Does Causal Meaning Depend on Models? A Critique of Mental Model Theory of Causation

Pengfei Yin

Shaanxi Key Laboratory of Behavior and Cognitive Neuroscience, School of Psychology

KEYWORDS

causal determinism
mental model theory
insufficient
infinite regress
causal relation

ABSTRACT

The mental model theory (MMT) provides a unified account of causal representation and inference. The theory claims that a causal assertion "A causes B" has a deterministic meaning that refers to three temporally ordered possibilities: A and B, not A and B, not A and not B. Furthermore, MMT proposes that causal relations depend only on these possibilities, and not on causal powers or mechanisms. In this paper, the MMT account of causation is critiqued by arguing that mental models alone are not sufficient to define the meaning of causal relations, and that if MMT adhered to its own principles, then its account of causation would fall into an infinite regress.

A SHORT INTRODUCTION TO THE MENTAL MODEL THEORY'S ACCOUNT OF CAUSATION

The Meaning of Causation in Mental Model Theory

The mental model theory (MMT) postulates that the meanings of causal relations (causes, prevents, and enables) refer to different sets of temporally ordered deterministic possibilities (Johnson-Laird & Khemlani, 2017). According to the original version of MMT, "A causes B" corresponds to three disjunctive possibilities: (A & B), (¬A & B), or (¬A & ¬B), with the temporal constraint that B does not precede A (Goldvarg & Johnson-Laird, 2001). According to the revised version of MMT, "A causes B" refers to the conjunction of the above three possibilities, that is, possibly (A & B), possibly (¬A & B), and possibly (¬A & ¬B; Johnson-Laird & Khemlani, 2017). The distinction between the two versions concerns the connective joining of the three possibilities. In the earlier version of MMT, it is a disjunction, whereas in the revised version of MMT, it is a conjunction. What remains unchanged are the three possibilities and the temporal constraint on cause and effect.

Are There Other Components to the Meanings of Causal Relations?

The mental model theory claims that any factors beyond temporally ordered possibilities such as properties, forces, mechanisms, intervention, and so forth, are not part of the core meaning of causation (Goldvarg & Johnson-Laird, 2001; Johnson-Laird & Khemlani, 2017; Khemlani et al., 2014). For example: " mechanisms and their cognates, such as forces and powers, cannot be part of the core meaning of causal assertions" (Khemlani et al., 2014, p. 2); "The phenomena, however, do not call for the introduction of causal properties, powers, or mechanisms, into the meaning of causal relations" (Goldvarg & Johnson-Laird, 2001, p. 574–575); "In sum, knowledge of any of the factors that theorists invoke—force, power, means of production, interventions, explanatory principles, and mechanisms, …, are not part of the core meanings of such claims" (Johnson-Laird & Khemlani, 2017, p. 176).

However, although MMT denies that these factors are part of the core meaning of causal relations, the proviso is added that these factors can be incorporated into the model by a process of modulation. In other theories, these factors are core to the meaning of causation, for example, "power" in the causal powers theory (White, 2005), "force" in force dynamics theory (Wolff, 2007), "mechanisms" in mechanistic accounts of causation (Ahn & Bailenson, 1996; Ahn & Kalish, 2000; Ahn et al., 1995; Stephan & Waldmann, 2022), "positive dependencies" in probabilistic accounts of causal conditionals (Skovgaard-Olsen, Collins et al., 2019; Skovgaard-Olsen, Kellen et al., 2019), and "explanatory quality" in semantic accounts of causal conditionals (Douven et al., 2018, see also Oaksford & Chater, 2020a). Given these factors' centrality in these other accounts, by virtue of what criterion can MMT expel them from the core meaning of causation?

Corresponding author: Pengfei Yin, Shaanxi Key Laboratory of Behavior and Cognitive Neuroscience, School of Psychology, Shaanxi Normal University, 199 Chang'an Road, Yanta, Xi'an 710062, China
Email: 752638238@qq.com

DOI • 10.5709/acp-0349-5

Meaning Depends on Models

The mental model theory argues that meaning depends on models. For example, Goldvarg et al. declared that “The meanings depend on possibilities, …, and does not depend on causal powers or mechanisms” (Goldvarg and Johnson-Laird, 2001, p. 565). For example, the meaning of “A prevents B” depends on the set of models “A&¬B, ¬A&B, ¬A&¬B”; the meaning of “A causes B” depends on the set of models “A&B, ¬A&B, ¬A&¬B”. Similar statements are as follows: “The models fix the appropriate causal relation” (Goldvarg & Johnson-Laird, 2001, p. 577); “Possibilities underlie the meanings of causal relations” (Johnson-Laird & Khemlani, 2017, p. 172); “Possibilities yield the distinct causal relations” (Johnson-Laird & Khemlani, 2017, p. 172.).

PROBLEMS FOR THE “CAUSAL RELATIONS DEPEND ON MODELS” ARGUMENT

In this section, I present some problems for the MMT. These problems arise at three levels: (a) MMT’s ability to distinguish causal from noncausal sequences; (b) MMT’s ability to distinguish causes and enabling conditions; and (c) the problem that the same mental model can have multiple causal descriptions.

Causal and Noncausal Sequences

Suppose a man hears two clocks striking the hours. One of the clocks is faster than the other by one second or so. Consequently, he hears the first clock striking the hours always a fraction earlier than the second. Suppose also that the first clock will power off one day before the second one. Consequently, the day after the first clock powers off, the man would hear the second clock strike 24 times, without any sound from the first clock. Finally, the next day, when the second clock also powers down, neither clock will strike. If we denote the above scenario as possibilities with “A” representing the striking of the first clock and “B” representing the striking of the second clock, we get Set 1:

A B
¬A B
¬A ¬B

where the symbol ‘¬’ denotes negation. Set 1 seems to fully specify the possible events I have described and their temporal dependence. But this representation could not warrant the assertion that “A causes B”, the first clock striking causes the second clock to strike. Take another example, Mike is very hygienic and the first thing he does after getting up is to brush his teeth. This yields two kinds of cases which are represented as Set 2 as follows:

Get up brush teeth
¬Get up ¬brush teeth

The above corresponds to the models of a strong causal relationship (Johnson-Laird &Khemlani, 2017, p. 173). However, it would not be warranted to say that getting up strongly causes Mike to brush his teeth. These examples result in temporally ordered models, but are they causal or noncausal?

Sets 1 and 2 show that many pairs of events unfolding in time can fully satisfy the constraints on the core meaning of cause defined by MMT. But these pairs have little or no causal relation. Thus, something more than the mere temporally ordered models must be in place to exclude noncausal relations before these models can be used in inference. Consequently, the models themselves are not sufficient to underlie the causal relations of interest. These examples also suggest that some factor beyond the models is required to distinguish causal from noncausal relations.

Mental Model Theory’s Ability to Distinguish Causes and Enabling Conditions

How does MMT distinguish different causal relations from each other based on different sets of models? For example, under what condition can “causes” be distinguished from “enabling conditions?”

The mental model theory would appear to have to assume that the causal relation should be deterministic rather than probabilistic to make this distinction. As an example, what is the relationship between having an injection and loss of consciousness? The following two sets of models can distinguish causes and enabling conditions but only if causation is both objectively and subjectively deterministic whether it is generic or singular.

Set 3:
injection loss-of-consciousness
¬injection loss-of-consciousness
¬injection ¬loss-of-consciousness

Set 4:
injection loss-of-consciousness
injection ¬loss-of-consciousness
¬injection ¬loss-of-consciousness

If causation is not objectively deterministic, then Set 3 will not be consistently observed even if injection does cause loss of consciousness. Subjectively, if people’s views of causation are not deterministic, then they would also hesitate to endorse that injection enables loss of consciousness, even if they consistently and objectively observed Set 4.

However, the models themselves cannot provide the information that a certain relation is deterministic or probabilistic. Thus, at this level, the models are again insufficient to distinguish causal relations. Furthermore, the prerequisite that causation is deterministic held by MMT is not uncontroversial (for a specific critique of MMT’s deterministic view of causation see Yin and Sun, 2021; for the probabilistic view of causal relations see, e.g., Oaksford & Chater, 2017; for more general arguments concerning the probabilistic new paradigm and MMT in reasoning see Oaksford et al., 2019).

The Same Mental Model Can Have Multiple Causal Descriptions

Even if information that two events are causally related and causation is deterministic is given, how an exclusive relation be specified for a given set of models since each set of models can correspond to multiple causal relations?

Table 1 shows that each set of models corresponds to six different forms of relations. These formulations are expansions of or derivations from MMT's proposals. In Goldvarg and Johnson-Laird's (2001) article, MMT already provided two different causal relations (the first two rows in bold) for each set of models. The remaining four relations are expanded or derived from MMT's proposals.

Thus, given that multiple relations can be chosen for a fixed set of models, by virtue of what cue can reasoners fix the appropriate description? Mental model theory claimed that "three possibilities are hard to hold in mind at the same time, but one description can focus on one possibility and another description focus on another possibility" (Goldvarg & Johnson-Laird, 2001, p. 573). But this cannot fully answer this question, because in some cases, even when focusing on one possibility, there is still more than one description corresponding to it. For example, in the second column of Table 2, there are two candidate descriptions for the set of possibilities, even if focusing just on the possibility "A & B": "A enables B," and "A does not cause B." What cue allows MMT fix one of them? A further question is: do "A enables B" and "A does not cause B" have the same meaning, provided they have the same models? This is disputable, as some theories will give a negative answer. For example, in the probabilistic contrast model (Cheng & Novick, 1991, 1992), "A enables B" means that the probability of B in the presence of A is greater than the probability of B in the absence of A, whereas "A does not cause B" means that the probability of B in the presence of A is not greater than the probability of B in the absence of A. From a force-dynamic perspective (Wolff, 2007, 2014), the answer should also be negative, because there can exist at least one different dimension among the three dimensions representing the two kinds of relations (see Table 2). As Table 2 shows, "A enables B" means that B is approached under the force of A, whereas "A does not cause B" means that B is not approached despite the configurations of the former two dimensions.[1]

Thus, if different causal relations describing the same set of models have different meanings, then the models themselves are not sufficient to make the distinction. This is because models themselves only concern the extrinsic co-occurrences between two events in a deterministic format, but do not concern the intrinsic configurations or mechanisms underlying causal relations.

THE REGRESS AND INFINITE LOOP DILEMMA

An immediate question is: what, in addition to mere temporally ordered models, distinguishes between causal relations? This unknown is labeled "X", and question is formulated as F1:

$$X + \text{Models} \rightarrow C \qquad \text{(F1)}$$

Where C represents a certain causal relation, and → represents sufficient condition. It should be emphasized again that there should be no causal components in the models, otherwise the definition will be circular.

This question is difficult to answer under the MMT framework if it takes its own principles seriously. Other theories can easily give an answer to what X is. For example, it is a mechanism (Ahn and Kalish, 2000), configuration of forces (Wolff, 2007, 2014), or positive dependency (Skovgaard-Olsen, Collins et al., 2019).

Of course, MMT can still say X may be knowledge of mechanisms or a similar factor which modulates the models, but such a response renders MMT too flexible (Bonatti, 1994) and ad hoc. It also risks MMT giving up its own principles, as it cannot arbitrarily assert that the contribution of X is less than the contribution of the models. And if it cannot, then the core status of the models may be taken over by the mechanisms.

To answer what X is, it seems more feasible to first answer what X is not. First, X should not be a certain causal relation itself. Otherwise what is known is already what is wanted to be known, and the models would serve no point.

Second, X should also not be the mechanisms between the pair of evens, otherwise the causal relationship between them will be directly inferred with no reference to the models, as it has already been established that people prefer information about causal mechanisms rather than covariation in singular causation judgments (Ahn et al., 1995, see also Stephan & Waldmann, 2022). If X is a mechanism, then the models are not playing a core role in determining the causal relation.

Third, X should not be the information about causal powers or properties of events under investigation because they will override informa-

TABLE 1.
The Corresponding Relationship Between a Fixed set of Models and Different Causal Relations

Temporally ordered possibilities	A B ¬A ¬B ¬A B	A B A ¬B ¬A ¬B	A ¬B ¬A ¬B ¬A B	A ¬B ¬A B A B
Causal relations	**A causes B**	**A enables B**	**A prevents B**	**A enables ¬B**
	¬A enables ¬B	**¬A prevents B**	**¬A enables B**	**¬A causes B**
	A prevents ¬B	¬A causes ¬B	A causes ¬B	¬A prevents ¬B
	A does not enable ¬B	¬A does not enable B	A does not enable B	¬A does not enable ¬B
	¬A does not prevent B	A does not prevent ¬B	¬A does not prevent ¬B	A does not prevent B
	¬A does not cause ¬B	A does not cause B	¬A does not cause B	A does not cause ¬B

TABLE 2.
Representations of Different Causal Relations (Descripting the Same Set of Models) in Three Dimensions According to Force Dynamic Theory

Models	Causal relations	Patient tendency for endstate	Affector-patient concordance	Endstate approached
A B	A enables B	Y	Y	Y
A ¬B				
¬A ¬B	A does not cause B	N?	N?	N

Note. Y = Yes, N = No; N? represents this dimension cannot be certainly specified, but at least there exist some magnitudes or directions that simultaneously satisfy the causal relation and the "N" configuration of this dimension.

tion about the covariation of cause and effect (the noncausal models), which is recognized by MMT (Goldvarg & Johnson-Laird, 2001, p. 574, see also White, 1995).

With possibilities about X excluded, the only choice remaining for MMT is that X is something about the higher or more abstract level of causal relation (contrasted with the hierarchically lower recourse to mechanisms) covering (contrasted with underlying) the events of interest.

But how is this higher causal relation mentally represented? We argue that if MMT sticks to its own principle that "the meanings depend on possibilities, …, and do not depend on causal powers or mechanisms" (Goldvarg & Johnson-Laird, 2001, p. 565), the higher level causal relation X must be represented as a set of higher level models termed Models^{h1}. Unfortunately, the Models^{h1} themselves are not sufficient to underlying X once more. Then something more than the models again is in needed, which again should not be mechanism and their cognates. Otherwise the purported core status of models will again be given away to these factors. Therefore, like formulation F1, the representation of X should be F2:

$$X' + \text{Models}^{h1} \rightarrow X \qquad \text{(F2)}$$

Where X' represents a higher or more abstract level causal relation covering X.

Continually, following the same logic, X' should be further represented as F3:

$$X'' + \text{Modelsh2} \rightarrow X' \qquad \text{(F3)}$$

Thus, just as MMT argues against mechanisms by saying that "these factors are impossible to define without referring to causation itself" (Johnson-Laird & Khemlani, 2017, p. 176), the above analyses exactly reveal that MMT's approach to causation also cannot escape from the charge of infinite (upward) regress.

The solution to the above infinite (upward) regress demands a point where Xn does not further appeal to its higher level causal relation – the top out point (corresponding to the bottom out point MMT charges mechanism). But at that point, the Modelshn must embody causal cues intrinsically, otherwise it still will not suffice to distinguish between causal and noncausal sequences. Interestingly, if MMT had to adopt this solution, then it would go to its own exact opposite, on pain of the regress faced by the mechanistic accounts of causation.

SUMMARY

By claiming that "some of these factors (force, power, mechanisms, explanatory principles, scientific laws, etc.) are impossible to define without referring to causation itself" (Johnson-Laird & Khemlani, 2017, p. 176), MMT excludes them but only endorses the temporally ordered possibilities as the core meaning of causation. Thus, as core meaning, these possibilities should embody no further causal clues, otherwise this would commit an immediate circularity in definition as MMT's own critique of other theories. Thus, the models should be hollow-causation models whether they are yielded from the meaning of the premises in verbal reasoning or observed from the co-occurrences of events (e.g., the two clocks or Mike's daily routine).

It is natural and easy to follow MMT's claims in other articles (e.g., Khemlani & Johnson-Laird, 2021) that in reasoning, people first interpret the meanings (intensions) of premises and then construct mental models (extensions) based on the meanings. That is, models depend on meanings. For example, from the premises "A causes B" and "B prevents C," one might build models "A & B, ¬A & B, ¬A & ¬B" and "B & ¬C, ¬B & C, ¬B & ¬C" for them, respectively (but just under the presupposition that causal relations are deterministic). Furthermore, one might combine the two sets of models into an integrated set of models "A & B & ¬C, ¬A & B & ¬C, ¬A & ¬B & C, ¬A & ¬B & ¬C".[2] However, questions arise when MMT declared that "the meanings depend on possibilities" (Goldvarg & Johnson-Laird, 2001, p. 565) in their articles concerning causal representation and reasoning. We revealed that the hollow-causation models themselves are not sufficient (a) to discriminate causal from noncausal sequences,[3] (b) to distinguish "cause" from "enabler," and (c) to fix a particular causal assertion among several legitimate options. Thus, something more than the mere temporally ordered possibilities must be in place to play the above functions in determining a certain causal meaning. Arguably, this "something more" cannot be the factors that already have been rejected by MMT such as force, power, and mechanisms, otherwise they will take over the core status of temporally ordered possibilities. Rather the "something more" should be models of higher causal concepts. But the higher models are still insufficient to underpin a causal relation at its own level, which leads to an infinite regress if MMT takes its own principle seriously.

Mental model theory may resort to modulation processes derived from prior knowledge to meet the above challenges. But "knowledge modulation" is itself a question-begging proposal (for the specific analyses see Yin, under review). The theory claimed that "when one model is based on knowledge, it takes precedence over a model based on premises" (Johnson-Laird & Khemlani, 2017, p. 181). Here, we put forward two points for MMT to consider: (a) When the mechanism (or power) theories are just the knowledge of individuals, do mechanisms take precedence over the core meaning defined by temporally ordered possibilities or are mechanisms still "not part of the meaning of causal assertions" (Goldvarg & Johnson-Laird, 2001, p. 603)? (b) Is knowledge globally consistent or full of both gaps and contradictions? If it is the latter case (Oaksford & Chater, 2020b), then which piece of knowledge plays the modulatory role and how does MMT deal with the possible contradictions between them?

FOOTNOTES

[1] For the subtle difference between the mental model theory and force dynamic theory, see also Wolff and Song (2003). Wolff et al. state that "the predictions of the model theory coincide with those of the force dynamic model only for the simplest kinds of force dynamic interactions: namely, those in which the forces associated with the affector and the patient are either diametrically opposed or fully concordant.

[2] Again, no causal information should remain in each of the models, otherwise (a) the definitions are circular, and (b) the remnant causal components should be represented as further models, which (if they still contain causal components) in turn should be further represented as further models--this will be an infinite recursion. So, the integrated set of models itself still faces the challenges described here.

[3] In verbal reasoning, participants already know there are some undefined causal relations between A and C from the premises "A causes B" and "B prevents C," but this knowledge is hinted at by language rather than by models.

DATA AVAILABILITY

Because this article is a short theoretical comment, there was no available data.

REFERENCES

Ahn, W., & Bailenson, J. (1996). Causal attribution as a search for underlying mechanism: An explanation of the conjunction fallacy and the discounting principle. *Cognitive Psychology, 31*, 82–123. doi: 10.1006/cogp.1996.0013

Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition, 54*, 299–352. doi: 10.1016/0010-0277(94)00640-7

Ahn, W., & Kalish, C. W. (2000). The role of mechanism beliefs in causal reasoning. In F. C. Keil & R. A. Wilson (Eds.) *Explanation and Cognition* (pp. 199–225). Cambridge, MA: MIT Press.

Bonatti, L. (1994). Propositional reasoning by model? *Psychology Review, 101*, 725–733. doi: 10.1037/0033-295X.101.4.725

Cheng, P. W., & Novick, L. R. (1991). Causes versus enabling conditions. *Cognition, 40*, 83–120. doi: 10.1016/0010-0277(91)90047-8

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review, 99*, 365–382.

Douven, I., & Mirabile, P. (2018). Best, second-best, and good-enough explanations: How they matter to reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*, 1792–1813. doi:10.1037/ xlm0000545

Goldvarg, E., & Johnson-Laird, P. N. (2001). Naive causality: A mental model theory of causal meaning and reasoning. *Cognitive Science, 25*, 565–610. doi: 10.1207/s15516709cog2504_3

Johnson-Laird, P. N., & Khemlani, S. (2017). Mental models and causation. In: M. Waldmann (Ed.). *Oxford handbook of causal reasoning* (pp. 1–42). Oxford University Press.

Khemlani, S., & Johnson-Laird, P. N. (2021). Reasoning about properties: A computational theory. *Psychological Review. Advance online publication*. https://doi.org/10.1037/rev0000240

Khemlani, S., Barbey, A. K., & Johnson-Laird, P. N. (2014). Causal reasoning with mental models. *Frontiers in Human Neuroscience, 8*, 849. doi: 10.3389/fnhum.2014.00849.

Oaksford, M., & Chater, N. (2017). Causal models and conditional reasoning. In: M. Waldmann (Ed.), *Oxford handbook of causal reasoning* (pp. 327–346). Oxford University Press.

Oaksford, M., & Chater, N. (2020a). Integrating causal Bayes nets and inferentialism in conditional inference. In: S. Elqayam, I. Douven, J. St. B. T. Evans, & N. Cruz (Eds.). *Logic and uncertainty in the human mind: A tribute to David E. Over* (pp. 116–132). London: Routledge.

Oaksford, M., & Chater, N. (2020b). New paradigms in the psychology of reasoning. *Annual Review of Psychology, 71*, 305–330. doi: 10.1146/annurev-psych-010419-051132

Oaksford, M., Over, D. E., & Cruz, N. (2019). Paradigms, possibilities, and probabilities: Comment on Hinterecker, Knauff, and Johnson-Laird (2016). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*, 288–297. doi: 10.1037/xlm0000586

Skovgaard-Olsen, N., Collins, P., Krzyżanowska, K., Hahn, U., & Klauer, K. C. (2019). Cancellation, negation, and rejection. *Cognitive Psychology, 108*, 42–71. doi: 10.1016/j.cogpsych.2018.11.002

Skovgaard-Olsen, N., Kellen, D., Hahn, U., & Klauer, K. C. (2019). Norm conflicts and conditionals. *Psychological Review, 126*, 611–633. doi: 10.1037/ rev0000150.supp

Stephan, S., & Waldmann, M. R. (2022). The role of mechanism knowledge in singular causation judgments. *Cognition, 218*, 104924. Advance online publication. doi: 10.1016/j.cognition.2021.104924

White, P. A. (1995). Use of prior beliefs in the assignment of causal roles: causal powers versus regularity-based accounts. *Memory and Cognition, 23*, 243–254. doi: 10.3758/BF03197225

White, P. A. (2005). The power PC theory and causal powers: Comment on Cheng (1997) and Novick and Cheng (2004). *Psychological Review, 112*, 675–684. doi: 10.1037/0033-295X.112.3.675

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General, 136*, 82–111. doi: 10.1037/0096-3445.136.1.82

Wolff, P. (2014). Causal pluralism and force dynamics. In: B. Copley, F. Martin, & N. Duffield (Eds.), *Forces in grammatical structures: Causation between linguistics and philosophy.* Oxford Scholarship Online. doi: 10.1093/acprof:oso/9780199672073.003.0005

Wolff, P., & Song, G. (2003). Models of causation and the semantics of causal verbs. *Cognitive Psychology, 47*, 276–332. doi: 10.1016/S0010-0285(03)00036-7

Yin, P., & Sun, J. (2021). Is causation deterministic or probabilistic? A critique of Frosch and Johnson-Laird (2011). *Journal of Cognitive Psychology*, *33,* 899–918. doi: 10.1080/20445911.2021.1963265

Yin, P. (under review). A comment on arguments of mental model theory of causation. Journal of Cognitive Psychology.

RECEIVED 30.10.2021 | ACCEPTED 18.02.2022