**Supplementary Note A**

A Turing machine is a hypothetical machine developed by Alan Turing in 1936. Turing machine is designed to simulate any computer algorithm, no matter how complicated it is (Ashrafian et al. 2015). Consider a universal Turing Machine . For any binary string , we define Kolmogorov complexity as the length of the shortest program that generates , denoted as , using universal prefix Turing machine that outputs and then stops (Peters et al. 2017; Kolmogorov 1965). Therefore, we have , where denotes the number of bits of a binary string. Intuitively, the Kolmogorov complexity measures the minimal amount of information required to generate by any effective process. Similar to conditional probability, we can also define conditional Kolmogorov complexity. The conditional Kolmogorov complexity of string given , is defined as the length of the shortest program that can generate from and then stops. The Kolmogorov complexity of the concatenation of two strings and is defined as the length of the shortest program that generate string , where is the prefix code of .

Now we introduce “additivity of complexity” property. It can be shown that (Grunwald and Vitanyi 2004):

, (A1)

where denotes the first shortest prefix program that generates and then stops and is in general uncomputable.

Algorithmic mutual information is defined as

. (A2)

Substituting in equation (A1) into equation (A2), we obtain

, (A3)

where the symbol ⩲ implies that the equation can hold for up to constants. Equation (A3) states that this information is symmetrical: . Therefore, is called algorithmic mutual information between and . The algorithmic mutual information quantifies the amount of information two strings or objects have in common, or the amount of bits saved when compressing jointly rather than compressing independently.

Similar to mutual information between two random variables where mutual information of zero implies independence of two variables, the algorithmic mutual information of zero indicates algorithmically independence of two distributions of random variables. We also can define algorithmic conditional mutual information as

. (A4)

In statistics, although dependence between two random variables can be measured, there are no measures to quantify dependence between two distributions. We use algorithmic mutual information to measure independence between two distributions which can be used to assess causal relationships between two variables. Consider two variables and and assume causes (). Let the marginal distribution of cause and conditional distribution of effect given be and , respectively. The independence of cause and mechanism (ICM) states that the distributions and are independent and hence and are algorithmically independent, which implies that their algorithmic mutual information should be equal to zero (Peters et al. 2017):

, (A5)

or , equivalently,

). (A6)

In other words, distributions and have no common information. If causes , then the conditional distribution of the effect given cause contains no information about cause . Thus, the algorithmic mutual information can be used to infer whether or . If then . Similarly, if then . Cause and effect cannot be identified from their joint distribution. Cause and effect are asymmetric. The joint distribution is symmetric. It can be factorized to .

**Supplementary Note B**

We show that for every joint distribution of real-valued variables and , there is a nonlinear model:

,

where are functions and is a real-valued noise variables.

Proof.

Define the conditional cumulative distribution function:

(B1)

and let

. (B2)

Define its inverse function

inf . (B3)

Define function

. (B4)

Now we make changes of variables:

(B5)

and

. (B6)

The Jacobian matrix of the transformation is given by

, (B7)

where ).

Using equation (B2), we obtain that is uniformly distributed on [0, 1]. If we assume that is independent of , then using the distribution transform theorem, we obtain

. (B8)

**Supplementary Note C**

To illustrate application of the algorithmic mutual information, we show that independence of cause and mechanism will imply that the cause and error in the nonlinear function model (1) are independent.

Independence of cause and mechanism states that the conditional distribution contains no information about the distribution of causal . In other words, and are algorithmically independent:

. (C1)

Assume a nonlinear function model:, but we do not assume that and are independent. We now show that independence of cause and mechanism implies that and are independent in the ANM (1).

The principle independence of cause and mechanism implies that . Therefore, from equation (C1) we obtain

, (C2)

which implies

. (C3)

Mutual information of zero between the cause and residual variable shows that and are independent. It is also well known that if and are independent, then or (Janzing and Schölkopf 2010). Therefore, algorithmic independence between the distribution of cause and conditional distribution of effect given the cause is equivalent to the independence of two random variables and in the ANM.