

Causal reasoning with mental models

1

2

3 **Sangeet S. Khemlani^{1*}, Aron K. Barbey², P.N. Johnson-Laird^{3,4}**4 ¹Navy Center for Applied Research in Artificial Intelligence, Naval Research Laboratory, Washington, DC, USA5 ²Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL, USA6 ³Department of Psychology, Princeton University, Princeton, NJ, USA7 ⁴Department of Psychology, New York University, New York, NY, USA

8 * **Correspondence:** Sangeet S. Khemlani, Navy Center for Applied Research in Artificial Intelligence, Naval Research
9 Laboratory, Washington, DC, 20375, USA; Aron K. Barbey, Beckman Institute for Advanced Science and Technology,
10 University of Illinois at Urbana-Champaign, Urbana, IL, 61801, USA; P. N. Johnson-Laird, Department of Psychology, New
11 York University, New York, NY, 10003, USA
12 skhemlani@gmail.com; barbey@illinois.edu; phil @princeton.edu

13 **Keywords:** causal reasoning, mental models, explanations, enabling conditions, lateral prefrontal cortex.

14

15 Abstract

16 This paper outlines the model-based theory of causal reasoning. It postulates that the core meanings
17 of causal assertions are deterministic and refer to temporally-ordered sets of possibilities: *A causes B*
18 *to occur* means that given *A*, *B* occurs, whereas *A enables B to occur* means that given *A*, it is
19 possible for *B* to occur. The paper shows how mental models represent such assertions, and how
20 these models underlie deductive, inductive, and abductive reasoning yielding explanations. It reviews
21 evidence both to corroborate the theory and to account for phenomena sometimes taken to be
22 incompatible with it. Finally, it reviews neuroscience evidence indicating that mental models for
23 causal inference are implemented within lateral prefrontal cortex.

24 1. Introduction

25 All reasonings concerning matter of fact seem to
26 be founded on the relation of Cause and Effect.
27 David Hume (1748/1988)

28 In *An Enemy of the People*, the protagonist, Dr. Stockmann, discovers that waste runoff from the
29 town tanneries is contaminating the water supply at the public baths, a municipal project that he
30 himself has led with his brother, the mayor. He exclaims:

31 “The whole Bath establishment is a whited, poisoned sepulchre, I tell you—the gravest
32 possible danger to the public health! All the nastiness up at Molledal, all that stinking
33 filth, is infecting the water in the conduit-pipes leading to the reservoir; and the same
34 cursed, filthy poison oozes out on the shore too...” (Act I, *An Enemy of the People*)

35 Dr. Stockmann acts on his conviction by alerting the mayor to the threat of contamination – and
36 suffers as a result. His actions are based on his causal beliefs:

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE 08 AUG 2014		2. REPORT TYPE		3. DATES COVERED 00-00-2014 to 00-00-2014	
4. TITLE AND SUBTITLE Causal Reasoning with Mental Models				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Research Laboratory ,Navy Center for Applied Research in AI,Washington,DC,20375				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution unlimited					
13. SUPPLEMENTARY NOTES Frontiers in Human Neuroscience, vol. 8, pp. 1-15, 2014.					
14. ABSTRACT This paper outlines the model-based theory of causal reasoning. It postulates that the core meanings 17 of causal assertions are deterministic and refer to temporally-ordered sets of possibilities: A causes B 18 to occur means that given A, B occurs, whereas A enables B to occur means that given A, it is 19 possible for B to occur. The paper shows how mental models represent such assertions, and how 20 these models underlie deductive, inductive, and abductive reasoning yielding explanations. It reviews 21 evidence both to corroborate the theory and to account for phenomena sometimes taken to be 22 incompatible with it. Finally, it reviews neuroscience evidence indicating that mental models for 23 causal inference are implemented within lateral prefrontal cortex.					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a REPORT unclassified	b ABSTRACT unclassified	c THIS PAGE unclassified			

- 37 • The waste from the tanneries causes contamination in the baths.
 38 • The townspeople are going to allow tourists at the baths to be at risk.
 39 • It is necessary to try to prevent further contamination.

40 Ibsen's play examines how these beliefs and Stockmann's consequent actions lead him to become a
 41 pariah – an enemy of the people – much as Ibsen perceived himself to be, as a result of his revealing
 42 depictions of Norwegian society.

43 Our research is more prosaic: it examines how individuals interpret and represent causal relations,
 44 how they reason from them and use them in explanations, and how these mechanisms are
 45 implemented in the brain. This paper brings together these various parts in order to present a unified
 46 theory of causal reasoning in which mental models play a central role. The theory of mental models
 47 – the “model theory”, for short – ranges over various sorts of reasoning – deductive, inductive, and
 48 abductive, and it applies to causal reasoning and to the creation of causal explanations.

49 The organization of the paper is straightforward. It begins with a defense of a deterministic theory
 50 of the meaning of causal assertions. It explains how mental models represent the meanings of causal
 51 assertions. It shows how the model theory provides a framework for an account of causal reasoning
 52 at three levels of analysis (Marr, 1982): what the mind computes, how it carries out these
 53 computations, and how the relevant mechanisms are realized in the brain, that is, the functional
 54 neuroanatomy of the brain mechanisms underlying causal reasoning.

55

56 2. The meaning of causal relations

57 One billiard ball strikes another, which moves off at speed. If the timing is right, we see a causal
 58 relation even when the billiard balls are mere simulacra (Michotte, 1946/1963). Many causal
 59 relations, however, cannot be perceived, and so the nature of causation is puzzling. Indo-European
 60 languages, such as English, contain many verbs that embody causation. They are highly prevalent
 61 because, as Miller and Johnson-Laird (1976) argued, causation is an operator that, like time, space,
 62 and intention, occurs in verbs across all semantic domains. Each of the verbs in the following
 63 sentences, for example, embodies the notion of cause and effect:

- 64 The wind pushed the fence down (caused it to fall down).
 65 His memory of his behavior embarrassed him (caused him to feel embarrassed).
 66 She showed the ring to her friends (caused it to be visible to them).

67 Scholars in many disciplines have studied causation, but they disagree about its philosophical
 68 foundations, about its meaning, and about causal reasoning. For Hume (1748/1888), causation was an
 69 observed regularity between the occurrence of the cause and the occurrence of the effect. As he
 70 wrote (p. 115): “We may define a cause to be an object followed by another, and where all the
 71 objects, similar to the first, are followed by objects similar to the second.” For Kant (1781/1934),
 72 however, a necessary connection held between cause and effect, and he took this component to be a
 73 part of an innate conception of causality. What is common to both views is that causal relations are,
 74 not probabilistic, but deterministic, and the same claim is echoed in Mill (1874). Our chief concern
 75 rests not in philosophical controversies, but rather the everyday psychological understanding of
 76 causal assertions, and reasoning from them. The psychological literature is divided on whether the

77 meanings of causal assertions are deterministic or probabilistic. Our aim is to decide between the
78 two accounts.

79 **2.1. Do causes concern possibilities or probabilities?**

80 For many proponents of a deterministic psychological conception of causality, causal claims
81 concern what is possible, and what is impossible (Goldvarg & Johnson-Laird, 2001; Frosch &
82 Johnson-Laird, 2011). The assertion:

83 Runoff causes contamination to occur.

84 means that runoff suffices for contamination to occur, though it may occur for other reasons; and the
85 relation is false in case there is runoff without contamination. Hence, the claim can be paraphrased in
86 a conditional assertion that would be false in case its antecedent is true and its consequent is false:

87 If runoff occurs then contamination occurs.

88 A categorical assertion such as:

89 Runoff caused contamination to occur.

90 can also be paraphrased in a conditional, but one that is counterfactual:

91 If runoff hadn't occurred then contamination wouldn't have occurred.

92 The conditional refers to the case in which neither the cause nor its effect occurred. At one time this
93 state was a future possibility, but after the fact it is a possibility that did not occur – it is
94 counterfactual possibility (Johnson-Laird & Byrne, 2002; Byrne, 2005). A more plausible and
95 weaker claim is expressed in a counterfactual conditional allowing that the contamination might have
96 occurred for other reasons:

97 If runoff hadn't occurred then there mightn't have been contamination.

98 Not all conditionals express causal relations, so we can ask what else is at stake. One prerequisite is
99 that causes precede their effects, or at least do not occur after them. The two states might be
100 simultaneous in the case of a billiard ball causing a dent in the cushion that it rests on. But, physical
101 contact is not part of the core meaning of a causal relation (cf. Michotte, 1946/1963; Geminiani,
102 Carassa, and Bara, 1996), because causal assertions can violate it, as in: The moon causes tides.
103 Claims about action at a distance may be false, but their falsity is not merely because they are
104 inconsistent with the meaning of *A causes B*. Likewise, contiguity seems irrelevant to causal
105 assertions about psychological matters, such as: His memory of his behavior embarrassed him.

106 Many factors – the existence of known mechanisms, causal powers, forces, structures – can be
107 important in inferring a cause (e.g., Ahn & Bailenson, 1996; Koslowski, 1996; White, 1995), and
108 they can be incorporated into the interpretation of a causal assertion or its conditional paraphrase (see
109 Johnson-Laird & Byrne, 1991, for an account of this process, which they refer to as modulation).
110 None of them, however, is part of the core meaning of *A causes B*. Consider mechanistic accounts of
111 causal systems, e.g., how sewing machines work (Miyake, 1986). Experts who use sewing machines

112 can explain their underlying components. However, there comes a point in any such explanation,
113 when everyone must make an assertion equivalent to:

114 A causes B, and that's that.

115 This cause has no support. Mechanisms cannot go all the way down – no more than the turtles
116 supporting the earth in primitive cosmology can go all the way down. Hence, mechanisms and their
117 cognates, such as forces and powers, cannot be part of the core meaning of causal assertions.

118 Granted that causal assertions and their corresponding conditionals concern possibilities, their
119 meaning is deterministic rather than probabilistic. However, some twentieth century theorists, from
120 Russell (1912-13) to Salsburg (2001, p. 185-6), denied the existence of a coherent notion of
121 causation. Russell was influenced by quantum mechanics, and argued that causation should be
122 replaced by probabilistic considerations. One reason for such skepticism is a failure to divorce
123 beliefs from meanings. Beliefs about causation are often incoherent. For example, some people
124 believe that *it is possible to initiate a causal chain*, and that *every event has a cause*. Both beliefs
125 can't be right, because if every event has a cause, an action to initiate a causal chain has itself a
126 cause, and so it doesn't really initiate the chain. Such beliefs, however, should not be confused with
127 the core meaning of causes, which does not legislate about them: we understand both the preceding
128 assertions that yield the inconsistency. Neither of them seems internally inconsistent.

129 Other theorists, also inspired by quantum mechanics, have maintained causation but rejected
130 determinism (e.g., Reichenbach, 1956; Salmon, 1980; Suppes, 1970). A cause and its effect are
131 related probabilistically. Reichenbach (1956) argued that a causal assertion, such as:

132 Runoff causes contamination to occur

133 means that contamination is more probable given that runoff occurs than given that it does not occur.
134 Hence, a causal claim holds provided that the following relation holds between the two conditional
135 probabilities:

136 $P(\text{contamination} \mid \text{runoff}) > P(\text{contamination} \mid \text{no runoff})$

137 The philosophical controversy between determinism and probabilism has spilled over into
138 psychology. Some psychological theories are probabilistic both for causation (e.g., Cheng, 1997,
139 2000) and for conditionals (Oaksford & Chater, 2007). The case for probabilistic meanings rests in
140 part on causal assertions such as:

141 Cars cause accidents.

142 Such assertions tolerate exceptions, which do not refute them, and which therefore imply a
143 probabilistic relation. But, it is the form of the generalization rather than its causal content that
144 enables it to tolerate exceptions. It is a generic assertion akin to:

145 Cars have radios.

146 A generic assertion is defined as a generalization with a subject, such as a noun phrase or a gerund,
147 lacking an explicit quantifier (Leslie, 2008). Certain sorts of generic, e.g., *snow storms close schools*,

148 imply a causal connection between their subject, snow storms, and their predicate, close schools. The
 149 meaning of the verb, “close,” is causal, and individuals readily infer that snow storms cause an agent
 150 to act to close schools (see Prasada, Khemlani, Leslie, & Glucksberg, 2013). Hence, generics tolerate
 151 exceptions. In contrast, if the subjects of assertions contain explicit quantifiers as in:

152 Some snow storms cause schools to close.

153 and:

154 All snow storms cause schools to close.

155 then the assertions have a deterministic meaning, and the first of these assertions is true as a matter of
 156 fact and the second of them is false.

157 **2.2. Evidence against probabilistic accounts of causation**

158 Several arguments count against probabilistic meanings for everyday causal assertions. A major
 159 historical problem is to explain why no one proposed such an analysis prior to the formulation of
 160 quantum mechanics. Moreover, a singular claim about causation, such as:

161 The runoff caused contamination to occur

162 is false if the runoff occurred without contamination. This factual relation is deterministic, and to
 163 introduce probabilities into the interpretation of counterfactual conditionals is problematic.

164 Individuals, as we show later, recognize the difference in meaning between causes and enabling
 165 conditions, such as, *The runoff allowed contamination to occur*. But, both increase the conditional
 166 probability of an effect given the antecedent, and so the difference in meanings between causes and
 167 enabling conditions is impossible to make in probabilistic accounts (Wolff, 2007; pace Cheng, 2000;
 168 Cheng & Novick, 1991). The same problem arises in implementing causation in Bayesian networks
 169 (Glymour, 2001; Gopnik et al., 2004; Pearl, 2000; Tenenbaum & Griffiths, 2001).

170 Reasoners often infer a causal relation from a single observation (e.g., Ahn & Kalish, 2000;
 171 Schlottman & Shanks, 1992; Sloman, 2005; White, 1999). But, if causal assertions are probabilistic,
 172 no single observation could suffice to establish cause and effect, because probabilistic interpretations
 173 tolerate exceptions. Lien and Cheng (2000) proposed instead that single observations can refer to
 174 previously established causal relations. Repeated observations of billiard balls, for example, establish
 175 causal relations about their collisions, which individuals can then use to infer a causal relation from a
 176 single new observation. However, as Fair (1979) anticipated, this proposal implies that individuals
 177 could never establish causal relations contrary to their expectations.

178 Interventions that initiate a causal chain are a feature of Bayesian networks (see, e.g., Pearl, 2000;
 179 Woodward, 2003), and evidence corroborates their psychological importance (Sloman, 2005; Sloman
 180 & Lagnado, 2005). As an example, suppose that the following claim is true:

181 Overeating causes indigestion.

182 If we then observe that Max doesn't have indigestion, we can infer that he hasn't overeaten. But,

183 Max could have intervened to prevent indigestion: he could have taken an anti-indigestion pill. In
 184 this case, we would no longer make the inference. No special logic or probabilistic considerations are
 185 needed to handle these cases (pace Sloman, 2005). Our initial claim is an idealization expressed in a
 186 generic, and so it tolerates exceptions.

187 In summary, the evidence seems to be decisive: causal relations in everyday life have
 188 deterministic meanings unless they make explicit reference to probabilities, as in:

189 Keeping to this diet probably causes you to lose weight.

190 Moreover, if causation were intrinsically probabilistic, there would be no need for the qualification in
 191 this example. Its effect is to weaken the causal claim. Studies of inferences from causal assertions,
 192 which we describe below, further bolster their deterministic meanings.

193

194 3. Mental models of causal assertions

195 We now turn to the model theory of mental representations, which we outline before we consider
 196 its application to reasoning. The theory goes back to Craik (1943) and has still earlier antecedents in
 197 philosophy. Its more recent development gives a general account of how individuals understand
 198 assertions, how they represent them, and how they reason from them (see, e.g., Johnson-Laird, 1983;
 199 Johnson-Laird & Byrne, 1991; Johnson-Laird & Khemlani, 2014). The theory has been implemented
 200 computationally, its predictions have been corroborated in psychological experiments and in recent
 201 neuroimaging results (e.g., Kroger, et al., 2008). And it is of sufficient maturity that given the
 202 semantics of a domain such as causation, it calls for few new assumptions in order to account for
 203 representation and reasoning.

204 The first step in understanding an assertion is to parse it in order to construct a representation of its
 205 meaning. The theory postulates that the parser's output (an intensional representation) is composed
 206 out of the meanings of its parts according to the grammatical relations amongst them. The
 207 intensional representation is used to construct, to update, to manipulate, or to interrogate, mental
 208 models of the situation under description (an extensional representation). The theory rests on three
 209 fundamental principles:

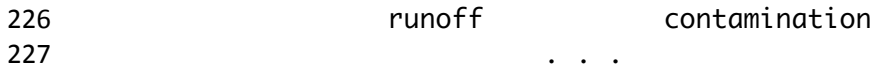
- 210 1. Mental models represent *possibilities*: each model captures a distinct set of possibilities to
 211 which the current description refers.
- 212 2. Mental models are *iconic*: the structure of a model corresponds to the structure of what it
 213 represents (see Peirce, 1931-1958, Vol. 4). Hence, kinematic models unfold in time to
 214 represent a temporal sequence of events (Khemlani, Mackiewicz, Bucciarelli, & Johnson-
 215 Laird, 2013). However, models can also include certain abstract symbols, such as one for
 216 negation (Khemlani, Orenes, & Johnson-Laird, 2012).
- 217 3. The principle of truth: Mental models represent only what is true, not what is false, in each
 218 possibility. They yield rapid intuitions. In contrast, *fully explicit* models represent what is
 219 false too, but their construction calls for deliberation and access to working memory.

220 The model theory implements the deterministic meanings of causal relations described in the

221 previous section. An assertion such as:

222 Runoff causes contamination to occur

223 has two mental models, one is an explicit model representing the case in which the cause and its
 224 effect both occur, and the other is an implicit mental model representing at least one other possibility
 225 in which the cause does not occur:



228 The rows in this schematic diagram represent two distinct possibilities. In fact, mental models do not
 229 consist of words and phrases, which we use for convenience, but of representations of the objects and
 230 events to which the words refer. The ellipsis denotes the other possibilities in which the cause does
 231 not occur. These possibilities are not immediately accessible, i.e., one has to work them out. We
 232 have omitted from the diagram the temporal relation between cause and effect: the cause cannot
 233 come after the effect, and by default comes before it.

234 The model theory draws a distinction in meaning between causes and enabling conditions
 235 (contrary to a tradition going back to Mill, 1874). An enabling condition makes its effect possible: it
 236 allows it to happen. The assertion:

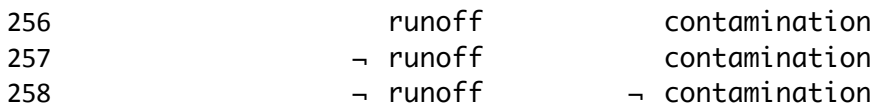
237 Runoff allows contamination to occur.

238 has a core meaning that is a tautology in which all things are possible provided they are in the correct
 239 temporal sequence. Like its corresponding conditional:

240 If runoff occurs then contamination may occur.

241 it is possible for runoff to occur, or not to occur, and in either case, with or without contamination.
 242 Such assertions are nearly vacuous, and so an obvious implication – an implicature from Grice’s
 243 (1975) conversational conventions – is that only runoff allows contamination to occur. There are
 244 then just three possibilities: with runoff, contamination does or does not occur; but without it, runoff
 245 does not occur. The mental models of an enabling assertion are identical to those of a causal
 246 assertion. One mental model represents the possibility in which both runoff and contamination occur,
 247 and the implicit model represents the other possibilities. A consequence of this identity is that people
 248 have difficulty in grasping that causal and enabling assertions differ in meaning. This difficulty has
 249 infected the legal systems of both the US and the UK, which make no distinction between the two
 250 sorts of causal relation (Johnson-Laird, 1999), though people judge those who cause harmful
 251 outcomes as more culpable than those who enable them (Frosch, Johnson-Laird, & Cowley, 2007).

252 When reasoners have to enumerate the possibilities consistent with an assertion, they are able to
 253 deliberate and to flesh out their mental models into fully explicit models. The difference between
 254 causing and enabling now becomes evident. The fully explicit models of the causal assertion, *runoff*
 255 *causes contamination to occur*, are:



259 where “¬” is a symbol corresponding to a mental token for negation (Khemlani, Orenes, & Johnson-
 260 Laird, 2012). What the assertion rules out is the possibility that runoff occurs without contamination.
 261 In contrast, the fully explicit models of the enabling assertion, *runoff allows contamination to occur*,
 262 and its implicature are:

263	runoff	contamination
264	runoff	¬ contamination
265	¬ runoff	¬ contamination

266 Some causal claims are stronger than the one above: they assert that the cause is the only way to
 267 bring about the effect. The only way to get cholera, for example, is to be infected by the bacterium
 268 *Vibrio cholerae*. The corresponding assertion has only two fully explicit models, one in which the
 269 cause and effect both occur – the bacterium and the infection, and one in which neither of them
 270 occurs. There are also weaker enabling assertions than the one above, that is, ones in which all
 271 appropriately temporally-ordered possibilities occur, including the possibility that the effect occurs in
 272 the absence of the enabling condition, i.e., the implicature does not occur.

273 When individuals have to list what is possible, and what is impossible, given each of the main
 274 sorts of causal relation, their listings tend to corroborate the model theory (Goldvarg & Johnson-
 275 Laird, 2001). Participants list either the three possibilities for *causes* or the two for its stronger
 276 interpretation. They are more confused by *enables*, but list the three possibilities above more often
 277 than chance, and likewise the four possibilities for its weaker interpretation. They list the three
 278 possibilities and the two possibilities for *A prevents B from occurring*, which is synonymous with *A*
 279 *causes B not to occur*.

280 One attempt to distinguish between causing and enabling in a probabilistic framework is to argue
 281 that an enabling condition is constant in the situation, whereas a cause is not (Cheng & Novick,
 282 1991). This difference does occur, but it is not essential according to the model theory. A crucial test
 283 used scenarios in which neither the causes nor the enabling conditions were constant (Goldvarg &
 284 Johnson-Laird, 2001). Readers may like to try to identify the cause and the enabler in each of the
 285 following scenarios:

286 Given that there is good sunlight, if a certain new fertilizer is used on poor flowers,
 287 then they grow remarkably well. However, if there is not good sunlight, poor
 288 flowers do not grow well even if the fertilizer is used on them.

289 and:

290 Given the use of a certain new fertilizer on poor flowers, if there is good sunlight
 291 then the flowers grow remarkably well. However, if the new fertilizer is not used on
 292 poor flowers, they do not grow well even if there is good sunlight.

293 In the first scenario, sunlight is the enabling condition, and the fertilizer is the cause; in the second
 294 scenario, the two swap roles. These roles derive from the possibilities to which the respective
 295 scenarios refer. In the first scenario, the possibilities are as follows:

296	sunlight:	fertilizer	growth
297		¬ fertilizer	growth
298		¬ fertilizer	¬ growth
299	¬ sunlight:		¬ growth

300 As they show, sunlight enables the fertilizer to cause the flowers to grow. Their roles swap in the
 301 possibilities for the second scenario. In an experiment, the participants were told that a cause brings
 302 about an event whereas an enabling condition makes it possible, and that they had to identify the
 303 cause and the enabling condition in sets of scenarios. The order of mention of the cause and enabler
 304 was counterbalanced over the scenarios, and each participant saw only one of the four versions of
 305 each content. The twenty participants made correct identifications on 85% of the trials, and each of
 306 them was right more often than not (Goldvarg & Johnson-Laird, 2001).

307 These phenomena account against rival accounts of the difference between causes and enabling
 308 conditions. The distinction between them is neither capricious nor unsystematic (Mill, 1874;
 309 Kuhnmünch & Beller, 2005). It is contrary to the claim that a cause violates a norm assumed by
 310 default whereas an enabling condition does not (Einhorn and Hogarth, 1986; Kahneman and Miller,
 311 1986). And the cause need not be conversationally relevant in explanations (Hilton and Erb, 1996;
 312 Mackie, 1980; Turnbull and Slugoski, 1988). In sum, the difference in meaning between the two
 313 principal sorts of causal assertion is real (see also Sloman, Barbey, & Hotaling, 2009; and Wolff &
 314 Song, 2003).

315

316 4. Models and causal deductions

317 How do naïve individuals make causal deductions? One answer is that they rely on the laws of
 318 thought, that is, on formal rules of inference akin to those of logic. Indeed, Rips (1994, p. 336) has
 319 proposed that formal rules could be extended to deal with causal reasoning. Pure logic makes no
 320 reference to specific contents, and so its application to causation depends on the introduction of
 321 axioms (or “meaning postulates”), such as:

322 If A causes B, and B prevents C, then A prevents C

323 where A, B, and C, are variables that take states or events as their values (von Wright, 1973). Logic,
 324 however, has several critical problems in coping with everyday reasoning. One is that infinitely
 325 many conclusions follow in logic from any set of premises, and most of them are trivial or silly, such
 326 as conjunction of a premise with itself. Another problem is that logic means never having to
 327 withdraw the conclusion of a valid inference, even if its conclusion turns out to be false. In jargon,
 328 logic is *monotonic* – as you accrue more premises, so you are able to draw more conclusions and
 329 never have a warrant for withdrawing any of them. In contrast, everyday reasoning is *nonmonotonic*.
 330 You withdraw a conclusion if the facts show it to be wrong.

331 Another theory is that causal inferences depend on *pragmatic reasoning schemas* (e.g. Cheng,
 332 Holyoak, Nisbett, and Oliver, 1986). In other words, the axiom above is framed instead as a rule of
 333 inference:

334 A causes B.
 335 B prevents C.
 336 Therefore, A prevents C.

337 This idea goes back to Kelley’s (1973) theory of causal attribution, which postulates such schemas
 338 for checking causal relations. Similarly, Morris and Nisbett (1993) proposed a schema including the
 339 following two rules:

340 If cause C is present then effect E occurs.
 341 Cause C is present.
 342 Therefore, Effect E occurs.

343 and:

344 If cause C is present then effect E occurs.
 345 Effect E does not occur.
 346 Therefore, Cause C is not present.

347 In contrast, the model theory makes no use of formal rules of inference, and no use of axioms,
 348 meaning postulates, or schemas concerning causation. It simply applies its general principles of
 349 reasoning to mental models of causal assertions.

350 Theorists distinguish among three main sorts of reasoning: deduction, induction, and abduction,
 351 which creates hypotheses or explanations. We shall do so too, but with the caveat that human
 352 reasoners make inferences without normally concerning themselves about such niceties. To make
 353 deductions, individuals draw conclusions that hold in all their models of the premises. To make
 354 inductions, they use their knowledge to build models going beyond the information given in the
 355 premises, and then infer corresponding conclusions, such as generalizations (Johnson-Laird, 2006).
 356 To make abductions, they use their knowledge to incorporate new concepts – those not in the
 357 premises – in order to yield causal explanations of everyday events (Johnson-Laird, et al., 2004).
 358 We will describe the model theory for each of these three sorts of reasoning, starting with deduction
 359 here, and we will show that the evidence corroborates its account rather than the alternatives.

360 At the computational level, the model theory postulates three constraints on everyday reasoning
 361 (Johnson-Laird & Byrne, 1991, Ch. 2). First, inferences do not throw away semantic information (see
 362 Bar-Hillel & Carnap, 1953). That is, people do not spontaneously make inferences, such as:

363 Runoff causes contamination.
 364 Therefore, runoff causes contamination or inoculations prevent disease, or both.

365 The inference is *valid*, because its conclusion must be true if its premise is true. But, its conclusion is
 366 less informative (e.g., by a measure of semantic information) than its premise, because the former is
 367 compatible with more possibilities than the latter. In contrast, induction and abduction increase
 368 semantic information. Second, inferences are parsimonious. For example, a conclusion does not
 369 merely consist of a conjunction of all the premises, even though such a conclusion is valid and
 370 maintains semantic information. Third, a conclusion should assert something new, and not repeat
 371 what is explicit in the premises. If no conclusion meets these three constraints, then individuals
 372 respond that nothing follows from the premises – a response that violates logic, but that is perfectly
 373 rational. Consider this inference, for instance:

374 Runoff causes contamination to occur.
 375 Three is a prime number.
 376 What follows?

377 A logician should respond: infinitely many conclusions, including a conjunction of the first premise
 378 with itself 101 times. A more sensible response is: nothing. In short, human reasoners aim not to lose
 379 information, to simplify where possible, and to infer something new whether they are making

380 deductive, inductive, or abductive inferences.

381 The model theory copes with the main sorts of non-monotonicity. It allows for information to be
 382 assumed by default, and to be overruled by subsequent information, as when individuals infer that a
 383 dog has four legs only to discover that a particular pet is three-legged. It also allows for deductions to
 384 be made in an experimental mode ignorant of the facts of the matter, so that when a conclusion turns
 385 out to be false, it can be withdrawn without cost. We illustrate such cases in the section below on
 386 explanations. It also diverges slightly from logic in its basic assumption about validity. In logic, a
 387 valid deduction is one that holds in every case in which the premises hold (Jeffrey, 1981, p. 1).
 388 Hence, any conclusion whatsoever follows from inconsistent premises, because there is no case in
 389 which the premises hold. The model theory adds a rider for everyday reasoning: there is at least one
 390 non-null model in which the premises hold. This proviso blocks valid inferences from inconsistent
 391 premises.

392 At the algorithmic level, the theory postulates that individuals build mental models of premises –
 393 they simulate the world under description. They use the information in the premises, their general
 394 knowledge, and their knowledge of the context. The system searches for a conclusion that holds in
 395 the models and that doesn't merely echo an explicit premise – a principle that holds for conversation
 396 in general (Grice, 1975). But, the system can also evaluate given conclusions. A conclusion that
 397 holds in all the models of the premises follows of necessity, but if there is a model of the premises in
 398 which it does not hold – a counterexample – it does not follow of necessity. Yet, if it holds in most
 399 models, it is probable. And if it holds in at least one model, it is possible. Because inferences are
 400 based on models of the premises, the resulting conclusions cannot throw semantic information away
 401 by adding disjunctive alternatives, or consist of a premise conjoined with itself,

402 Mental models can be three-dimensional in order to represent spatial relations, and they can be
 403 kinematic, unfolding in time to represent a temporal sequence of events (Johnson-Laird, 1983).
 404 Evidence supports these hypotheses in the use of mental simulations to deduce the consequences of
 405 informal algorithms (Khemlani, et al., 2013). Temporal order, however, can also be represented by
 406 an axis in a static model.

407 The “force dynamics” theory of causal reasoning (Barbey & Wolff, 2007; Wolff, 2007) makes
 408 analogous claims. It assumes that individuals envisage interacting entities in iconic models in which
 409 vectors represent the directions and magnitudes of forces. The theory explains the interpretations of
 410 such assertions as:

411 Pressure will cause the water to remain below 0°C.
 412 Small ridges cause water to stand on the concrete.
 413 The pole will prevent the tent from collapsing.

414 Each assertion refers to a configuration of forces. The third assertion, for instance, refers to a
 415 configuration in which the pole acts against the tendency of the tent to collapse. These tendencies are
 416 represented in a vector model. We simplify the diagrams illustrating these models: arrows denote
 417 vectors corresponding to the direction and magnitude of forces, and the box denotes the point of
 418 stasis, which is the origin of all vectors. The tendency of the tent to collapse is diagramed here,
 419 where the two overlaid vectors represent the tent (one vector) heading towards collapse (another
 420 vector):

421 $\square \text{---} \rightarrow \text{-----} \rightarrow \text{collapse}$

422 tent

423 The pole provides a countervailing force, and so its vector is in the opposite direction:

424 <-----□
425 pole

426 Because the magnitude of the pole's vector is larger than the magnitude of the tent's vector, the
427 combination of the two yields a small magnitude in the direction away from collapse:

428 <----□
429 pole+tent

430 So, the diagram representing all the interacting vectors is as follows:

431 pole+tent
432 <-----<-----□-----> collapse
433 pole tent

434 Such diagrams represent a relation in which *A* prevents *B*. Hence, the force theory, like the model
435 theory, postulates that reasoners build up a mental model of causal relations, which can then be
436 scanned to yield inferences. The model theory has not hitherto been formulated to represent forces or
437 the interactions amongst them, and so the force theory contributes an important and hitherto missing
438 component. The resulting models can also underlie kinematic mental simulations of sequences of
439 events.

440 The model theory can represent probabilities. It uses proportions of models to draw conclusions
441 about *most* entities or *few* of them. These proportions are used to make inferences about
442 probabilities. Individual models can also be tagged with numerals to represent their relative
443 frequencies or probabilities. This algorithmic account unifies deductive and probabilistic reasoning,
444 and it is implemented in a computer program, *mReasoner*, which we continue to develop, and its
445 source code is available at: <http://mentalmodels.princeton.edu/models/mreasoner/>.

446 In broad terms, three strands of evidence corroborate the model theory of causal deductions. The
447 first strand of evidence bears out the difference in the possibilities referred to in assertions about
448 causes and assertions about enabling conditions. Readers might like to consider what response they
449 would make to this problem:

450 Eating protein will cause her to gain weight.
451 She will eat protein.
452 Will she gain weight?
453 Yes, No, and Perhaps yes, perhaps no.

454 Most participants in an experiment (Goldvarg & Johnson-Laird, 2001) responded: yes. But, when
455 the first premise was instead:

456 Eating protein will allow her to gain weight

457 as its fully explicit models predict, the majority rejected the “yes” response. The opposite pattern of
458 results occurred when the second assertion and question were changed to:

459 She will not gain weight.
460 Will she not eat protein?

461 The results therefore bear out the difference in meaning between causing and enabling.

462 The second strand of evidence supports the deterministic interpretation of causal assertions
463 embodied in the model theory. It rests on the fact that reasoners grasp the force of a counterexample.
464 When they evaluate given inferences, they tend to justify their rejection of an invalid inference by
465 citing a counterexample to its conclusion (Johnson-Laird & Hasson, 2003). Likewise, consider an
466 assertion, such as:

467 Following this diet causes a person with this sort of metabolism to lose weight.

468 Participants in experiments were asked about what evidence would refute such claims and similar
469 ones about enabling conditions (Frosch & Johnson-Laird, 2011). In several experiments, every
470 single participant chose a single observation to refute the assertions more often than not, but as the
471 model theory predicts they were more likely to do so for causal assertions than enabling assertions.
472 For both sorts of relation, they chose refutations of the form *A and not-B*, e.g.:

473 A person with this sort of metabolism followed this diet and yet did not lose
474 weight.

475 But, as the theory predicts, they chose refutations of the form *not-A and B*, e.g.:

476 A person with this sort of metabolism did not follow this diet and yet lost weight
477 more often to refute enabling assertions than causes.

478 The third strand of evidence concerns the principle of truth and the difference between mental
479 models and fully explicit models. Most of us usually rely on our intuitions, and they are based on a
480 single mental model, which represents only what is true in the corresponding possibility. The
481 following problem illustrates one consequence of this bias:

482 One of these assertions is true and one of them is false:
483 Marrying Evelyn will cause Vivien to relax.
484 Not marrying Evelyn will cause Vivien to relax.
485 The following assertion is definitely true:
486 Vivien will marry Evelyn.
487 Will Vivien relax? Yes/No/It's impossible to know.

488 The initial rubric is equivalent to an exclusive disjunction between the two causal assertions. It
489 yields the following two mental models:

490 Vivien: marries Evelyn relaxes
491 ¬marries Evelyn relaxes

492 The final categorical assertion eliminates the second possibility, and so most reasoners infer that
493 Vivien will relax. It seems plausible, but the intuition is wrong. The fully explicit models of the
494 disjunction of the two assertions yield only two possibilities, one in which the first assertion is true
495 and the second assertion is false, and one in which the first assertion is false and the second assertion

496 is true. But, in the first case, the second assertion is false, and so Vivien doesn't marry Evelyn and
 497 doesn't relax; and, in the second case, the first assertion is false and so Vivien marries Evelyn but
 498 doesn't relax. So, the fully explicit and correct models are respectively:

499 Vivien: ¬ marries Evelyn ¬ relaxes
 500 marries Evelyn ¬ relaxes

501 The final categorical assertion eliminates the first of them, and it follows that Vivien will not relax.
 502 None of the participants in an experiment drew this correct conclusion. The majority inferred that
 503 Vivien will relax, and the remainder inferred that it was impossible to know (Goldvarg & Johnson-
 504 Laird, 2001).

505 The model theory makes predictions about causal reasoning that have yet to be tested, though they
 506 have been corroborated in other domains. The most important of these predictions are that the more
 507 models that have to be taken into account, the more difficult an inference should be, and that a
 508 common source of error should be to overlook the model of a possibility. Yet, the evidence we have
 509 described here illustrates the case for the model theory, and the alternative theories that we reviewed
 510 at the start of this section offer no account of it.

511

512 **5. The induction of causal relations**

513 The vessel, *The Herald of Free Enterprise*, was a roll-on roll-off car ferry. Its bow doors were
 514 opened in the harbor to allow cars to drive into the ship, and at its destination, the cars drove off the
 515 ship the same way. When it sailed from Zeebrugge in Belgium on March 6th, 1987, the master made
 516 the plausible induction about a causal relation, namely, that the assistant bosun had closed the bow
 517 doors. The chief officer made the same inference, and so did the bosun. But, the assistant bosun
 518 hadn't closed the bow doors: he was asleep in his bunk. Shortly after the ferry left the calm waters of
 519 the harbor, the sea poured in through its open doors, and it capsized with the loss of nearly two
 520 hundred lives. Inductions are risky. There is no guarantee that they yield the truth, and, as this
 521 example also illustrates they can concern an individual event, not just generalizations of the sort in
 522 textbook definitions of induction.

523 The risk of inductions arises in part because they go beyond the information in the premises, such
 524 as that no-one has reported that the bow doors are open. As a result, they can eliminate possibilities
 525 that the premises imply, and they can add relations, such as a temporal order of events within a model
 526 of a situation (Johnson-Laird, 1988). In all these cases, the inductive operation depends on
 527 knowledge or beliefs. And beliefs are sometimes wrong.

528 Students of induction from Polya (1973) onwards have postulated formal rules of inference to
 529 underlie it – to parallel the formal rules of inference used in logic. These systems have grown ever
 530 more sophisticated in programs for machine learning (e.g., Michalski & Wojtusiak, 2007). The
 531 model theory, however, assumes that knowledge and beliefs can themselves be represented in
 532 models, and so the essential inductive operation is to conjoin two sets of models: one set represents
 533 the possibilities derived from the premises, which may be direct observations, and the other set is part
 534 of long-term knowledge and beliefs. A simple but common example occurs when knowledge
 535 modulates the core interpretation of causality, just as it can do in the interpretation of conditionals
 536 (Johnson-Laird & Byrne, 2002). The core meaning of *A causes B*, as we argued earlier, is consistent

537 with three possibilities. Hence, an assertion such as:

538 A deficiency of some sort causes rickets

539 refers to three possibilities in which there is a temporal order from cause to effect:

540 deficiency rickets

541 – deficiency rickets

542 – deficiency – rickets

543 Many people know, however, that rickets has a unique cause – a deficiency in vitamin D, and this
 544 knowledge blocks the construction of the second model above in which rickets arise in a person with
 545 no deficiency. Modulation in the interpretation of assertions is a bridge from deduction to induction.
 546 The resulting models allow one to infer that if a patient has no dietary deficiency, then the patient
 547 doesn't have rickets. Logicians can argue that the inference is an enthymeme, that is, it is a valid
 548 deduction granted the provision of the missing premise that no other cause for rickets exists. But,
 549 one could just as well argue that the inference is an induction, since the conclusion rests on more
 550 information than the premises provide. The reasoning system is not concerned with the correct
 551 analysis. It relies on whatever relevant knowledge is available to it.

552 Observations of contingencies can lead to inductive inferences in daily life and in science. Hence,
 553 many theories concern inductions from the frequencies of contingencies (e.g., De Houwer & Beckers,
 554 2002; Hattori & Oaksford, 2007; Perales & Shanks, 2007; Shanks, 1995). The analogy with classical
 555 conditioning is close. The analyses of frequencies can also yield inductions about causation at one
 556 level that feed into those at a higher or more abstract level in a hierarchical Bayesian network (e.g.,
 557 Gopnik et al., 2004; Griffiths & Tenenbaum, 2005; Lu et al., 2008). Once its structure is established,
 558 it can assign values to conditional probabilities that interrelate components in the network, e.g., it can
 559 yield the conditional probability of lung cancer given that coughing occurs, and the conditional
 560 probability of smoking given lung cancer (see Tenenbaum, Griffiths, & Kemp, 2006, for a review).

561 In contrast, observations can lead to inductions without probabilities. For instance, Kepler
 562 analyzed Tycho Brahe's astronomical observations, and used them to induce his three laws of
 563 planetary motion, of which the best known is his first law: a planet moves in an elliptical orbit around
 564 the sun with the sun at one focus. But, the mind prepared with knowledge can also make an
 565 induction from a single observation – a claim supported by considerable evidence (see, e.g., White,
 566 2014). One source of such inferences is knowledge of a potential mechanism, and this knowledge
 567 may take the form of a model.

568 Adults induce new concepts throughout their life. Some are learned from knowledge by
 569 acquaintance, others from knowledge by description. You cannot acquire the full concept of a color,
 570 a wine, or a sculpture without a direct acquaintance with them, but you can learn about quarks, genes,
 571 and the unconscious, from descriptions of them. Likewise, the induction of a generalization is
 572 equivalent to the induction of a concept or of a change to a concept, as in Kepler's change to the
 573 concept of a planetary orbit. Novel concepts can be put together out of existing concepts. Hence,
 574 causal inductions are part of the acquisition of concepts. Causes are more important than effects in
 575 the features of a concept. This difference explains why the constituents of natural kinds are
 576 important, whereas the functions of artifacts are important (Ahn, 1998). A genetic code is
 577 accordingly more critical to being a goat than that it gives milk, whereas that a mirror reflects an
 578 image is more important to a mirror than that it is made of glass.

579 Knowledge of a category's causal structure is important in categorization. Objects are classified as
 580 members of a category depending on how well their features fit our intuitive theory, or model, of the
 581 causal relations amongst the category's features (see, e.g., Waldmann, Holyoak, & Fratianne, 1995).
 582 Reasoners judge an exemplar as a better instance of a category when its features fit the causal
 583 structure of the category (Rehder, 2003). Figure 1 illustrates two contrasting causal structures. In the
 584 common-cause structure, one feature is a common cause of three effects, such as the symptoms of a
 585 disease, whereas in the common-effect structure, one feature is a common effect of each of three
 586 causes, such as a disease that has three independent etiologies. In Rehder's experimental study,
 587 which used sensible everyday features, the participants rated category-membership depending on an
 588 instance's features, pairs of its features, and high-order relations among its features. The results
 589 showed that the participants were indeed sensitive to the difference between the two sorts of causal
 590 structure in Figure 1.

591 At the center of the model theory is the hypothesis that the process of understanding yields a
 592 model. In deduction, if a mental model yields a conclusion, its validity can be tested in a search for
 593 alternative models. In induction, however, the construction of models increases semantic
 594 information. In the case of inductions about specific events in everyday life, this process is part of the
 595 normal effort to make sense of the world. Human reasoning relies, wherever possible, on general
 596 knowledge. Hence, when the starter won't turn over your car's engine, your immediate inference is
 597 that the battery is dead. Another role that knowledge plays is to provide interstitial causal relations
 598 that make sense of assertions hitherto lacking them – a process that is case of what Clark (1975)
 599 refers to as “bridging” inferences. We demonstrated the potency of such inferences in a series of
 600 unpublished studies. One study included a condition in which the participants were presented with
 601 sets of assertions for which in theory they could infer a causal chain, such as:

602 David put a book on the shelf.
 603 The shelf collapsed.
 604 The vase broke.

605 In another condition, the participants were presented with sets of assertions for which they could not
 606 infer a causal chain, such as:

607 Robert heard a creak in the hall closet.
 608 The faucet dripped.
 609 The lawn sprinklers started.

610 The theory predicts that individuals should infer the causal relations, and the experiment corroborated
 611 this hypothesis. When a further assertion contradicted the first assertion in a set, the consequences
 612 were quite different between the two conditions. In the first condition, the contradictory assertion:

613 David didn't put a book on the shelf

614 led to a decline in the participants' belief in all the subsequent assertions, and so only 30% of them
 615 believed that the vase broke. In the second case, the contradictory assertion:

616 Robert did not hear a creak in the hall closet

617 had no effect in the participants' belief in the subsequent assertions. All of them continued to believe
 618 that the lawn sprinklers started. This difference in the propagation of doubt is attributable to the

619 causal interpretation of the first sort of scenario, and the impossibility of such an interpretation for
 620 the second scenario. This example is close, if not identical, to an abduction, because the attribution
 621 of causes explains the sequence of events in the causal scenarios. It leads us to consider abduction in
 622 general.

623

624 6. Abduction of causal explanations

625 A fundamental aspect of human rationality is the ability to create explanations. Explanations, in
 626 turn, depend on understanding: if you don't understand something, you can't explain it. It is easier
 627 to state criteria for what counts as understanding than to define it. If you know what causes
 628 something, what results from it, how to intervene to initiate it, how to guide or to govern it, how to
 629 predict its occurrence and the course of its events, how it relates to other phenomena, what internal
 630 structure it has, how to fix it if it malfunctions, then to some degree you understand it. According to
 631 the model theory, "if you understand inflation, a mathematical proof, the way a computer works,
 632 DNA or a divorce, then you have a mental representation that serves as a model of an entity in much
 633 the same way as, say, a clock functions as a model of the earth's rotation" (Johnson-Laird, 1983, p.
 634 2). And you can use your model to formulate an explanation. Such explanations can help others to
 635 understand – to make sense of past events and to anticipate future events. Many psychological
 636 investigations have focused on explanatory reasoning in the context of specific, applied domains,
 637 such as fault diagnosis (e.g., Besnard & Bastien-Toniazzo, 1999) and medical decision-making (e.g.,
 638 Ramoni, Stefanelli, Magnani, & Barosi, 1992). But, as Hume (1748/1988) remarks in the epigraph to
 639 this paper, most reasoning about factual matters is founded on cause and effect. To illustrate the role
 640 of models in causal abductions, consider this problem:

641 If someone pulled the trigger, then the gun fired.
 642 Someone pulled the trigger, but the gun did not fire.
 643 Why not?

644 Most people presented with the problem offered a causal explanation, such as:

645 Someone unloaded the gun and so there were no bullets in it.

646 They even rated such an explanation as more probable than either the cause alone or the effect alone
 647 (Johnson-Laird et al., 2004). In daily life, explanations tend to explain only what needs to be
 648 explained (Khemlani, Sussman, & Oppenheimer, 2011), but, as the case above illustrates, causal
 649 relations take priority over parsimony (pace Lombrozo, 2007). In science, Occam's razor calls for
 650 parsimonious explanations.

651 When the preceding problem is couched in these terms:

652 If someone pulled the trigger, then the gun fired.
 653 The gun did not fire.
 654 Why not?

655 many individuals preferred a causal explanation to a simple deductive one:

656 No one pulled the trigger.

657 The bias does not appear to reflect cultural background, and it is much the same for Westerners and
658 East Asians (Lee & Johnson-Laird, 2006), but it is sensitive to personality. Individuals who are, or
659 who feel, open to experience and not so conscientious tend to make the causal explanation, whereas
660 their polar opposites tend to make the deductive explanation (Fumero, Santamaría, & Johnson-Laird,
661 2010).

662 The nonmonotonic retraction of a conclusion and modification of beliefs is a side effect of
663 explanation. When individuals explain what's going on in a scenario, they then find it harder to
664 detect an inconsistency it contains than when they have not formulated an explanation (Khemlani &
665 Johnson-Laird, 2012). Conversely, they are faster to revise assertions to make them consistent when
666 they have explained the inconsistency first (Khemlani & Johnson-Laird, 2013). And they rate
667 explanations as more plausible and probable than modifications to the premises that remove the
668 inconsistency – a pattern of judgments that occurs both in adults (Khemlani & Johnson-Laird, 2011)
669 and in children (Legare, 2012). In short, the priority in coping with inconsistencies is to find a causal
670 explanation that resolves them. Explanations first, nonmonotonic modifications after.

671

672 7. The lateral prefrontal cortex and mental models for causal inference

673 A critical brain region underlying mental models for causal inference is the lateral prefrontal
674 cortex, which is known to encode causal representations and to embody the three foundational
675 assumptions of the model theory (see the earlier account of the theory): mental models represent
676 possibilities; their structure can be iconic, mimicking the structure of what they represents; and they
677 represent what is true at the expense of what is false. We now turn to a review of the neuroscience
678 evidence linking each assumption of these principles to core functions of lateral prefrontal cortex.

679 7.1. Mental models represent possibilities

680 The lateral prefrontal cortex is known to play a central role in the representation of behavior-
681 guiding principles that support goal-directed thought and action (Miller & Cohen, 2001). Such top-
682 down representations convey information about possible states of the world, representing what goals
683 are available in the current environment and what actions can be performed to achieve them.

684 The lateral prefrontal cortex represents causal relations in the form of learned task contingencies
685 (causal relations, which neuroscientists refer to as if-then rules). Asaad and colleagues trained
686 monkeys to associate each of two cue objects (*A* and *B*) with a saccade to the right or a saccade to the
687 left (Asaad et al., 1998). The authors found relatively few lateral prefrontal cortex neurons whose
688 activity simply reflected a cue (e.g., *A*) or response (e.g., a saccade to the right). Instead, the modal
689 group of neurons (44% of the population) showed activity that reflected the current association
690 between a visual cue and the directional saccade it instructed. For example, a given cell might be
691 strongly activated only when object *A* instructed “saccade left” and not when object *B* instructed the
692 same saccade or when object *A* instructed another saccade. Likewise, lateral prefrontal cortex
693 neurons acquire selectivity for features to which they are initially insensitive but that are behaviorally
694 important. For example, Watanabe trained monkeys to recognize that certain visual and auditory
695 stimuli signaled whether or not a reward, a drop of juice, would be delivered (Watanabe, 1990;
696 1992). He found that neurons in the lateral prefrontal cortex came to reflect specific cue-reward
697 dependencies. For example, a given neuron could show strong activation to one of the two auditory
698 (and none of the visual) cues, but only when it signaled reward.

699 Studies of monkeys and humans with lateral prefrontal cortex damage also suggest that this region
700 is critical for representing causal principles (if-then rules) that underlie goal-directed thought and
701 adaptive behavior. Early studies of the effects of prefrontal cortex damage in humans suggested its
702 role in goal-directed behavior (e.g., Ferrier, 1876) and since then broad consensus in the literature
703 implicates this region in the ability to control lower-level sensory, memory, and motor operations in
704 the service of a common goal (Shallice, 1982; Duncan, 1986; Passingham, 1993; Grafman, 1994;
705 Wise et al., 1996). Contemporary lesion mapping studies in large populations of patients with focal
706 brain damage further indicate that selective damage to the lateral prefrontal cortex produces
707 impairments in the ability to acquire and use behavior-guiding rules (causal relations) that are central
708 to higher cognitive functions, including general intelligence (Barbey et al., 2012b), fluid intelligence
709 (Barbey et al., 2012a, 2014a), cognitive flexibility (Barbey et al. 2013), working memory (Barbey et
710 al., 2011; 2012c), and discourse comprehension (Barbey et al., 2014b). In monkeys, damage to
711 ventrolateral prefrontal cortex also impairs the ability to learn causal relations in tasks (Halsband &
712 Passingham, 1985; Murray et al., 2000; Petrides, 1982, 1985). Most, if not all, tasks that are
713 disrupted following prefrontal cortex damage rely on mental models that capture the causal structure
714 of experience (cf. Passingham, 1993).

715 Further evidence implicating the lateral prefrontal cortex in causal inference is provided by the
716 fMRI literature (for reviews, see Barbey & Patterson, 2011; Patterson & Barbey, 2012). An important
717 study by Satpute and colleagues demonstrates activity within the dorsolateral prefrontal cortex for the
718 processing of causal versus associative relations (Satpute et al., 2005). Selective activity within the
719 dorsolateral prefrontal cortex for causal (rather than associative) inference provides evidence against
720 associationist accounts of causal representation and instead supports the mental models framework.

721 In sum, the reviewed findings indicate that the lateral prefrontal cortex represents causal relations
722 that establish mappings between possible states of the world, providing the links that bind situations,
723 actions and consequences necessary for goal-directed behavior. These mappings are believed to bias
724 competition in other parts of the brain responsible for task performance (Miller & Cohen, 2001).
725 Thus, signals in the lateral prefrontal cortex guide activity along pathways that connect task-relevant
726 sensory inputs, memories, and motor outputs, which can be naturally represented in the form of
727 mental models of causal relations.

728 **7.2. Mental models are iconic**

729 The information processing architecture of the lateral prefrontal cortex supports the iconic nature
730 of mental models: the structure of a model corresponds to the structure of what it represents in the
731 visual, spatial, auditory, motor and kinematic domains. The cytoarchitectonic areas that comprise
732 lateral prefrontal cortex are often grouped into three regional subdivisions that emphasize processing
733 of particular information based on their interconnections with specific cortical sites. Ventrolateral
734 prefrontal cortex is heavily interconnected with cortical regions for processing information about
735 visual form and stimulus identity (inferior temporal cortex), supporting the categorization of
736 environmental stimuli in the service of goal-directed behavior. Dorsolateral prefrontal cortex is
737 interconnected with cortical areas for processing auditory, visuospatial and motor information
738 (parietal cortex), enabling the regulation and control of responses to environmental stimuli. Finally,
739 anterolateral prefrontal cortex is indirectly connected (via the ventromedial prefrontal cortex) with
740 limbic structures that process internal information, such as emotion, memory and reward (Fuster,
741 2008; Goldman-Rakic, 1995; Petrides et al., 2012). The lateral prefrontal cortex is therefore
742 connected with virtually all sensory neocortical and motor systems and a wide range of subcortical
743 structures, supporting the iconic nature of mental models in the visual, spatial, auditory, motor and

744 kinematic domains. The lateral prefrontal cortex integrates information across this broadly distributed
 745 set of systems and is known to support higher-order symbolic representations, such as negation
 746 (Tettamanti et al., 2008), that go beyond modality-specific knowledge (Ramnani & Owen, 2004).

747 **7.3. Mental models represent only what is true**

748 A third property of lateral prefrontal cortex function is that it represents directly experienced (i.e.,
 749 “true”) events and actively maintains these representations over time in a highly accessible form (i.e.,
 750 storage of information via sustained neuronal activity patterns). The capacity to support sustained
 751 activity in the face of interference is a distinguishing feature of the lateral prefrontal cortex. Sustained
 752 neural activity within the lateral prefrontal cortex was first reported by Fuster (1973), who
 753 demonstrated that neurons within the lateral prefrontal cortex remain active during the delay between
 754 a presented cue and the later execution of a contingent response. Such sustained neural activity often
 755 represents a particular type of information, such as the experienced location or identity of a stimulus
 756 (di Pellegrino and Wise, 1991; Funahashi et al., 1989; Fuster, 1973; Fuster & Alexander, 1971;
 757 Kubota & Niki, 1971) or a particular relation between a stimulus and its corresponding response
 758 (Asaad et al., 1998).

759 **7.4. Summary**

760 In summary, mental models for causal inference critically depend on the lateral prefrontal cortex,
 761 and neuroscience evidence indicates that this region extracts goal-relevant features of experience
 762 (causal relations or if-then rules), it can construct iconic representations, and they represent only what
 763 is true.

764 **8. General discussion**

765 In Ibsen’s play, Dr. Stockmann sought to prevent further contamination of the public bath facility
 766 by alerting the town to the problem. To *prevent* an outcome is to cause it not to occur, and so he
 767 acted in the hope that his causes would have consequences. The meaning of a causal relation
 768 according to the model theory concerns possibilities: a cause suffices to bring about the effect, which
 769 does not precede the cause; an enabling condition makes such an effect possible; and a preventative
 770 action causes the effect not to occur. We have argued that reasoners interpret causal assertions by
 771 simulating the situation, i.e., by building a mental model, to which the assertions refer, and then they
 772 inspect that model to draw conclusions from it. Their initial mental models reflect intuitive
 773 interpretations of causal relations, e.g., their initial model of *runoff causes contamination to occur* is
 774 identical to that of *runoff enables contamination to occur*, i.e.:

	runoff	contamination
	. . .	

777 The first row of the diagram represents a possibility in which runoff occurs concurrently with
 778 contamination, and the second row of the diagram represents that other possibilities are consistent
 779 with the assertion. The theory therefore explains why reasoners often conflate causes and enabling
 780 conditions, i.e., the mental models of the assertions are the same. When prompted to deliberate about
 781 alternative possibilities, however, reasoners are able to flesh out the models and can distinguish
 782 causes from enabling conditions (Goldvarg & Johnson-Laird, 2001).

783 The model theory is deterministic. It posits that causal assertions are used to build discrete models

784 of possibilities. The construction of these discrete models excludes continuous probabilistic
785 information. Three overarching phenomena support a deterministic interpretation of causality:

- 786 • reasoners can infer causal relations from single observations;
- 787 • they distinguish causes from enabling conditions
- 788 • they refute causal assertions with single instances.

789 None of these effects is consistent with a probabilistic interpretation of causality.

790 Reasoners make deductions, inductions, and abductions from causal premises. They base their
791 causal deductions on mental models of the premises; they infer conclusions from the possibilities
792 corresponding to those of the premises. Models can include information about the dynamics of
793 forces. The evidence corroborating the model theory shows that individuals succumb to fallacies –
794 illusory inferences – because mental models do not represent what is false in a possibility (Goldvarg
795 & Johnson-Laird, 2001). Causal induction depends on the use of background knowledge to build
796 models that go beyond the information in the premises. And causal abduction is a complex process in
797 which knowledge is used to introduce new causal relations, which are not part of the premises, in
798 order to provide explanations. Explanation takes priority over the nonmonotonic retraction of
799 conclusions and the editing of propositions to eliminate inconsistencies.

800 The evidence from neuroscience strongly implicates lateral prefrontal cortex as the site of causal
801 processing, and corroborates the principal assumptions of the theory. Just as there are untested
802 behavioral claims of the theory, so too many aspects of causal processing in the brain have yet to be
803 investigated. Inferences from causal assertions, for example, should yield a time course reflecting the
804 successive activation of linguistic areas and then prefrontal activation – a time course that has been
805 observed in studies of deduction in other domains (Kroger et al., 2008). Similarly, materials that
806 elicit visual imagery as opposed to spatial representations impede reasoning, because they elicit
807 irrelevant activity in visual cortex (Knauff, Fangmeier, Ruff & Johnson-Laird, 2003). Analogous
808 effects may also occur in causal reasoning. Likewise, recent evidence to support the hierarchical
809 organization of lateral prefrontal cortex function may reflect the complexity of causal representations
810 for goal-directed thought and behavior (for reviews, see Ramnani & Owen, 2004; Badre, 2008).

811 In sum, the model theory provides a comprehensive account of causal reasoning: what causal
812 assertions mean, how they are interpreted to build models, how these models underlie deductive
813 conclusions; how they incorporate background information in inductive inferences and abductive
814 explanations.

815

816 **9. Acknowledgement**

817 This research reported herein was supported by a National Science Foundation Graduate Research
818 Fellowship to the first author, and by National Science Foundation Grant No. SES 0844851 to the
819 second author to study deductive and probabilistic reasoning. We are grateful for Max Lotstein for
820 help in all aspects of the research, including the computational modeling. We thank Paul Bello, Ruth
821 Byrne, Sam Glucksberg, Adele Goldberg, Catrinel Haught, Max Lotstein, Marco Ragni, and Greg
822 Trafton for helpful criticisms.

823

824 **10. References**

- 825 Ahn, W. (1998). Why are different features central for natural kinds and artifacts? The role of causal
826 status in determining feature centrality. *Cognition*, *69*, 135–178.
- 827 Ahn, W., & Bailenson, J. (1996). Causal attribution as a search for underlying mechanism: An
828 explanation of the conjunction fallacy and the discounting principle. *Cognitive Psychology*, *31*,
829 82–123.
- 830 Ahn, W., & Kalish, C. W. (2000). The role of mechanism beliefs in causal reasoning. In F. C. Keil &
831 R. A. Wilson (Eds.). *Explanation and cognition* (pp. 199–225). Cambridge, MA: MIT Press.
- 832 Asaad, W.F., Rainer, G., & Miller, E.K. (1998). Neural activity in the primate prefrontal cortex
833 during associative learning. *Neuron*, *21*, 1399-1407.
- 834 Badre, D. (2008). Cognitive control, hierarchy, and the rostro-caudal organization of the frontal
835 lobes. *Trends in Cognitive Sciences*, *12*, 193–200.
- 836 Barbey, A.K., Colom, R., & Grafman, J. (2012). Dorsolateral prefrontal contributions to human
837 intelligence. *Neuropsychologia*, *51*, 1361-1369.
- 838 Barbey, A.K., Colom, R., & Grafman, J. (2013). Architecture of cognitive flexibility revealed by
839 lesion mapping. *Neuroimage*, *82*, 547-554.
- 840 Barbey, A.K., Colom, R., Paul, E.J., & Grafman, J. (2014a). Architecture of fluid intelligence and
841 working memory revealed by lesion mapping. *Brain Structure & Function*, *219*, 485-494.
- 842 Barbey, A.K., Colom, R., & Grafman, J. (2014b). Neural mechanisms of discourse comprehension: a
843 human lesion study. *Brain*, *137*, 277-287.
- 844 Barbey, A.K., Colom, R., Solomon, J., Krueger, F., Forbes, C., & Grafman, J. (2012b). An
845 integrative architecture for general intelligence and executive function revealed by lesion
846 mapping. *Brain*, *135*, 1154-1164.
- 847 Barbey, A.K., Koenigs, M., & Grafman, J. (2011). Orbitofrontal contributions to human working
848 memory. *Cerebral Cortex*, *21*, 789-795.
- 849 Barbey, A.K., Koenigs, M., & Grafman, J. (2013). Dorsolateral prefrontal contributions to human
850 working memory. *Cortex*, *49*, 1195-1205.
- 851 Barbey, A.K., & Patterson, R. (2011). Architecture of explanatory inference in the human prefrontal
852 cortex. *Frontiers in Psychology*, *2*, 162.
- 853 Barbey, A.K., & Wolff, P. (2007). Learning causal structure from reasoning. *Proceedings of the 29th*
854 *Annual Conference of the Cognitive Science Society* (pp. 713–718). Mahwah, NJ: Lawrence
855 Erlbaum.
- 856 Besnard, D., & Bastien-Toniazzo, M. (1999). Expert error in trouble-shooting: An exploratory study
857 in electronics. *International Journal of Human Computer Studies*, *50*, 391–405.
- 858 Byrne, R.M.J. (2005). *The rational imagination: How people create alternatives to reality*.
859 Cambridge, MA: MIT.
- 860 Cheng, P. W. (1997). From covariation to causation: a causal power theory. *Psychological Review*,
861 *104*, 367–405.
- 862 Cheng, P.W. (2000). Causal reasoning. In R. Wilson & F. Keil (Eds.), *The MIT Encyclopedia of*

- 863 *Cognitive Sciences* (pp. 106–108). Cambridge, MA: MIT Press.
- 864 Cheng, P. W., Holyoak, K. J., Nisbett, R. E., & Oliver, L. M. (1986). Pragmatic versus syntactic
865 approaches to training deductive reasoning. *Cognitive Psychology*, *18*, 293–328.
- 866 Cheng, P. W., & Novick, L. R. (1991). Causes versus enabling conditions. *Cognition*, *40*, 83–120.
- 867 Clark, H. H. (1975). Bridging. In R. C. Schank & B. L. Nash-Webber (Eds.), *Theoretical issues in*
868 *natural language processing*. New York: Association for Computing Machinery.
- 869 Craik, K. (1943). *The nature of explanation*. Cambridge: Cambridge University Press.
- 870 De Houwer, J., & Beckers, T. (2002). A review of recent developments in research and theory on
871 human contingency learning. *Quarterly Journal of Experimental Psychology*, *55B*, 289–310.
- 872 Di Pellegrino, G., & Wise, S.P. (1991). A neurophysiological comparison of three distinct regions of
873 the primate frontal lobe. *Brain*, *114*, 951–978.
- 874 Duncan, J. (1986). Disorganization of behavior after frontal lobe damage. *Cognitive*
875 *Neuropsychology*, *3*, 271–290.
- 876 Einhorn, H.J., & Hogarth, R.M. (1986). Judging probable cause. *Psychological Bulletin*, *99*, 3–19.
- 877 Fair, D. (1979). Causation and the flow of energy. *Erkenntnis*, *14*, 219–250.
- 878 Ferrier, D. (1876). *The Functions of the Brain*. London: Smith, Elder & Co.
- 879 Frosch, C.A., & Johnson-Laird, P.N. (2011). Is everyday causation deterministic or probabilistic?
880 *Acta Psychologica*, *137*, 280–291.
- 881 Frosch, C.A., Johnson-Laird, P.N., & Cowley, M. (2007). It's not my fault, your Honor, I'm only the
882 enabler. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Meeting of the*
883 *Cognitive Science Society*, 1755.
- 884 Fumero, A., Santamaría, C., & Johnson-Laird, P. N. (2010). Reasoning and autobiographical memory
885 for personality. *Experimental Psychology*, *57*, 215–220.
- 886 Funahashi, S., Bruce, C.J., & Goldman-Rakic, P.S. (1989). Mnemonic coding of visual space in the
887 monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology*, *61*, 331–349.
- 888 Fuster, J.M. (1973). Unit activity in prefrontal cortex during delayed-response performance: neuronal
889 correlates of transient memory. *Journal of Neurophysiology*, *36*, 61–78.
- 890 Fuster, J.M. (2008). *The prefrontal cortex*. Amsterdam ; Boston: Academic Press/Elsevier.
- 891 Fuster, J.M., & Alexander, G.E. (1971). Neuron activity related to short-term memory. *Science*, *173*,
892 652–654.
- 893 Geminiani, G.C., Carassa, A., & Bara, B.G. (1996). Causality by contact. In J. Oakhill, & A.
894 Garnham (Eds.) *Mental models in cognitive science* (pp. 275–303). Hove, East Sussex:
895 Psychology Press.
- 896 Glymour, C. (2001). *The mind's arrows*. Cambridge, MA: The MIT Press.
- 897 Goldman-Rakic, P.S. (1995). Architecture of the prefrontal cortex and the central executive. *Annals*
898 *of New York Academy of Science*, *769*, 71–83.
- 899 Goldvarg, Y., & Johnson-Laird, P.N. (2001). Naive causality: a mental model theory of causal
900 meaning and reasoning. *Cognitive Science*, *25*, 565–610.
- 901 Gopnik, A., Glymour, C., Sobel, D. M., Schulz, D. E., Kushnir, T., & Danks, D. (2004). A theory of
902 causal learning in children: Causal maps and Bayes nets. *Psychological Review*, *111*, 1–31.
- 903 Grafman, J. (1994). Alternative frameworks for the conceptualization of prefrontal functions. In F.
904 Boller, & J. Grafman (Eds.) *Handbook of Neuropsychology*, pp. 187. Amsterdam: Elsevier.

- 905 Grice, H. P. (1975). Logic and conversation. In P. Cole, & J.L. Morgan (Eds.) *Syntax and semantics*,
906 *Vol. 3: speech acts*. New York: Academic Press.
- 907 Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive*
908 *Psychology*, *51*, 354–384.
- 909 Halsband, U., & Passingham, R.E. (1985). Premotor cortex and the conditions for movement in
910 monkeys (*Macaca fascicularis*). *Behavioral and Brain Research*, *18*, 269-277.
- 911 Hattori, M., & Oaksford, M. (2007). Adaptive non-interventional heuristics for covariation detection
912 in causal induction: Model comparison and rational analysis. *Cognitive Science*, *31*, 765–814.
- 913 Hilton, D.J., & Erb, H-P. (1996). Mental models and causal explanation: Judgements of probable
914 cause and explanatory relevance. *Thinking & Reasoning*, *2*, 273–308.
- 915 Hume, D. (1988). *An Enquiry Concerning Human Understanding*. Ed. A. Flew. La Salle, IL: Open
916 Court. (Originally published 1748.)
- 917 Jeffrey, R. (1981). *Formal logic: Its scope and limits* (2nd ed.). New York, NY: McGraw-Hill.
- 918 Johnson-Laird, P.N. (1983). *Mental models*. Cambridge: Cambridge University Press. Cambridge,
919 MA: Harvard University Press.
- 920 Johnson-Laird, P.N. (1988). A taxonomy of thinking. In R.J. Sternberg & E.E. Smith, (Eds.) *The*
921 *psychology of human thought* (pp. 429–457). New York: Cambridge University Press.
- 922 Johnson-Laird, P.N. (1999). Causation, mental models, and the law. *Brooklyn Law Review*, *65*,
923 67–103.
- 924 Johnson-Laird, P.N. (2006). *How we reason*. New York: Oxford University Press.
- 925 Johnson-Laird, P. N., & Byrne, R.M.J. (1991). *Deduction*. Hillsdale, NJ: Erlbaum.
- 926 Johnson-Laird, P.N., & Byrne, R.M.J. (2002). Conditionals: A theory of meaning, pragmatics, and
927 inference. *Psychological Review*, *109*, 646–678.
- 928 Johnson-Laird, P. N., & Hasson, U. (2003). Counterexamples in sentential reasoning. *Memory &*
929 *Cognition*, *31*, 1105–1113.
- 930 Johnson-Laird, P.N., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to
931 consistency. *Psychological Review*, *111*, 640-661.
- 932 Johnson-Laird, P. N., & Khemlani, S. (2014). Toward a unified theory of reasoning. In B. Ross (Ed.),
933 *The psychology of learning and motivation* (pp. 1-42). Elsevier, Inc.: Academic Press.
- 934 Kahneman, D., & Miller, D. T. (1986). Norm theory: comparing reality to its alternative.
935 *Psychological Review*, *93*, 75–88.
- 936 Kant, I. (1934). *Critique of pure reason*. Trans. J.M.D. Meiklejohn, New York: Dutton. (Originally
937 published 1781.)
- 938 Kelley, H.H. (1973). The processes of causal attribution. *American Psychologist*, *28*, 107-128.
- 939 Khemlani, S., & Johnson-Laird, P.N. (2011). The need to explain. *Quarterly Journal of Experimental*
940 *Psychology*, *64*, 2276–2288.
- 941 Khemlani, S. & Johnson-Laird, P. N. (2012). Hidden conflicts: Explanations make incon-
942 sistencies harder to detect. *Acta Psychologica*, *139*, 486–491.
- 943 Khemlani, S., & Johnson-Laird, P. N. (2013). Cognitive changes from explanations. *Journal of*
944 *Cognitive Psychology*, *25*, 139–146.
- 945 Khemlani, S., Mackiewicz, R., Bucciarelli, M., & Johnson-Laird, P. N. (2013). Kinematic mental
946 simulations in abduction and deduction. *Proceedings of the National Academy of Sciences of the*

- 947 *United States of America, 110, 16766–16771.*
- 948 Khemlani, S., Orenes, I., & Johnson-Laird, P. N. (2012). Negation: A theory of its meaning,
949 representation, and use. *Journal of Cognitive Psychology, 24, 541–559.*
- 950 Khemlani, S., Sussman, A., & Oppenheimer, D. (2011). Harry Potter and the sorcerer's scope: Scope
951 biases in explanatory reasoning. *Memory & Cognition, 39, 527–535.*
- 952 Knauff, M., Fangmeier, T., Ruff, C. C. & Johnson-Laird, P. N. (2003). Reasoning, models, and
953 images: Behavioral measures and cortical activity. *Journal of Cognitive Neuroscience, 4, 559–573.*
- 954 Koslowski, B. (1996). *Theory and evidence: The development of scientific reasoning.* Cambridge,
955 MA: MIT Press.
- 956 Kroger, J.K., Nystrom, L.E., Cohen, J.D., & Johnson-Laird, P.N. (2008). Distinct neural
957 substrates for deductive and mathematical processing. *Brain Research, 1243, 86–103.*
- 958 Kubota, K., & Niki, H. (1971). Prefrontal cortical unit activity and delayed alternation performance
959 in monkeys. *Journal of Neurophysiology, 34, 337–347.*
- 960 Kuhnmüch, G. & Beller, S. (2005). Distinguishing between causes and enabling conditions –
961 through mental models or linguistic cues? *Cognitive Science, 29, 1077–1090.*
- 962 Lee, N. Y. L., & Johnson-Laird, P. N. (2006). Are there cross-cultural differences in reasoning? In
963 *Proceedings of the 28th Annual Meeting of the Cognitive Science Society* (pp. 459–464).
964 Vancouver, BC: Cognitive Science Society.
- 965 Lee, N.Y.L., & Johnson-Laird, P.N. (2013). A theory of reverse engineering and its application to
966 Boolean systems. *Journal of Cognitive Psychology, 25, 365–389.*
- 967 Legare, C.H. (2012). Exploring explanation: Explaining inconsistent information guides hypothesis-
968 testing behavior in young children. *Child Development, 83, 173–185.*
- 969 Leslie, S.J. (2008). Generics: Cognition and acquisition. *Philosophical Review, 117, 1–47.*
- 970 Lien, Y., & Cheng, P. (2000). Distinguishing genuine from spurious causes: A coherence hypothesis.
971 *Cognitive Psychology, 40, 87–137.*
- 972 Lombrozo, T. (2007). Simplicity and probability in causal explanations. *Cognitive Psychology, 55,*
973 *232–257.*
- 974 Lu, H., Yuille, A., Liljeholm, M., Cheng, P. W., Holyoak, K. J. (2008). Bayesian generic priors for
975 causal learning. *Psychological Review, 115, 955–984.*
- 976 Mackie, J.L. (1980). *The cement of the universe: A study in causation.* Second edition. Oxford:
977 Oxford University Press.
- 978 Marr, D. (1982). *Vision: A computational investigation into the human representation and*
979 *processing of visual information.* New York: Freeman.
- 980 Michalski, R. S., & Wojtusiak, J. (2007). Generalizing data in natural language. In *Rough sets and*
981 *intelligent systems paradigms* (pp. 29–39). Berlin: Springer.
- 982 Michotte, A. (1963). *The perception of causality.* London: Methuen. (Originally published 1946.)
- 983 Mill, J.S. (1874). *A system of logic, ratiocinative and inductive: being a connected view of the*
984 *principles of evidence and the methods of scientific evidence.* (8th ed.) New York: Harper. (First
985 edition published 1843.)
- 986 Miller, E.K., & Cohen, J.D. (2001). An integrative theory of prefrontal cortex function. *Annual*
987 *Review of Neuroscience, 24, 167–202.*
- 988 Miller, G. A., & Johnson-Laird, P. N. (1976). *Language and perception.* Cambridge, MA: Harvard

- 989 University Press.
- 990 Miyake, N. (1986). Constructive interaction and the iterative process of understanding. *Cognitive*
- 991 *Science*, 10, 151–177.
- 992 Morris, M.W., & Nisbett, R.E. (1993). Tools of the trade: Deductive schemas taught in psychology
- 993 and philosophy. R.E. Nisbett (Ed.) *Rules for Reasoning* (pp. 228–256). Hillsdale, NJ: Lawrence
- 994 Erlbaum Associates.
- 995 Murray, E.A., Bussey, T.J., & Wise, S.P. (2000). Role of prefrontal cortex in a network for arbitrary
- 996 visuomotor mapping. *Experimental Brain Research*, 133, 114-129.
- 997 Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human*
- 998 *reasoning*. Oxford: Oxford University Press.
- 999 Passingham, R.E. (1993). *The frontal lobes and voluntary action*. New York: Oxford University
- 1000 Press.
- 1001 Patterson, R., & Barbey, A.K. (2012). A cognitive neuroscience framework for causal reasoning. In
- 1002 J. Grafman & F. Krueger (Eds.) *The Neural Representation of Belief Systems* (pp. 76-120). New
- 1003 York, NY: Psychology Press.
- 1004 Pearl, J. (2000). *Causality*. New York: Cambridge University Press.
- 1005 Peirce, C.S. (1931–1958). *Collected papers of Charles Sanders Peirce*, C. Hartshorne, P. Weiss, &
- 1006 A. Burks (Eds.) Cambridge, MA: Harvard University Press.
- 1007 Perales, J. C., & Shanks, D. R. (2007). Models of covariation-based causal judgment: A review and
- 1008 synthesis. *Psychonomic Bulletin & Review*, 14, 577–596.
- 1009 Petrides, M. (1982). Motor conditional associative-learning after selective prefrontal lesions in the
- 1010 monkey. *Behavioral & Brain Research*, 5, 407-413.
- 1011 Petrides, M. (1985). Deficits in non-spatial conditional associative learning after periarculate lesions
- 1012 in the monkey. *Behavioral & Brain Research*, 16, 95-101.
- 1013 Petrides, M., Tomaiuolo, F., Yeterian, E.H., & Pandya, D.N. (2012). The prefrontal cortex:
- 1014 comparative architectonic organization in the human and the macaque monkey brains. *Cortex*, 48,
- 1015 46-57.
- 1016 Polya, G. (1973). *How to solve it: A new aspect of mathematical methods*. 2nd ed. Princeton:
- 1017 Princeton University Press. (Originally published 1945.)
- 1018 Prasada, S., Khemlani, S., Leslie, S.-J., & Glucksberg, S. (2013). Conceptual distinctions amongst
- 1019 generics. *Cognition*, 126, 405–422.
- 1020 Ramnani, N., & Owen, A.M. (2004). Anterior prefrontal cortex: insights into function from anatomy
- 1021 and neuroimaging. *Nature Reviews Neuroscience* 5, 184-194.
- 1022 Ramoni, M.F., Stefanelli, M., Magnani, L., & Barosi, G.(1992). An epistemological framework for
- 1023 medical knowledge based system. *IEEE Transactions on Systems, Man, and Cybernetics*, 22,
- 1024 1361–1375.
- 1025 Rehder, B. (2003). Categorization as causal reasoning. *Cognitive Science*, 27, 709–48.
- 1026 Reichenbach, H. (1956). *The direction of time*. Berkeley: University of California Press.
- 1027 Rips, L.J. (1994). *The psychology of proof*. Cambridge, MA: MIT Press.
- 1028 Russell, B.A.W. (1912-13). On the notion of cause. *Proceedings of the Aristotelian Society*, 13, 1–
- 1029 26.
- 1030 Salmon, W. C. (1980). Probabilistic causality. *Pacific Philosophical Quarterly*, 61, 50–74.

- 1031 Salsburg, D. (2001). *The lady tasting tea: How statistics revolutionized science in the twentieth*
1032 *century*. New York: W.H. Freeman.
- 1033 Satpute, A.B., Fenker, D.B., Waldmann, M.R., Tabibnia, G., Holyoak, K.J., & Lieberman, M.D.
1034 (2005). An fMRI study of causal judgments. *European Journal of Neuroscience*, *22*, 1233-1238.
- 1035 Schlottman, A., & Shanks, D. R. (1992). Evidence for a distinction between judged and perceived
1036 causality. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*,
1037 *44(A)*, 321–342.
- 1038 Shanks, D.R. (1995). *The psychology of associative learning*. Cambridge: Cambridge University
1039 Press.
- 1040 Shallice, T. (1982). Specific impairments of planning. *Philosophical Transactions of the Royal*
1041 *Society: Biological Sciences*, *298*, 199–209.
- 1042 Sloman, S. A. (2005). *Causal models: How we think about the world and its alternatives*. New York:
1043 Oxford University Press.
- 1044 Sloman, S. A., Barbey, A. K., & Hotaling, J. (2009). A causal model theory of the meaning of cause,
1045 enable, and prevent. *Cognitive Science*, *33*, 21–50
- 1046 Sloman, S. A., & Lagnado, D. A. (2005). Do we ‘do’? *Cognitive Science*, *29*, 5–39.
- 1047 Suppes, P. (1970). *A probabilistic theory of causality*. Amsterdam: North-Holland.
- 1048 Tenenbaum, J. B., & Griffiths, T. L. (2001). Structure learning in human causal induction. In T.
1049 Leen, T. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems 13* (pp.
1050 59–65). Cambridge, MA: MIT Press.
- 1051 Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive
1052 learning and reasoning. *Trends in Cognitive Sciences*, *10*, 309–318.
- 1053 Tettamanti, M., Manenti, R., Della Rosa, P.A., Falini, A., Perani, D., Cappa, S.F., & Moro, A.
1054 (2008). Negation in the brain: modulating action representations. *Neuroimage*, *43*, 358-367.
- 1055 Turnbull, W., & Slugoski, B.R. (1988). Conversational and linguistic processes in causal attribution.
1056 In D. Hilton (Ed.) *Contemporary science and natural explanation: Commonsense conceptions of*
1057 *causality* (pp. 66–93.) Brighton, Sussex: Harvester Press.
- 1058 von Wright, G.H. (1973). On the logic and epistemology of the causal relation. In P. Suppes, (Ed.)
1059 *Logic, methodology and philosophy of science, IV* (pp. 293–312). Amsterdam: North-Holland.
- 1060 Waldmann, M. R., Holyoak, K. J., & Fratianne, A. (1995). Causal models and the acquisition of
1061 category structure. *Journal of Experimental Psychology: General*, *124*, 181–206.
- 1062 Watanabe, M. (1990). Prefrontal unit activity during associative learning in the monkey.
1063 *Experimental Brain Research*, *80*, 296-309.
- 1064 Watanabe, M. (1992). Frontal units of the monkey coding the associative significance of visual and
1065 auditory stimuli. *Experimental Brain Research*, *89*, 233-247.
- 1066 White, P.A. (1995). Use of prior beliefs in the assignment of causal roles: Causal powers versus
1067 regularity-based accounts. *Memory & Cognition*, *23*, 243–254.
- 1068 White, P.A. (1999). Toward a causal realist account of causal understanding. *American Journal of*
1069 *Psychology*, *112*, 605–642.
- 1070 White, P.A. (2014). Singular cues to causality and their use in human causal judgment. *Cognitive*
1071 *Science*, *38*, 38–75.
- 1072 Wise, S.P. (1996). The frontal-basal ganglia system in primates. *Critical Review of Neurobiology*, *10*,

- 1073 317–356.
1074 Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, *136*,
1075 82–111.
1076 Wolff, P., & Song, G. (2003). Models of causation and causal verbs. *Cognitive Psychology*, *47*,
1077 276–332.
1078 Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford, UK: Oxford
1079 University Press.

1080 **11. Figure legends**

1081 Figure 1 The common-cause and common-effect causal schemas. Reproduced with permission from
1082 Rehder (2003).