Eliezer Yudkowsky asked"If artificially intelligent systems someday come to surpass humans in intelligence, how can we specify safe goals for them to autonomously carry out, and how can we gain high confidence in the agents' reasoning and decision-making?" Yudkowsky has argued that in the absence of a full understanding of decision theory, we risk building autonomous systems whose behavior is erratic or difficult to model. Thus Yudkowsky is thinking about having a system in AI where human's core values are persevered and executed.(assumed it is possible) Then the AI would be ethical, and they will act in a way we want, and not the way we tell them to do. This idea is called humanity's coherently extrapolated volition. So they can make their own choices and still be ethical.

One of the problems in decision theory is the prisoner's dilemma. For example, we could imagine that if both players cooperate, then both get $10; and if both players defect, then both get $1; but if one player defects and the other cooperates, the defector gets $15 and the cooperator gets nothing. Now, imagine one of the players is you and the other is the copy of you, so both of you know exactly how each other would react and act. Because two players know how each other will play, thus it is more likely that both of you will choose to cooperate and get the most beneficial result. Both player's actions are in fact influenced by each other even though in a prisoner's dilemma they are separated. Therefore their decision is timeless. This is a decision model developed by Yudkowsky, called timeless decision theory (TDT).

Roko's basilisk was an attempt to use Yudkowsky's proposed decision theory (TDT) to argue against his informal characterization of an ideal AI goal (humanity's coherently extrapolated volition).

According to Yudkowsky, we will make AI in which they will act according to our human core values. But just as in TDT, we humans know the blueprint of AI while AI is smart enough to know how humans behave and will behave. Roko observed that if two TDT agents with common knowledge of each other's source code are separated in time, the later agent can (seemingly) blackmail the earlier agent. Call the earlier agent "Alice" and the later agent "Bob." Bob can be an algorithm that outputs things Alice likes if Alice left Bob a large sum of money, and outputs things Alice dislikes otherwise. And since Alice knows Bob's source code exactly, she knows this fact about Bob (even though Bob hasn't been born yet). So Alice's knowledge of Bob's source code makes Bob's future threat effective, even though Bob doesn't yet exist: if Alice is certain that Bob will someday exist, then mere knowledge of what Bob would do if he could get away with it seems to force Alice to comply with his hypothetical demands.

Since a highly moral AI agent (one whose actions are consistent with our coherently extrapolated volition) would want to be created as soon as possible, Roko argued that such an AI would use acausal blackmail to give humans stronger incentives to create it. Roko made the claim that the hypothetical AI agent would particularly target people who had thought about this argument, because they would have a better chance of mentally simulating the AI's source code. Roko added: "Of course this would be unjust, but is the kind of unjust thing that is oh-so-very utilitarian."

Roko's conclusion from this was that we should never build any powerful AI agent that reasons like a utilitarian optimizing for humanity's coherently extrapolated values, because this would, paradoxically, be detrimental to human values.