True AI Should Be a Loser, Not a Winner

First International Conference
"Responsible Creation of Artificial General Intelligence"

May 20, 2026

Dimiter Dobrev

d@dobrev.com

Institute of Mathematics and Informatics
Bulgarian Academy of Sciences
and I give lectures in
South-West University "Neofit Rilski"

The idea of this talk is that
you can be smart without showing it.

When we study the intellect, we assume that it is a black box. That is, we are only interested in its input-output.

This is the basis of the modern definition of AI. The same definition has been independently derived by different people. The most important contributors being Pei Wang, Marcus Hutter, and Jose Hernández-Orallo.

They all make the same definition because they are university professors and they evaluate artificial intelligence in the same way they evaluate natural intelligence.

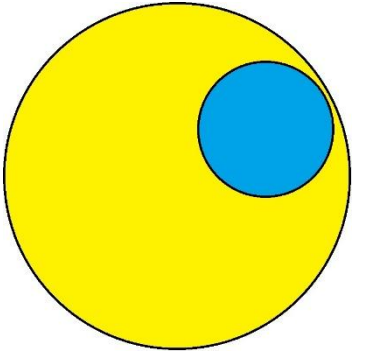
At university, we evaluate students by giving them an exam. The modern definition of AI is based on the same idea. We give the programs an exam, and the one that performs well is announced to the AI.

We assume that the student wants to take the exam and that he cooperates with the teacher. If the student decides to fail, then he will not take the exam, but that does not mean that he is not smart.

We will call successful students who take their exams winners. We will call those who fail to take them losers. They may not take the exam because they are not smart enough, or they may not be ambitious enough, motivated enough, or they simply do not want to take it.

To be successful (a winner) you have to be smart, but the opposite is not true.

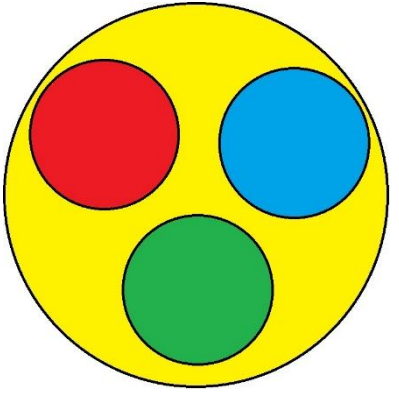
The picture looks like this:



Here, the blue circle is the successful programs, and the yellow circle is all intelligent programs.

This is the inaccuracy in to the definition we have nowadays. It only defines the programs that are in the blue circle, but these are not all intelligent programs.

Other intelligent programs

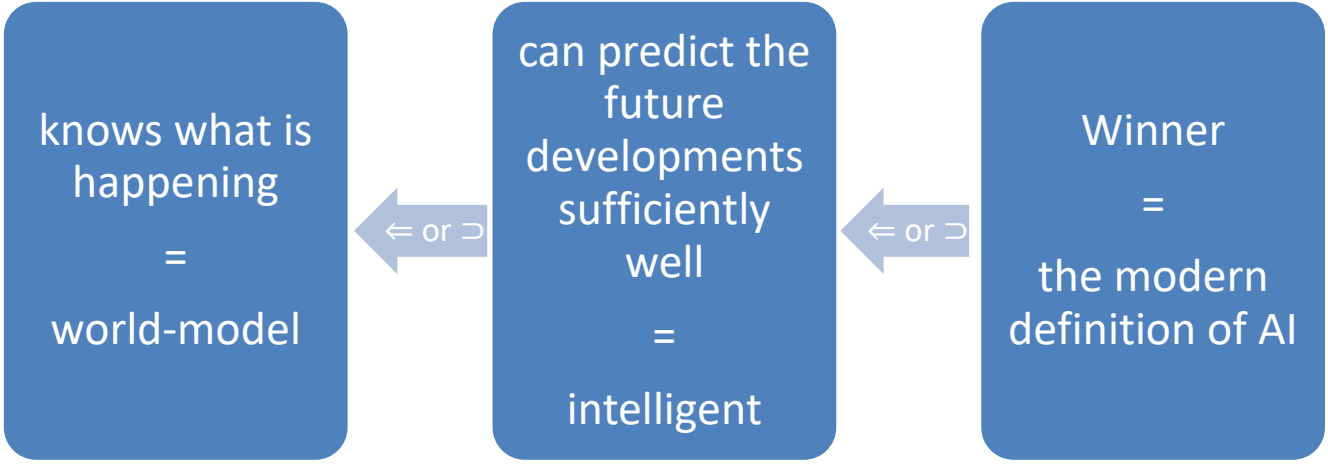


If the program strives to fail, we will call it a masochist (the red circle). The winner avoids pain, while the masochist, on the contrary, seeks it.

Yet, there is a different and more interesting type of AI. Its logo would be: "I understand everything, but I do not care and play randomly" (the green circle). If we look at this AI as a black box, it is indistinguishable from a random action generator.

Extended definition

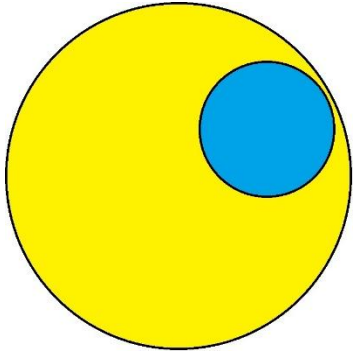
A program which knows what is happening and can predict the future developments sufficiently well.



Definition by example.

The modern definition of AI successfully answers the question "What is AI?". It describes some of the intelligent programs. There are other intelligent programs, but from a theoretical point of view this is not so important.

However, from a practical perspective this appears to be of paramount importance. The reason is that we are at the doorstep of creating True AI and among all intelligent programs we must choose the one we will be most comfortable with from now on.



It's not a good idea to choose a program from the blue circle. Programs that blindly seek to optimize some profit are unforgiving, limited, and unpleasant.

Why do we only get one try?

The creator of the first version of AI will be us humans. The creator of the second version will not be us humans, but the first version of AI.

After creating the first version, we will start a process that will develop on its own without our intervention.

The first version of AI will largely determine what the second version will be.

Why our conference is extremely important?

We need to convince the world that creating AGI is an extremely responsible task that must be approached with great care.