— Fairness —

Ulle Endriss
Institute for Logic, Language and Computation
University of Amsterdam

[Guest Lecture for "Fairness, Accountability, Confidentiality and Transparency in AI" — MSc AI, 9 January 2026]

Outline

The term "fairness" has been used to refer to lots of different concepts.

We'll focus on fairness in decision making scenarios affecting multiple individuals, such as *allocating resources* or *elections*. We'll see:

- many *different definitions* of the notion of "fairness"
- some *unintended consequences* of seemingly reasonable definitions
- a glimpse at a *principled approach* towards choosing a definition

This is a huge area of research. While we'll see but a few examples, we'll extract some general *take-home messages* from these examples.

Fairness in Machine Learning

In the context of ML, fairness is often taken to be about *classification:*

> *The predicted value for a given variable should be statistically independent of sensitive attributes such as gender or ethnicity.*

This boils down to ensuring that people who *should be treated equally* (based on *relevant* attributes) really *are treated equally*.

While this is addressing a difficult and important problem, it's rather narrow an interpretation of the much broader concept of fairness.

Fairness in Philosophy and Economics

We are going to look into fairness in decision making more generally:

> *Decisions affecting a group of individuals should adequately balance the interests of those individuals.*

This is fairness as studied in *philosophy* and *economics*, which entered AI through *computational social choice* and *algorithmic game theory*.

Of course, this first attempt at a definition is highly underspecified: what is "adequate"?, how do we model "interests"?

Case Study: Resource Allocation

Let G be a finite set of *goods*. So 2^G is the set of all *bundles* of goods.

Let $N = \{1, \ldots, n\}$ be a set of *agents*. Need to decide *who* gets *what*.

An *allocation* is a mapping $A : N \to 2^G$ from agents to bundles that satisfies $A(1) \cup \cdots \cup A(n) = G$ and $A(i) \cap A(j) = \varnothing$ for $i \neq j$.

Each agent $i \in N$ has a *utility function* $u_i : 2^G \to \mathbb{R}_{\geqslant 0}$, indicating how much she likes any given bundle $S \subseteq G$.

We shall assume that utility functions are also defined on individual goods $x \in G$ (not just on bundles $S \subseteq G$) and that they are *additive:*

$$u_i(S) = \sum_{x \in S} u_i(x)$$

Exercise: *Is this assumption of additivity reasonable? Why (not)?*

The Additivity Assumption

The assumption of additivity for utility functions clearly is a simplifying assumption that won't always be justified. Examples:

- *Superadditivity:* Suppose G includes a left and a right shoe. Then you might have $u_i(L) = u_i(R) = 0$ but $u_i(\{L, R\}) = 50$.
- *Subadditivity:* Suppose G includes (non-resellable) tickets for a cinema and a theatre show that take place at the same time. Then you might have $u_i(C) = 15$ and $u_i(T) = 35$ but $u_i(\{C, T\}) = 35$.

> Abstraction is useful when trying to understand a new concept, but simplifying assumptions should always be questioned.

Anyway, for today, additivity will do for us . . .

Perfect Equality

Maybe the most natural interpretation of the term “fairness” would be to require that all agents enjoy the exact same level of utility.

So: find an allocation A with $u_i(A(i)) = u_j(A(j))$ for all $i, j \in N$!

Exercise: *Is this a good solution? Why (not)?*

Utilitarian Social Welfare

What makes for a good allocation? A natural approach to measuring the quality of an allocation is to compute its *utilitarian social welfare:*

$$usw(A) = \sum_{i \in N} u_i(A(i))$$

This idea can be traced back to the philosophy of *utilitarians* such as Jeremy Bentham (1748–1832) and John Stuart Mill (1806–1873).

Observation: Increasing social welfare means increasing *average utility*.

So: look for an allocation that maximises utilitarian social welfare!

Exercise: *Describe an algorithm to do this. What is its runtime?*

Exercise: *Is maximising USW a good social objective? Why (not)?*

Egalitarian Social Welfare

The *egalitarian social welfare* of an allocation is defined as the utility of the worst-off agent for that allocation:

$$esw(A) = \min_{i \in N} u_i(A(i))$$

ESW is inspired by the work of John Rawls (1921–2002), one of the most important moral and political philosophers of the 20th century.

So: look for an allocation that maximises egalitarian social welfare!

Exercise: *Describe an algorithm to do this. What is its runtime?*

Intractability

In fact, finding an allocation with maximal egalitarian social welfare is computationally intractable. Here's the corresponding decision problem:

EgalSW
Input: Allocation problem $\langle N, G, (u_i)_{i \in N} \rangle$ and target value α
Question: Is there an allocation A such that $esw(A) \geqslant \alpha$?

Proposition 1 (Bouveret et al., 2005) EgalSW *is NP-complete.*

Recall: *NP-completeness* $=$ *NP-hardness* $+$ *NP-membership*

Proving NP-membership is straightforward: we simply need to observe that, when someone guesses an allocation A they claim does the job, we can verify in polynomial time that A really meets our conditions.

So we are left with having to prove NP-hardness . . .

S. Bouveret, M. Lemaître, H. Fargier, and J. Lang. Allocation of Indivisible Goods: A General Model and Some Complexity Results. AAMAS-2005.

Proof of NP-Hardness

We get *NP-hardness* of $\textsc{EgalSW}$ even when there are just *two agents*. To show this, use a *reduction* from this problem known to be NP-hard:

$\textsc{Partition}$
Input: Numbers $(w_1, \ldots, w_m) \in \mathbb{N}^m$ and threshold $k > 0$
Question: Is there an $I \subseteq \{1, \ldots, m\}$ s.t. $|\sum_{i \in I} w_i - \sum_{i \notin I} w_i| < k$?

Take any given instance of $\textsc{Partition}$ and construct an instance of $\textsc{EgalSW}$ with $N = \{1, 2\}$ and $G = \{x_1, \ldots, x_m\}$ as follows:

$$u_1(x_i) := w_i \text{ for all } i \in \{1, \ldots, m\}, \text{ and } u_2 := u_1$$

Then every attempt to solve $\textsc{Partition}$ corresponds to an allocation (by giving item x_i to agent 1 iff $i \in I$) and *vice versa*, and we can obtain the partition quality as a linear combination of the egalitarian social welfare: $|\sum_{i \in I} w_i - \sum_{i \notin I} w_i| = (w_1 + \cdots + w_m) - 2 \cdot esw(A)$.

Thus, solving $\textsc{EgalSW}$ is at least as hard as solving $\textsc{Partition}$. ✓

Take-Home Message

Definitions that are elegant and normatively attractive may turn out to have unintended consequences in some other respect.

Discussion

So do we need to abandon the idea of using egalitarian social welfare?

No, of course not:

- NP-hardness makes things difficult, but not necessarily impossible.
- ESW is still an improvement over the naïve “perfect equality” idea.
- Btw, if we drop the additivity assumption, USW is NP-hard as well.

Nash Social Welfare

Recall that utilitarian social welfare was about *adding* utilities. What if we *multiply* them instead? This is known as *Nash social welfare:*

$$nsw(A) = \prod_{i \in N} u_i(A(i))$$

This idea goes back to Nobel laureate John Nash (1928–2015).

Seems counterintuitive at first. Some intuition for why it makes sense:

- NSW favours increases in overall utility (just like USW)
- NSW favours inequality-reducing redistributions ($2 \cdot 6 < 4 \cdot 4$)

Exercise: *Good solution concept, but not perfect. Do you see why?*

The Axiomatic Method

So how to pick the right solution concept (definition of “fairness”)?

The central approach in social choice theory is the *axiomatic method:*

- identify and formalise normatively appealing properties (“axioms”)
- systematically check which solution concepts satisfy your axioms

Let’s restrict attention to possible definitions of *social welfare:*

$$sw : (N \to 2^G) \to \mathbb{R}$$

Examples include the functions usw, esw, and nsw we just discussed.

We’ll review two examples for simple axioms encoding basic properties of social welfare that we might care about . . .

Axiom: Scale Independence

Suppose one agent changes the "currency" she uses to measure her own utility (say, from euros to dollars). We wouldn't want judgments of relative allocation quality to be affected by such a change:

> Let A and A' be two allocations with $sw(A) \leqslant sw(A')$.
>
> Then $sw(A) \leqslant sw(A')$ should remain true if, for one $i \in N$, we replace u_i by u'_i, where $u'_i(S) := c \cdot u_i(S)$ for all $S \subseteq G$, for some fixed "conversion factor" $c > 0$.

Findings: NSW satisfies this axiom of *scale independence*, but neither USW nor ESW does. So this is helpful to differentiate. *Nice!*

Axiom: The Pigou-Dalton Principle

A notion of social welfare sw respects the *Pigou-Dalton Principle* if it encourages pairwise inequality-reducing utility redistributions:

> $sw(A) \leqslant sw(A')$ should hold for any two allocation A and A' that satisfy these conditions for two specific agents $i, j \in N$:
>
> - $u_k(A(k)) = u_k(A'(k))$ for all $k \in N \setminus \{i, j\}$
> - $u_i(A(i)) + u_j(A(j)) \leqslant u_i(A'(i)) + u_j(A'(j))$
> - $|u_i(A(i)) - u_j(A(j))| > |u_i(A'(i)) - u_j(A'(j))|$
>
> Thus, *only two agents are involved*, the change from A to A' is (at least) *mean-preserving*, and it is *inequality-reducing*.

Findings: All three concepts satisfy this (though USW trivially so)!

It can be difficult to get formal results that match our intuitions (often, because those intuitions are not entirely correct).

Exercise: *Idea for a "reasonable" SW measure that fails Pigou-Dalton?*

Revisiting the Basic Model

Ultimately, agents have *preferences over allocations* (induced by their preferences over bundles). So can think of the fair allocation problem as a *voting problem*, where the allocations are the candidates.

Today we modelled those *preferences* as *utility functions*. *Good idea?*

- Using utility functions presupposes that *preference intensity* makes sense ("I like A *twice as much* as B", not just "*more than*").
- Also: *interpersonal comparison* ("I like A *as much* as you like B").

> Not just abstract models that are too simplistic can be inadequate. The same is true for models that are too expressive.

Alternative idea: *preferences as orders* ($A \succcurlyeq_i B$ iff A no worse than B)

Exercise: *How do you construct $\succcurlyeq_i$ given u_i? Other direction?*

Last Slide

This has been an introduction to thinking about the rich concept of fairness in a principled manner. Final take-home messages:

There are many different ways to define “fairness”, depending on application context and issues you want to emphasise.

If you need to use fairness in your work, look for (and adapt) definitions in the literature, rather than reinventing the wheel!

There are mathematical tools for analysing and comparing different notions of fairness in a principled manner. Use them!

Want to learn more about this? Relevant courses:

- *Game Theory* (Apr/May, by me), should be taken first
- *Algorithmic Game Theory* (Sept/Oct, by Guido Schäfer)
- *Computational Social Choice* (Nov/Dec, by me)

Also: *Handbook of Computational Social Choice* (`bit.ly/HBcomsoc`)