Chapter 1

Theoretical Frameworks of Causal Induction

> "That the sun will not rise tomorrow implies no more contradiction than that it will rise."
>
> *David Hume*

People are intrinsically motivated to learn how our world works. However, as David Hume argues, the reality is inherently uncertain. We can often only learn by collecting limited and noisy evidence and then inferring general rules to guide our prediction and control in the future. Causal induction theories are concerned with how people infer the nature of the causal relationships between entities or phenomena on the basis of information they have obtained. Here are four computational theories that have attempted to describe human causal induction over the last five decades. Each theory builds on earlier approaches, with later theories addressing issues that arose or could not be solved by the preceding theories.

1.1 Rescorla-Wagner Rule

Empirical research about the recognition of relationships between events can be traced back to Behaviourism, one of the most influential approaches in 20^{th} century psychology. Behaviouristic scientists focus on how human and non-human animals would associate binary variables based on the experience of statistical contingency as well as spatiotemporal contiguity (see Pavlov, 1928; Skinner, 1938, for review). Their experimental paradigms often include training stages that display associated or disassociated evidence and testing stages that examine associative strengths by measure subjects' behaviour patterns.

Rescorla and Wagner (1972) formalise experienced association strengths on the basis of co-occurrence evidence (i.e., subjects experienced evidence trial by trial, and in each trial different stimuli may be present or absent) with a simple mechanical rule – RW rule (Eq 1.1-1.2). This suggests that learning occurs to the extent that a learner feels "surprised" about a new observation. For instance, if a learner believes there is no relationship between A and B but then observes A and B's co-occurrence (which is denoted as $A = B = 1$), she should slightly increase her association between A and B. However, with repeated exposures to $A = B = 1$, the association starts to asymptote (i.e., acquisition curves) because the evidence becomes less surprising to the learner. RW can also be applied to situations when multiple events are associated with a target event (Saavedra, 1975). If the learner already believes A and B are associated, the later observation of $A = C = B = 1$ will be not surprising and therefore cannot increase the association between B and C. This phenomenon is called *forward blocking* (Kamin, 1967; Le Pelley, Griffiths, & Beesley, 2017).

$$V_{C_1}^{t} = V_{C_1}^{t-1} + \Delta V_{C_1}^{t} \tag{1.1}$$

$$\Delta V_{C_1}^{t} = \alpha\beta(\lambda - \sum_{C \in \mathbb{C}_t} V_C^{t-1}) \tag{1.2}$$

RW depicts how beliefs change dynamically as trial-based information flows in. Eq 1.1 states that the associative strength at the current trial t depends on the original strength plus the change due to trial t. Eq 1.2 specifies that in a given trial, the strength for particular cause C_1 is updated according to both whether the effect co-occurs with C_1 ($\lambda = 1$) or not ($\lambda = 0$) and the existing association strength based on how many causes occur in the current trial as well as their predictive strengths respectively. Two fixed learning rate parameters α and β are added which depend on the salience of C_1 or outcomes.

RW rule assumes that subjects learn a network of associations rather than setting out to learn a model of the world. Despite the fact that RW has proven a successful predictor for many aspects of human and non-human animals' behaviour (Allan, 1993), it fails to predict some phenomena especially in human subjects that raise doubt as to whether causal learning could be equal to associative learning. That is, whether people learn relationships between events without the support of structured mental representation. I list three of them below.

Firstly, although *forward blocking* that a new cause would hardly be experienced as strongly causal if it always co-occurs with existing causes that have well explained the effect, is perfectly shown in non-human animals, it is relatively weak or even failed to observe in the human learning process (Kamin, 1967; Shanks, 1985; Cheng & Lu, 2017; Le Pelley et al., 2017). As a contrast to forward blocking, the second phenomenon is called *backward blocking*, which describes the finding that when people experience co-occurrences of Cause A, Cause B, and an effect, and are then trained on co-occurrences of only A and the effect, their causal strength judgment of B will decrease (Le Pelley & McLaren, 2001; Shanks, 1985; Wasserman & Berglan, 1998). However, since there is no information about B at the second stage, RW does not predict this updating of B's causal strength. A third phenomenon is inferences of *unobserved causes* found in both humans (e.g., Lipp & Vaitl, 1992) and non-human animals (e.g., Hall & Honey, 1989). These are at odds with RW that only considers observed variables. For example, Hall and Honey (1989) train rats on the association between a cue and outcome, then they experience no association at the second phase, and finally experience the association again. The learning speed in the third phase is much quicker than RW predicts. It suggests rats may not directly unlearn the association at the second phase, but assume the latent inhibitory cause, and hence at the third phase they learn the inhibitory cause is removed where the situation is treated as returning back to the initial phase than running into a new phase (see Gershman, Blei, & Niv, 2010; Redish, Jensen, Johnson, & Kurth-Nelson, 2007, for review and computational explainations). All three phenomena suggest

Table 1.1: Two by Two Causal Tabular Data

	Cause=1	Cause=0
Effect=1	a	b
Effect=0	c	d

that causal learning is not simple reflections of covariation but are additionally sensitive to one's mental representation of the underlying causal structure. More psychological assumptions are waiting to be built into causal induction theories.

1.2 Delta-P

Historically, human cognition researchers focus on how humans make causal inferences from descriptions such as whether a certain fertiliser will cause plants to bloom. In these studies, human participants no longer personally experience event associations, but they read the summarised statistical information and then judge causal strength on scales. The information of two binary variables is usually presented as *2 by 2 tables* (Table 1.1). As shown in Eq 1.3, Delta-P rule (Allan, 1980; Jenkins & Ward, 1965) assumes that people infer causal strength by comparing cases that effect occurs with the cause present, with cases that effect occurs with the cause absent. Generative causal judgments equal to ΔP and Preventative causal judgments equal to $-\Delta P$.

$$\Delta P = P(E|C) - P(E|\neg C) = \frac{a}{a+c} - \frac{b}{b+d} \tag{1.3}$$

As with RW, Delta-P is also an associative quantity that does not address any mental causal representation. It provides a solution for prevailing scenarios in human life and scientific discovery. Indeed, large attention in causal reasoning research is drawn to inferences upon contingency tables from then on. Delta-P performs better than many other calculations in predicting causal strength judgments (Allan & Jenkins, 1983), but it is insensitive to the "density bias" found in humans (Allan & Jenkins, 1983; Baker, Berbrier, & Vallee-Tourangeau, 1989; Buehner, Cheng, & Clifford, 2003; Shanks & Dickinson, 1991): If a set of scenarios are constructed in which ΔP is fixed while other aspects of the contingencies are varied, human generative causal strength judgments do not remain constant. Specifically, they tend to increase as $(a+b)/(a+b+c+d)$ increases, and preventative judgments decrease as $(a+b)/(a+b+c+d)$ decreases. Causal Power theory described below successfully solves this problem.

1.3 Causal Power

As Delta-P, Causal Power theory (aka. power PC, see Cheng, 1997; Cheng & Lu, 2017, for review) also aims to extrapolate causal strength between binary variables from 2 by 2 tabular data (Table 1.1). Compared to associative theories, power PC demonstrates four assumptions about human causal reasoning:

- There is an unobserved cause A that can produce the effect E but not prevent it.
- The evaluated cause C and the unobserved cause A influence E independently.

- The power of a cause is independent of the frequency of occurrence of the cause.
- E does not occur unless it is caused.

The core feature of power PC is assuming an unobserved hidden cause that accounts for the effect's presence when the observed cause is absent. Accordingly, when C is a generative cause, the effect could be caused by either C or A, and therefore the probability of observing E follows a *noisy-OR* function in Eq 1.4, where $c, e \in [0, 1]$ represent the absence or presence of C and E, q_c represents the causal strength of c, and w_a represents $q_a \cdot a$. When C is present, $P(e = 1|c = 1) = q_c + w_a - q_c \cdot w_a$; when C is absent, $P(e = 1|c = 0) = w_a$. Therefore, we can derive Eq 1.5 as the calculation of causal strength q_c.

$$P(e = 1|c; w_a, q_c) = q_c \cdot c + w_a - q_c \cdot c \cdot w_a \tag{1.4}$$

$$q_c = \frac{P(e = 1|c = 1) - P(e = 1|c = 0)}{1 - P(e = 1|c = 0)} \tag{1.5}$$

If C is a preventative cause, the effect could be caused by A but then possibly presented by C. The probability of observing E follows a *noisy-AND-NOT* function in Eq 1.6. Accordingly, the preventative causal strength can be represented as Eq 1.7.

$$P(e = 1|c; w_a, q_c) = w_a(1 - q_c \cdot c) \tag{1.6}$$

$$q_c = \frac{P(e = 1|c = 0) - P(e = 1|c = 1)}{P(e = 1|c = 0)} \tag{1.7}$$

We could find that power PC can be seen as Delta-P adjusted by the effect's base rate. It considers the proportion of cases that E is already present or absent and therefore C would have no chance to affect E. Power PC successfully predicts the phenomenon that ratings of generative (or preventative) strengths become conservative when the effect's base rate is already high (or low; Buehner et al., 2003). It also predicts that when the base rate equals to 1 in generative causal judgments (or 0 in preventative causal judgments), people would consider evidence to be uninformative to infer causal relationships (Wu & Cheng, 1999). In a meta-analysis, Lu et al. (2008) show that power PC obtains .96 correlation with human causal strength judgments, which indicates that generally people do use inner causal representation when inferring about the outside relationships.

1.4 Causal Bayesian Networks

Although power PC well captures causal judgments in two binary variables, real situations are frequently more complex. For example, it may be reasonable to presume that depression causes insomnia, and insomnia causes anxiety (i.e., a chain structure), or alternatively that depression causes insomnia and anxiety independently (i.e., a fork structure). People need to distinguish which structure is correct given some observations (See Figure 1.1 for how many possible generative structures on the basis of three variables). Causal Bayesian Networks (CBNs, Pearl, 2000) provide a principled approach to formalise complex causal relationships among multiple variables. It was developed for modelling large datasets in computer science contexts, but subsequent research

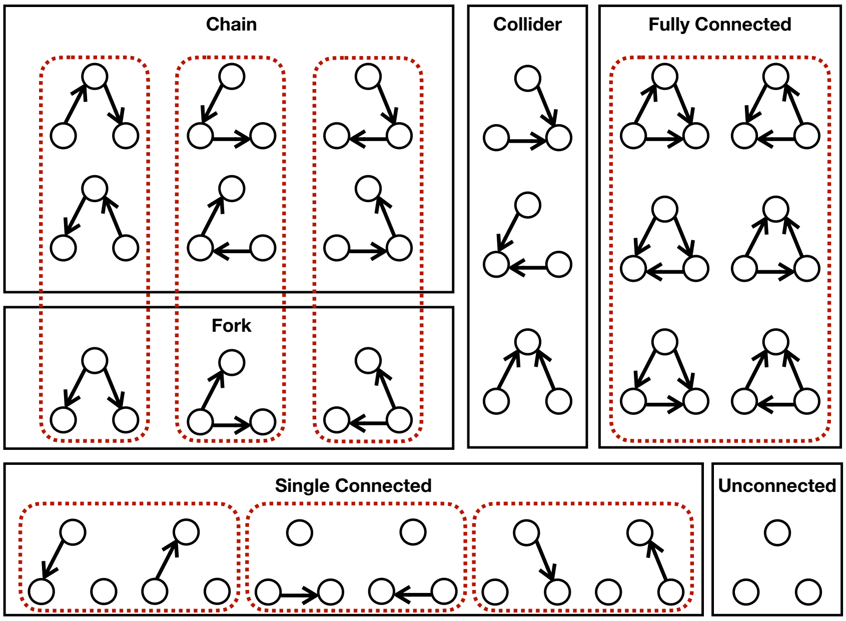

Figure 1.1: Causal Bayesian Networks on Three Variables
Red boxes indicate Markov equivalence structures.

shows that laypeople have a commensurate intuitive understanding of complex causal structures and both learn and make inferences that broadly reflect the predictions of CBNs theory (Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003; Griffiths & Tenenbaum, 2005, 2009).

CBNs are kinds of probabilistic directed acyclic graphical models. They represent variables as nodes and causal relationships as arrows and are defined by the "Markov assumption" that once all the direct causes of a variable X are controlled for, X must be statistically independent of other variables in the causal network that are not its direct or indirect effects. CBNs provide ideas for causal structure learning from both qualitative and quantitative aspects. The qualitative aspect is that causality is represented as a network of direct dependence between variables. For example, if we want to confirm that depression causes insomnia and anxiety independently (insomnia ← depression → anxiety), there should be statistical dependence between insomnia and depression, and depression and anxiety, but insomnia and anxiety should be irrelevant once the state of depression is known.

Quantitatively, it follows the Bayes' rule (see Box 2.1 for more introduction of Bayes' rule) that one can incorporate a prior belief with the likelihood of newly observed data to form an updated "posterior" belief. In this case, this is a posterior over all hypothetical causal structures and then choose the most likely causal structure. The likelihood $P(\{x_1, ..., x_i\}|H)$ is calculated according to Eq 1.8 where $\{x_1, ..., x_i\}$ represents the data in one trial. The $genPa(x_i)$ and $prePa(x_i)$ are data sets of generative or preventative parent nodes associated with x_i. The calculation of $P(x_i|Pa(x_i))$ refers to noisy-OR and noisy-AND-NOT functions (Eq 1.9), which is similar to power PC despite that now one effect can be influenced by multiple causes (and the base rate could be regarded as one cause in $genPa(x_i)$). The causal power parameters q could be learned via instructions or jointly learned with causal structures (Griffiths & Tenenbaum, 2005, 2009).

$$P(\{x_1, ..., x_i\}|H) = \prod_i P(x_i|genPa(x_i), prePa(x_i)) \tag{1.8}$$

$$P(x_i|genPa(x_i), prePa(x_i)) = [1 - \prod_{g\in genPa(x_i)} (1 - q_g \cdot g)] \prod_{p\in prePa(x_i)} (1 - q_p \cdot p) \tag{1.9}$$

CBNs provide a comprehensive theoretical framework that can handle a wide range of human causal reasoning questions across different domains (see Rottman, 2017, for review). Compared to power PC, CBNs not only can reflect the inner representation of unobserved hidden causes, but also incorporate Bayesian priors to capture potential human inductive biases. Griffiths and Tenenbaum (2005) encode two-variable causal strengths usually demonstrated under power PC into CBNs. This is motivated by the finding that when $\Delta P = 0$, judged causal strength still increases as observed data points decreases, which is inconsistent with power PC that predicts a constant judgment at zero. Griffiths and Tenenbaum (2005) explain this "frequency illusion" by assuming that what people actually do is to distinguish between two causal hypotheses: a "Graph 1" where both unobserved "background causes" and the target cause are linked to the effect, and "Graph 0" where only the unobserved background causes are linked to the effect. Under limited observed data, people are uncertain about both graphs, and therefore the causal judgment – the normalized probability of Graph 1 – is larger than that after people gather enough evidence to support Graph 0. Essentially, they argue that what people infer in these settings is not the "strength" of association between cause-effect pairs, but the "probability" that a causal link exists. Thus, they sometimes replace the term "causal strength" with "causal support" to describe the belief about the link between a putative cause and effect.

Box 1.1: Probabilistic Inference – Bayes' Rule. Bayes' rule provides a solution to how we can use evidence to revise our beliefs, i.e., the problem of induction. According to the property of conditional probability that $P(A, B) = P(A|B)P(B) = P(B|A)P(A)$, we can get Eq 1.10, where h represents hypothesis and d represents data.

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)} \tag{1.10}$$

Since we often do not update single hypothesis (e.g, if I tell you here is a weighted coin and flip it several times, you will update degrees of beliefs "this coin is more likely to land heads" and "this coin is more likely to land tails" simultaneously, with one increase and another decrease), we can form $H = \{h1, h2, h3, \ldots\}$ as the set to include all possible hypothesis we consider, which is often called *hypothesis space*. Meanwhile, we also often receive a number of data points, for which we form $D = \{d1, d2, d3, \ldots\}$ as the dataset we use to update our beliefs. Then we can revise Eq 1.10 as:

$$P(H|D) = \frac{P(D|H)P(H)}{\sum_{h\in H} P(D|H)P(H)} \tag{1.11}$$

For most cases, we do not need to know the absolute probability of our hypotheses given the data, but relatively which hypothesis wins, so we can ignore $\sum_{h\in H} P(D|H)P(H)$ since it is constant:

$$P(H|D) \propto P(D|H)P(H) \tag{1.12}$$

Eq 1.12 is the most commonly used formula in induction problems. $P(H)$ is called *prior distribution* (or inductive bias) that reflect people's degrees of different beliefs before observing data. $P(D|H)$ is called *likelihood* where we calculate the probability of observing all data in D if a hypothesis is true. $P(H|D)$ is called *posterior distribution* that combines the prior and likelihood to know the revised degrees of each belief, where we can finally choose the most likely belief to be the answer of the induction problem.

The *likelihood function*, i.e., how we gain $P(D|H)$, would be an important piece that researchers need to illustrate in their works since it depends on the specific task. If the number of potential hypotheses is small, researchers can simply define a uniform prior distribution that all hypotheses are treated as equal possible before looking at the data, whereas if the hypothesis space is large and contain unfitted parameters researchers also need to carefully consider the prior setting in their models. Finally, when the posterior calculation is intractable due to large hypothesis spaces or complex likelihood functions, researchers need to use some algorithms (e.g., Monte Carlo methods, particle filtering) to approximate the posterior distribution. Since the posterior calculation in learning problems of this thesis as well as many causal induction papers I mentioned in this thesis can be done by simple enumeration, I do not plan to introduce these approximate algorithms here (see Griffiths, Chater, Kemp, Perfors, and Tenenbaum (2010) for a short overview of probabilistic inference in human cognition and Oaksford and Chater (2007) for a detail one).

References

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society*, *15*(3), 147–149.

Allan, L. G. (1993). Human contingency judgments: Rule based or associative? *Psychological Bulletin*, *114*(3), 435–448.

Allan, L. G., & Jenkins, H. (1983). The effect of representations of binary variables on judgment of influence. *Learning and Motivation*, *14*(4), 381–405.

Baker, A. G., Berbrier, M. W., & Vallee-Tourangeau, F. (1989). Judgements of a 2× 2 contingency table: Sequential processing and the learning curve. *The Quarterly Journal of Experimental Psychology Section B*, *41*(1b), 65–97.

Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From covariation to causation: a test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(6), 1119.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367.

Cheng, P. W., & Lu, H. (2017). Causal invariance as an essential constraint for creating a causal representation of the world: Generalizing. In M. Waldmann (Ed.), *The oxford handbook of causal reasoning* (p. 65-84). New York: Oxford University Press.

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review*, *117*(1), 197.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–384.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*(4), 661-716.

Hall, G., & Honey, R. C. (1989). Contextual effects in conditioning, latent inhibition, and habituation: Associative and retrieval functions of contextual cues. *Journal of Experimental Psychology: Animal Behavior Processes*, *15*(3), 232.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological monographs: General and Applied*, *79*(1), 1.

Kamin, L. J. (1967). Predictability, surprise, attention, and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior* (pp. 279–296). New York: Appleton-Century-Crofts.

Le Pelley, M. E., Griffiths, O., & Beesley, T. (2017). Associative accounts of causal cognition. In M. Waldmann (Ed.), *The oxford handbook of causal reasoning* (p. 13-28). New York: Oxford University Press.

Le Pelley, M. E., & McLaren, I. P. L. (2001). Retrospective revaluation in humans: Learning or memory? *The Quarterly Journal of Experimental Psychology Section B*, *54*(4b), 311–352.

Lipp, O. V., & Vaitl, D. (1992). Latent inhibition in human pavlovian differential conditioning: Effect of additional stimulation after preexposure and relation to schizotypal traits. *Personality and Individual Differences*, *13*(9), 1003–1012.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *115*(4), 955.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning.* New York: Oxford University Press.

Pavlov, I. P. (1928). *Lectures on conditioned reflexes.* New York: W H Gantt International Publishers.

Pearl, J. (2000). *Causality.* New York: Cambridge University Press (2009 reprint).

Redish, A. D., Jensen, S., Johnson, A., & Kurth-Nelson, Z. (2007). Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. *Psychological Review*, *114*(3), 784–805.

Rescorla, R. A., & Wagner, A. R. (1972). A theory on pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In A. Black & W. Prokasy (Eds.), *Classical conditioning ii: Current theory and research* (pp. 64–99). New York: Appleton Century Crofts.

Rottman, B. M. (2017). The acquisition and use of causal structure knowledge. In M. Waldmann (Ed.), *The oxford handbook of causal reasoning* (pp. 85–114). New York: Oxford University Press.

Saavedra, M. A. (1975). Pavlovian compound conditioning in the rabbit. *Learning and Motivation*, *6*(3), 314–326.

Shanks, D. R. (1985). Forward and backward blocking in human contingency judgement. *The Quarterly Journal of Experimental Psychology Section B*, *37*(1b), 1–21.

Shanks, D. R., & Dickinson, A. (1991). Instrumental judgment and performance under variations in action-outcome contingency and contiguity. *Memory & Cognition*, *19*(4), 353–360.

Skinner, B. F. (1938). *The behaviour of organisms: An experimental analysis.* New York: D. Appleton-Century Company Incorporated.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive science*, *27*(3), 453–489.

Wasserman, E. A., & Berglan, L. R. (1998). Backward blocking and recovery from overshadowing in human causal judgement: The role of within-compound associations. *The Quarterly Journal of Experimental Psychology: Section B*, *51*(2), 121–138.

Wu, M., & Cheng, P. W. (1999). Why causation need not follow from statistical association: Boundary conditions for the evaluation of generative and preventive causal powers. *Psychological Science*, *10*(2), 92–97.