Causal Reasoning and Meno's Paradox

Melvin Chen¹ · Lock Yue Chew²

Received: 9 December 2019 / Accepted: 6 August 2020
© The Author(s) 2020

Abstract

Causal reasoning is an aspect of learning, reasoning, and decision-making that involves the cognitive ability to discover relationships between causal relata, learn and understand these causal relationships, and make use of this causal knowledge in prediction, explanation, decision-making, and reasoning in terms of counterfactuals. Can we fully automate causal reasoning? One might feel inclined, on the basis of certain groundbreaking advances in causal epistemology, to reply in the affirmative. The aim of this paper is to demonstrate that one still has good skeptical grounds for resisting any conclusions in favour of the automation of causal reasoning. If by causal reasoning is meant the entirety of the process through which we discover causal relationships and make use of this knowledge in prediction, explanation, decision-making, and reasoning in terms of counterfactuals, then one relies besides on tacit knowledge, as might be constituted by or derived from the epistemic faculty virtues and abilities of the causal reasoner, the value systems and character traits of the causal reasoner, the implicit knowledge base available to the causal reasoner, and the habits that sustain our causal reasoning practices. While certain aspects of causal reasoning may be axiomatized and formalized and algorithms may be implemented to approximate causal reasoning, one has to remain skeptical about whether causal reasoning may be fully automated. This demonstration will involve an engagement with Meno's Paradox.

Keywords Causal epistemology · Causal reasoning · Automation question · Meno's Paradox

1 Introduction

Are we in a position to fully automate causal reasoning? Given recent groundbreaking advances in causal epistemology, one might be tempted to reply in the affirmative. In this paper, it will be demonstrated how, notwithstanding a number of significant philosophical and computational developments, one still has good skeptical grounds for resisting any conclusions in favour of the automation of causal reasoning. Our critical examination of the central opening question (hereafter: the automation question) will rely on a delineation of these philosophical and computational advances in causal epistemology, a careful articulation of the automation question, Meno's Paradox and Polanyi's treatment of it, and

an application of Polanyi's insights about the nature of tacit knowledge to the automation question.

1.1 Main discussion: philosophical advances in causal epistemology

Before we commence our critical examination of the automation question, an acknowledgment of the state-of-the-art in causal epistemology will first be in order.¹ After all, one could argue that certain philosophical and computational developments have made it plausible (or at least more plausible than hitherto has been the case) to answer the automation question in the affirmative. In the first half of the twentieth

✉ Melvin Chen
melvinchen@ntu.edu.sg

¹ Philosophy, Nanyang Technological University, 50 Nanyang Ave, Singapore S639798, Singapore

² Physics, Nanyang Technological University, 50 Nanyang Ave, Singapore S639798, Singapore

¹ The metaphysics of causation is concerned with the intrinsic nature of causal relationships and causal relata. Causal epistemology, on the other hand, is concerned with how certain methods might be developed from assumptions and operationalized into procedures that enable us to successfully and reliably acquire causal knowledge, learn about causality, make causal inferences, and provide justification for these causal inferences. For the more formally inclined, a lot of the relevant logical and computational notation will be provided in the succeeding footnotes. Nothing will be lost argumentatively, if these footnotes are glossed over in a more general attempt to understand how a skeptical response to the automation question might be motivated.

century, regularity theory was the dominant philosophical approach to causal epistemology. According to the regularity camp, we infer the existence of type-level causal relationships from identified regularities in sequences of event types (Hume 1748; Mill 1843). For example, we infer that C causes E if every event of type C is routinely followed by an event of type E .² Naïve regularity theorists traditionally maintain that C causes E if certain Humean criteria (viz. temporal succession, spatiotemporal contiguity, necessary connection) are satisfied. However, these criteria may not be sufficient for relationships to count as causal in nature. Even though day regularly follows the night, both relata are contiguous, and there is the idea of a necessary connection between day and night, day does not cause the night (Reid 1785). Likewise, even though umbrella vendors are always around before it starts raining, we do not conclude that the presence of umbrella vendors causes rain (Kleinberg 2013). In other instances, these criteria may be unnecessary.³ Even more sophisticated regularity theorists face the problem of distinguishing between type-level causation (i.e. regular occurrences at population level) and token-level causation (i.e. the occurrence of an effect in a particular scenario).

Philosophical advances in the second half of the twentieth century led to the development of the counterfactual and probabilistic approaches to causation and extended our understanding of how we make causal inferences, learn about causality, and acquire causal knowledge. According to the counterfactual camp, we can infer the existence of token-level causal relationships from the use of counterfactual conditionals and possible world semantics (Stalnaker 1968; Lewis 1973). For example, one might learn from the regularity camp that C causes E , since a material implication relation exists at type-level between C and E . When reasoning about a possible world that is sufficiently similar to the actual world in which ' $C \rightarrow E$ ' is true, one might conclude

that if the event-token c had not occurred, then the event-token e would not have occurred.⁴

According to the probabilistic camp, both the regularity and counterfactual camps erroneously hold that causes produce their effects without fail at type- or token-level.⁵ Rather, the causal relationship between variables is probabilistic rather than deterministic in nature: C is a positive cause of E if the probability of E occurring is still (positively) altered by C , after common causes or confounders have been conditioned on (Reichenbach 1956; Suppes 1970; Eells 1991).⁶ It should be noted that these philosophical advances have since been superseded by certain computational advances in the field of computer science and AI research (Pearl 2000; Kleinberg 2013). Nonetheless, while computational methods may or may not be philosophically informed, they tend to rely on certain philosophical assumptions as axioms for developing causal models of reality.

² Formally: $C \rightarrow E$. According to the naïve regularity theorist, C causes E if every event of type C is followed by an event of type E . More sophisticated regularity theorists are keener on describing causal relationships in terms of necessity and sufficiency. According to these theorists, C causes E if C is at least an INUS (viz. insufficient but non-redundant part of an unnecessary but sufficient) condition of E (Mackie 1965; see also Rothman 1976).

³ Some causal relationships may be characterized by a lack of spatiotemporal contiguity: the absence (rather than the presence) of an event could give rise to an effect and an event unfolding in one spatiotemporal region of the world could cause another event in another spatiotemporal region. Some causal relationships may be characterized in terms of a lack of temporal succession of events: physics has discovered simultaneous causal influence.

⁴ Formally: $\sim c \square \rightarrow \sim e$. Following the analysis of the truth-conditions of counterfactual conditionals in terms of possible world semantics and causation in terms of counterfactual dependence, we infer that c causes e .

⁵ Even after the causes have been fully specified and we possess perfect information about the world, we still cannot take it for granted that causes will produce their effects without fail.

⁶ Formally: $P(E|C) > P(E|\sim C)$.

According to the Common Cause Principle, C is causally relevant to E iff (i) C temporally precedes E ; (ii) $P(E|C) > P(E)$; and (iii) there does not exist a set of events S such that C and E are uncorrelated, conditional upon S (i.e. S 'screens off' C from E) (Reichenbach 1956).

According to the two-step approach of Suppes (1970):

Step 1: $X_{t'}$ is a *prima facie* cause of Y_t iff (i) $t > t'$; (ii) $P(X_{t'}) > 0$; and (iii) $P(Y_t|X_{t'}) > P(Y_t)$ ($X_{t'}$ is a *prima facie* positive cause of Y_t) or $P(Y_t|X_{t'}) < P(Y_t)$ ($X_{t'}$ is a *prima facie* negative cause of Y_t).

Step 2: $X_{t'}$ is a *prima facie* but *e-spurious* cause of Y_t iff $\exists t'' < t'$ and $Z_{t''}$ such that: (i) $P(X_{t'} \wedge Z_{t''}) > 0$; (ii) $P(Y_t|X_{t'} \wedge Z_{t''}) = P(Y_t|Z_{t''})$; and (iii) $|P(Y_t|X_{t'} \wedge Z_{t''}) - P(Y_t|Z_{t''})| < \epsilon$.

(iv) Otherwise, $X_{t'}$ is an *e-significant* cause of Y_t .

According to the more sophisticated probabilistic approach of Eells (1991), which takes into account the background contexts or knowledge bases (viz. K_i):

At type level and for each i :

(i) C is a positive causal factor of E iff $P(E|K_i \wedge C) > P(E|K_i \wedge \sim C)$;

(ii) C is a negative causal factor of E iff $P(E|K_i \wedge C) < P(E|K_i \wedge \sim C)$;

(iii) C is a neutral causal factor of E iff $P(E|K_i \wedge C) = P(E|K_i \wedge \sim C)$;

or.

(iv) Otherwise, C is a mixed causal factor of E .

At token level:

ADCS (E, C) = $\sum_i P(K_i) [P(E|K_i \wedge C) - P(E|K_i \wedge \sim C)]$, where ADCS (E, C) is the average degree of causal significance of C for E .

1.2 Main discussion: key philosophical assumptions

A key philosophical assumption that has emerged from these philosophical approaches to causation is the No Backwards Causation assumption. According to this assumption, causes cannot be temporally preceded by their effects.⁷ Another key philosophical assumption is the Alteration of Probabilities assumption, according to which causes alter the probabilities of their effects. Positive causes raise the probability of their effects, whereas negative causes lower the probability of their effects.⁸ Without the Alteration of Probabilities assumption, we will not be able to rely on observed statistical correlations as a guide to type-level causal relationships. Last but not least, the Causal Sufficiency assumption is a key philosophical assumption according to which there are no hidden common causes. Alternatively, the set of measured variables must be causally sufficient. By causal sufficiency is meant the following: for any set of variables V , every common cause (relative to V) of any pair of variables in V is also contained in V . Absent the Causal Sufficiency assumption, and hidden common causes (or confounders) can lead to spurious correlations between causally unrelated variables (or confounding bias) and result in erroneous causal inferences.⁹

⁷ The Temporal Priority assumption is the assumption according to which causes must temporally precede their effects. This assumption rules out *sensu stricto* both simultaneous causation (causes being temporally simultaneous with their effects) and backwards causation (causes being temporally preceded by their causes). The No Backwards Causation assumption, on the other hand, rules out only backwards causation. Both the Temporal Priority and the No Backwards Causation assumptions are heavily influenced by classical physics. Recent work on quantum switches appears however to suggest that at the quantum level, it might not be defined whether an event is a cause or an effect of another event. This renders the causal relationship between the two events fuzzy. See Castro-Ruiz et al. (2018).

⁸ See fn. 6 and the discussion about Eells (1991). The Alteration of Probabilities assumption is a matter of methodological expediency rather than conceptual necessity, since there are instances in which causal relationships exist, without any change in probability. See Hesslow (1976) for a celebrated example of multiple pathways between causal relata that exactly cancel each other out and Cartwright (1989), Hoover (2001), and Zhang and Spirtes (2008) for further critical discussion.

⁹ To be certain, there are other philosophical assumptions that are generally made in causal epistemology. Among the assumptions that are made in support of induction in general are the Uniformity of Nature and *Ceteris Paribus* assumptions. According to the former assumption, the world is structured in a certain orderly way and its underlying causal laws are preserved in an induction-friendly manner. According to the latter assumption, all other things being equal, we are justified in expecting certain patterns between variables in the past to continue into the future. Other assumptions that support causal inference in particular include the Parsimony assumption and the Connection by Mechanism assumption. According to the former assumption, all other things being equal, we are justified in ruling in favour of the most parsimonious causal hypothesis. According to the latter assumption, there must exist a mechanism or set of mechanisms connecting causes to their effects.

Fig. 1 Reid's (1785) Day-and-night example

To illustrate the Causal Sufficiency assumption by means of a toy example, let us return to Reid's (1785) day-and-night example (Fig. 1):

We observe a marginal dependence relation between the relata X and Y : day regularly follows the night and vice versa. If we fail to observe the Causal Sufficiency assumption, we end up omitting the common cause or confounder Z from our set of variables. This will result in our incorrectly inferring, on the basis of the marginal dependence relation between X and Y , a causal dependence relation between X and Y . Conversely, if we observe the Causal Sufficiency assumption, no common causes (including Z) will be omitted. The joint effects of Z (viz. X and Y) will no longer erroneously appear to be dependent. In accordance with Reichenbach's (1956) Common Cause Principle (described in fn. 6), X and Y will be conditionally independent, given their common cause Z (formally: $X \perp Y \mid Z$) (Fig. 1).

1.3 Main discussion: computational advances in causal epistemology

The No Backwards Causation, Alteration of Probabilities, and Causal Sufficiency assumptions are central features of the probabilistic camp and undergird the most prominent computational approaches to causal reasoning. Nonetheless, it is important to bear in mind that philosophical and computational advances in causal epistemology are related though distinct. While they might be guided by the theoretical considerations of philosophical approaches to causation, computational approaches are also characterized by a distinct set of concerns: the ability to translate theory into methods, practices, and procedures, the ability to operationalize definitions about causation and causal relationships, the ability to infer causal relationships from data, and the ability to render feasible the degree of computational complexity of the task.¹⁰

¹⁰ Some philosophical advances are computationally inclined though ultimately computationally infeasible. Consider the more computationally inclined probabilistic approach of Eells (1991) (see fn. 6): if n variables are under consideration, then there will be 2^n background contexts. Testing for all contexts (viz. $K_1, K_2, \dots, K_{2^n}$) to determine $\text{ADCS}(E, C)$ will be a task of considerable computational complexity. For a critical overview of the key differences between the philosophical and computational approaches to causation, see Kleinberg (2013).

Philosophical advances provide theoretical guidance about how causal relationships may be inferred. In turn, computational methods aim to translate theory into computationally feasible methods, practices, and procedures, operationalize general definitions about causation and causal relationships, and infer causal relationships from data.

An early computational advance in the second half of the twentieth century was the Granger Causality approach, which takes two time series, determines whether one time series is useful for forecasting the other, and subsequently uses these determinations as a basis for making causal inferences (Wiener 1956; Granger 1969, 1980). This approach relies on the Temporal Priority assumption (a variant of the No Backwards Causation assumption), the Alteration of Probabilities assumption, and a more method-specific No Redundant Information assumption.¹¹ While the pragmatically-oriented Granger Causality approach is recognized for its ability to operationalize definitions about causal relationships, incorporation of temporal information, ability to handle continuous variables, and applicability across such diverse fields as finance, neuroscience, and physics, there remains much debate about whether Granger Causality is necessary or sufficient for causality (Cartwright 1989; Hoover 2001).

A more recent computational advance has been the Graphical Model approach, which generates a Directed Acyclic Graph (or DAG) from a set of data (Pearl 2000; Spirtes et al. 2000).¹² These DAGs, also known as Bayesian networks, represent the causal structure of the system under investigation.¹³ The Graphical Model approach has been designed to produce one or more graphs to represent the independence relations that are consistent with a given set of data and does not require the incorporation of temporal information, unlike the Granger Causality approach. This approach relies on the Alteration of Probabilities

assumption, the Causal Sufficiency assumption, and a couple of other more method-specific axioms.¹⁴

While the Graphical Model approach has enjoyed much popularity, it is arguably less complete than its more recent computational rival: the Probabilistic Temporal Logic approach of Kleinberg (2013, 2016). Like the Granger Causality approach, Kleinberg's aim is to develop a computationally feasible method for reliably inferring causal relationships from time series data. Kripke structures represent the causal structure of the system under investigation and the DAGs of the Graphical Model approach can be subsumed under these Kripke structures.¹⁵ Guided by the more computationally inclined philosophical considerations of Eells (1991), Kleinberg has developed an ADCS-inspired measure of causal significance to determine the average difference a cause makes to the probability of its effect.¹⁶ Kleinberg's approach relies on the No Backwards Causation assumption, the Alteration of Probabilities assumption, and the Causal Sufficiency assumption.

Nonetheless, certain issues remain. In the first instance, it has been argued that mechanistically mediated effects are hybrids of causal relationships (i.e. between causes and their effects) and constitutive relationships (i.e. between parts and the whole). Consider the causal claim that myocardial infarction is the cause of death. Death is a mechanistically mediated effect: when the heart stops beating, it stops transporting oxygen and nutrients to other tissues of the body, these tissues stop functioning, and the non-functioning

¹¹ The former assumption is a variant of the No Backwards Causation assumption (see fn. 7). According to the latter assumption, all the knowledge in the universe available at time n (denoted by Ω_n) contains no redundant information. Formally, X Granger-causes Y if $P(Y_{n+1}|\Omega_n) \neq P(Y_{n+1}|\Omega_n - X^*(n))$.

¹² While the Graphical Model approach typically draws on the probabilistic theory about causation to make type-level causal inferences from the data, more recent advances have sought the counterfactual camp to cope with token-level causal inferences. See the work of Hopkins and Pearl (2007) on situation calculus.

¹³ DAGs are visual representations of Nonparametric Structural Equation Models (or NPSEMs), which are representations of the causal relationships between variables in the form of equations. Formally, $V = f(pa(V), U_v)$, where V denotes a variable, f denotes an arbitrary deterministic function, and U_v is the idiosyncratic error term for V .

¹⁴ The more method-specific axioms or conditions are (variously) the Causal Markov Condition, the Causal Faithfulness Condition, and the Frugality Condition. According to the Causal Markov Condition (or CMC), any variable x in a variable set V is conditionally independent of its non-effects, given its direct causes or parents. According to the Causal Faithfulness Condition (or CFC), if the causal structure does not entail a conditional independence relation according to the CMC, then the conditional independence relation does not hold of the true probability distribution. The CFC implies that any conditional independence relations that have been identified are due to causal structure, rather than coincidence, unmeasured variables, or other factors. According to the Frugality Condition, given a set of causal hypotheses that are compatible with the probability distribution, one ought to choose from among those hypotheses that postulate the lowest number of causal arrows. Compare the Frugality Condition with the Parsimony assumption, as described in fn. 9.

¹⁵ More specifically, whereas the DAGs in the Graphical Model approach forbid cycles (hence the titular acyclicity), the probabilistic Kripke structures in the Probabilistic Temporal Logic approach naturally admit cycles. With respect to systems that exhibit repeated or homeostatic behaviour and behavioral feedback loops (e.g. biological systems), the wider expressive capabilities of probabilistic Kripke structures would constitute a distinct advantage. Probabilistic Kripke structures are also called Discrete Time Markov Chains (or DTMCs).

¹⁶ Let X denote the set of *prima facie* causes of e and $\epsilon_{avg}(c, e)$ denote the significance of cause c for an effect e . Furthermore, let $\epsilon_x(c, e) = P(e|c \wedge x) - P(e|\sim c \wedge x)$. Formally: $\epsilon_{avg}(c, e) = \frac{\sum_{x \in X} \epsilon_x(c, e)}{|X|}$.

constitutes the death of that individual (Craver and Bechtel 2007). Where causal reasoning in certain domains (e.g. neurobiology, molecular biology) typically necessitates an appeal to mechanisms, the details of scientific practice, and a multi-level character of explanations (both causal and constitutive), skepticism remains about whether our best-going computational models can capture and automate causal reasoning in these domains (Machamer et al. 2000). In the second instance, there is an ongoing debate about whether mental properties can in principle function as efficacious causes of both physical and mental properties. This is a debate about the validity and soundness of causal exclusion arguments, which typically conclude that mental properties supervening on physical properties cannot cause physical or other mental properties (Kim 2003). In addition, if mental properties count as causally efficacious, can the causal efficacy of these mental properties be accounted for on empirical grounds? We still lack a complete understanding of whether and how mental states (e.g. beliefs, thoughts, intentions, etc.) causally interact with one another and with physical states. Where causal explanation in certain domains (e.g. law, the social sciences) appeals to reasons, motives, intentions, desires, and beliefs, skepticism likewise remains about whether causal reasoning can be fully automated in these domains (Knobe 2009). The mechanistic and mental aspects of causal reasoning have not yet been successfully automated, and it is unclear whether they can be subsumed under the best-going computational models.¹⁷

In the third instance, consider various dilemmas that arise in causal reasoning. Does the chicken cause the egg or does the egg cause the chicken? This dilemma is at least as old as Aristotle and may be structured as an infinite regression in causal reasoning. The causal relationship between the chicken and the egg is considered a strange loop in direct relation to Gödel's incompleteness theorem (Hofstadter 1999). In the context of complete causal self-referencing, it is Turing-incomputable insofar that the actual cause is undecidable (Kampis 1995). This chicken-and-egg-type dilemma cannot be resolved by a definite procedure. One may perform a root-cause analysis (or RCA), uncovering the underlying biological mechanisms when determining whether it is the chicken or the egg that is causally prior. However, the complex and non-linear nature of relationships over an evolutionary time-scale fosters enormous difficulties with respect to an RCA approach.¹⁸ Retrodiction of the initial causes of

complex and non-linear systems is impossible due to the intrinsic chaos and complexity of the systems in question (Li and Yorke 1975).

In the fourth instance, the best-going computational models aim to tell us what it is for something to count as a cause or how we can learn of causal relationships through datasets. However, computational models and methods may not be appropriate for all cases. As has been conceded by Kleinberg (2013), all theories of and methods for inferring causality—including Kleinberg's own method—fail to handle at least some scenarios, potential challenges, and counterexamples. The lack of a unified solution has lent weight to philosophical arguments in favour of causal pluralism (Anscombe 1971; Cartwright 2007). If causal pluralism is correct, then there may be several different ways to learn about causal relationships and several methods, each of which may be suitable for different situations. In these (viz. the multi-level character of explanations, the causal efficacy of mental properties, chicken-and-egg-type dilemmas of causal reasoning, the complexity of systems, causal pluralism) and other instances, we maintain that causal reasoning cannot be fully automated and remain skeptical about the existence of a general and universal program capable of causal reasoning in a fully automated sense.

1.4 Main discussion: the automation question

Given these philosophical and computational advances in the second half of the twentieth and first decades of the twenty-first century, we might have good reason to believe that it has become plausible for us to answer the automation question in the affirmative. After all, might we not have good reason to expect further philosophical and computational advances to be made in the remainder of the twenty-first century, advances that could well deliver the hoped-for outcome of fully automated causal reasoning?

One might raise the caveat that handling real-world data is challenging: there might be biases, errors, or noise in the data, there is the ever-growing computational complexity of analyzing multidimensional data to be reckoned with, there could be gaps in the recording of the relevant variables, certain variables might be missing, and our knowledge could still remain incomplete. Suppose for the sake of argument that we encounter no problems with our data. Suppose further that we have in place a sufficiently advanced computational method that is guided by the state-of-the-art in philosophical theory and will allow us to infer causal relationships relative to this set of problem-free data. Are we therefore in a position to fully automate causal reasoning?¹⁹

¹⁷ We are grateful to Judea Pearl for motivating us to develop this portion of our critique.

¹⁸ The order of difficulty involved here is comparable to the order of difficulty associated with the task of inferring a cause-effect relationship between the flap of the wings of a butterfly in Brazil and a tornado in Texas (Lorenz 1963).

¹⁹ It should be stressed that the target of our critique is the aspiration toward full automation. We may distinguish between partial automation and full automation. Whereas the former may be possible (e.g. the development of automatic tools, such as automated theorem-prov-

1.5 Main discussion: Meno's Paradox

To address the automation question, we must proceed by way of a paradox, first introduced in a dialogue by Plato, and a response offered to this paradox by Michael Polanyi. Consider the following argument, adapted from Plato's *Meno* (n.d., 80d-e):

P1: For any p , either one knows that p or one does not know that p .

P2: If one knows that p , then inquiry about p is unnecessary.

P3: If one does not know that p , then inquiry about p is impossible.

C: Therefore, for any p , inquiry about p is either unnecessary or impossible.²⁰

This argument is known as Meno's Paradox. Plato's resolution of this paradox involves the invocation of a theory of recollection. According to Plato's theory of recollection, we begin neither in a state of complete knowledge about p nor in a state of complete ignorance about p . We come to know more about p by recollecting something that was learnt about p in a previous life.²¹ There is much controversy surrounding Plato's resolution of Meno's Paradox via a theory of recollection (or the discovery of innate ideas). While the problem posed by Meno's Paradox is an acute one, potentially throwing into question all inquiry as absurd (i.e. either unnecessary or impossible), an alternative response is required if Plato's theory of recollection proves to be unsatisfactory.

Polanyi's (1966) response is that if all knowledge is explicit (i.e. capable of being clearly stated in the propositional form p), then either inquiry is unnecessary (i.e. you know what you are searching for) or impossible (i.e. you do not know what you are searching for). In other words, if all knowledge is explicit, then Meno's Paradox stands. However, if at least some knowledge is tacit (i.e. non-explicit),

then we come to know more about something by relying on tacit knowledge.²² Tacit knowledge, in turn, consists in the 'intimation of something hidden, that we may yet discover' (Polanyi 1966, pp 22–3). Given the fact that problems exist and discoveries (causal, scientific, etc.) can be made by solving these problems, the search for knowledge is clearly not absurd. The implication, if one accepts Polanyi's (1966) response to Meno's Paradox, is therefore that tacit knowledge exists: we can know things that we cannot yet tell.

One could raise the objection that the connection between the automation question and Meno's Paradox is tenuous and unconvincing at best, insofar that the trade-off between implicit and explicit knowledge has typically been a topic of much AI research. How might causal reasoning be uniquely hampered by this particular trade-off?²³ In response to this objection, we must first distinguish between the nature of causation (ontic) and the nature of our knowledge about causation (epistemic). What Polanyi's resolution of Meno's Paradox suggests is that inquiry into the nature of both causation and our knowledge about causation is neither unnecessary nor impossible, since the original dilemma is false. The nature of causation is neither fully known nor fully unknown. We have at least some knowledge about certain aspects of the ontology of causation (e.g. causal relations may be characterized in terms of events, objects, processes, states of affairs, etc.) and the formal properties of causal relationships (e.g. transitivity, asymmetry, etc.). Likewise, the nature of our knowledge about causation is neither fully known nor fully unknown. We have some tacit knowledge of the nature of our knowledge about causation, which could be made explicit in the process of rational inquiry. We may even come to formalize and axiomatize certain salient aspects of causal reasoning. However, the nature of both causation and our knowledge about causation continues to remain contested. Where this contestation remains, we must conclude that we can know things that we cannot yet tell. Until the contested nature of both causation and our knowledge about it has been removed, our response to the automation question must remain a skeptical one. Last but not least, consider the paradox of self-reference that arises with respect to such statements as 'This sentence is false' and 'All Cretans are liars' (uttered by a Cretan). The ability to resolve paradoxes and decipher statements of this nature involves *inter alia* an ability to discern the implicit and explicit knowledge that

Footnote 19 (continued)

ing programs, that tend to be useful and even amplify the intellect of the theoretician), the latter may well be impossible on grounds that we will proceed to discuss in our investigation of Meno's Paradox. The objection that causal reasoning is already automated (at least to some extent in certain domains) is therefore not one that applies to our position. We are grateful to George Kampis for this comment.

²⁰ Formally:

P1: $P \vee \sim P$ (where ' $\vee$ ' is the logical symbol for exclusive disjunction).

P2: $P \rightarrow Q$.

P3: $\sim P \rightarrow R$.

C: $\therefore Q \vee R$.

This argument is logically valid.

²¹ Plato's theory of recollection is a controversial one, and its proof involves Socrates successfully eliciting the correct response to a geometrical problem from an uneducated slave through the posing of certain leading questions.

²² The distinction between explicit knowledge and tacit knowledge may be usefully compared with the German distinction between *wissen* and *können* and the Rylean distinction between knowing-that and knowing-how (Polanyi 1966, pp 6–7). One could also introduce Russell's (1910) distinction between knowledge by description and knowledge by acquaintance.

²³ We are again grateful to Judea Pearl for raising this particular objection.

is contained within these paradoxes and linguistic tricks. Before a response to the automation question can even be attempted, we must first understand the cognitive processes and causal reasoning procedures that bring about this discernment (Hofstadter 1999). While our understanding of these processes and procedures remains incomplete, we have further reasons to maintain a healthy skepticism when responding to the automation question.

1.6 Main discussion: tacit knowledge

In the context of causal epistemology, any philosophical and computational advances (e.g. in assumptions, general or operational definitions for causal relationships, supporting formal notation, axioms, equations, graphical representations, or data) pertain to the domain of explicit knowledge. Causal reasoning is not exhausted by the explicit knowledge that can be secured by the most robust and appropriate assumptions, philosophical theories, computational methods, and data. Rather, causal reasoning relies on the use of at least some tacit knowledge.

This tacit knowledge is in turn constituted by or derived from the following:

- (A) The epistemic faculty virtues and abilities of the causal reasoner:
  - a Reliable perception;
  - b Reliable memory;
  - c Observation skills;
  - d Procedural knowledge of and competence in how to skillfully conduct experiments, undertake statistical analysis, and imagine counterfactually;
  - e Procedural knowledge of and competence in how to carry out critical evaluation, with a view to developing better theories, methods, and approaches²⁴;
- (B) The value systems and character traits of the causal reasoner²⁵:

²⁶ See the related work of Zagzebski (2003) and Battaly (2008) in virtue epistemology.

²⁴ See the related work of Sosa (1985) and Greco (1993, 2009) in virtue epistemology.

²⁵ One often easily forgets how one's reasoning about causes and their effects is value-laden and interest-driven. Values and interests determine the domains of interest, the range of what counts as causally relevant, and the grade of priority to be given to each problem in the use of intellectual and computational resources. Polanyi (1969) refers to the last-mentioned of these determinations as the 'strategic intuition'.

a The goals, ends, values, interests, and purposes (ends) toward which causal reasoning (means) is directed;

b Epistemic trait virtues and virtuous motivations (e.g. love of truth, avoidance of error, conscientiousness, open-mindedness, intellectual humility)²⁶;

- C The implicit knowledge base available to the causal reasoner:
  - a Implicit background knowledge of experts (e.g. knowledge about possible mechanisms);
- D The habits that sustain our causal reasoning practices:
  - a Heuristics and other cognitive shortcuts developed from experience;
  - b Epistemic practices (e.g. deferring to and trusting epistemic superiors and peer review processes).

If at least some knowledge is tacit in nature and if not all of this tacit knowledge can ultimately be reduced to explicit knowledge, then we must recognize the tacit dimension in causal reasoning.²⁷ This tacit knowledge allows us inter alia to make certain refinements and further determinations (e.g. the relevant domain of application, the level of sophistication required, Kleinberg's ϵ -value threshold as discussed in fn. 16, etc.) within each philosophical theory and computational method, critically evaluate competing theories and methods, recognize certain variables as genuine causes despite the violation of certain axioms, assumptions, and other related parameters, and potentially discover and develop new and alternative theories and methods.

An implication of this is that the idea that causal reasoning can be fully automated rests on a mereological mistake. What AI systems are capable of performing on the causal reasoning front is complex information processing, which is a part of causal reasoning but not the whole of it.²⁸ To assert that causal reasoning can be fully automated is to mistake the part (complex information processing) for the whole (causal reasoning, which presupposes in addition the skills, abilities, faculties, value systems, character traits, and cultural habits that give rise to tacit knowledge). At the same time, one could offer the following olive branch to AI researchers: what is denied is the possibility of full automation rather than that of partial automation (see fn. 19 for the distinction). The commonplace *ex nihilo nihil sunt*, first defended by Parmenides and later installed as a dictum in the AI research tradition, tells us that we get nothing from nothing. While not all assumptions can be made explicit in causal reasoning, that is also true for other domains in which AI has been applied. One must assume something

²⁷ This is ultimately Polanyi's position: the process of formalizing all knowledge to the exclusion of any tacit knowledge is a self-defeating one. See Sen (2009).

²⁸ It should not therefore surprise us that at least some of the AI pioneers (viz. Newell and Simon) disliked the phrase 'AI' *malgré eux* and preferred the term 'complex information processing'. See Nilsson (2010, pp 53–4). Perhaps the time has come for us to call a spade a spade and revive the usage of the term 'complex information processing'.

from the outset to get a formal theory. However, once that assumption is in place, partial automation may be possible relative to our assumptions (both stated and implicit). Causal reasoning may not be that radically different after all from other AI-relevant domains, and one is reminded that while a part of the whole can be automated, the whole itself (e.g. causal reasoning in its broadest and most representative sense) cannot be.²⁹ Finally, it is known according to the Church–Turing thesis that there exist uncomputable functions and uncomputable real numbers that cannot be computed by any algorithm. This constitutes a fundamental mathematical constraint on the issue of full automation. To gain an understanding about the nature of this mereological mistake is to attain wisdom about causal reasoning, where wisdom is a higher rung on the epistemic hierarchy than (in descending order) knowledge, reasoning, data, and information.³⁰

2 Conclusion

To conclude, if by an algorithm for causal reasoning is meant a finite procedure, explicitly stated in pseudo- or actual code, through which causal reasoning is replicated, then an affirmative response to the automation question implies the existence (actual or possible) of such an algorithm.³¹ Such an algorithm might be constructed on the basis of the state-of-the-art theoretical frameworks in philosophy, computational methods in econometrics and computer science, and the data relative to which causal relationships might be inferred. The most sophisticated algorithm would, however, remain strictly in the domain of explicit knowledge.

Given the nature of Meno's Paradox and the possibility of scientific discovery (causal, scientific, etc.), one plausible philosophical response has been to invoke a distinction between tacit and explicit knowledge. If by causal reasoning is meant the entirety of the process through which we discover causal relationships between variables and make use of this knowledge in prediction, explanation, decision-making, and reasoning in terms of counterfactuals, then we have good grounds for maintaining that we do not merely rely on explicit knowledge (i.e. knowledge that is capable of being clearly stated in the propositional form). In addition, we rely on tacit knowledge, as might be constituted by or

derived from the epistemic faculty virtues and abilities of the causal reasoner, the value systems and character traits of the causal reasoner, the implicit knowledge base available to the causal reasoner, and the habits that sustain our causal reasoning practices.

If the foregoing analysis is correct, then the answer to the automation question must be a resounding 'no'. If the foregoing analysis is correct, then any implementation of algorithms that approximate causal reasoning ought to be corroborated where possible with other methods (e.g. the use of background knowledge, the reliance on domain experts, the undertaking of experimental studies and RCTs, etc.), given the probabilistic and defeasible nature of causal reasoning. If the foregoing analysis is correct, then in the event of confusion and uncertainty (e.g. when dealing with complex cases), any final appeal should be made to our traditional storehouses of tacit knowledge, the domain experts themselves.

Acknowledgements We would like to take this opportunity to thank a number of individuals and organizations: Satinder Gill, for her indefatigable good humour and patience; the attendees at the 2019 AI & Society CRASSH Conference (University of Cambridge) for their feedback and encouragement; George Kampis and Judea Pearl, for their invaluable and illuminating remarks; Nanyang Technological University and the Accelerating Creativity & Excellence (ACE) programme, for having the boldness to select a philosopher-physicist duo among the awardees at the inaugural ACE grant call; Quek Wei Liang, Suryadi, and Mark Muthiah, for sharing our aspiration of partially automating (or 'assisting', as is our preferred term) medical diagnosis for better healthcare outcomes; Daniel Huang and Eunice Tan, for having recently joined the fray; and (last but certainly not least) Barbora and Henrik for their constant love, support, and encouragement.

Funding This research is part of a Medical AI project that is generously supported by the intramural Nanyang Technological University ACE Grant (Award Number: NTU-ACE2018-05).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Anscombe GEM (1971) Causality and determination. In: Sosa E, Tooley M (eds) *Causation*. Oxford University Press, Oxford, pp 88–104
- Battaly H (2008) Virtue epistemology. *Philos Compass* 3:639–663
- Cartwright N (1989) *Nature's capacities and their measurement*. Clarendon Press, Oxford

²⁹ Again, we would like to thank George Kampis for this insight.

³⁰ For a neo-Luddite critique of this tendency to conflate epistemic hierarchies within the AI tradition, see Roszak (1986).

³¹ The automation question could also be understood in terms of the feasibility of an adequate formalization that will allow for a parity in performance levels between human professionals and state-of-the-art AI systems in causal reasoning tasks. See Chen (2019) for a discussion of the causality deficit.

- Cartwright N (2007) *Hunting causes and using them: approaches in philosophy and economics*. Cambridge University Press, Cambridge
- Castro-Ruiz E, Giacomini F, Brukner C (2018) Dynamics of quantum causal structures. *Phys Rev X* 8:011047
- Chen M (2019) A tale of two deficits: causality and care in medical AI. *Philos Technol* 33:245–267
- Craver C, Bechtel W (2007) Top-down causation without top-down causes. *Biol Philos* 22:547–563
- Eells E (1991) *Probabilistic causality*. Cambridge University Press, Cambridge
- Granger CWJ (1969) Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37:424–438
- Granger CWJ (1980) Testing for causality: a personal viewpoint. *J Econ Dyn Control* 2:329–352
- Greco J (1993) Virtues and vices of virtue epistemology. *Can J Philos* 23:413–432
- Greco J (2009) Knowledge and success from ability. *Philos Stud* 142:17–26
- Hesslow G (1976) Two notes on the probabilistic approach to causality. *Philos Sci* 43:290–292
- Hofstadter DR (1999) *Gödel, Escher, Bach: an eternal golden braid*. Basic Books, New York
- Hoover KD (2001) *Causality in macroeconomics*. Cambridge University Press, Cambridge
- Hopkins M, Pearl J (2007) Causality and counterfactuals in the situation calculus. *J Logic Comput* 17:939–953
- Hume D (1748) *An enquiry concerning human understanding*. Clarendon, Oxford, pp 134–198
- Kampis G (1995) Computability, self-reference, and self-amendment. *Commun Cogn Artif Intell* 12:91–109
- Kim J (2003) Blocking causal drainage and other maintenance chores with mental causation. *Philos Phenomenol Res* 67:151–176
- Kleinberg S (2013) *Causality, probability, and time*. Cambridge University Press, Cambridge
- Kleinberg S (2016) *Why: a guide to finding and using causes*. O'Reilly Media, Massachusetts
- Knobe J (2009) Folk judgments of causation. *Stud Hist Philos Sci* 40:238–242
- Lewis D (1973) Causation. *J Philos* 70:556–567
- Li TY, Yorke J (1975) Period three implies chaos. *Am Math Mon* 82:985–992
- Lorenz EN (1963) Deterministic nonperiodic flow. *J Atmos Sci* 20:130–141
- Machamer PK, Darden L, Craver C (2000) Thinking about mechanisms. *Philos Sci* 67:1–25
- Mackie JL (1965) Causes and conditions. *Am Philos Q* 2:245–264
- Mill JS (1843) *A system of logic, ratiocinative and inductive*. University of Toronto Press, Toronto
- Nilsson N (2010) *The quest for artificial intelligence*. Cambridge University Press, Cambridge
- Pearl J (2000) *Causality: models, reasoning, and inference*. Cambridge University Press, Cambridge
- Plato (n.d.) *Meno*. Grube GMA (trans), 2nd edn. Hackett Publishing
- Polanyi M (1966) *The tacit dimension*. University of Chicago Press, Chicago, London
- Polanyi M (1969) The creative imagination. *Psychol Issues* 6:53–91
- Reichenbach H (1956) *The direction of time*. University of Los Angeles Press, Berkeley
- Reid T (1785) *Essays on the intellectual powers of man*. In: Brooks D (ed), Pennsylvania State University Press, University Park
- Roszak T (1986) *The cult of information*. Pantheon, New York
- Rothman KJ (1976) Causes. *Am J Epidemiol* 104:587–592
- Russell B (1910) Knowledge by acquaintance and knowledge by description. *Proc Aristotel Soc* 11:108–128
- Sen A (2009) Foreword. In: Polanyi M (ed) *The tacit dimension*. University of Chicago Press, Chicago, London, pp vii–xvi
- Sosa E (1985) The coherence of virtue and the virtue of coherence: justification in epistemology. *Synthese* 64:3–28
- Spirtes P, Glymour C, Scheines R (2000) *Causation, prediction and search*. MIT Press, Cambridge
- Stalnaker R (1968) A theory of conditionals. In: Rescher N (ed) *Studies in logical theory*. Blackwell, Oxford, pp 98–112
- Suppes P (1970) *A probabilistic theory of causality*. North-Holland, Amsterdam
- Wiener N (1956) The theory of prediction. In: Beckenbach E (ed) *Modern mathematics for the engineer*. McGraw-Hill, New York, pp 165–190
- Zagzebski L (2003) The search for the source of epistemic good. *Metaphilosophy* 34:12–28
- Zhang J, Spirtes P (2008) Detection of unfaithfulness and robust causal inference. *Mind Mach* 18:239–271

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.