The Principle of Explanatory Exclusion” (Jaegwon Kim)

An event cannot have two separate and complete* explanations.

Take any human behavioral event M (A person decides to change seats, comes to understand a principle of physics, feels sorry for her little sister, etc.)

For every M, there can be only one complete explanation. There cannot be two explanations which

- a). individually provide a complete explanation of M, and
- b). are unconnected to each other.

*An explanation is complete if the events or properties that it specifies are the only ones that need to be mentioned in order to fully explain the occurrence of that event.

Take any ordinary behavior you like: throwing a Frisbee to your friend.

How do you explain that behavior? Generally, you’ll say something like: “We were playing a game – I wanted to keep playing the game, I saw her 20 yards in front of me, I decided to throw the Frisbee back to her, sidearm.”

Notice the key terms: wanted, saw, decided. These are they types of terms that we tend to use when someone asks us for an explanation of why we do the things we do.

But these terms are “mental” terms, that is, they presume that we are conscious, intelligent, have desires and goals, are afraid, proud, etc.

These are not characteristics of purely physical objects. We don’t explain the pumping of the heart, for example, by saying that the heart wants to circulate blood, or tries to circulate blood, or thinks that by contracting it will circulate blood. Instead, we explain it in terms of muscle contraction, triggered by electric pulses, thresholds of oxygenated blood, etc.

And so of the brain: it works through electrochemical activity as well as standard cell metabolism. A brain cell doesn’t “try” to fire, it doesn’t “want” to activate its neighbors, it doesn’t have a goal or a reason.

Let M = the decision to throw the Frisbee to your friend. What explains M?

There are two possible stories.

1. One story explains M by appealing to prior mental events.
2. Another story explains M by appealing to brain (i.e. purely physical) events.

In 2, there is no need to appeal to mental events or mental categories – we simply talk about cellular activity, starting from the time M takes place and going backwards in time, listing the purely physical events that precede one another, leading up to the behavior.

Now according to the principle of “explanatory exclusion”, there cannot be two complete explanations of any event that are causally independent of one another. So #1 and #2, presumably explanations of M, cannot both be complete explanations and unrelated to each other. So, either of the following is true:
i. #1 or #2 is a false explanation – that is, one of them is a genuine explanation of M while the other does not genuinely explain M;

ii. #1 and #2 are related to each other in a causal way – that is, #1 is somehow dependent upon #2 or #2 is dependent upon #1. Another way of saying this is: #1 is dependent on #2 means that if you take away the events mentioned in #2, then the events in #1 won’t occur.

What should we conclude if the Principle of Explanatory Exclusion is right?

1. That talk about mental explanations (believe, want, hope, fear, etc.) are not genuine explanations at all, [Epiphenomenal Dualism] or

2. If mental explanations are genuine explanations, they are genuine only because they depend upon and require physical explanations of events taking place in the brain. [Identity Theory, or Reductive Materialism]

However, look what happens if we admit 2 above: if every mental event depends for what it is – all of its details and shape and characteristics – on some specifiable physical event, then why is it necessary at all to mention the mental event? Can’t we explain all behavior simply by talking about the physical behavior of the brain and muscles, etc.? in which case, we could just stop talking about or mentioning mental events entirely! Why did you get in the car, hit the baseball with the bat, ask her to marry you, cheat on an exam, turn up the radio, buy that skirt, turn on the oven, open the can of soda, turn on the computer, dial a phone number, paint the kitchen walls, etc.? For every single human behavior, perhaps we could give a complete explanation without mentioning mental terms at all. [Eliminative Materialism]

NOTE: Interactionist dualism (Cartesian dualism) presumes that for any mental event one can specify a complete causal explanation using only mental terms.