What is a black box algorithm?

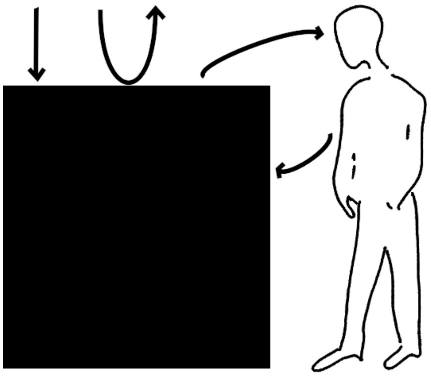
Erwan Le Merrer, Gilles Trédan

What is a black box algorithm?

Tractatus of algorithmic decision-making

v1.0 EN

Erwan Le Merrer, Inria
Gilles Trédan, CNRS



Companies or institutions (Divinities) disintermediate their relations with users via decision-making algorithms (Pythias): users (mortals) are offered arbitrary decisions (oracles) during their interactions. How to apprehend these black box algorithms, from the point of view of a user or regulator?

Algorithms

1. A data is the numerical representation of an object, or the product of computation on numerical variables.
2. An algorithm describes a set of operations to be executed, to solve a given problem.
3. An algorithm takes data (its inputs), and performs operations on them, to produce other data (its outputs).
4. An algorithm is implemented (*i.e.*, put into operation) via a description by a high level language, called source code, intelligible (*i.e.*, *compilable*, then executable by a machine).
5. The same algorithm can be implemented in several different but equivalent ways (*i.e.*, these variants produce the same outputs).
6. The disclosure of the source code implementing an algorithm amounts to exposing the entirety of the operations performed by it.
7. Algorithms can make use of variables. These variables are set at the beginning of its execution, or during its execution.
8. Algorithms can use randomness; they are then said to be non-deterministic or randomized. They are deterministic in the opposite case.

(a) A deterministic algorithm produces the same outputs for the same inputs, all variables being equal.

9. The so-called non-explicit algorithms perform complex operations on data, that are not describable by a high-level language.
10. An algorithm or an implementation is said to be biased if it produces an output which diverges from the expectation. It is said to be *buggy* if this bias is unintentional.
11. The interaction between algorithms via their inputs/outputs composes a complex system.

(a) The dynamics of this interaction and its output are said to be emergent.

Algorithmic decisions

12. The input (resp. output) space is the set of combinations of values that the data can take as input (resp. output).

(a) The size of the input (resp. output) space is the number of such possible combinations.

13. Executing a deterministic algorithm on the whole input space (with fixed variables and randomness), allows to fully characterize this algorithm (correspondence table).

(a) However, this method does not indicate in any way how to recover the source code of (*i.e.*, retro-engineer) this algorithm.

14. Releasing the source code of an algorithm allows to define its input and output spaces.

(a) An algorithm whose source code is open is not sufficient to describe a precise computation, in the absence of its input data, variables and sources of randomness.

15. A so-called decision-making algorithm is like any other processing algorithm: its output takes a value among a predefined set; but this value is then called a decision and given to the user.

(a) A decision-making algorithm can then try to approach as closely as possible (*i.e.*, to encode) a rational human judgment leading to a decision.
(b) A decision-making algorithm can very well be purely random and take no input data; it is then a random generator of decisions (binary for example).

16. A decision is said to be biased if it is produced causally by data which should not lead to it.

Explanation of algorithmic decisions

17. The possibility of freely manipulating an algorithm source code, and thus of knowing each of its operations, and of interacting via its inputs and observing its outputs, allows it to be qualified as a white box.
18. The implementation of an algorithm can only be "transparent" by making it available as a white box. Indeed, the source code alone does not explain the decisions but constrains them.
19. The explanation of a decision is a projection on the language of the high level instructions of an algorithm having led to this particular decision.
20. A decision is said to be non-explicit, if the operations leading from the input data to this decision cannot be described by a high-level language.
21. An explanation is underdetermined if it describes only part of the data and computations that led to the decision.
22. Varying the inputs in order to vary the decisions makes it possible to quantify the relative impact of these inputs.
23. Decisions are identical within boundaries, each of which leads to another decision.

(a) A point on a decision boundary is defined as a specific location where a minute change in the input causes the decision to flip.
(b) The decision boundaries of an algorithm are made up of all the points previously defined.

Black box algorithms

24. A black box algorithm is said to be observable if any change of its variables is reflected in its outputs.

25. A decision-making algorithm is said to be opaque to an individual if the latter cannot access to at least one of the following components: source code, inputs or variables.

26. The execution of an algorithm on a machine controlled by an observer is in no way a guarantee of its white box nature for this same observer (closed source code, enclaves, encryption).

(a) This is all the more obvious for an execution on a third party machine (cf "*There is no cloud, just someone else's computer*").

27. A black-box algorithm is an algorithm for which a user can only propose data as input and observe the outputs that result *a priori* causally from these inputs.

(a) *a priori*, because an algorithm can quite possibly respond with a fixed or random value regardless of its inputs.

(b) In other words, the veil that turns an algorithm from white box to black box is the execution by a third party, or the multiple techniques of code obfuscation. In other words, even a known algorithm (white box) when executed by a third party becomes de facto a black box.

28. An explanation by a black-box algorithm is falsifiable, because there is always an explanation validating the output, which is not the explanation of the operations actually performed on the inputs (the bouncer problem).

29. Let us call the stability assumption the fact that a black-box algorithm and its variables remain fixed during a period of interaction.

30. Under the stability assumption, two black-box algorithms are indistinguishable if all their outputs are equal for all their respective inputs.

(a) Two black box objects, which are indistinguishable, can be the same algorithm (their source code may indeed diverge), or two different algorithms (*e.g.*, a bubble sort algorithm and a fusion sort algorithm); their outputs are expected to be identical.

(b) Two algorithms are distinguishable if there exist two different outputs for at least one given input.

(c) The distinguishability of two algorithms may then depend only on one particular input.

31. Just as in the quantum domain, a simple interaction with a black box algorithm in order to observe its output may well modify its variables. For instance by the re-training of this algorithm, incorporating the input used for the interaction.

(a) Any hope of reiterating an observation is thus potentially vain. In the worst case, the observation is "one shot".
(b) The more favorable case of stable black boxes because unchanged by the observation is also possible.
(c) The act of voluntarily providing particular data as input, so as to modify in a certain direction the internal state (or variables) of the black-box algorithm –and consequently, its outputs–, is called data poisoning.

32. The size of the input space is directly related to the difficulty of observing properties of a black-box algorithm. The more this size increases, the more the curse of dimensionality is prevalent, and the more complex it is to observe these properties.

33. Learning the result of several pairs of inputs/outputs can allow an extra/interpolation in the prediction of other outputs of the black box.

34. An algorithm can very well specialize its outputs (*e.g.*, create a bias voluntarily) by taking into account solely one fraction of the input ("sandboxing").

35. Neural networks, the pillars of modern advances in machine learning, are black boxes because they produce non-explicit decisions.

(a) A learning process proceeds through billions of trial-and-errors, and concludes with the positioning of the algorithm's variables at the values that have minimized its final error on known cases.
(b) These variables being in gigantic quantity, and in the absence of algorithmic steps, there is thus no possible direct explanation of the resulting decisions.
(c) Thus is achieved a trial and error pinpoint of the numerical elements that led to a given decision, on a given input (interpretation).

The algorithm (in white box) on the third party side

36. We call third party the entity that runs an algorithm that faces users (which then appears as a black box to them).

(a) The relative novelty and generalization of the white-box/black-box dichotomy stems from the placement in production of user-facing decision-making algorithms. The previous use of algorithms was rather to advise an institution, which in turn was the intermediary between the algorithm and the users. Disintermediation thus motivates the understanding of what is feasible or intelligible in this new type of interaction.
(b) The placement of source code by a third party on a public server for inspection cannot imply any conclusion of transparency. Indeed, and trivially, nothing proves that it is indeed the algorithm that is actually running in the service in question.

Audit of algorithms

37. A class of algorithmic audits seeks to establish the validity or not in a black box algorithm of predicates defined beforehand; *e.g.*, for this precise input, this particular output must not be returned.

38. Under the hypothesis of stability, another class seeks to establish the trends at work in the input-output relationship (substitute model).

(a) The greater the quantity of input-output relations observed, the more it is possible to construct a precise representation (surrogate) of the algorithm (*e.g.*, to approximate it finely). Thus, for a budget of inputs tending towards the size of the input space, the algorithm can be approximated with an arbitrarily small error.

39. Shannon's maxim (or Kerckhoffs's), which specifies that one must make the assumption that "the adversary (*i.e.*, the observer) knows the system (*i.e.*, the algorithm here)", seems to be little used in practice in the domain: third parties rely largely on security through obscurity.

Shortcuts

40. What is a black box algorithm for which I can predict all outputs? An algorithm for which I have a perfect surrogate, in the sense of its indistinguishability on all possible inputs (**??**) from this surrogate.

41. A black-box algorithm falls under the definition of proprietary software, because it does not exhibit a single of the four points defining free software.

(a) Either the impossibility to execute, to distribute, to study, or to modify this software.

Implications and societal distortions

42. An algorithm is also said to be biased if it produces decisions considered as "unfair" by a group of individuals.

(a) The notion of fairness for a decision-making algorithm would like to measure the balance of decisions for groups of individuals represented by specific inputs.

(b) The very principle of some classification algorithms is to discriminate individuals from their inputs.

(c) A classification algorithm will be considered fair or unfair according to the intersection of its boundaries with a set of socially constructed boundaries.

43. The doubt that the possibility of sandboxing casts on any audit attempt is glaringly illustrated.

(a) "Diesel gate": a qualified auditor was confronted with an alternative reality (discrimination on a "auditor" flag in the input), *i.e.*, a specialization of the outputs in view of obtaining a certification, while the rest of the users experienced a completely different reality.
(b) Shadow banning practices offer the illusion to a user that he is being treated normally, while the third party has decided to lower his visibility (on a social network for example).
(c) The same applies to proposals from third parties to open up interfaces (APIs) to audit their algorithms. There is no guarantee in practice that the outputs are those that a standard user may observe.

44. A biased algorithm is not wrong as such. It has been placed in front of the users on the decision of the third party, who ordered it. Thus the fault falls transitively and naturally on the latter.

45. The practices of whistleblowers can find a societal legitimacy simply in the exhibition of an algorithm outputs judged improper (*e.g.*, Tay bot). Even if some had voluntarily and abusively manufactured the inputs that would lead to this improper behavior.

(a) The problem here is not in the input-output correspondence, but rather in the absence of regulation/censorship on certain outputs by the algorithm (*accountability*, *i.e.*, third party responsibility).
(b) In other words, some outputs should never be proposed, by construction or constraint.

46. Ranking or recommendation systems are decision-making algorithms that filter and rank objects from a large catalog for a user (personalization).

47. A filter bubble is the result of the over-personalization of a recommendation algorithm with regard to its outputs for a particular user.

(a) A corollary is the collapse of the diversity of proposals made by the algorithm to this user.
(b) A particular case (*rabbit-hole*) is the result of the increasing interaction of a user with the personalization made to him by an algorithm, thus creating a phenomenon of amplification of the algorithmic reaction.

48. The so-called predictive algorithms group together the techniques whose aim is to maximize a metric (such as the appreciation of the users of what is proposed to them) on the future behavior of the users in front of their proposals (*i.e.*, outputs).

(a) An algorithm does not know what a user wants. To produce this illusion, it can exploit what other users –whom it considers similar– have chosen in the past. In other words: so-called predictive algorithms do not predict the future, they encode the past to replay it.

(b) Knowing the many biases that can afflict decision-making algorithms, a common flaw is to take the algorithmic decision as a reliable prediction and to impose it on everyone by justifying it as scientifically indisputable.

49. Proponents of privacy online suggest that it is possible to interact with black boxes via one part of a user's personal data, but that the other part will remain secret, or even non-inferrable.

(a) Due to the masses of data available for training algorithms and the accuracy of their inferences, the possibility of providing inputs that do not allow for correct inference of the rest of the "hidden" data becomes negligible.