Trustworthy AI Needs Unbiased Dictators!

Kian Abolfazlian

Trustworthy AI Needs Unbiased Dictators!

Kian Abolfazlian[0000−0001−7826−547X]

SPIDER Business Ideas Architects, Stockholm, Sweden
kian.abolfazlian@spider-bia.com

Abstract. EU Draft Ethics guidelines for Trustworthy AI [8] has been proposed to promote ethical, lawful and robust AI solutions. In this article, we entertain the systemic issues and challenges of any development of the proposed guidelines.

Keywords: Trustworthy AI · Ethical AI · Dictatorial decision-makers · Manipulation · Cognitive biases.

1 Is Trustworthy AI achievable?

The accelerated enthusiasm of all walks of life for Artificial Intelligence and its applications has been both blessing and preoccupying. The AI solutions, backed by private and public sector interests have opened different and interesting avenues for how to address societal as well as business challenges, which were not thought possible.

As usual for any new technological advances, there have been many lessons learned, as a direct result of application of AI solutions in varied contexts. The lessons have shown us that, AI solutions, much more than any other type of solutions, are affected by the cognitive biases, inherent in the way, their human designers, developers and implementers among others, interact with their end-users. Every solution is biased in that the design choices, underlying its creation, naturally includes and excludes some groups of end-users. In the case of AI solutions, their inherent *biased-by-design* scope have been much faster detected and criticized, due to their rapid advances, achieved reach and application contexts. As a novel approach, adjectives such as Trustworthy and Ethical have been used in describing what AI solutions must be.

As a step towards this, European Union (EU) commission's High-level Expert Group on AI (AI HLEG) has formulated a framework for Trustworthy AI [8]. They define Trustworthy AI as follows:

> *Trustworthy Al has three components: (1) it should be lawful, ensuring compliance with all applicable laws and regulations (2) it should be ethical, demonstrating respect for, and ensure adherence to, ethical principles and values and (3) it should be robust, both from a technical and social perspective, since, even with good intentions, AI systems can cause*

> *unintentional harm. Trustworthy Al concerns not only the trustworthiness of the Al system itself but also comprises the trustworthiness of all processes and actors that are part of the system's life cycle.*

As well AI HLEG defines Ethical AI as:

> *The development, deployment and use of Al that ensures compliance with ethical norms, including fundamental rights as special moral entitlements, ethical principles and related core values. It is the second of the three core elements necessary for achieving Trustworthy Al.*

Furthermore, they define the following non-exhaustive list of requirements, needed to ensure Trustworthy AI. The requirements include systemic, individual and societal factors:

1. ***Human agency and oversight****, including fundamental rights, human agency and human oversight.*
2. ***Technical robustness and safety****, including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility.*
3. ***Privacy and data governance****, including respect for privacy, quality and integrity of data, and access to data.*
4. ***Transparency****, including traceability, explainability and communication.*
5. ***Diversity, non-discrimination and fairness****, including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation.*
6. ***Societal and environmental wellbeing****, including sustainability and environmental friendliness, social impact, society and democracy.*
7. ***Accountability****, including auditability, minimization and reporting of negative impact, trade-offs and redress.*

These seven requirements are all interrelated in that they are *"all of equal importance, support each other, and should be implemented and evaluated throughout the Al system's life-cycle"* [8]. This is depicted in Figure 1. As well, the AI HLEG has presented technical and non-technical methods to realize Trustworthy AI (Figure 2).

The focus on the design and design choices behind the development of AI solutions is inherent in the way the concept of Trustworthy AI has been handled in AI HLEG guidelines.

In our opinion, here lies the biggest challenge to achieving these guidelines' goals. The guidelines develop a set of check-lists and principals, which must be observed, whenever any taken design decision is to be examined for its ethical character and trustworthiness. This could be the task of a group of developers or internal "AI auditors". In any case, the dynamics of how such groups act have been studied by pioneers such as Nobel laureate, Kenneth Arrow [1], and philosophers and economists such as Allan Gibbard [2] [3] [4], Mark Satterwaithe [7], and Aanund Hylland [6], as well as Cognitive Scientists such as Gärdenfors [5] to mention a few.

For instance, if we have, for a specific design factor, n alternatives, A_1 to A_n, where the members of the committee (or development team) are asked to present an ordering $\succ$ to show their preference, e.g. $A_1 \succ \ldots \succ A_n$, [1] has shown (Theorem 3 in section 2, below) that, as long as;

1. There are at least 3 possible outcomes

2. If everybody in the group prefers, for instance, an alternative A_i to another alternative A_j, then the result of the voting would also reflect this unanimity of preferences

3. Each voter's preference between any two alternatives A_i and A_j, is indifferent to the preferences of other alternatives

Then *the voting process would be dictatorial*, i.e. among the voters, there would be a *dictator*, in the sense of a person, whose choices and preferences dictates what the committee would decide, regardless of how other members of the committee are voting.

Even if we change the scheme, so the alternatives are ordered from most preferred to the least one as above, but, now, the alternatives are *ranked using an ordinal system*, we would still have challenges. Here, the alternatives are ordered, so the most preferred alternative (rank 1) is given the highest point (from an ordinal system), and so forth. Then the points are counted in a voting process to declare a winner. Gibbard [2] and Satterwaithe [7] showed (Theorem 2, section 2, below) that *as long as there are at least 3 possible outcomes, then the voting process is either manipulable or dictatorial.*

A manipulable voting process is one, where the voters can vote tactically. This means that they, realizing that their preferred or sincere ordering of the alternatives would not result in what *serves them best*, would change their vote

(i.e. an insincere vote) so to achieve the result that is closest to their preference.

In above situations, the alternatives are ordered strictly, so the voters cannot order them in a way that there are ties among the alternatives (i.e. not a strict order). A more realistic situation would allow a not-strict ordering of alternatives. Even for these situations, Gibbard [2] showed (Theorem 1, section 2, below) that *if there are at least 3 outcomes, and the committee does not have a dictator, then the voting is manipulable.*

Furthermore, Gärdenfors [5] has shown (Theorem 4, section 2 below), that *any democratic voting process with at least 3 voters is manipulable*!

Here a voting process is democratic, if it is *anonymous* (i.e. the voting process treats every voter the same way), it is *neutral* (i.e. the voting process treats every alternative the same way), and it satisfies the *Condorcet criterion* (i.e. majority rules).

The results on the systemic issues of such decision making processes are many. One has even looked at situations where some of the alternatives are dependent of or discovered by chance (e.g. [4] and [6]). In the next section the above mentioned results are described in a more technical terms.

2 Dictators and tactical voters in a decision-making context

In the following, we follow the definitions and technical format of [2].

Definition 1. Game form. *A game form is characterized by:*

*(i) A set X, whose members are called **possible outcomes**, or simply **outcomes**. Unless otherwise stated, variables x, y, and z will range over outcomes.*

*(ii) A positive integer n, called the **number of players**. The n players will be denoted by the integers 1 to n, and variables i, j, and k will range over these integers.*

*(iii) n sets S_i, one for each i. For each i the members of S_i are called strategies for i. The word "strategy," then, refers here to what in game theory is usually called a "pure strategy." An n-tuple $(s_1, \ldots, s_n)$, with $s_1 \in S_1, \ldots, s_n \in S_n$ will be called a **strategy** n-tuple. Strategy n-tuples will be indicated by bold-face small letters on the pattern $\boldsymbol{s} = (s_1, \ldots, s_n)$, $\boldsymbol{s'} = (s'_1, \ldots, s'_n)$, and so forth.*

(iv) A function g, defined for every strategy n-tuple, whose range is X. Strictly speaking, a game form is simply a function g which can be characterized as above. We can define a game form, then, as a function whose domain is the Cartesian product $S_1 \times \ldots \times S_n$ of a finite number of finite non-empty sets. Its values are called ***outcomes****, its arguments are called strategy n-tuples, and a member of a set S_i is called a* ***strategy*** *for i.*

Definition 2. Orderings

(i) An ***ordering*** *of a set Z is a two-place relation R between members of Z, (i.e. $R \subseteq Z \times Z$) such that for all x, y and z in Z*
- $\neg(xRy \ \wedge \ yRx)$
- $xRz \Rightarrow (xRy \ \vee \ yRz)$

As such, we can define an ordering R of the set X of outcomes of a game form g

(ii) A ***preference ordering*** *P is an ordering of X, the set of outcomes of a game form g. The relation xPy then means "x* ***is preferred to*** *y" under ordering P. For distinct x and y in X, we may have that $(x, y) \notin P$, i.e. neither xPy nor yPx. In this case we say that x and y are* ***indifferent under ordering*** *P. As such P indicates* ***strict preference*** *between 2 elements of X.*

(iii) For any preference ordering P of set of outcomes X, we can define another binary relation $R \subseteq X \times X$, such that: $xRy \iff \neg yPx$. For any distinct x and y in X, the relation R indicates their ***preference or indifference****.*

(iv) For any preference ordering P of set of outcomes X, we can define another binary relation $I \subseteq X \times X$, such that: $xIy \iff (\neg yPx \wedge \neg xPy)$. For any distinct x and y in X, the relation I indicates their ***absolute indifference****.*

Remark 1. We will use the following notation. For any n-tuple $\boldsymbol{s} = (s_1, \ldots, s_n)$, we indicate the result of the altering of its k-th place as follows:

$$\boldsymbol{s}\langle k/t\rangle = (s_1, \ldots, s_{k-1}, t, s_{k+1}, \ldots, s_n)$$

In other words, $\boldsymbol{s}' = \boldsymbol{s}\langle k/t\rangle \quad \text{iff} \quad \{(s'_k = t) \ \wedge \ (\forall i)\,[i \neq k \ \Rightarrow \ s'_i = s_i]\}$

Definition 3. ***P*-dominance**. *Where P is a preference ordering, a strategy t is P-dominant for k if for every strategy n-tuple $\boldsymbol{s}$, we have that $g(\boldsymbol{s}\langle k/t\rangle)\ R\ g(\boldsymbol{s})$.*

In other words, t is P-dominant for k iff no matter what strategies are fixed for everyone else, strategy t for k produces an outcome at least as high in preference ordering P as does any other. As such, the player k, by choosing the strategy t, needs not to think of other players' strategies. Strategy t serves the player k's interests best.

Definition 4. Straightforward game form. *A game form is* ***straightforward*** *if, for every preference ordering P and player k, there is a strategy which is P-dominant for k.*

As such, in a straightforward game form, each player has a strategy, which serves his/her interests best, and it is independent of what any other player chooses as strategy. So the players need not play strategically (i.e. be attentive to what others do). Therefore, a straightforward game form is also called **strategy-proof**.

Definition 5. Manipulable game form. *A* ***manipulable*** *game form is one which is not straightforward. In a manipulable game form, there exists a player k, who, given a preference ordering P, cannot find any strategy t, which is P-dominant for him/her. So given a game form g, a preference ordering P, and a strategy n-tuple $\boldsymbol{s} = (s_1, \ldots, s_n)$, the player k needs to choose a strategy t, where even though $s_k P t$, but by choosing this strategy, he/she would achieve $g(\boldsymbol{s}\langle k/t\rangle)\ R\ g(\boldsymbol{s})$.*

Definition 6. Dictatorial game form. *A player k is a dictator for game form g if, for every outcome x, there is a strategy $s(x)$ for k such that for strategy n-tuple $\boldsymbol{s} = (s_1, \ldots, s_k \ldots, s_n)$, the $g(\boldsymbol{s}) = x$ whenever $s_k = s(x)$. A game form g is dictatorial if there is a dictator for g.*

Having a dictator in a game form means that there exists a player k whose choices are always the outcome of the game, no-matter what other players have chosen (i.e. their strategies).

Theorem 1 (Gibbard [2]). *Every straightforward game form with at least three possible outcomes is dictatorial.*

Definition 7. Voting scheme. *A voting scheme is a game form v with set of possible outcomes X, such that for some set Z or set of* ***alternatives*** *with $X \subseteq Z$, the set S_i of strategies open to each player i is the set of orderings of Z. We call this set Π_Z. Then a voting scheme is a single valued function from Π_Z^n to X, which given a n-tuple $\boldsymbol{P} = (P_1, \ldots, P_n)$, returns a single possible outcome $x \in X$.*

In a voting scheme, an ordering P_i represents the ballet that voter i casts. The orderings can be fixed by an ordinal scheme, so the voter i, in his/her ordering P_i, places his/her most preferred alternative as number 1, the second most preferred alternative as number 2, and so forth. Then, the voting scheme simulates a counting mechanisms of the ballets and their orderings, which results in naming an alternative as the winner.

Definition 8. Manipulable voting scheme. *A voting scheme v is manipulable if for some voter k, and for some n-tuple $\boldsymbol{P} \in \Pi_Z^n$, there exists some ordering $P^* \in \Pi_Z$, such that $v(\boldsymbol{P}\langle k/P^*\rangle)\ P_k\ v(\boldsymbol{P})$.*

This means that for voter k, in the situation, which is represented by n-tuple $\boldsymbol{P} = (P_1, \ldots, P_k, \ldots, P_n)$, there exists another voting possibility (represented by ordering P^*), such that if the voter k changes P_k to P^*, then the result of the voting would represent his/her interest best, in the sense of her original vote P_k.

Definition 9. Dictatorial voting scheme. *Player k is a dictator for a voting scheme v if, for every possible outcome $x \in X$, player k can choose an ordering $P(x)$, so $v(\boldsymbol{P}) = x$ whenever $P_k = P(x)$, for any voting $\boldsymbol{P} = (P_1, \ldots, P_k \ldots, P_n)$. A voting scheme v is dictatorial if there is a dictator for it.*

Theorem 2 (Gibbard [2], Satterthwaite [7]). *Every voting scheme with at least three outcomes is either dictatorial or manipulable.*

Definition 10. Social welfare function. *A preference n-tuple over a set X is an n-tuple $(P_1, \ldots, P_n)$ whose terms are preference orderings of X, the set of possible outcomes. Preference n-tuples will be designated in bold-face type on the pattern $\mathbf{P} = (P_1, \ldots, P_n)$, $\boldsymbol{P'} = (P'_1, \ldots, P'_n)$ and so forth. A **social welfare function** is a function whose arguments, for some fixed n (or number of voters) and the set of possible outcomes X are all preference n-tuples $\boldsymbol{P}$ over X, and whose values are preference orderings of X.*

Given n voters and a set of possible outcomes X, a social welfare function $f : \Pi_X^n \longrightarrow \Pi_X$, is a mapping that takes a n-tuple $\boldsymbol{P} = (P_1, \ldots, P_n) \in \Pi_X^n$ and delivers a preference ordering $P \in \Pi_X$. Here Π_X represents the set of preference orderings on X.

Before we go further, we choose a notation that makes it easier to visualize the orderings. Let us use $\succ$ for orderings of X. For alternatives x and y, the $x \succ y$ signifies x is preferred to y. So if $\succ$ is an ordering of $x_1, x_2, \ldots, x_m \in X$ of m alternatives may look like this: $x_1 \succ x_2 \succ \ldots \succ x_m$

Theorem 3 (Arrow [1]). *Every social welfare function defined for a set of possible outcomes X, and n voters, violates one of following **Arrow conditions**:*

(i) ***Scope**: X has at least three members.*

(ii) ***Unanimity** or **weak Pareto efficiency**: If $\mathbf{P} = (\succ_1, \ldots, \succ_n)$ and $\forall i.\ x \succ_i y$, and $f(\boldsymbol{P}) = \succ$, then $x \succ y$.*

(iii) ***Pairwise Determination**: If for $\mathbf{P} = (\succ_1, \ldots, \succ_n)$ and $\mathbf{P'} = (\succ'_1, \ldots, \succ'_n)$:*

- $\forall i.\ [x \succ_i y \iff x \succ'_i y] \quad and \quad \forall i.\ [y \succ_i x \iff y \succ'_i x]$
- $f(\boldsymbol{P}) = \succ \quad and \quad f\left(\boldsymbol{P}'\right) = \succ'$

then: $x \succ y \iff x \succ' y$

(iv) ***Non-dictatorship****: There is no dictator for f, where a dictator for f is a voter k such that for every preference ordering $\succ$, and every x, $y \in X$, if $x \succ_k y$ and $f(\boldsymbol{P}) = \succ$, then $x \succ y$.*

The Arrow conditions are a group of logical and fair conditions which can be put on any voting system. The *Scope* and *non-Dictatorship* conditions are self-explanatory. The *Unanimity* or *weak Pareto efficiency* condition says that if for any two possible outcomes x and y, every voter has the same preference, e.g. x is preferred to y, then the result of voting and the final order of preference of the candidates (i.e. what social welfare function returns) must also reflect that x is preferred to y. At the same time, the *Pairwise Determination* condition says that the order of preference between two candidates is not dependant of how the order of preference between other pairs of candidates is.

In the case of general game forms and voting schemes, we looked at the result of the game and/or scheme to be a single element x of X, the set of possible outcomes. In the case of social welfare functions, the result was a preference ordering $\succ$ on $X = \{x_1, \ldots, x_m\}$, where, for instance $x_1 \succ \ldots \succ x_m$.

As such, the Gibbard-Satterwaithe result of Theorems 2 gives an impression that it is never possible to have a voting mechanism, which is non-manipulable, have more than 2 possible outcomes, and is non-dictatorial. Further investigation shows that the proof of Theorem 2 is, very much dependent on the assumption that the voting process selects a single winner [5]. But this is not the case in the real world. Many times, we will have situations, where several alternatives are chosen, al par with each-other, as acceptable outcomes (i.e. tied winners). Then the tie is broken and a single winner is chosen by some extra-ordinary process (e.g. in alphabetic order, or by random chance).

In order to work with situations with possibility of tied winners, Gärdenfors [5] developed $\gg$ as a new ordering between sets (e.g. $\{x, y\} \gg \{y\}$ if $x \succ y$), which would be a generalization of the ordering $\succ$ from above. The definition can be found in [5]. In the following we will use it to find a very strong result (Theorem 4, below) on democratic voting processes. This result is very important, since such democratic voting processes define the ideal processes towards the development of Trustworthy AI.

Definition 11. Social choice function. *Let X be the set of possible outcomes, $i = 1, \ldots n$ the set of voters, and Π_X the set of preference orderings on X. A* ***social choice function (SCF)*** *is a function $\boldsymbol{F} : \Pi_X^n \longrightarrow \boldsymbol{2}^X - \emptyset$, where $\boldsymbol{2}^X$ denotes the set of all subsets of X.*

Definition 12. *A social choice function* $\boldsymbol{F}$ *is* ***manipulable*** *by voter* i *at situation* $\boldsymbol{s} = (P_1, \ldots, P_n)$ *iff there is an ordering* P_i' *such that* $\boldsymbol{F}(\boldsymbol{s}\langle i/P_i'\rangle) \gg_i \boldsymbol{F}(\boldsymbol{s})$, *where* $\gg_i$ *is the ordering derived from* P_i . $\boldsymbol{F}$ *is* ***non-manipulable*** *or* ***stable*** *iff* $\boldsymbol{F}$ *is nowhere manipulable.*

Definition 13. *A social choice function* $\boldsymbol{F}$ *is* ***anonymous*** *iff whenever two situations* $\boldsymbol{s}_1$ *and* $\boldsymbol{s}_2$ *are identical except that for some voters* i *and* j, *we have* $P_{i|\boldsymbol{s}_1} = P_{j|\boldsymbol{s}_2}$ *and* $P_{j|\boldsymbol{s}_1} = P_{i|\boldsymbol{s}_2}$ *then* $\boldsymbol{F}(\boldsymbol{s}_1) = \boldsymbol{F}(\boldsymbol{s}_2)$. *So a social choice function is annonymous if it treats every voter in the same way. Here* $P_{i|\boldsymbol{s}}$ *means the preference* P_i *of voter* i *in the situation* $\boldsymbol{s}$.

Definition 14. *A social choice function* $\boldsymbol{F}$ *is* ***neutral*** *iff whenever two situations* $\boldsymbol{s}_1$ *and* $\boldsymbol{s}_2$ *are identical except for two alternatives* x *and* y *that have changed places everywhere (i.e. in the preferences of each voters), then* $x \in \boldsymbol{F}(\boldsymbol{s}_1)$ *iff* $y \in \boldsymbol{F}(\boldsymbol{s}_2)$ *and* $y \in \boldsymbol{F}(\boldsymbol{s}_1)$ *iff* $x \in \boldsymbol{F}(\boldsymbol{s}_2)$. *So a social choice function is neutral if it treats every alternative in the same way.*

Definition 15. *A social choice function* $\boldsymbol{F}$ *satisfies the* ***Condorcet criterion*** *iff whenever there is an alternative* x *in a situation* $\boldsymbol{s}$ *such that, for every alternative* $y \neq x$, *the number of individuals who strictly prefer* x *to* y *is greater than the number of individuals who strictly prefer* y *ot* x, *then* $\boldsymbol{F}(\boldsymbol{s}) = \{x\}$. *Such an alternative is called a* ***majority alternative*** *in the situation* $\boldsymbol{s}$.

Definition 16. *A social choice function* $\boldsymbol{F}$ *which is anonymous and neutral, and satisfies the Condorcet criterion, is a* ***democratic*** *social choice function.*

Theorem 4 (Gärdenfors [5]). *Any democratic social choice function which is defined for at least three voters is manipulable.*

3 Conclusions

We conclude that any development of AI HLEG proposed guidelines, framework and requirements for Trustworthy and Ethical AI suffers from systemic problems. We would be either in a situation where there are dictator decision-makers, or we will have design choices which are manipulated by those who are responsible for ensuring the trustworthiness of AI solutions. In both cases, the cognitive biases of the decision maker (either the dictator or manipulator) would have significant effect on the result of the designed and developed AI solution. This is the main challenge here.

As mentioned above, it is a fact that most of the voting processes can suffer of either having a dictator or voters who vote strategically to serve their own benefits. The issue is that what if the cognitive biases of the dictator decision-maker, or the ones of the tactical voter is exactly the biases, which we are trying to root-out from the design and architecture of AI solutions. As such, the very challenge of biases, which the EU guidelines are developed to answer, remains unanswered.

References

1. Arrow, K. J.: Social Choice and Individual Values. Wiley, New York (1963)
2. Gibbard, A.: Manipulation of voting schemes: A general result. Econometrica. **41**(4), 587–601 (1973)
3. Gibbard, A.: Manipulation of Schemes that Mix Voting with Chance. Econometrica, 45, pp. 665 – 681 (1977)
4. Gibbard, A.: Straightforwardness of Game Forms with Lotteries as Outcomes. Econometrica, 46, pp. 595–614 (1978)
5. Gärdenfors, P.: Manipulations of social choice functions. Journal of Economic Theory **13**, pp. 217–228 (1976)
6. Hylland A.: Strategy-proofness of voting procedures with lotteries as outcomes and infinite sets of strateges. Unpublished (1980)
7. Satterthwaite, M. A.: (April 1975) Strategy-proofness and Arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. Journal of Economic Theory. **10**(2), 187–217 (1975)
8. Draft Ethics guidelines for trustworthy AI, http://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai. Last accessed 4 Oct 2017

Fig. 1. The interrelation of the 7 requirements of Trustworthy AI (adopted from [8])

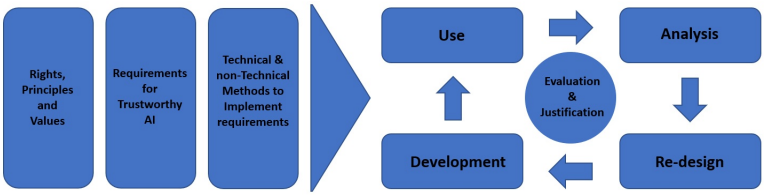

Fig. 2. Realizing Trustworthy AI throughout the system's entire life cycle (source [8])