Mark Kuang

PHIL185N

Profs. Keeley and Scott-Kakures

May 13, 2020

Integration Information Theory and the Causal Exclusion Problem

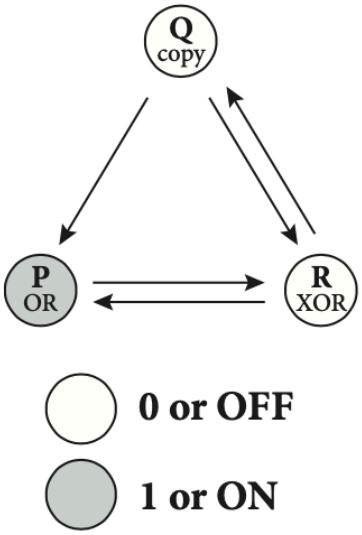
In the *Feeling of Life Itself*, Christof Koch introduces the Integrated Information Theory (IIT) of consciousness. IIT, along with other scientific theories of consciousness, usually claims to be very successful in their explanatory capacities. This paper engages IIT with the mental causation problem form philosophy of mind and philosophy of science and argues that IIT has potential theoretical problems to be resolved. This paper is further divided into three sections. The first section of the paper examines the five postulates of IIT and its central identity. The second section considers the causal exclusion problem for IIT as raised by Baxendale and Mindt (335). The final section evaluates Baxendale and Mindt's (346) interventionist solution of the causal exclusion problem for IIT and argues that it might be ultimately unsatisfactory.

What is IIT

IIT proposes five phenomenological axioms of experience. Any experience exists intrinsically and is structured, specific, unitary, and definite (Tononi et al. 450). From these axioms, five corresponding postulates about its physical substrates are derived, respectively 1) intrinsic existence, 2) composition, 3) information, 4) integration, and 5) exclusion (Tononi et al. 460).

Intrinsic Existence. The first axiom of IIT claims that every experience has an intrinsic, observer-independent existence. Translated to the corresponding proposition about its physical substrate, the intrinsic existence postulate states that the physical correlate of consciousness (PSC) must have "cause-effect power for itself" for the experience to exist intrinsically (Tononi et al. 450). The cause-effect power is characterized by "the extent to which the current state of, say, an electronic circuit or a neural network, causally constrains its past and future states," i.e. make a difference to itself and something else (Koch 79). A neuron in the brain exists intrinsically as it both makes a difference to itself (activate or inactivate) via input and make a difference to something else via output.

Composition. The second axiom of IIT claims that every experience has structure. The subparts of experience constitute several phenomenological distinctions within the experience (Tononi et al. 450). Translated to the corresponding proposition about its physical substrate, the composition postulate states that the subparts of PSC or any physical systems must have cause-effect power within the whole, composing of first-order and higher-order mechanisms, respectively (Tononi et al. 451). Suppose we have a triadic circuit composed of three elementary gates of (P), (Q), and (R) with “three possible pairs of gates, (PQ), (PR), and (QR), and the three-element circuit (PQR)” (82). Now based on IIT, the circuit board can have an integrated information of more than zero due to its intrinsic cause-effect structure. To measure the integrated information, denoted as Φ, of a triadic circuit composed of three elementary gates of (P), (Q), (R), both the first-order elementary mechanism of (P), (Q), and (R) and the higher-order mechanism of (PQ), (PR), and (QR) and the three-element circuit (PQR) has to be measured (Koch 82). These seven mechanisms in the triadic circuit match on to the existence of several phenomenological distinctions, which are part of the composition axiom.



t_n \ t_{n+1}	PQR 000	100	010	110	001	101	011	111
PQR 000	1	0	0	0	0	0	0	0
100	0	0	0	0	1	0	0	0
010	0	0	0	0	0	1	0	0
110	0	1	0	0	0	0	0	0
001	0	0	0	1	0	0	0	0
101	0	0	0	0	0	0	0	1
011	0	0	0	1	0	0	0	0
111	0	0	0	1	0	0	0	0

Fig. 1. The Triadic Circuit from: Christof Koch. *The Feeling of Life Itself: Why Consciousness is Widespread But Can't be Computed*. Mit Press, 2019. Page 83.

Information. The third axiom of IIT claims that every experience has a specific structure. Experience specifies "a particular set of phenomenal distinctions (qualia), which make it what it is and different from other experiences" (Tononi et al. 451). Translated to the corresponding proposition about its physical substrate, the information postulate states that the PSC of experience must have a specific cause-effect power defined as "the set of cause–effect repertoires (φ)" that specifies 'how a mechanism in its current state affects the probability distribution of past and future states of the system" (Tononi et al. 452). In the case of the triadic circuit, the possibility for the (P), (Q), and (R) elementary gates to be ON or OFF specifies the probabilities of possible cause states and effect states, respectively the past and future states of the system constrained by the current state (Koch 84).

Integration. The fourth axiom of IIT—the integration postulate—claims that every experience is unitary and irreducible. The specific structure of experience (information), or the specific set of phenomenal distinctions, is irreducible to "independent, noninteracting components without losing something essential" (Koch 85). Translated to the corresponding

proposition about its physical substrate, the integration postulate states that "the cause–effect structure specified by the PSC must be ... irreducible to the cause–effect structure specified by non-interdependent subsystems" (Tononi et al. 452). The integration postulate entails "the irreducibility of each cause-effect repertoire and the irreducibility of relations among overlapping cause-effect repertoires" (Tononi et al. 452). In the case of the triadic circuit, the integrated information Φ of the circuit cannot be reduced to all possible partitions and decompositions of the circuits (Koch 85-86).

Exclusion. The final and fifth axiom of IIT claims that every experience is definite. The specific structure and phenomenal distinctions that constitute the information is definite, nothing more or nothing less. The spatio-temporal grain of the experience is also definite (Tononi et al. 452). Translated to the corresponding proposition about its physical substrate, the exclusion postulate states that "the cause–effect structure ... must specify a definite set of cause–effect repertoires over a definite set of elements ... at a definite spatio-temporal grain" (Tononi et al. 452). The maximally irreducible integrated information (Φ^{max}) of the specific cause-effect structure over a definite set of elements and spatio-temporal grains "stipulates that only the circuit that is maximally irreducible exists for itself, rather than any of its supersets or subsets. All overlapping circuits with smaller values of Φ are excluded" (Koch 86). In the case of the triadic circuit, the whole system (PQR) is maximally irreducible to any of its subsets of seven mechanisms. The cause-effect repertoires, called concepts, are the subsets of the maximumly irreducible cause-effect structure, called the "conceptual structure" (Tononi et al. 452).

Taken together, the central identity of the IIT is the identity between the experience and the conceptual structure specified by its PSC. The identity is two-fold: "the quality or content of consciousness is identical to the form of the conceptual structure specified by the PSC, and

the quantity or level of consciousness corresponds to its irreducibility (integrated information Φ)" (Tononi et al. 460). Experience is identical to the maximally irreducible cause-effect structure under a definite information and spatio-temporal grain. Koch makes it clear that IIT asserts an ontological claim of the identity between experience and the conceptual structure (88). Given the five postulates and the central identity, IIT does seem to be a pretty complete and well-formulated theory of conscious experience. However, IIT is still subjected to the causal exclusion argument, which might inflict heavily on the established axioms and postulates.

The Causal Exclusion Problem

Like any position, the integrated information theory is must confront the mental causation problem, as it postulates its central identity between the experience and conceptual structure. The most prominent version of the causal exclusion is captured in the work of Jaegwon Kim. Given that the causal exclusion argument is sufficiently well-known, here I present a summary of Kim's causal exclusion argument by Baxendale and Mindt. Then, I will focus on the critique and solutions of IIT proposed by Baxendale and Mindt and points out their respective problems.

The first principle is that the mental supervenes on the physical (Baxendale & Mindt 335). If M_1 and M_2 are two mental states and they have the corresponding physical bases P_1 and P_2, then M_1 and M_2 are said to supervene on P_1 and P_2. The supervenience relationship should be understood as "P_1 is a realizer of M_1 but not identical with it and P_2 is a realizer of M_2 but is not identical with it" (Woodward 2).

The second principle is the principle of causal exclusion. For any phenomenon, "there can be no more than one distinct cause that is wholly responsible for the occurrence of that event, apart from in cases of 'genuine' over-determination ... whereas the over-determination is not systematic" (Baxendale & Mindt 335). For P_2, there might be two causes M_1 or P_1, but there has to be one distinct cause for P_2.

The causal exclusion can be satisfied by the third principle of causal closure. According to Baxendale and Mindt, "if a physical event has a cause, then it has a physical cause" (335). For P_2, the cause has to be P_1 instead of M_1. Now given that M_1 and M_2 has supervenience bases P_1 and P_2, the instantiation of P_1 and P_2 "necessitate M_1 and M_2's occurrence at time t" (335). Now M_2 seems to have two causes, respectively M_1 and P_2. The cause must be P_2 "because regardless of the occurrence of M_1, the very occurrence of P_2 necessitates the instantiation of M_2, thus the role of M_1 seems superfluous" (Baxendale & Mindt 335-336). Now if we consider both the relation between M_1 and M_2 and the relation between M_1 and P_2, neither time does M_1 has a potential causal role, leaving the status of M_1 epiphenomenal (Baxendale & Mindt 336).

Given the above description of causal relations, we can now ask the question whether IIT has the causal exclusion problem or not. According to Baxendale and Mindt, "the global experience (E_1) of a system at a particular time" can be identify as the mental properties such that $M_1 = E_1$ (Baxendale & Mindt 337). The physical correlates of consciousness (PSC_1) are identical to P_1 in Figure 2. The analogous inference leads to the following diagram about the relationship between E_1 and PSC_1:

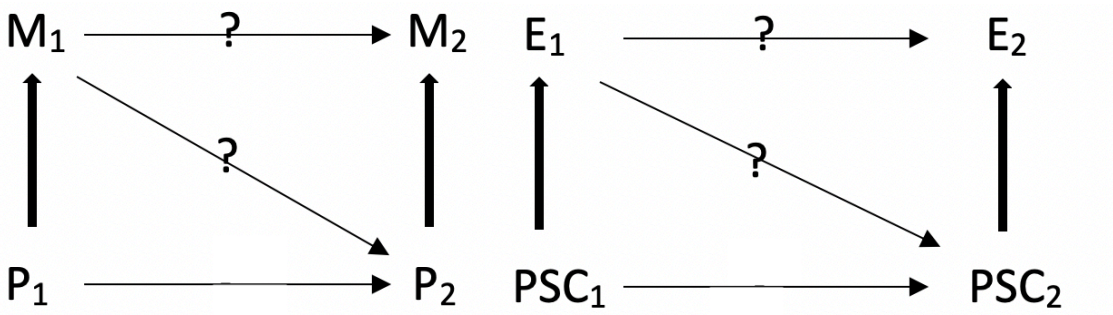


Fig. 2. The Exclusion Problem **Fig. 3.** The Exclusion Problem for IIT

Fig. 2. & Fig. 3. P_1, P_2, PSC_1, and PSC_2 are physical correlates and possess causal relationship, shown by the arrow. The arrow with a question mark might or might not indicate a causation. The thick arrow represents the supervenience relationship.

Now if this analogy holds, E_1 is epiphenomenal at best and the central identity between the phenomenological properties of experience and the cause-effect structure specified by PSC for IIT does not hold (Baxendale & Mindt 337).

There is at least one problem with Baxendale and Mindt's analogy. First, the first principle of supervenience might not work for IIT. supervenience for non-reductive physicalism is that "P_1 is a realizer of M_1 but not identical with it and P_2 is a realizer of M_2 but is not identical with it" (Woodward 2). The central identity of IIT is not supervenience. Whether or not identity entails supervenience is a topic for another day.

In any case, if we accept the first principle of supervenience, the problem of causal exclusion does seem to be genuine for integrated information theorists. IIT accounts for the intrinsic nature of experience by appealing to the cause-effect conceptual structures specified by the PSC. There is indeed a causation from PSC_1 to PSC_2, specified as the intrinsic cause-effect power. The causal status of E_1 and E_2 is nevertheless left as epiphenomenal. There might not be a causation from E_1 to E_2. Instead, the conscious experience will be an epiphenomenon with no cause-effect power upon itself under the causal exclusion argument.

If so, the first axiom of intrinsic existence is in serious jeopardy. A new account of causation will be needed to resolve the causal exclusion problem and its consequence on IIT's axioms.

The Interventionist Account of Causation

The problem of causal exclusion is not a new problem for physicalism. The rejection of nonreductive physicalism brings up the explanatory exclusion of higher-level psychological characterizations. Polger and Shapiro argue that an interventionist account of causal inference circumvents the problem by allowing for explanatory pluralism (206). Interventionalism is a theory of causal inference that allows for different explanations of the same phenomenon as long as different variables act as different interventions in a causal relation. Such plurality challenges the second principle of the causal exclusion argument, i.e. one event has only one distinct cause.

In this paper, rather than focusing on methodological pluralism, I focus on the resolution of causal exclusion under the interventionist account provided by Baxendale and Mindt. An event X is said? to cause an event Y if an intervention variable I:

I_1. I causes X

I_2. I acts as a switch for all other variables that cause X, such that X ceases to depend on the values of any other variables.

I_3. Any directed path that goes from I to Y goes through X.

I_4. I is statistically independent of any other variable Z that causes Y and that is on a directed path that does not go through X. (Woodward 98-100)

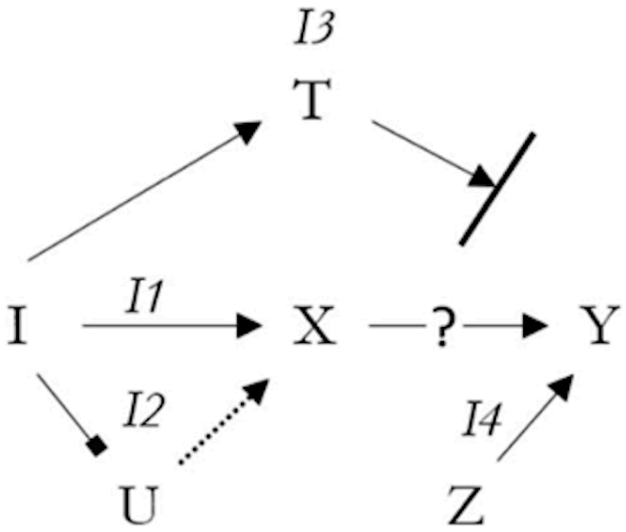


Fig. 4. Interventionalist account of causal relation. In the above figure, I_1 is the direct cause to X while in the same time cut off the dependence of X on U via I_2. I_3 variable cause some event T without affect either X or Y. I_4 is exogenous with respect to Y and is controlled. If such a condition holds, then there is a genuine causation from X to Y, exemplified by the change in Y due to the change in X.

The potential problem of the Interventionist account rests on the assumption that I_2 can cause X to cease to depend on U, in this case X could be the E_1 and U can be the PSC_1 as a physical substrate (Baxendale & Mindt 343). In other words, it is impossible "it is impossible to change M_1 (X) in a way that it is statistically independent of P_1 (U)" (Baumgartner's argument as cited in Woodward 5). The thrust of the matter still lies in the causal status of the supervenience relationship between E_1 and PSC_1.

One response Woodward has is "one should not control for (condition on) supervenience bases in assessing the causal efficacy of supervening properties" (9). In actual scientific practice,

> if scientists were to control for supervenience bases when attempting to establish causal relations between variables through experimental interventions, the results

> would have absurd consequences such that scientists would not be able to establish causal relations between variables representing anything other than fundamental properties (Baxendale & Mindt 344).

Woodward adds that "A full account of this sort (the causal efficacy of supervening properties) is beyond the scope of this essay but let me suggest a few principles which will play a role later in my discussion" (9). This problem is ultimately unresolved at least in Woodward's paper.

In any case, if we assume that the first principle of supervenience holds for Kim's causal exclusion argument, and if we also assume that X (E_1) can cease to be dependent on U (PSC_1), then it does seem X (E_1) can genuinely cause Y (E_2) in some cases. However, the real tension is between how both the principle of supervenience and the denial of supervenience —X's ceasing to be dependent on U—can hold at the same time. It might ultimately be a paradox as the principle of supervenience cannot be true and not true at the same time. Statistical causal modeling might resolve the problem by making X statistically independent of U, but that is a topic beyond the scope of this paper.

Conclusion

In this paper, I lay out the five postulates and the central identity of the Integrated Information Theory (IIT). I argue that IIT has the causal exclusion problem if the first principle of supervenience holds. The interventionist solution of causal exclusion provided by Baxendale and Mindt might be self-defeating, as the principle of supervenience cannot hold and not hold at the same time. More details on statistical causal modeling may need to be worked out for IIT to defend itself from the causal exclusion argument.

Now if we accept the causal exclusion problem for IIT, Koch's five axioms (intrinsic existence, composition, integration, information, exclusion) derived from phenomenology do seem to be genuinely threaten. If consciousness is an epiphenomenon, then every experience is not observer-independent as the first axiom of intrinsic existence suggests. In the last chapter of his book, Koch claims that "the experience is not identical to the Whole. My experience is not my brain" (163). Given that the status of conscious experience might be epiphenomenal, it is hard to see how the mental can have intrinsic cause-effect power upon itself. The first and the most important axiom of IIT is then left unresolved.

Integration Information Theorists now have several possible moves. The first move is to reject the first principle of supervenience. If the first principle of supervenience does not hold, the causal exclusion argument offered by Kim fails, as supervenience is a necessary assumption. Also, the denial of supervenience leave open the possibility for genuine interventionalist causation. However, the burden of proof would be on IIT theorists' shoulder to demonstrate identity without supervenience. The rejection of supervenience principle is built into the assumption of the first axiom of intrinsic existence which claims for the cause-effect power for every experience upon itself, including the cause-effect power upon its physical correlate. The job is then to set out to prove the first axiom of IIT, which is actually built into the assumption already. There is immediately a problem of circularity in IIT, and it should be attributed primarily to the seemingly self-evident but nevertheless potentially problematic phenomenological axioms.

The second move is to reject the second principle of causal exclusion of Kim's argument, i.e. every event has one distinct cause. This move is identical with the first move of the denial of supervenience. By denying supervenience, it is possible to change M_1 (X) in a way that it is independent of the supervenience basis P1 (U). Unless there are more effective

causal models than the interventionalist account, IIT theorists have to face the same problem as the first move.

The third move is rather not preferable, which is to reject the third principle of causal closure. This would require challenging physicalism altogether, which is an ambitious theoretical project. The possible ways out are therefore 1) revising the five phenomenological axioms and 2) proposing better models of mental causation. For Koch to properly title his book as the *Feeling of Life Itself*, it would require the resolution of the causal exclusion problem. As Koch asserts "the central role of feeling to a lived life" in the last paragraph of his book, it does seem urgent to resolve the epiphenomenal status of conscious experience, as experience cannot be central if it is already nonessential.

Work Cited

Baxendale, Matthew, and Garrett Mindt. "Intervening on the Causal Exclusion Problem for Integrated Information Theory." *Minds and Machines* 28.2 (2018): 331-351.

Koch, Christof. *The Feeling of Life Itself: Why Consciousness is Widespread But Can't be Computed*. Mit Press, 2019.

Polger, Thomas W., and Lawrence A. Shapiro. *The Multiple Realization Book*. Oxford University Press, 2016

Tononi, Giulio, et al. "Integrated Information Theory: from Consciousness to its Physical Substrate." *Nature Reviews Neuroscience* 17.7 (2016): 450-461.

Woodward, James. "Intervening in the Exclusion Argument." (2014)

Woodward, James. *Making Things Happen: A Theory of Causal Explanation*. Oxford university press, 2005.