Causal vs. evidential decision theory

Zola Donovan [1]

Lane Department of Computer Science and Electrical Engineering
West Virginia University, Morgantown, WV

October 25, 2016

Outline

1. Newcomb's problem

Outline

1. Newcomb's problem
2. Causal decision theory

Outline

1. Newcomb's problem
2. Causal decision theory
3. Evidential decision theory

Newcomb's problem (predictor's paradox)

Newcomb's problem (predictor's paradox)

A decision problem proposed by physicist William Newcomb in the 1960s

Newcomb's problem (predictor's paradox)

A decision problem proposed by physicist William Newcomb in the 1960s

- Imagine a being who is able to predict with 99% accuracy.

Newcomb's problem (predictor's paradox)

A decision problem proposed by physicist William Newcomb in the 1960s

- Imagine a being who is able to predict with 99% accuracy.
- You are offered a choice between two boxes, B_1 and B_2.

Newcomb's problem (predictor's paradox)

A decision problem proposed by physicist William Newcomb in the 1960s

- Imagine a being who is able to predict with 99% accuracy.
- You are offered a choice between two boxes, B_1 and B_2.
  - Box B_1 is transparent, and you can see that it contains \$1,000.

Newcomb's problem (predictor's paradox)

A decision problem proposed by physicist William Newcomb in the 1960s

- Imagine a being who is able to predict with 99% accuracy.
- You are offered a choice between two boxes, B_1 and B_2.
  - Box B_1 is transparent, and you can see that it contains \$1,000.
  - Box B_2 contains either \$0 or \$1M.

Newcomb's problem (predictor's paradox)

A decision problem proposed by physicist William Newcomb in the 1960s

- Imagine a being who is able to predict with 99% accuracy.
- You are offered a choice between two boxes, B_1 and B_2.
  - Box B_1 is transparent, and you can see that it contains \$1,000.
  - Box B_2 contains either \$0 or \$1M.
- You are invited to make a choice between the following pair of alternatives:

Newcomb's problem (predictor's paradox)

A decision problem proposed by physicist William Newcomb in the 1960s

- Imagine a being who is able to predict with 99% accuracy.
- You are offered a choice between two boxes, B_1 and B_2.
  - Box B_1 is transparent, and you can see that it contains \$1,000.
  - Box B_2 contains either \$0 or \$1M.
- You are invited to make a choice between the following pair of alternatives:

Alternative 1

Newcomb's problem (predictor's paradox)

A decision problem proposed by physicist William Newcomb in the 1960s

- Imagine a being who is able to predict with 99% accuracy.
- You are offered a choice between two boxes, B_1 and B_2.
  - Box B_1 is transparent, and you can see that it contains \$1,000.
  - Box B_2 contains either \$0 or \$1M.
- You are invited to make a choice between the following pair of alternatives:

Alternative 1

- Take box B_1 (\$1,000) and box B_2 (either \$0 or \$1M).

Newcomb's problem (predictor's paradox)

A decision problem proposed by physicist William Newcomb in the 1960s

- Imagine a being who is able to predict with 99% accuracy.
- You are offered a choice between two boxes, B_1 and B_2.
  - Box B_1 is transparent, and you can see that it contains \$1,000.
  - Box B_2 contains either \$0 or \$1M.
- You are invited to make a choice between the following pair of alternatives:

Alternative 1

- Take box B_1 (\$1,000) and box B_2 (either \$0 or \$1M).

Alternative 2

Newcomb's problem (predictor's paradox)

A decision problem proposed by physicist William Newcomb in the 1960s

- Imagine a being who is able to predict with 99% accuracy.
- You are offered a choice between two boxes, B_1 and B_2.
  - Box B_1 is transparent, and you can see that it contains \$1,000.
  - Box B_2 contains either \$0 or \$1M.
- You are invited to make a choice between the following pair of alternatives:

Alternative 1

- Take box B_1 (\$1,000) and box B_2 (either \$0 or \$1M).

Alternative 2

- Take only box B_2 (either \$0 or \$1M).

Newcomb's problem (predictor's paradox)

A decision problem proposed by physicist William Newcomb in the 1960s

- Imagine a being who is able to predict with 99% accuracy.
- You are offered a choice between two boxes, B_1 and B_2.
  - Box B_1 is transparent, and you can see that it contains \$1,000.
  - Box B_2 contains either \$0 or \$1M.
- You are invited to make a choice between the following pair of alternatives:

Alternative 1

- Take box B_1 (\$1,000) and box B_2 (either \$0 or \$1M).

Alternative 2

- Take only box B_2 (either \$0 or \$1M).

The catch

Newcomb's problem (predictor's paradox)

A decision problem proposed by physicist William Newcomb in the 1960s

- Imagine a being who is able to predict with 99% accuracy.
- You are offered a choice between two boxes, B_1 and B_2.
  - Box B_1 is transparent, and you can see that it contains \$1,000.
  - Box B_2 contains either \$0 or \$1M.
- You are invited to make a choice between the following pair of alternatives:

Alternative 1

- Take box B_1 (\$1,000) and box B_2 (either \$0 or \$1M).

Alternative 2

- Take only box B_2 (either \$0 or \$1M).

The catch

- You are told that the predictor will put \$1M in box B_2 *if and only if* she predicts that you will take just box B_2, and nothing in it otherwise.

Newcomb's problem (predictor's paradox)

Newcomb's problem (predictor's paradox)

Nozick's first way of reasoning

Newcomb's problem (predictor's paradox)

Nozick's first way of reasoning

- It is rational to take both boxes, because then you get the $1,000 in the first box, and whatever amount of money there is in the second.

Newcomb's problem (predictor's paradox)

Nozick's first way of reasoning

- It is rational to take both boxes, because then you get the $1,000 in the first box, and whatever amount of money there is in the second.
  - At the time of your choice, the $1M is either in the second box or not, so the fact that the predictor has made a prediction does not make any difference.

Newcomb's problem (predictor's paradox)

Nozick's first way of reasoning

- It is rational to take both boxes, because then you get the $1,000 in the first box, and whatever amount of money there is in the second.
  - At the time of your choice, the $1M is either in the second box or not, so the fact that the predictor has made a prediction does not make any difference.
- This line of reasoning can be seen as a straightforward application of the dominance principle: Taking two boxes dominates taking just one.

Newcomb's problem (predictor's paradox)

Nozick's first way of reasoning

- It is rational to take both boxes, because then you get the $1,000 in the first box, and whatever amount of money there is in the second.
  - At the time of your choice, the $1M is either in the second box or not, so the fact that the predictor has made a prediction does not make any difference.
- This line of reasoning can be seen as a straightforward application of the dominance principle: Taking two boxes dominates taking just one.

Nozick's second way of reasoning

Newcomb's problem (predictor's paradox)

Nozick's first way of reasoning

- It is rational to take both boxes, because then you get the $1,000 in the first box, and whatever amount of money there is in the second.
  - At the time of your choice, the $1M is either in the second box or not, so the fact that the predictor has made a prediction does not make any difference.
- This line of reasoning can be seen as a straightforward application of the dominance principle: Taking two boxes dominates taking just one.

Nozick's second way of reasoning

- It is rational to also consider the fact that the predictor has predicted your choice, and adjusted the amounts of money in the second box accordingly.

Newcomb's problem (predictor's paradox)

Nozick's first way of reasoning

- It is rational to take both boxes, because then you get the $1,000 in the first box, and whatever amount of money there is in the second.
  - At the time of your choice, the $1M is either in the second box or not, so the fact that the predictor has made a prediction does not make any difference.
- This line of reasoning can be seen as a straightforward application of the dominance principle: Taking two boxes dominates taking just one.

Nozick's second way of reasoning

- It is rational to also consider the fact that the predictor has predicted your choice, and adjusted the amounts of money in the second box accordingly.
  - If you take both boxes she has almost certainly predicted this, and hence put $0 in the second box.

Newcomb's problem (predictor's paradox)

Nozick's first way of reasoning

- It is rational to take both boxes, because then you get the $1,000 in the first box, and whatever amount of money there is in the second.
  - At the time of your choice, the $1M is either in the second box or not, so the fact that the predictor has made a prediction does not make any difference.
- This line of reasoning can be seen as a straightforward application of the dominance principle: Taking two boxes dominates taking just one.

Nozick's second way of reasoning

- It is rational to also consider the fact that the predictor has predicted your choice, and adjusted the amounts of money in the second box accordingly.
  - If you take both boxes she has almost certainly predicted this, and hence put $0 in the second box.
  - If you take only the second box, the predictor has almost certainly predicted that decision correctly, and consequently put $1M in the second box.

Newcomb's problem (predictor's paradox)

Nozick's first way of reasoning

- It is rational to take both boxes, because then you get the $1,000 in the first box, and whatever amount of money there is in the second.
  - At the time of your choice, the $1M is either in the second box or not, so the fact that the predictor has made a prediction does not make any difference.
- This line of reasoning can be seen as a straightforward application of the dominance principle: Taking two boxes dominates taking just one.

Nozick's second way of reasoning

- It is rational to also consider the fact that the predictor has predicted your choice, and adjusted the amounts of money in the second box accordingly.
  - If you take both boxes she has almost certainly predicted this, and hence put $0 in the second box.
  - If you take only the second box, the predictor has almost certainly predicted that decision correctly, and consequently put $1M in the second box.
- This line of reasoning can be seen as a straightforward application of the principle of maximizing expected utility.

Newcomb's problem (predictor's paradox)

Newcomb's problem (predictor's paradox)

The decision matrix

Newcomb's problem (predictor's paradox)

The decision matrix

	Second box contains $1M	Second box is empty
Take second box only	$1M (prob. 0.99)	$0 (prob. 0.01)
Take both boxes	$1M + $1,000 (prob. 0.01)	$1,000 (prob. 0.99)

Newcomb's problem (predictor's paradox)

The decision matrix

	Second box contains $1M	Second box is empty
Take second box only	$1M (prob. 0.99)	$0 (prob. 0.01)
Take both boxes	$1M + $1,000 (prob. 0.01)	$1,000 (prob. 0.99)

Expected utilities (assume utility of money is linear)

Newcomb's problem (predictor's paradox)

The decision matrix

	Second box contains $1M	Second box is empty
Take second box only	$1M (prob. 0.99)	$0 (prob. 0.01)
Take both boxes	$1M + $1,000 (prob. 0.01)	$1,000 (prob. 0.99)

Expected utilities (assume utility of money is linear)

- The expected utility of taking only the second box:

Newcomb's problem (predictor's paradox)

The decision matrix

	Second box contains $1M	Second box is empty
Take second box only	$1M (prob. 0.99)	$0 (prob. 0.01)
Take both boxes	$1M + $1,000 (prob. 0.01)	$1,000 (prob. 0.99)

Expected utilities (assume utility of money is linear)

- The expected utility of taking only the second box:
  $0.99 \cdot u(\$1M) + 0.01 \cdot u(\$0) = 0.99 \cdot 1,000,000 + 0.01 \cdot 0 = 990,000$

Newcomb's problem (predictor's paradox)

The decision matrix

	Second box contains $1M	Second box is empty
Take second box only	$1M (prob. 0.99)	$0 (prob. 0.01)
Take both boxes	$1M + $1,000 (prob. 0.01)	$1,000 (prob. 0.99)

Expected utilities (assume utility of money is linear)

- The expected utility of taking only the second box:
  $0.99 \cdot u(\$1M) + 0.01 \cdot u(\$0) = 0.99 \cdot 1,000,000 + 0.01 \cdot 0 = 990,000$
- The expected utility of taking both boxes:

Newcomb's problem (predictor's paradox)

The decision matrix

	Second box contains $1M	Second box is empty
Take second box only	$1M (prob. 0.99)	$0 (prob. 0.01)
Take both boxes	$1M + $1,000 (prob. 0.01)	$1,000 (prob. 0.99)

Expected utilities (assume utility of money is linear)

- The expected utility of taking only the second box:
  $0.99 \cdot u(\$1M) + 0.01 \cdot u(\$0) = 0.99 \cdot 1,000,000 + 0.01 \cdot 0 = 990,000$
- The expected utility of taking both boxes:
  $0.01 \cdot u(\$1.001M) + 0.99 \cdot u(\$1,000) = 0.01 \cdot 1,001,000 + 0.99 \cdot 1,000 = 11,000$

Newcomb's problem (predictor's paradox)

The decision matrix

	Second box contains $1M	Second box is empty
Take second box only	$1M (prob. 0.99)	$0 (prob. 0.01)
Take both boxes	$1M + $1,000 (prob. 0.01)	$1,000 (prob. 0.99)

Expected utilities (assume utility of money is linear)

- The expected utility of taking only the second box:
  $0.99 \cdot u(\$1M) + 0.01 \cdot u(\$0) = 0.99 \cdot 1,000,000 + 0.01 \cdot 0 = 990,000$
- The expected utility of taking both boxes:
  $0.01 \cdot u(\$1.001M) + 0.99 \cdot u(\$1,000) = 0.01 \cdot 1,001,000 + 0.99 \cdot 1,000 = 11,000$
- Since $990,000 > 10,000$, the principle of maximizing expected utility tells you that it is rational to only take the second box.

Newcomb's problem (predictor's paradox)

Newcomb's problem (predictor's paradox)

Conflicting recommendations

Newcomb's problem (predictor's paradox)

Conflicting recommendations

- Two of the most fundamental principles of decision theory – the dominance principle and the principle of maximizing expected utility – yield conflicting recommendations.

Newcomb's problem (predictor's paradox)

Conflicting recommendations

- Two of the most fundamental principles of decision theory – the dominance principle and the principle of maximizing expected utility – yield conflicting recommendations.

Consider decision matrix of an example analogous to Newcomb's problem

Newcomb's problem (predictor's paradox)

Conflicting recommendations

- Two of the most fundamental principles of decision theory – the dominance principle and the principle of maximizing expected utility – yield conflicting recommendations.

Consider decision matrix of an example analogous to Newcomb's problem

	Gene	No gene
Read Section 9.2	Pass exam & miserable life	Pass exam & normal life
Stop at Section 9.1	Fail exam & miserable life	Fail exam & normal life

Causal decision theory

Causal decision theory

Causal decision theory

Causal decision theory

Causal decision theory

- *Causal decision theory* is the view that a rational decision maker should keep all her beliefs about causal processes fixed in the decision-making process,

Causal decision theory

Causal decision theory

- *Causal decision theory* is the view that a rational decision maker should keep all her beliefs about causal processes fixed in the decision-making process, and always choose an alternative that is optimal according to these beliefs.

Causal decision theory

Causal decision theory

- *Causal decision theory* is the view that a rational decision maker should keep all her beliefs about causal processes fixed in the decision-making process, and always choose an alternative that is optimal according to these beliefs.

Decision Matrix

Causal decision theory

Causal decision theory

- *Causal decision theory* is the view that a rational decision maker should keep all her beliefs about causal processes fixed in the decision-making process, and always choose an alternative that is optimal according to these beliefs.

Decision Matrix

	Gene	No gene
Read Section 9.2	Pass exam & miserable life	Pass exam & normal life
Stop at Section 9.1	Fail exam & miserable life	Fail exam & normal life

Causal decision theory

Causal decision theory

Formulating causal decision theory

Causal decision theory

Formulating causal decision theory

The following statement:

Causal decision theory

Formulating causal decision theory

The following statement:

- Rational decision makers should do whatever is most likely to bring about the best expected result, while holding fixed all views about the likely causal structure of the world.

Causal decision theory

Formulating causal decision theory

The following statement:

- Rational decision makers should do whatever is most likely to bring about the best expected result, while holding fixed all views about the likely causal structure of the world.

Can be formalized as follows:

Causal decision theory

Formulating causal decision theory

The following statement:

- Rational decision makers should do whatever is most likely to bring about the best expected result, while holding fixed all views about the likely causal structure of the world.

Can be formalized as follows:

- Let $X \mathrel{\square\!\!\rightarrow} Y$ abbreviate the proposition 'If the decision maker were to do X, then Y would be the case', and let $p(X \mathrel{\square\!\!\rightarrow} Y)$ denote the probability of $X \mathrel{\square\!\!\rightarrow} Y$ being true.

Evidential decision theory

Evidential decision theory

Causal decision theory may yield counter-intuitive recommendations

Imagine that Paul is told that the number of psychopaths in the world is fairly low. The following scenario would then cast doubt on the causal analysis.

Evidential decision theory

Causal decision theory may yield counter-intuitive recommendations

Imagine that Paul is told that the number of psychopaths in the world is fairly low. The following scenario would then cast doubt on the causal analysis.

- Paul is debating whether to press the 'kill all psychopaths' button.

Evidential decision theory

Causal decision theory may yield counter-intuitive recommendations

Imagine that Paul is told that the number of psychopaths in the world is fairly low. The following scenario would then cast doubt on the causal analysis.

- Paul is debating whether to press the 'kill all psychopaths' button.
- He thinks it would be much better to live in a world with no psychopaths.

Evidential decision theory

Causal decision theory may yield counter-intuitive recommendations

Imagine that Paul is told that the number of psychopaths in the world is fairly low. The following scenario would then cast doubt on the causal analysis.

- Paul is debating whether to press the 'kill all psychopaths' button.
- He thinks it would be much better to live in a world with no psychopaths.
- Unfortunately, Paul is quite confident that only a psychopath would press such a button.

Evidential decision theory

Causal decision theory may yield counter-intuitive recommendations

Imagine that Paul is told that the number of psychopaths in the world is fairly low. The following scenario would then cast doubt on the causal analysis.

- Paul is debating whether to press the 'kill all psychopaths' button.
- He thinks it would be much better to live in a world with no psychopaths.
- Unfortunately, Paul is quite confident that only a psychopath would press such a button.
- Paul very strongly prefers living in a world with psychopaths to dying.

Evidential decision theory

Causal decision theory may yield counter-intuitive recommendations

Imagine that Paul is told that the number of psychopaths in the world is fairly low. The following scenario would then cast doubt on the causal analysis.

- Paul is debating whether to press the 'kill all psychopaths' button.
- He thinks it would be much better to live in a world with no psychopaths.
- Unfortunately, Paul is quite confident that only a psychopath would press such a button.
- Paul very strongly prefers living in a world with psychopaths to dying.
- Should Paul press the button?

Evidential decision theory

Causal decision theory may yield counter-intuitive recommendations

Imagine that Paul is told that the number of psychopaths in the world is fairly low. The following scenario would then cast doubt on the causal analysis.

- Paul is debating whether to press the 'kill all psychopaths' button.
- He thinks it would be much better to live in a world with no psychopaths.
- Unfortunately, Paul is quite confident that only a psychopath would press such a button.
- Paul very strongly prefers living in a world with psychopaths to dying.
- Should Paul press the button? Yes, according to causal decision theory.

Evidential decision theory

Causal decision theory may yield counter-intuitive recommendations

Imagine that Paul is told that the number of psychopaths in the world is fairly low. The following scenario would then cast doubt on the causal analysis.

- Paul is debating whether to press the 'kill all psychopaths' button.
- He thinks it would be much better to live in a world with no psychopaths.
- Unfortunately, Paul is quite confident that only a psychopath would press such a button.
- Paul very strongly prefers living in a world with psychopaths to dying.
- Should Paul press the button? Yes, according to causal decision theory.
  - $p(\text{press button} \mathrel{\square\!\!\rightarrow} \text{dead}) \ll p(\text{press button} \mathrel{\square\!\!\rightarrow} \text{live in a world without psychopaths})$

Evidential decision theory

Causal decision theory may yield counter-intuitive recommendations

Imagine that Paul is told that the number of psychopaths in the world is fairly low. The following scenario would then cast doubt on the causal analysis.

- Paul is debating whether to press the 'kill all psychopaths' button.
- He thinks it would be much better to live in a world with no psychopaths.
- Unfortunately, Paul is quite confident that only a psychopath would press such a button.
- Paul very strongly prefers living in a world with psychopaths to dying.
- Should Paul press the button? Yes, according to causal decision theory.
  - $p(\text{press button} \mathrel{\square\!\!\rightarrow} \text{dead}) \ll p(\text{press button} \mathrel{\square\!\!\rightarrow} \text{live in a world without psychopaths})$
  - This is because Paul either is or is not a psychopath, and the probability of the two possibilities does not depend on what he decides to do.

Evidential decision theory

Evidential decision theory

Evidential decision theorists

Evidential decision theory

Evidential decision theorists

- *Evidential* decision theorists, explicitly deny the causal analysis.

Evidential decision theory

Evidential decision theorists

- *Evidential* decision theorists, explicitly deny the causal analysis.
- They claim that it would be rational **not** to press.

Evidential decision theory

Evidential decision theorists

- *Evidential* decision theorists, explicitly deny the causal analysis.
- They claim that it would be rational **not** to press.
- The gist of their argument is that they think causal decision theorists calculate probabilities in the wrong way.

Evidential decision theory

Evidential decision theorists

- *Evidential* decision theorists, explicitly deny the causal analysis.
- They claim that it would be rational **not** to press.
- The gist of their argument is that they think causal decision theorists calculate probabilities in the wrong way.
  - Instead of asking yourself, "what is the probability that if I were to do X, then Y would be the case?",

Evidential decision theory

Evidential decision theorists

- *Evidential* decision theorists, explicitly deny the causal analysis.
- They claim that it would be rational **not** to press.
- The gist of their argument is that they think causal decision theorists calculate probabilities in the wrong way.
  - Instead of asking yourself, "what is the probability that if I were to do X, then Y would be the case?", a rational decision maker should ask, "what is the probability that if I were to do X, then Y would be the case given that I do X?"

Evidential decision theory

Evidential decision theorists

- *Evidential* decision theorists, explicitly deny the causal analysis.
- They claim that it would be rational **not** to press.
- The gist of their argument is that they think causal decision theorists calculate probabilities in the wrong way.
  - Instead of asking yourself, "what is the probability that if I were to do X, then Y would be the case?", a rational decision maker should ask, "what is the probability that if I were to do X, then Y would be the case given that I do X?"
  - So, they believe that it is not probabilities such as $p(X \boxdot\!\!\rightarrow Y)$ that should guide one's decision, but rather probabilities such as $p((X \boxdot\!\!\rightarrow Y) \mid X)$.

Evidential decision theory

Evidential decision theory

Objection to evidential decision theory

Evidential decision theory

Objection to evidential decision theory

- Evidential decision theory seems to require that the decision maker can somehow ascribe probabilities to his or her own choices.

Evidential decision theory

Objection to evidential decision theory

- Evidential decision theory seems to require that the decision maker can somehow ascribe probabilities to his or her own choices.
  - This is incoherent because one's own choices are not the kind of things one can reasonably ascribe probabilities to.