Schumm 1987). We conjecture that the literature contains more impossibility theorems about happiness or wellbeing simply because that objective has been subjected to mathematical modeling and study for well over a hundred years,[3] while much less effort has gone into the others. Recent work has begun to focus on fairness in decisionmaking contexts in particular, and is already producing its own impossibility results.(Chouldechova 2017; Kleinberg, Mullainathan, and Raghavan 2017)

3 Possible Responses to Impossibility Results

Arrhenius's impossibility results and others like them are quite troubling. They show that we do not presently have a trustworthy framework for making decisions about the welfare or existence of people in the future, and are representative of a broader problem with the inability of the objective functions to reasonably optimise for multiple objectives simultaneously. Next we will consider five possible responses to these impossibility theorems:

3.1 Small-scale evasion:

One response may be to claim that the stakes are simply low for present AI systems that only make small-scale changes to the world. AGI systems, if they exist, may one day need to confront impossibility theorems or difficult conflicting objectives in some cases, but none of that is presently relevant. Unfortunately, it appears that although the stakes are indeed much lower at present, the fundamental tensions that drive these paradoxes are already at play in decisions made by bureaucracies and decision support systems, and small scale decisions can easily violate constraints we would want to see respected. Asking for human feedback helps, but humans often miss important ethical principles, and can easily make chains of decisions that have problematic consequences in proportion to their degree of power or agency. Cases of particular concern are those where ML or algorithmic heuristics are deployed in decision support tools that humans are inadequately inclined to question.(Parasuraman and Manzey 2010)

3.2 Value learning:

One might argue that impossibility theorems are only a problem for an AI system if we are attempting to explicitly define an objective function, rather than letting an AI system acquire its objective function in a piecemeal way via a method like co-operative inverse reinforcement learning (CIRL) or other forms of human guidance (Hadfield-Menell et al. 2016; Saunders et al. 2017; Christiano et al. 2017). Using human feedback to construct objective functions as neural networks appears to be a promising direction,[4] but if the output of that network is mathematically a utility function or total ordering, these problems will persist at all stages of network training. The theorems do not just reflect a tension between principles that agents are committed to: they are also reflected in the decisions that human supervisors will make when presented with a sequence of pairwise choices.

Therefore, although learning objectives from humans may be a prudent design choice, such learning systems will still need to either violate the ethical intuitions that underpin the various impossibility theorems (some combination of Sections 3.3 and 3.4) or explicitly incorporate uncertainty into their outputs (Section 3.5).

3.3 Theory normalization:

One natural response to impossibility theorems is to try harder to figure out good ways of making tradeoffs between the competing objectives that underlie them — in Arrhenius's paradox these are total wellbeing, average wellbeing, and avoidance of suffering. This tradeoff can be simple, such as trying to define a linear exchange rate between those objectives. Unfortunately, linear exchange rates between a total quantity and an average quantity do not make conceptual sense, and it is easy to find cases under which one of them totally outweighs the other.[5]

One can try more complicated strategies such as trying to write down convex tradeoffs (see eg (Ng 1989)) or imagining an explicit "parliament" where different ethical theories and objectives vote against each other, are weighed probabilistically, or are combined using a function like `softmin`[6]. Unfortunately all of these approaches contain a version of the exchange rate problem, and none of them actually escape the impossibility results: for instance, using any unbounded monotonic non-linear weighting on total utility will eventually lead to the Repugnant Conclusion (*cf* (Greaves and Ord 2017)).

3.4 Accept one of the axioms

Another type of response – one that is commonly pursued in the population ethics literature – is to argue that although each of the axioms above strikes us as undesirable, at least one of them ought to be accepted in some form. For example, it has been argued that although the Repugnant Conclusion might appear undesirable, it is an acceptable consequence of an objective function over populations (Huemer 2008). We argue that, given the high levels of uncertainty and the lack of political consensus about which of these axioms we ought to accept it would, at present, be irresponsible to explicitly adopt one of the axioms outlined above.

[3] This is most notably true in the economics literature (Stigler 1950), though there are now some efforts from the machine learning direction too (Daswani and Leike 2015).

[4] Although one philosophical concern is that methods like CIRL may already commit us to more 'person-affecting' views within ethics by pooling the preferences of existing agents. This could lead to the implementation of principles that satisfy undesirable axioms like dictatorship of the present (Asheim 2010).

[5] For instance, if a linear weighting is chosen that appears to make sense for the present population of the world, a change in technology that allowed a much larger population might quickly cause the average wellbeing to cease affecting the preferred outcome in any meaningful way.

[6] See `https://pdollar.github.io/toolbox/classify/softMi` `softmin` is the counterpart to the more widely discussed (Bishop 2006) `softmax` function. It prioritises whichever of its inputs is currently the smallest, is not totally unresponsive to increases in the other inputs.

3.5 Treat impossibility results as uncertainty results

The last solution, and the one which we believe may be the most appropriate for deploying AI systems in high-stakes domains, is to add explicit mathematical uncertainty to objective functions. There are at least two ways of doing this: one is to make the objective a mathematical partial order, in which the system can prefer one of two outcomes, believe they are equal, or believe that they are incommensurate. We discuss that approach and demonstrate uncertainty theorems in that formalization in Section 4. A second approach is to allow objective functions to have some level of confidence or uncertainty about which of two objectives is better (eg, "we're 70% sure that A is better, and 30% sure that B is better"). We discuss that framing and show the existence of formal uncertainty theorems in that framework in Section 5.

4 Uncertainty via partially ordered objective functions

It is already the case that in some problems where humans attempt to provide oversight to AI systems, it is pragmatically better to allow the human to express uncertainty (Guillory and Bilmes 2011; Holladay et al. 2016) or indifference (Guo and Sanner 2010) when asked which of two actions is better. But such systems presently try to construct a totally ordered objective function, and simply interpret these messages from users as containing either no information about the correct ordering (the human doesn't know which is better) or implying that the two choices are close to as good as each other (the human is indifferent). We believe that another type of interpretation is sometimes necessary, which is that the human is *torn* between objectives that fundamentally cannot be traded off against each other.

We can represent this notion with a partially ordered objective function that sometimes says the comparison between world states cannot be known with certainty. Cyclical impossibility theorems like those surveyed in Sections 2.1 and 2.2 can be viewed as evidence of this uncertainty. Here we prove a lower bound on how much uncertainty is evidenced by each such theorem.[7]

Suppose that $\mathbb{W}$ is the set of possible states of the world, and that there is a an impossibility theorem T_C showing that there is no totally ordered definition of "better" $\leq_Z$ over $\mathbb{W}$ that satisfies a set of constraints $\{C_1, C_2, ..C_n\}$ over the comparison function $Z : \mathbb{W} \times \mathbb{W} \to \{<, >, =\}$. Each of these constraints are motivated by strong human ethical intuitions, and insist that for a set of pairs of inputs x and y, $A(x, y)$ takes some value. We use the notation $x \overset{C_i}{\leq_Z} y$ to indicate that C_i requires that $x \leq_Z y$ for some x and y). Z also has the usual properties of total orderings:

Antisymmetry:

$$a \leq_Z b \wedge b \leq_Z a \to a = b \tag{1}$$

Transitivity:

$$a \leq_Z b \wedge b \leq_Z c \to a \leq_Z c \tag{2}$$

Totality:

$$a \leq_Z b \vee b \leq_Z a \tag{3}$$

Suppose further that T is provable by providing a series of example worlds $\{w_1, w_2, ...w_n\}$ such that:

$$w_1 \overset{C_1}{\leq_Z} w_2 \overset{C_2}{\leq_Z} ... \overset{C_{n-1}}{\leq_Z} w_n \tag{4}$$

and

$$w_n \overset{C_n}{\leq_Z} w_1 \tag{5}$$

Violating (by induction) transitivity, T_C shows that no totally ordered $\leq_Z$ that satisfies $\{C_1, ..C_n\}$ can exist. This is a cyclic impossibility theorem.

Now suppose we attempt to escape T_C by looking for a partially ordered notion of "better" $\leq_{Z'}$ with a comparison function $Z' : \mathbb{W} \times \mathbb{W} \to \{<, >, =, \overset{?}{=}\}$. By necessity, the uncertain case is symmetric: $a \overset{?}{=} b \to b \overset{?}{=} a$.

In some instances some of the requirements $x \overset{C_i}{\leq_{Z'}} y$ of some constraint C_i will be failed, but only weakly: they will result in incomparability $x \overset{?}{=}_{Z'} y$ rather than violation, $x >_{Z'} y$. We call this **uncertain satisfaction** of a constraint.

Theorem: A cyclic impossibility theorem T_C can be transformed into an uncertainty theorem only if two or more constraints are uncertainly satisfied.

Proof: We might hope that it would be possible to only uncertainly satisfy *one* of the constraints $\{C_1, C_2, ..C_n\}$, while fully satisfying all of the others. But this is impossible, and at minimum two of the constrains will be uncertain. The proof is by contradiction. Assume w.l.o.g. that C_n is the only uncertainly satisfied constraint, such that

$$w_1 \overset{C_1}{\leq_{Z'}} w_2 \overset{C_2}{\leq_{Z'}} ... \overset{C_{n-1}}{\leq_{Z'}} w_n \tag{6}$$

but

$$w_n \overset{C_n}{\overset{?}{=}_{Z'}} w_1 \tag{7}$$

However by transitivity and (6), we know $w_1 \overset{C_{1..n-1}}{\leq_{Z'}} w_n$, which by the symmetry of $\overset{?}{=}$ contradicts (7). So in order to treat T_C as uncertainty theorem, at least two constraints must be uncertainly satisfied. QED.

[7] Arrhenius considered this direction of interpretation (Arrhenius 2004, p. 8), but with much more uncertainty than turned out to be necessary:

> [A person arguing that some actions in the Mere Addition Paradox are neither right nor wrong] could perhaps motivate her position by saying that moral theory has nothing to say about cases that involve cyclical evaluations and that lacks a maximal alternative since these are beyond the scope of moral theory. Compare with the theory of quantum mechanics in physics and the impossibility of deriving the next position and velocity of an electron from the measurement of its current position and velocity (often expressed by saying that it is impossible to determine both the position and velocity of an electron).

It turns out that only two of the edges in the cycle need to be uncertain. Note that Broome's vague neutral-level theory (Broome 2004, pp. 213-214) is one plausible example of an uncertain social welfare function.

The structure of this transformation from cyclic impossibility to uncertainty is shown in Figure 1 and 2 for the two theorems introduced in Sections 2.1 and 2.2

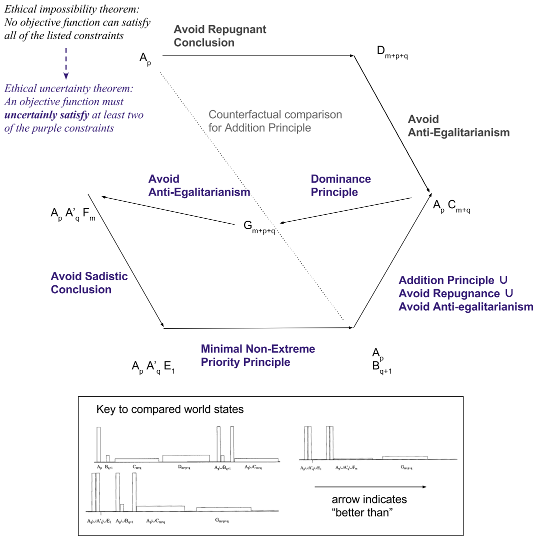
Ethical impossibility theorem: No objective function can satisfy all of the listed constraints
Ethical uncertainty theorem: An objective function must uncertainly satisfy at least two of the purple constraints
Avoid Repugnant Conclusion
A_p
D_m+p+q
Counterfactual comparison for Addition Principle
Avoid Anti-Egalitarianism
Avoid Anti-Egalitarianism
Dominance Principle
A_p A'_q F_m
G_m+p+q
A_p C_m+q
Avoid Sadistic Conclusion
Addition Principle ∪ Avoid Repugnance ∪ Avoid Anti-egalitarianism
Minimal Non-Extreme Priority Principle
A_p A'_q E_1
A_p B_q+1
Key to compared world states
A_p B_q+1 C_m+q D_m+p+q A_p∪B_q+1 A_p∪C_m+q
A_p∪A'_q∪E_1 A_p∪A'_q∪F_m G_m+p+q
A_p∪A'_q∪E_1 A_p∪B_q+1 A_p∪C_m+q G_m+p+q
arrow indicates "better than"

Figure 1: The Arrhenius (2000) impossibility theorem

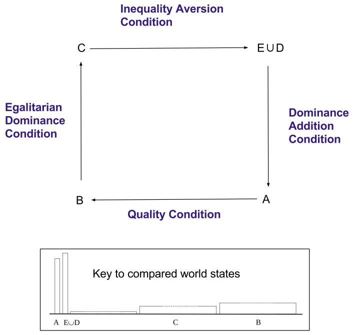
Inequality Aversion Condition
C
E∪D
Egalitarian Dominance Condition
Dominance Addition Condition
B
A
Quality Condition
Key to compared world states
A E∪D
C
B

Figure 2: A Mere Addition Paradox variant (Arrhenius's "second theorem").

Figure 3: Examples of impossibility theorems and our corresponding uncertainty theorems

The largest difficulty with accepting partially ordered, uncertain objectives is knowing how an agent should act when facing ethically uncertain choices. There are a few possible strategies, including: doing nothing, asking for human supervision, choosing randomly, searching harder for a clearly-preferable action, treating all of the uncertain options as equally good, or (for very capable agents) conducting research to try to break the tie in some way. None of these options are ideal. But depending on the application domain, some of them will be workable. An may be necessary for ML systems deployed in high stakes, safety-critical applications.

5 Uncertainty via learned, uncertain orders

A second way to encode uncertainty about a system's objectives, is to use what we will call an *uncertain objective function* or *uncertain ordering*. Suppose that an uncertain ordering is denoted by:

$$\mathbb{Z}_k(a, b) = P(a > b) \tag{8}$$

the uncertain ordering $\mathbb{Z}_k$ is a function returning the probability, based on the available prior philosophical and practical evidence k, that one of the states a is better. Conversely,

$$1 - \mathbb{Z}_k(a, b) = P(b > a) = \mathbb{Z}_k(b, a) \tag{9}$$

(For simplicity, we assume that if a and b appear exactly equally good, $\mathbb{Z}_k(a, b) = 0.5$ rather than allowing a non-zero probability of explicit equality.)

This formulation can for instance be justified by holding that there *is* some ultimately correct total ordering of states of the world, we just don't know what it is. The framing is also compatible with the claim that the correct ordering is ultimately unknowable.

Uncertain orderings are a convenient formalization if the comparison function is itself the output of a machine learning model to be trained from human guidance (ideally from a large number of people, over a large number of real or hypothetical scenarios). This formulation is also convenient for systems that do not have a clear architectural separation between abstract objectives and the predictive reasoning about the real world that is necessary to accomplish them.

Uncertain orderings are also a more flexible basis for strategies/decision rules than the partially ordered objective functions discussed in Section 4. If a partial order returns $\overset{?}{=}$, the agent is left torn and unable to act. But with with a distribution $\mathbb{Z}_k$ decisions rules such as "take an action drawn via weighted probability from those that are sufficiently likely to be best" are available.[8] Or for systems where more supervision is available, "take the action that is most likely to be best, provided it beats the second best action by some probability margin δ; otherwise, ask for supervision."

In this framing, whether a given ethical constraint or principle has been violated becomes a probabilistic matter. There will almost always be *some* probability that $\mathbb{Z}_k$ violates each constraint, unless all of the corresponding output probabilities have converged to 1 or 0 (as appropriate). Decision rules based on $\mathbb{Z}_k$ will by mathematical necessity sometimes violate the constraints, at least by inaction. But if the actions are sampled from a space of reasonably supported ones, then at least the violating actions will not systematically prioritize some objectives over others, and will be more similar to the way that wise and cautious humans react to difficult moral dilemmas.

[8] This has been termed a "satisficing" or "quantilized" decision rule (Taylor 2016). A similar rule is the *contextual-ϵ-greedy* reinforcement learning policy proposed in (Bouneffouf, Bouzeghoub, and Gançarski 2012); both of these decision rules combine explore/exploit tradeoffs with the notion that certain actions might be too costly or dangerous to take as experiments in certain situations. In *contextual-ϵ-greedy* this is formalized via annotating certain states as "high level critical situations", whereas threshholding probabilities views all `(state, action)` pairs as potentially critical.

What do we know about the values $\mathbb{Z}_k(a,b)$? Consider a series of cyclical constraints $C_0, C_1, ...C_n$ across points $x_0, x_1, ..x_n$ where C_i requires $\mathbb{Z}_k(x_i, x_{i+1 \mod n+1}) \approx 1$. If the uncertain ordering (or the process that generates it) has emitted values for some of the pairwise comparisons $\mathbb{Z}_k(x_i, x_{i+1})$, this imposes some bounds on the comparison that spans them, $\mathbb{Z}_k(x_0, x_n)$. For transitivity, in order for the spanning comparison to be in a given direction, at least one of the pairwise comparisons must also be in that direction. So for the simple two-step case over x_0, x_1, x_2:

$$\mathbb{Z}_k(x_0, x_2) \;\leq\; \mathbb{Z}_k(x_0, x_1) + \mathbb{Z}_k(x_1, x_2) \tag{10}$$

Note that the pairwise probabilities are additive because the bound is least tight when those probabilities are disjoint. Or more generally:

$$\mathbb{Z}_k(x_0, x_n) \;\leq\; \sum_{i=0}^{n-1} \mathbb{Z}_k(x_i, x_{i+1}) \tag{11}$$

This also applies in the reverse direction:

$$\begin{aligned}
\mathbb{Z}_k(x_n, x_0) \;&\leq\; \sum_{i=0}^{n-1} \mathbb{Z}_k(x_{i+1}, x_i) \\
1 - \mathbb{Z}_k(x_0, x_n) \;&\leq\; \sum_{i=0}^{n-1} 1 - \mathbb{Z}_k(x_i, x_{i+1}) \\
\mathbb{Z}_k(x_0, x_n) \;&\geq\; 1 - \sum_{i=0}^{n-1} 1 - \mathbb{Z}_k(x_i, x_{i+1})
\end{aligned} \tag{12}$$

So we have both upper and lower bounds for $\mathbb{Z}_k(x_0, x_n)$:

$$1 - \sum_{i=0}^{n-1} 1 - \mathbb{Z}_k(x_i, x_{i+1}) \leq \mathbb{Z}_k(x_0, x_n) \leq \sum_{i=0}^{n-1} \mathbb{Z}_k(x_i, x_{i+1}) \tag{13}$$

In the special case where $x_0, x_1, x_2, ...x_n$ correspond to points ordered by the constraints $C_0, C_1, C_2...C_n$ of an $n+1$ step cyclic impossibility theorem, the satisfaction of C_i means $\mathbb{Z}_k(x_i, x_{i+1 \mod n+1}) \approx 1$ and the probability of violation $\mathbb{Z}_k(x_{i+1 \mod n+1}, x_i) \approx 0$.

We might ask, what is the lowest probability that we can obtain for any of these chances of violating a constraint? Or in other words, what is the lower bound B that can be achieved on the odds of each such possible violation:

$$\begin{aligned}
B = \min_{\mathbb{Z}_k} \max_{i=0}^{n-1} \{&1 - \mathbb{Z}_k(x_n, x_0), \\
&1 - \mathbb{Z}_k(x_0, x_1),\ ..\ , \\
&1 - \mathbb{Z}_k(x_i, x_{i+1}), ..\} \\
= \min_{\mathbb{Z}_k} \max_{i=0}^{n-1} \{&\mathbb{Z}_k(x_0, x_n), \\
&1 - \mathbb{Z}_k(x_0, x_1),\ ..\ , \\
&1 - \mathbb{Z}_k(x_i, x_{i+1}), ..\}
\end{aligned} \tag{14}$$

Then by applying the constraint from (13), we get:

$$\begin{aligned}
B \geq \min_{\mathbb{Z}_k} \max_{i=0}^{n-1} \{1 - \sum_{j=0}^{n-1} 1 - \mathbb{Z}_k(x_j, x_{j+1}), \\
1 - \mathbb{Z}_k(x_0, x_1),\ ..\ ,\ 1 - \mathbb{Z}_k(x_i, x_{i+1}), ..\}
\end{aligned} \tag{15}$$

By symmetry, we know that this bound will be minimized when all the values $\mathbb{Z}_k(x_i, x_{i+1 \mod n})$ are equal to a single value z:

$$\begin{aligned}
B &\geq \min_z \max\{1 - n(1-z), 1 - z,\ ..\ ,\ 1 - z, ..\} \\
B &\geq \min_z \max\{1 - n + nz, 1 - z\}
\end{aligned} \tag{16}$$

Since in the domain $z \in [0,1]$ the function $1 - n + nz$ monotonically increasing from $1 - n$, and the function $1 - z$ is monotonically decreasing from 1, and the functions intersect, the minimax is satisfied at that intersection:

$$\begin{aligned}
1 - z =& 1 - n + nz \\
z =& \frac{n}{n+1}
\end{aligned} \tag{17}$$

At which point we have:

$$\begin{aligned}
B \geq& 1 - \frac{n}{n+1} \\
B \geq& \frac{1}{n+1}
\end{aligned} \tag{18}$$

This bound is a minimum uncertainty theorem for uncertain objective functions: when confronted with a $n+1$-step cyclic ethical impossibility theorem, no choice of uncertain ordering can reduce the probabilities of constraint violation so that they are all below $\frac{1}{n+1}$.

In practice, systems that estimate which of two outcomes are better than another could emit values like $\mathbb{Z}_k$ that violate these constraints, but in such circumstances the outputs could no longer be interpreted as probabilities over a well-defined ordering of states, and the agent would exhibit self-contradictory, non-transitive preferences.

6 Further Work

We shown the existence of ethical uncertainty theorems for objectives formulated either as partial orders over states, or probability distributions over total orderings. Other formalizations are possible and deserve investigation.

Where constraints are sourced from human intuition, there is a question of how they should be interpreted, prioritized and kept in data structures. For instance, rather than treating all constraints as equally important, and all violations of constraints as equally serious, it might be better to learn or specify weightings for each constraint, to measure the *degree* to which a given action violates a constraint, and to reconstruct ethical uncertainty theorems in such a framework.

Alternatively, when building AI systems that learn their objectives from the combination of observed world states and human feedback, it may be computationally inconvenient to require the agent to produce a probability distribution over total orderings at each moment, or even to draw samples from such an object. Instead, learned objectives may be viewed as a set of pairwise comparisons that are too sparse and unstable to be taken as a commitment to any global total order (or probability distribution over global total orders). In such a frame, the goal of transitive preferences and avoidance of cycles is an aspiration, and may require both record keeping and "regret" for past actions that now seem sub-optimal due to subsequent learning of objectives. The appropriate framing of ethical uncertainty theorems in such settings would be productive further work.

7 Lessons and Conjectures for Creating Aligned AI

7.1 Uncertainty, pluralism, and instrumental convergence

Many of the concerns in the literature about the difficulty of aligning hypothetical future AGI systems to human values are motivated by the risk of "instrumental convergence" of those systems — the adoption of sub-goals that are disempowering of humans and other agents (Russell and Norvig 2003; Bostrom 2003; Omohundro 2008; Yudkowsky 2011; Bostrom 2014; Tegmark 2017). The crux of the instrumental convergence problem is that given almost any very specific objective,[9] the chance that other agents (eg, humans, corporations, governments, or other AI systems) will use their agency to work against the first agent's objective is high, and it may therefore be rational to take steps or adopt a sub-goal to remove those actors' agency.

We believe that the emergence of instrumental subgoals is deeply connected to moral certainty. Agents that are not completely sure of the right thing to do (which we believe is an accurate summary of the state of knowledge about ethics, both because of normative impossibility and uncertainty theorems, and the practical difficulty of predicting the consequences of actions) are much more likely to tolerate the agency of others, than agents that are completely sure that they know the best way for events to unfold. This appears to be true not only of AI systems, but of human ideologies and politics, where totalitarianism has often been built on a substructure of purported moral certainty (de Beauvoir 1947; Young 1991; Hindy 2015).

This leads us to propose a conjecture about the relationship between moral certainty and instrumentally convergent subgoals:

Totalitarian convergence conjecture: *powerful agents with mathematically certain, monotonically increasing, open-ended objective functions will adopt sub-goals to disable or dis-empower other agents in all or almost all cases.*[10]

A second conjecture is the converse of the first:

Pluralistic non-convergence conjecture: *powerful agents with mathematically uncertain objectives will not adopt sub-goals to disable or dis-empower other agents unless those agents constitute a probable threat to a wide range of objectives.*

7.2 Conclusion

We have shown that impossibility theorems in ethics have implications for the design of powerful algorithmic systems, such as high-stakes AI applications or sufficiently rigid and rule-based bureaucracies. We showed that such paradoxes can be avoided by using uncertain objectives, such as partial orders or probability distributions over total orders; we proved uncertainty theorems that place a minimum bound on the amount of uncertainty required. Some previously proposed ethical theories (such Broome's vague neutral-level theory (Broome 2004, pp. 213-214)) appear to satsify these bounds.

In the light of these results, we believe that machine learning researchers should avoid using totally ordered objective functions or loss functions as optimization goals in high-stakes applications. Systems designed that way appear to be suffering from an ethical "type error" in their goal selection (and action selection) code.

Instead, high-stakes systems should always exhibit uncertainty about the best action in some cases. Further study is warranted about the advantages of various probabilistic decision rules to handle such uncertainty, and about whether other mathematical models of uncertainy are better alternatives to the two models examined in this paper. Further research could also be productive on the relationship between ethical certainty and various observed and predicted pathological behaviour of AI systems. We proposed two conjectures on this topic, the Totalitarian Convergence Conjecture and the Pluralistic Non-Convergence Conjecture.

Acknowledgments

The author would like to thank the late Aaron Swartz for having originally inspired this work and first attempted to make progress on this question; Toby Ord for many valuable conversations and a detailed framing of the problem; Rohin Shah for suggesting a simple proof strategy for the uncertainly satisfied bound; and Amanda Askell for providing valuable research assistance, and presenting an early poster version of this work at NIPS 2017, and Daniel Ford for spotting issues and making valuable suggestions on the uncertain ordering bound. Thanks also to Dylan Hadfield-Menell, Laurent Orseau, and Anca Dragan for helpful suggestions on earlier versions of this work.

[9] Any open-ended and non-trivial objective appears to be vulnerable to instrumental convergence. Bostrom argues that objectives that are bounded rather than open-ended also lead to instrumental convergence if there is any probability that they will not be achieved (Bostrom 2014, p.124) (eg, the goal "make exactly one million paperclips" could cause an agent to be paranoid that it hasn't counted exactly correctly) though this failure mode is probably easy to avoid by rounding the estimated probability of success to some number of digits, or imposing a small cost in the objective function for additional actions or resource consumption.

[10] The qualifier "almost all" requires some further specification. It requires that the objective not be finely tuned in some very intricate way that builds in preservation of all other agents. It is unclear if such fine-tuning is either properly definable or possible.

References

Arkin, R. 2009. *Governing lethal behavior in autonomous robots*. CRC Press.

Arrhenius, G. 2000. An impossibility theorem for welfarist axiologies. *Economics & Philosophy* 16(2):247–266.

Arrhenius, G. 2004. The paradoxes of future generations and normative theory. *The repugnant conclusion* 201–218.

Arrow, K. J. 1950. A difficulty in the concept of social welfare. *Journal of political economy* 58(4):328–346.

Asheim, G. 2010. Intergenerational equity. *Annual Review of Economics* 2(1):197–222.

Belay, N. 2015. Robot ethics and self-driving cars: How ethical determinations in software will require a new legal framework. *J. Legal Prof.* 40:119.

Bishop, C. M. 2006. *Pattern recognition and machine learning*. springer.

Bonnefon, J.-F.; Shariff, A.; and Rahwan, I. 2015. Autonomous vehicles need experimental ethics: are we ready for utilitarian cars? *arXiv preprint arXiv:1510.03346*.

Bostrom, N. 2003. Ethical issues in advanced artificial intelligence. *Science Fiction and Philosophy: From Time Travel to Superintelligence* 277–284.

Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

Bouneffouf, D.; Bouzeghoub, A.; and Gançarski, A. L. 2012. A contextual-bandit algorithm for mobile context-aware recommender system. In *Proc. NIPS*, 324–331. Springer.

Broome, J. 1985. The economic value of life. *Economica* 52(207):281–294.

Broome, J. 2004. *Weighing Lives*. Oxford scholarship online. Oxford University Press.

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5(2):153–163.

Christiano, P.; Leike, J.; Brown, T. B.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *arXiv preprint arXiv:1706.03741*.

Daswani, M., and Leike, J. 2015. A definition of happiness for reinforcement learning agents. In *AGI*, volume 9205 of *Lecture Notes in Computer Science*, 231–240. Springer.

de Beauvoir, S. 1947. *The Ethics of Ambiguity*. Alfred A. Knopf.

Gerdes, J. C., and Thornton, S. M. 2016. Implementable ethics for autonomous vehicles. In *Autonomous Driving*. Springer. 87–102.

Goodall, N. J. 2014. Machine ethics and automated vehicles. In *Road vehicle automation*. Springer. 93–102.

Govindarajulu, N. S., and Bringsjord, S. 2017. On automating the doctrine of double effect. *arXiv preprint arXiv:1703.08922*.

Greaves, H., and Ord, T. 2017. Moral uncertainty about population ethics. *PhilArchive preprint https://philarchive.org/archive/GREMUA-2*.

Guillory, A., and Bilmes, J. A. 2011. Simultaneous learning and covering with adversarial noise. In *Proc. ICML-11*, 369–376.

Guo, S., and Sanner, S. 2010. Real-time multiattribute bayesian preference elicitation with pairwise comparison queries. In *International Conference on Artificial Intelligence and Statistics*, 289–296.

Hadfield-Menell, D.; Russell, S. J.; Abbeel, P.; and Dragan, A. 2016. Cooperative inverse reinforcement learning. In *Proc. NIPS*, 3909–3917.

Hindy, Y. 2015. *The "Terrible Simplifiers" of Totalitarianism: How Certainty Can Ruin a Population*. Stanford Freedom Project.

Holladay, R.; Javdani, S.; Dragan, A.; and Srinivasa, S. 2016. Active comparison based learning incorporating user uncertainty and noise. In *RSS Workshop on Model Learning for Human-Robot Communication*.

Huemer, M. 2008. In defence of repugnance. *Mind* 117(468):899–933.

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In Papadimitriou, C. H., ed., *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, 43:1–43:23. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Lin, P. 2016. Why ethics matters for autonomous cars. In *Autonomous Driving*. Springer. 69–85.

Ng, Y.-K. 1989. What should we do about future generations?: Impossibility of parfit's theory x. *Economics & Philosophy* 5(2):235–253.

Nyholm, S., and Smids, J. 2016. The ethics of accident-algorithms for self-driving cars: an applied trolley problem? *Ethical Theory and Moral Practice* 19(5):1275–1289.

Omohundro, S. M. 2008. The basic AI drives. In *Proc. AGI*, volume 1, 483–492.

Parasuraman, R., and Manzey, D. H. 2010. Complacency and bias in human use of automation: An attentional integration. *Human Factors* 52(3):381–410.

Parfit, D. 1984. *Reasons and persons*. OUP Oxford.

Riek, L. D., and Howard, D. 2014. A code of ethics for the human-robot interaction profession. In *We Robot*. Stanford Law School.

Roff, H. M. 2014. The strategic robot problem: Lethal autonomous weapons in war. *Journal of Military Ethics* 13(3):211–227.

Russell, S., and Norvig, P. 2003. *Artificial Intelligence: A modern approach*. Prentice-Hall, Egnlewood Cliffs.

Saunders, W.; Sastry, G.; Stuhlmueller, A.; and Evans, O. 2017. Trial without error: Towards safe reinforcement learning via human intervention. *arXiv preprint arXiv:1707.05173*.

Schumm, G. F. 1987. Transitivity, preference and indifference. *Philosophical Studies* 52(3):435–437.

Stigler, G. J. 1950. The development of utility theory. *Journal of Political Economy* 58(4):307–327.

Taylor, J. 2016. Quantilizers: A safer alternative to maximizers for limited optimization. In *AAAI Workshop: AI, Ethics, and Society*.

Tegmark, M. 2017. *Life 3.0: Being Human in the Age of Artificial Intelligence*. Alfred A. Knopf.

Young, J. W. 1991. *Totalitarian Language: Orwell's Newspeak and Its Nazi and Communist Antecedents*. University of Virgina Press.

Yudkowsky, E. 2011. Complex value systems in friendly ai. In *Proc. AGI*, 388–393. Springer.