A Representation Theorem for Decisions about Causal Models

Daniel Dewey

Future of Humanity Institute

Abstract. Given the likely large impact of artificial general intelligence, a formal theory of intelligence is desirable. To further this research program, we present a representation theorem governing the integration of causal models with decision theory. This theorem puts formal bounds on the applicability of the *submodel hypothesis*, a normative theory of decision counterfactuals that has previously been argued on *a priori* and practical grounds, as well as by comparison to theories of counterfactual cognition in humans. We are able to prove four conditions under which the submodel hypothesis holds, forcing any preference between acts to be consistent with some utility function over causal submodels.

1 Introduction

Artificial general intelligence will likely have a large impact on the world. It is plausible that the course of AGI research will influence the character of this impact significantly, and therefore that researchers can take an active role in managing the impact of AGI. For example, Arel [1] argues that reinforcement learning is likely to cause an "adversarial" dynamic, and Goertzel [8] proposes ways to bias AGI development towards "human-friendliness."

A particularly large impact is predicted by I. J. Good's intelligence explosion theory [9,3,4], which argues that repeated self-improvement could yield super-intelligent (and hence super-impactful) AGIs. A few recent accounts of how an intelligence explosion could come about, what its effects could be, or how it could be managed include Schmidhuber [17], Hutter [10], Legg [13], Goertzel [7], Norvig [16, pp. 1037], Chalmers [3,4], Bostrom [2], Muehlhauser and Salamon [14], and Yudkowsky [23].

With this in mind, a formal theory of intelligence is preferable to a less formal understanding. First, though we won't be able to prove what the final result of an AGI's actions will be, we may be able to prove that it is pursuing a desirable goal, in the sense that it is Pareto-optimal, maximizes expected value, or is the best approximation possible given space and time constraints [11]; this appears to be the highest level of certainty available to us [24,2]. Second, we may be able to design an AGI that has a formal understanding of its own intelligence, which could then execute a series of provably goal-retaining self-improvements, where an equally long series of heuristic self-modifications would carry a high risk of "goal drift" [22]. Indeed, the theory of provably optimal self-improvement

J. Bach, B. Goertzel, and M. Iklé (Eds.): AGI 2012, LNAI 7716, pp. 60–68, 2012.

has been under investigation for some time by Schmidhuber, under the name of "Gödel machines" (e.g. [18]).

In searching for a formal theory of intelligence, this paper focuses on decision theory as it applies to causal models. If an agent holds its beliefs in the form of a causal model, is there a provably valid way that it should use that model to make decisions?

We consider the submodel hypothesis: "If an agent holds its beliefs in the form of a causal model, then it should use submodels as decision counterfactuals." We are able to show that the submodel hypothesis holds over a sharply defined set of decision problems by proving a representation theorem: an agent's preferences can be represented by a utility function over submodels if and only if they are complete, transitive, function-independent, and variable-independent.

2 Causal Models

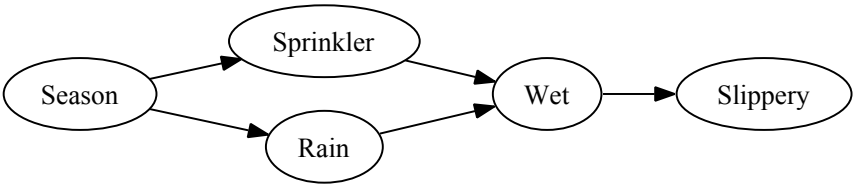
A causal model represents *events* and the *relationships* between them as *variables* and *functions*, respectively. For each variable, a model contains up to one function that calculates the value of that variable from the values of a set of other variables, representing the way that event depends on other events[1]. This allows a causal model to implicitly encode a joint distribution over values of the variables in the model; if a particular set of variable values is compatible with the functions between the variables, then it has a non-zero probability in the joint distribution. If an agent has observed a certain joint distribution of events in the world, it may be able in some cases to infer an underlying causal structure, and thereafter to represent its world using a causal model. For a full exposition of causal models and their properties, see [15].

In this paper, causal models will be written M or M', variables X or Y, and values of variable X will be written x or x' (except in concrete cases, e.g. variable "Switch" with values "on" and "off"). If X's value in M is given by function f applied to values of variables Y, this is written $X = f(Y)$. If X's value is given by a constant function with value x, this is written $X = x$.

Causal models can be pictured in two complementary ways: as a set of *structural equations* representing the functions, or as a *causal diagram*, a directed graph representing the dependencies and conditional independencies that hold between the variables.

The canonical example of a causal model (from [15]) is shown in Figure 1. It is a rudimentary model of the relationships between the *Season*, whether *Rain* is falling, whether a *Sprinkler* is on, whether the sidewalk is *Wet*, and whether the sidewalk is *Slippery*. In the causal diagram, an arrow from Season to Sprinkler indicates that the season plays an unmediated role in determining whether the sprinkler is on, though the graph does not show precisely what the relationship is. In the set of functional equations, the second equation shows the

[1] To simplify this work, error factors are left out of our account of causal models; reintroducing them should not interfere with our representation theorem or conclusions.



Rain = (Season = winter ∨ Season = fall) ? yes : no
Sprinkler = (Season = spring ∨ Season = summer) ? on : off
Wet = (Rain = falling ∨ Sprinkler = on) ? yes : no
Slippery = Wet ? yes : no

Fig. 1.

full relationship: in spring and summer, the sprinkler is on, and otherwise, it is off.

A *submodel* is a kind of causal model. Let M be a causal model, X be a variable, and x be a value of that variable: submodel M_x is derived from M by replacing X's function with the constant function $X = x$. Submodels may more generally replace a whole set of variables' functions with a set of constant functions, but this generalization will not be needed here. We use one non-standard notation: let $M_{X=f(Y)}$ denote the model derived by replacing X's function with f over values of Y in M.

3 The Submodel Hypothesis

The submodel hypothesis asserts that if an agent holds its beliefs in the form of a causal model, then it ought to use submodels as decision counterfactuals. A *decision counterfactual* is an agent's predictions of what would happen if it were to take a particular action. Thus, the submodel hypothesis can be restated as follows: "If an agent holds its beliefs in the form of a causal model, then it ought to predict the consequences of potential actions by replacing particular functions in that model with constants, and then choose the action whose consequences are most desirable."

In [15], Pearl argues for the submodel hypothesis by demonstrating how it avoids evidentialist decision errors, and by showing how it is formally very similar to Lewis' "closest world" theory of human counterfactual cognition [6]. He also argues that agents should model their own actions as uncaused "objects of free choice", and that the submodel method is the natural formalization of this idea.

Yudkowsky [25] builds on this work, arguing that decisions should be treated as *abstract computations*, representing them with variables that explain correlations in uncertainty stemming from bounded reasoning time and ability. Yudkowsky shows that agents who use submodels (of these kinds of models) as decision counterfactuals outperform other agents on many difficult decision theoretic problems, including Newcomb-like problems (where agents are simulated or predicted by their environments) and Prisoner's-dilemma-like problems

(where certain types of coordination between agents are required to reach more desirable equilibria). Yudkowsky also asserts in [26] that his framework "explains why the counterfactual surgery can have the form it does".

In this paper, we seek formal justification: what kinds of agents, in what kinds of decision problems, *must* use submodels (or an equivalent procedure) as decision counterfactuals? Conversely, what do the necessary and sufficient conditions for the submodel hypothesis tell us about its plausibility as a normative theory of decision-making?

4 Integrating Causal Models with Decision Theory

Causal models are not a standard part of decision theory, so we begin with a simple, naturalistic integration of causal-model-based beliefs into decision theory.

Suppose that an agent holds its beliefs in the form of a causal model M. So that the model can guide the agent in making a choice, let some variable X in M represent the current decision, and let the rest of the model represent the decision's relationships to other events.

Though values of X represent different choices, a single variable value does not contain the beliefs the agent uses to make its decision. In order to state an agent's preferences, it will be convenient to bundle beliefs and choices together into *acts*. Each act is a pair $\langle M, x\rangle$, where X taking value x represents the choice of this act, so that all of the information an agent has about an act is contained within the act itself. We can therefore define a *decision problem* to be a set of acts; an agent solves a decision problem by choosing one of the acts. Since beliefs are bundled with acts, a weak preference between acts, $\succsim$, can be used to characterize all of the agent's decisions in all possible states of belief. We can now state the submodel hypothesis formally:

> *An agent should act according to a preference over acts $\succsim$ that is representable by a utility function over submodels; i.e., there should exist a U from submodels to reals such that*
>
> $$\langle M, x\rangle \succsim \langle M', y\rangle \iff U(M_x) \geq U(M'_y).$$

5 The Conditions

We have found four conditions on preferences over acts that are jointly equivalent to representability by a utility function over submodels. The first and second can be plausibly argued for by assuming that the agent is consequentialist; the third and fourth are novel, and whether they are justified is still an open question.

Suppose that the agent is consequentialist: it chooses one act or another for the sake of achieving a more desirable eventual outcome. If this is so, then even acts that could never appear in the same decision problem, such as $\langle M, x\rangle$ and $\langle M', y\rangle$, should be comparable according to the desirability of the eventual

outcomes they are expected to bring about. Consequentialism, then, implies that an agent's preference over acts should be complete:

$$(A \succsim B) \vee (B \succsim A) \qquad \text{(Completeness.)}$$

Likewise, unless the agent's concept of desirability has cycles (in which outcome 1 is better than 2, 2 is better than 3, and 3 is better than 1), its preference over outcomes, and hence over acts, should be transitive:

$$(A \succsim B) \wedge (B \succsim C) \Rightarrow (A \succsim C) \qquad \text{(Transitivity.)}$$

It thus seems plausible that a consequentialist agent must have a complete and transitive preference over acts.

The third and fourth conditions are novel, and apply specifically to agents whose beliefs are held as causal models. Recall that each act specifies a particular variable to represent the decision event; if the agent is naturalistic, meaning that it represents its own decision process in the same way that it represents other cause-effect relationships, then the decision variable's function must represent the agent's decision process. *Function-independence* states that if two acts differ only in the function representing the decision process, they must be equally preferable:

$$\langle M, x\rangle \sim \langle M_{X=f(Y)}, x\rangle. \qquad \text{(Function-independence)}$$

The fourth condition, variable-independence, also requires certain indifferences between acts. In particular, variable-independence applies to acts that model the agent's decision as uncaused, representing it as a variable with no parents. Formally, variable-independence states that if a pair of acts share a model, and if each act represents the agent's decision process as a function of no inputs, then the two acts must be equally preferable:

$$X = x \wedge Y = y \text{ in } M \Rightarrow \langle M, x\rangle \sim \langle M, y\rangle. \qquad \text{(Variable-independence)}$$

We have found function-independence and variable-independence to be necessary for the submodel hypothesis, but attempts to discover whether and how they are generally justified have not been successful. This could be a fruitful area for future work.

6 The Representation Theorem

We are now ready to show that the four conditions together are necessary and sufficient for the submodel hypothesis:

Theorem 1. *If and only if a preference $\succsim$ over acts is complete, transitive, function-independent, and variable-independent, then $\succsim$ can be represented by a utility function over submodels, i.e. there exists a U from submodels to reals such that*

$$\langle M, x\rangle \succsim \langle M', y\rangle \iff U(M_x) \geq U(M'_y).$$

Proof. First, it is easy to show that each condition is necessary. Assuming that U represents $\succsim$, $\succsim$ must be:

Complete: Any two real utilities are comparable with $\geq$, so if U is complete and represents $\succsim$, then any two acts must be comparable with $\succsim$.

Transitive: Any three real utilities obey transitivity, so if U is complete and represents $\succsim$, then any three acts must be transitive under $\succsim$.

Function-independent:

$$\begin{aligned} M_x &= (M_{X=f(Y)})_x \\ &\Rightarrow U(M_x) = U((M_{X=f(Y)})_x) \\ &\Rightarrow \langle M, x\rangle \sim \langle M_{X=f(Y)}, x\rangle. \end{aligned}$$

Variable-independent:

$$\begin{aligned} X = x \wedge Y = y \text{ in } M \\ &\Rightarrow M = M_x = M_y \\ &\Rightarrow U(M_x) = U(M_y) \\ &\Rightarrow \langle M, x\rangle \sim \langle M, y\rangle. \end{aligned}$$

Second, we show that the conditions are sufficient for the existence of a utility representation over submodels; from here on, we assume that all conditions hold. Let α be any function from submodels "back to corresponding acts", meaning that $\alpha(S) = \langle M, x\rangle \Rightarrow S = M_x$. The following lemmas will be useful:

Lemma 1. $\forall M, x : \langle M, x\rangle \sim \alpha(M_x)$.

Proof. Let $\langle M', y\rangle = \alpha(M_x)$. By definition of α, $M_x = M'_y$.

$$\begin{aligned} \langle M, x\rangle &\sim \langle M_x, x\rangle && \text{by function-independence,} \\ &\sim \langle M'_y, x\rangle && \text{since } M_x = M'_y; \end{aligned}$$

because $M_x = M'_y$, we know that $X = x$ in M'_y, and trivially $Y = y$ in M'_y, and so by variable-independence,

$$\begin{aligned} &\sim \langle M'_y, y\rangle \\ &\sim \langle M', y\rangle && \text{by function-independence,} \\ &\sim \alpha(M_x), \end{aligned}$$

and so $\langle M, x\rangle \sim \alpha(M_x)$. □

Lemma 2. $\langle M, x\rangle \succsim \langle M', y\rangle \iff \alpha(M_x) \succsim \alpha(M'_y)$.

Proof. $\Rightarrow$: Assume $\langle M, x\rangle \succsim \langle M', y\rangle$. By Lemma 1,

$$\alpha(M_x) \sim \langle M, x\rangle \succsim \langle M', y\rangle \sim \alpha(M'_y),$$

and since $\succsim$ is transitive, $\alpha(M_x) \succsim \alpha(M'_y)$.
$\Leftarrow$: Assume $\alpha(M_x) \succsim \alpha(M'_y)$. By Lemma 1,

$$\langle M, x\rangle \sim \alpha(M_x) \succsim \alpha(M'_y) \sim \langle M', y\rangle,$$

and since $\succsim$ is transitive, $\langle M, x\rangle \succsim \langle M', y\rangle$.

□

Now we can construct a utility function on submodels and to show that it represents $\succsim$. Let v be an injective function from submodels to the set $\{2^{-n} : n \in \mathbb{N}\}$, and let U be defined as

$$U(S) = \sum_{S' : \alpha(S) \succsim \alpha(S')} v(S').$$

Since the sum of $\{2^{-n} : n \in \mathbb{N}\}$ converges, the utility function is defined even when the set of submodels is (countably) infinite [21].

First, we will show that every preference over acts is represented in utilities. Assume that one act is weakly preferred over another, so that $\langle M, x\rangle \succsim \langle M', y\rangle$. By Lemma 2, $\alpha(M_x) \succsim \alpha(M'_y)$. Since $\succsim$ is transitive, any $\alpha(S)$ weakly dispreferred to $\alpha(M'_y)$ is also dispreferred to $\alpha(M_x)$, and so

$$\{S : \alpha(M_x) \succsim \alpha(S)\} \supseteq \{S : \alpha(M'_y) \succsim \alpha(S)\}.$$

By definition of U, we conclude that $U(M_x) \geq U(M'_y)$.

Second, we will show that every utility difference represents a preference. Let $U(M_x) \geq U(M'_y)$. To draw a contradiction, assume that $\alpha(M_x) \not\succsim \alpha(M'_y)$. By completeness, $\alpha(M'_y) \succ \alpha(M_x)$. It follows by transitivity that

$$\{S : \alpha(M'_y) \succsim \alpha(S)\} \supset \{S : \alpha(M_x) \succsim \alpha(S)\}.$$

By definition of U, this means that $U(M'_y) > U(M_x)$, a contradiction; therefore, $\alpha(M_x) \succsim \alpha(M'_y)$. By Lemma 2, $\langle M, x\rangle \succsim \langle M', y\rangle$.

Thus, we have shown that the conditions given are necessary and sufficient for the existence of a representative utility function over submodels; the submodel hypothesis is confirmed over the class of problems defined by the conditions. □

7 Conclusion

In this paper, we have shown a set of four conditions under which the submodel hypothesis is confirmed, i.e. an agent whose beliefs are held as a causal model must have preferences that can be represented by a utility function over submodels. This puts sharply-defined boundaries on where the submodel hypothesis, which has previously been argued by Pearl [15] and Yudkowsky [25], is justified and required. More broadly, we have aimed to contribute to a formal theory of intelligence, with the goal of shaping the impact of AGI to be safe and beneficial.

Acknowledgements. Thanks to Vladimir Slepnev, Benja Fallenstein, and Luke Muehlhauser for their comments on earlier versions of the paper.

References

1. Arel, I.: Reward Driven Learning and the Risk of an Adversarial Artificial General Intelligence. Talk at "The Future of AGI Workshop Part 1 - Ethics of Advanced AGI," The Fourth Conference on Artificial General Intelligence (2011)
2. Bostrom, N.: The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. Minds and Machines 22, 71–85 (2012)
3. Chalmers, D.: The Singularity: A Philosophical Analysis. Journal of Consciousness Studies 17, 7–65 (2010)
4. Chalmers, D.: The Singularity: a Reply. Journal of Consciousness Studies 19 (2012)
5. Drescher, G.: Good and real: Demystifying paradoxes from physics to ethics. Bradford Books, MIT Press, Cambridge, MA (2006)
6. Galles, D., Pearl, J.: An Axiomatic Characterization of Counterfactuals. Foundations of Science III, 151–182 (1998)
7. Goertzel, B.: Should Humanity Build a Global AI Nanny to Delay the Singularity Until It's Better Understood? Journal of Consciousness Studies 19, 96–111 (2012)
8. Goertzel, B.: Nine Ways to Bias Open-Source AGI Toward Friendliness. Journal of Evolution and Technology 22, 116–131 (2012)
9. Good, I.J.: Speculations Concerning the First Ultraintelligent Machine. In: Alt, F.L., Rubinoff, M. (eds.) Advances in Computers, vol. 6, pp. 31–88 (1965)
10. Hutter, M.: Can Intelligence Explode? Journal of Consciousness Studies 19, 143–166 (2012)
11. Hutter, M.: Universal algorithmic intelligence: A mathematical top-down approach. In: Artificial General Intelligence, pp. 227–290. Springer, Berlin (2007)
12. Legg, S.: Is there an Elegant Universal Theory of Prediction? IDSIA Technical Report No. IDSIA-12-06 (2006)
13. Legg, S.: Machine Super Intelligence. PhD dissertation, University of Lugano (2008)
14. Muehlhauser, L., Salamon, A.: Intelligence Explosion: Evidence and Import. In: The Singularity Hypothesis: A Scientific and Philosophical Assessment. Springer, Berlin (2012)
15. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press (2000)
16. Russell, S., Norvig, P.: AI: A Modern Approach, 3rd edn. Prentice-Hall, Englewood Cliffs (1995)
17. Schmidhuber, J.: Philosophers & Futurists, Catch Up! Journal of Consciousness Studies 19, 173–182 (2012)
18. Schmidhuber, J.: Gödel machines: Fully Self-Referential Optimal Universal Self-Improvers. In: Artificial General Intelligence, pp. 119–226 (2006)
19. Solomonoff, R.: A Formal Theory of Inductive Inference, Part I. Information and Control 7(1), 1–22 (1964)
20. Solomonoff, R.: A Formal Theory of Inductive Inference, Part II. Information and Control 7(2), 224–254 (1964)
21. Voorneveld, M.: Mathematical Foundations of Microeconomic Theory: Preference, Utility, Choice (2010), `https://studentweb.hhs.se/CourseWeb/CourseWeb/Public/PhD501/1001/micro1.pdf`

22. Yudkowsky, E.: Artificial intelligence as a positive and negative factor in global risk. In: Global Catastrophic Risks. Oxford University Press, Oxford (2008)
23. Yudkowsky, E.: Complex Value Systems are Required to Realize Valuable Futures. In: The Proceedings of the Fourth Conference on Artificial General Intelligence (2011)
24. Yudkowsky, E., et al.: Reducing Long-Term Catastrophic Risks from Artificial Intelligence. The Singularity Institute, San Francisco (2010)
25. Yudkowsky, E.: Timeless decision theory. The Singularity Institute, San Francisco (2010)
26. Yudkowsky, E.: Ingredients of Timeless Decision Theory (2009), http://lesswrong.com/lw/15z/ingredients_of_timeless_decision_theory/