The Invariance Trap

How Conceptual Transfers Generate Philosophical Errors

Wolfgang Stegemann

Abstract

Philosophical fallacies arise not only from false premises or ambiguous words. They can also arise when an expression, a function, a structure, or a description is transferred from one context to another. In such transformations, something is typically preserved while something else changes. This paper calls a property invariant relative to a particular transformation when it is preserved across that transition. Depending on the case, what remains invariant may be a linguistic expression, a semantic function, a material structure, a formal relation, or some other feature. The central error occurs when the preservation of one dimension is taken, without independent justification, to establish the preservation of another. This is termed unwarranted invariance transfer. The approach is developed through six types of transformation: expression and meaning, semantic function and causal realization, levels of description, perspectives of attribution, model and object, and metaphorical transfer. The analysis does not claim that every such case is a category mistake in Ryle's classical sense. Rather, it identifies a general inferential form through which equivocations, level confusions, category mistakes, and unwarranted ontological attributions can arise. Functionalism, the Free Energy Principle, and Integrated Information Theory serve as critical test cases for locating the burden of justification and analyzing concrete invariance errors.

1. The Basic Problem

Science and philosophy depend on transfer. Concepts are abstracted, models are applied to new domains, mathematical structures are identified across different systems, and functions are described independently of their concrete realization. Without such operations, neither theory construction nor technical design would be possible. The same word can be used in different disciplines, the same mathematical operation can be realized in different materials, and the same structure can occur in different models.

This very productivity, however, creates a systematic risk. A transfer is rarely complete. Something remains the same while something else changes. The expression "information" may remain the same while its meaning changes. In computation, by contrast, the mathematical operation may remain the same while the material and causal organization change. A model may preserve the same formal relation while the status of the claim changes when one moves from the model description to the object itself.

The approach proposed here begins with this relation between preservation and change. It does not ask primarily whether a concept has been defined correctly, but what actually remains the same in a particular transfer and what does not. This yields a simple guiding question: Which equivalence is secured by the transformation, and which additional equivalence is merely carried along with it?

The central thesis is this: a conceptual or theoretical transformation can preserve certain properties without preserving others. The preservation of one property therefore does not entail the preservation of another. When that additional step is nevertheless made without an independently justified bridge, an unwarranted invariance transfer occurs.

2. Expression, Concept, and Meaning

The analysis must first distinguish linguistic expression, concept, and meaning. This distinction is necessary because the formulation "the same concept with a different meaning" already presupposes a particular theory of concepts. The present approach is intended to avoid that prior commitment.

By expression I mean the linguistically identifiable form, for example the word "information". By meaning I mean the relevant content that this expression acquires in a particular theoretical or practical context. "Concept" is used below as an umbrella term for a conceptual unit or conceptual use, not as a formally identical entity that must remain unchanged independently of shifts in meaning. The formal variable for linguistic identity is therefore not C for concept, but E for expression.

This caution is especially appropriate in light of Wittgenstein and Ryle. For a large class of cases, Wittgenstein ties meaning to the use of a word (Wittgenstein 1953, § 43). Ryle emphasizes the "systematic ambiguity" of natural language: the same expression can convey ideas of different logical types in different contexts and thereby facilitate type transgressions (Ryle 1938; Ryle 1945).

For the simplest case, we can therefore write:

$$(E_1, M_1) \rightarrow (E_2, M_2)$$

If the same expression is used in two contexts, the following may hold:

$$E_1 = E_2, \qquad M_1 \neq M_2$$

There is no contradiction here. What has remained the same is the linguistic form, not necessarily the concept in a strong intensional sense. This distinction prevents the theory from reproducing its own target error in its notation.

3. Transformation and Invariance

The term invariance is used here in a deliberately weak and relational sense. A property is not invariant in itself. It is invariant relative to a particular transformation if it is preserved across that transition. Invariance therefore initially means nothing more than remaining the same under a specified change.

The example of computation makes this especially clear. If the same mathematical operation is realized once neurally and once electronically, the semantically specified computational operation may be preserved while the material and causal organization change. In this transition, the computational operation is invariant, whereas the material realization is not. In another transformation, the material might remain invariant while the description or function changes.

Invariance is therefore not a privileged semantic, material, or formal property. Semantic, material, structural, or linguistic features can be invariant or variant depending on the transformation. What matters is which dimension is actually preserved.

For the transitions examined here, a conceptual configuration can be analyzed along several dimensions. The following notation does not claim to constitute an autonomous formal calculus or a proof-theoretic derivation. It serves as a formal paraphrase that makes explicit which dimension is preserved in a given case and which changes:

$$K = \langle E, M, F, C, L, P, O \rangle$$

Here E denotes the linguistic expression, M the meaning, F the function or operation, C the causal or material realization, L the level of description, P the perspective of attribution, and O the theoretical or ontological status. These dimensions are separated analytically without thereby asserting their ontological independence. Function, meaning, and causal embedding in particular may be closely related. The purpose of the separate notation is precisely to make visible which of these dimensions an argument treats as preserved.

A transformation is then a transition:

$$T: K_1 \to K_2$$

For each dimension d, one can ask whether it is preserved under T according to the criterion appropriate to that dimension:

$$\mathrm{Inv}_T(d) \Leftrightarrow d_1 \sim_d d_2$$

The symbol ∼ does not denote a single relation of equality, but the dimension-specific criterion of preservation. For E, this may be identity of the same linguistic expression type; for M, sufficient semantic equivalence in the specified context; for F, functional equivalence with respect to a particular operation or role; for C, the relevant equivalence of causal or material organization; for L, the same level of description; for P, the same perspective of attribution; and for O, the same theoretical status. In structural transfers, the relevant criterion may, for example, be isomorphism. The definition therefore does not determine in advance what "the same" means in every case. It requires the relevant criterion of preservation to be made explicit.

Invariance is thus not understood as absolute identity. A function may be invariant even though its material realization changes; a structure may be invariant even though the domain of objects is different. What matters is that the equivalence criterion being used is fixed before the inference is made and is not selected retrospectively so as to guarantee the desired transfer.

4. When Is an Invariance Transfer Unwarranted?

The mere observation that a property has not been demonstrated must not place every scientific abstraction under general suspicion. Science must be able to work with justified transfers. The theory therefore needs a criterion for determining when the move from an established invariance to a further invariance is legitimate.

An additional invariance is justified when its preservation follows from the definition of the transformation or is supported by an independent bridge justification. Such a bridge may rest on empirical evidence or theoretical derivation. It must, however, be justified independently of the invariance it is meant to establish, it must not already presuppose that invariance, and it must be relevant to the transformation under consideration. Otherwise any disputed transfer could be rescued by an ad hoc bridge.

The invariance of one dimension x does not by itself entail the invariance of another dimension y:

$$\mathrm{Inv}_T(x) \quad \text{does not by itself entail} \quad \mathrm{Inv}_T(y)$$

A legitimate transition requires a bridge justification R that specifically establishes the connection between the two invariances:

$$\mathrm{Inv}_T(x) \wedge R_T(x, y) \Rightarrow \mathrm{Inv}_T(y)$$

An unwarranted invariance transfer therefore occurs when the invariance of y is required or asserted in an argument even though only the invariance of x has been justified and an independent, non-circular bridge between them is missing. The error is not that y could not also be invariant. The error is that its invariance is inferred from a different invariance that does not support that conclusion without additional justification.

The theory is therefore fallibilist. A transfer that is initially unjustified may later be legitimized by empirical or theoretical work. The analytical model does not automatically evaluate the truth of a bridge. It locates the point at which additional justification is required and excludes only circular or merely stipulated bridges as justification.

5. Category Mistakes, Equivocation, and Fallacy

The term category mistake must therefore be used with care. Ryle's classical category mistake concerns a confusion of logical or grammatical types. More recent discussion shows that category mistakes themselves can be analyzed in semantic, syntactic, or pragmatic terms; Magidor (2013) provides a systematic reassessment. Not every unwarranted invariance transfer has this form. Some cases are equivocations, others are non sequitur-like inferences, level confusions, or unwarranted ontological attributions.

Unwarranted invariance transfer should therefore not be introduced as a synonym for category mistake. It denotes a more general inferential form through which category mistakes can arise. In a level shift, for example, a property of the organism may be attributed to a part of the organism as if the same logical agent were involved. In a model case, a property of the description may be treated as a property of the object. In both cases the result may be a category mistake even though the underlying transition differs.

The objection that the approach merely redescribes equivocation applies only to part of the typology. Type I can indeed be analyzed as systematic equivocation or polysemy. Type II, by contrast, presupposes that the semantic function remains the same. Type III concerns the level of description, Type IV the perspective of attribution, Type V the status of model and object, and Type VI the scope of a structural analogy. Their common structure does not consist in a shift of meaning, but in the apparent carrying over of a further, not yet established equivalence by a dimension that is genuinely preserved.

5.1 Claim to Novelty

What is new is not the logical insight that one property does not entail another without further premises. Nor is the bridge justification presented as a new general theory of valid inference. The distinctive contribution lies in the systematic analysis of conceptual transformations in terms of which dimension is preserved, which changes, and which additional equivalence an argument requires.

This makes comparable several forms of error that are usually treated separately: equivocation, multiple realizability, level shifts, perspective shifts, model transfer, and metaphor. The typology connects these cases not by reducing them to the same semantic error, but by making visible in each case the preserved dimension, the changed dimension, and the point at which additional justification is required. The formalization serves to represent this structure unambiguously.

6. A Typology of Conceptual Transformations

The following typology distinguishes six basic forms of conceptual transformation. It does not claim deductively demonstrable completeness, but it captures the principal types of transition that can be distinguished in the philosophical arguments relevant here. An additional type would be required if a relevant case could be assigned neither to one of these forms nor to a combination of them. The types are therefore analytically distinguishable but not necessarily exclusive. A single argument may, for example, involve both a shift of level and a shift of perspective.

Type	Invariant	Changed	Additional inference / required bridge	Example
I Expression / meaning	Expression	Meaning	Same expression ⇒ same meaning; semantic equivalence would be required.	Information
II Realization	Function	Material, causality	Same function ⇒ same further property; a theory connecting F with that property is required.	Calculation
III Level	Expression / relation	Level of description	Level shift ⇒ same agent or content; a justified relation between levels is required.	Person / brain decides
IV Perspective	Goal relation	Perspective of attribution	Externally assigned goal ⇒ intrinsic goal or relevance; a theory of intrinsic goal relations is required.	Algorithm pursues a goal
V Model / object	Formal relation	Theoretical status	Model relation ⇒ property of the object; a justified model-object bridge is required.	Prediction, Free Energy
VI Metaphor	Partial structure	Domain of objects	Analogy ⇒ further identity; evidence is required that the additional property is transferred as well.	Code, memory

7. Type I: Same Expression, Different Meaning

The first type is the most familiar from philosophy of language. The same expression is preserved while its meaning changes. "Information" is a particularly suitable example. In Shannon's mathematical theory of communication, information is formalized in terms of probabilities or possibilities of selection; semantic aspects of the message do not belong to the technical problem (Shannon 1948). In biological or cognitive contexts, by contrast, "information" may denote semantic, functional, or organismic relations.

$$E_1 = E_2, \qquad M_1 \neq M_2$$

The error does not consist in using the same expression. Ambiguity and shifts of meaning are normal features of natural and scientific language. The transition becomes problematic only when a property of the first meaning is transferred to the second meaning solely on the basis of identity of expression.

This type is close to classical equivocation and to Ryle's analysis of systematic ambiguity. The theory of invariance transfer therefore does not claim a new discovery here, but places this familiar case within a broader schema.

8. Type II: Same Function, Different Causal Realization

The second type is more important for functionalism and computationalism. A human being can calculate, and so can an electronic computer. For a clearly defined mathematical operation, the semantic function can indeed remain the same. Two plus two equals four regardless of whether the operation is realized by neural activity, silicon circuitry, or written symbol manipulation.

$$\mathrm{Inv}_T(F), \qquad C_1 \neq C_2$$

The transfer here is precisely not an equivocation. The relevant operation can be semantically invariant. Its causal and material realization is not. It follows legitimately that this function is multiply realizable. The classical debate over multiple realizability made this possibility a central argument against simple psychophysical type identities (Putnam 1967; Block and Fodor 1972).

It does not follow that differences in realization are irrelevant to every further property. Anyone who wants to infer the equivalence of an additional property Y from functional equivalence requires a bridge:

$$\mathrm{Inv}_T(F) \wedge R_T(F, Y) \Rightarrow \mathrm{Inv}_T(Y)$$

Without this bridge, only the function has been transferred. This is why the thesis "the same function can be realized differently" must be distinguished from the stronger thesis "all properties relevant to Y are already captured by the functional description".

9. Type III: Shift of Level of Description

In the third type, the linguistic expression or a relation often remains the same while the level of description changes. We say, for example, "The person decides" and "The brain decides". Both statements may be intelligible in certain contexts. They are not automatically equivalent in meaning.

At the level of the organism, "deciding" may denote an action among possibilities. At the neurophysiological level, neuronal processes involved in that action are described. If the shift of level is ignored, an alternative description may appear to introduce a second acting instance. The error does not consist in discussing the neural prerequisites of decisions, but in transferring the logical role of the organism concept unchanged to a partial description.

This case directly touches on the mereological fallacy diagnosed by Bennett and Hacker (2003): psychological predicates that are meaningfully ascribed to the whole living being or person are attributed to a part, such as the brain. The analysis proposed here is somewhat more general. It additionally asks which meaning and logical role of a predicate are actually preserved when moving from an organism-level description to a neurophysiological description. The mereological fallacy thus appears as a particularly important case of an insufficiently justified level transfer.

This case shows why the theory extends beyond pure semantics. Even if both statements are acceptable in everyday language, a philosophical argument must clarify which properties of the action level are preserved at the neurophysiological level and which are merely redescribed.

10. Type IV: Shift of Attributional Perspective

A fourth type arises when a relation defined from the perspective of an observer or designer is treated as an intrinsic property of the system. An algorithm can have an objective function. For the designer, it is specified which state is to be maximized or minimized. This can easily give rise to the statement that the algorithm "pursues a goal".

The formal goal relation may remain invariant. What has changed, however, is the perspective of attribution. A goal for the designer becomes a goal for the system. In technical descriptions this shorthand may be harmless. Philosophically, it becomes relevant as soon as intrinsic relevance, meaning for the system, or even intentionality is derived from it.

$$\mathrm{Inv}_T(\text{goal relation}) \quad \text{does not by itself entail} \quad \mathrm{Inv}_T(\text{intrinsic relevance})$$

Here too, an additional theory could in principle justify the bridge. The present approach does not exclude that possibility. It requires only that the shift in perspective not itself function as the justification.

11. Type V: From Model to Object

Scientific models contain quantities, relations, and operations used to describe and predict observed processes. The success of a model concept, however, does not by itself make that concept a property of the modeled object in the same sense. The transition from model to object changes the theoretical status of a claim.

Predictive Processing provides an illustrative field. Sprevak and Smith explicitly distinguish computational, algorithmic, and implementation levels and describe the relations among predictive coding, Bayesian inference, and variational free energy (Sprevak and Smith 2023). This differentiation itself shows that moving from a formal quantity to a claim about neural implementation requires an additional explanatory step.

Here the analytical model takes the following form:

$$\mathrm{Inv}_T(\text{formal relation}) \quad \text{does not by itself entail} \quad \mathrm{Inv}_T(\text{ontological status})$$

A formal relation may be preserved across the transition while the question changes as to whether it is understood as a description, a mechanism, or an intrinsic property of the object.

12. Type VI: Metaphorical Transfer

Metaphorical transfer typically preserves a partial structure while changing the domain of objects. Terms such as "code", "memory", "network", or "decision" can thereby be productively applied to new domains. Theories of conceptual metaphor have systematically described such mappings between source and target domains (Lakoff and Johnson 1980).

A metaphor, however, does not assert complete identity between the domains. Only the structure relevant to the particular transfer is preserved. The error arises when further properties of the source domain are carried over without additional justification.

$$\mathrm{Inv}_T(S) \quad \text{does not by itself entail} \quad \mathrm{Inv}_T(Q)$$

Here S denotes the structure actually transferred and Q a further property of the source domain. A metaphor can thus be heuristically successful and ontologically misleading at the same time.

13. Functionalism as an Application

Functionalism is not a single unified position, and the error described here should not be attributed to it wholesale. Analytic functionalism, machine-state functionalism, computationalism, and non-reductive variants differ in their theoretical commitments. The relevant point therefore concerns not "functionalism" as such, but a particular argumentative use of functional abstraction.

Functions can be abstracted from a biological organism. Learning, calculating, classifying, storing, or controlling behavior can be described in ways that initially bracket their concrete material realization. If corresponding functions are technically realized, invariance has thereby been demonstrated for the defined function:

$$\mathrm{Inv}_T(F)$$

At the same time, the following may hold:

$$C_1 \neq C_2$$

The legitimate conclusion is that the function in question can be realized on different material bases. The stronger conclusion that a further property Y is therefore also independent of realization requires additional justification.

This distinction is decisive for the debate on consciousness. From the fact that a biological and a technical system realize the same computational operation, classification performance, or learning function, it does not yet follow that experience is also among the invariant properties of this transformation. Anyone who makes that claim must show why the conditions relevant to experience are fully captured by the functional abstraction.

The criticism is therefore not directed against functional description. It is directed against the unjustified transition from demonstrated functional equivalence to a more extensive ontological or phenomenal equivalence.

14. Test Case: The Free Energy Principle

For the present analysis, the Free Energy Principle is not merely a neutral test case. It shows particularly clearly how a formally correct relation can, through several conceptual transformations, acquire the status of a biological explanation. Friston defines surprisal in information-theoretic terms as the negative log probability of an outcome and Variational Free Energy as a quantity that upper-bounds surprisal (Friston 2010). At this stage, these are relations within a probabilistic model.

$$\mathrm{Surprisal}(o) = -\ln p(o)$$

Variational Free Energy can be decomposed such that its difference from negative log evidence is a Kullback-Leibler divergence. Because this divergence is non-negative, Variational Free Energy provides an upper bound on surprisal:

$$F[q] = D_{\mathrm{KL}}\big(q(s) \parallel p(s \mid o)\big) - \ln p(o) \geq -\ln p(o)$$

Up to this point there is no category mistake. The critical step begins when this formal relation becomes a claim about how an organism causally maintains its existence. The fact that the stationary states of a self-organizing system can be described probabilistically as if a variational quantity were minimized initially shows a property of the model description. It does not yet show that the organism minimizes Variational Free Energy as a physically efficacious quantity or that such minimization explains the mechanism of its self-organization. At precisely this point, the theoretical status shifts from a formal description to a claim about a biological mechanism.

$$\mathrm{Inv}_T(\text{formal model relation}) \quad \text{does not by itself entail} \quad \mathrm{Inv}_T(\text{biological mechanism})$$

This distinction corresponds to the problem Williams (2022) develops as the difference between a descriptive and an explanatory Free Energy Principle. The descriptive reading can specify how persistent systems may be formally described under certain assumptions. It does not follow that the same formalization explains the concrete processes by which an organism maintains its states. Colombo and

Wright (2021) similarly point out that the Free Energy Principle abstracts from biophysical details and therefore cannot straightforwardly assume the role of a mechanistic explanation.

A second transformation already concerns the expression "free energy". In Friston's definition, Variational Free Energy is an information-theoretic quantity. Thermodynamic free energy, by contrast, is a physical state quantity. The linguistic form is similar, but meaning, dimension, and causal status are not thereby invariant. A relation between the two quantities may be derived within a specific theory, but it does not arise from the shared terminology. If the thermodynamic meaning of energy is transferred to the variational quantity, we have precisely the transition described in Type I, from constancy of expression to unwarranted constancy of meaning.

A third transformation concerns "surprise". Surprisal formally means negative log probability. Psychological surprise, by contrast, is a state of an experiencing organism. The mathematical definition therefore entails no psychological property. Formulations such as "the organism avoids surprises" may be useful as intuitive shorthand. But once they are meant to explain psychological or biological processes, an additional bridge is required.

A fourth transformation lies in the language of minimization. In an optimization problem, "minimizing" is a formal relation. It does not follow that the described system itself has a goal, seeks to avoid something, or is directed toward a target state. When mathematical optimization becomes purposiveness attributed to the organism, the perspective shifts from the modeler to the system. This is the perspective shift described in Type IV.

This yields a sharper criticism. If the Free Energy Principle remains a general formal description of persistent systems, it may be useful as a mathematical organizing principle, but by itself it explains neither concrete biological self-organization nor a neural mechanism. If, by contrast, it is intended to provide a causal biological or neurophysiological explanation, the transitions from the probabilistic model relation to material dynamics must be independently justified empirically or mechanistically. The same formalization cannot supply that bridge merely by redescribing biological processes in its own vocabulary. The critical issue therefore lies not in the mathematics of Variational Free Energy, but in the unwarranted transfer of invariance from model to mechanism, from information-theoretic to thermodynamic meaning, and from formal optimization to organismic purposiveness.

The very broad scope of the principle sharpens this point. If the Free Energy Principle is formulated as a necessary condition for any persistent or self-organizing system, it cannot by itself discriminate among competing biological mechanisms. Its empirical and causal content must then come from specific process models. Here the theory of invariance transfer identifies not merely an open burden of justification but the concrete source of error: a general invariance of description is treated as an invariance of causal explanation.

15. Test Case: Integrated Information Theory

In Integrated Information Theory, the critical transition is explicit. IIT 4.0 begins with phenomenal axioms. Experience is said to be intrinsic, specific, unitary, definite, and structured. IIT assigns corresponding physical postulates to these phenomenal properties and describes this step as an "inference to a good explanation". It then proposes a fundamental "explanatory identity": an experience is to be identical with the unfolded cause-effect structure of a maximal substrate (Albantakis et al. 2023). Precisely because these transitions are stated explicitly, the present model can identify where the theoretical burden lies.

A first invariance error already threatens through the parallel terminology. "Intrinsicality", "information", and "integration" designate properties of experience on the phenomenal side and properties of causal structures on the physical side. Phenomenal unity means that an experience is experienced as a whole. Physical integration in IIT, by contrast, means causal irreducibility with respect to a partition. Phenomenal intrinsicality means that experience exists for the experiencer. Physical intrinsicality is operationalized as the cause-effect power of a system over itself. Shared terminology therefore does not by itself establish semantic or ontological identity.

$$\text{phenomenal unity} \quad \neq \quad \text{causal irreducibility}$$

$$\text{phenomenal intrinsicality} \quad \neq \quad \text{intrinsic causal power}$$

This is a combination of Type I and Type V. A phenomenal concept is first transformed into a physical-operational concept. But the claim that both concepts play the same theoretical role requires independent justification. The phenomenological fact that experience is unitary contains, by itself, no information about the concrete physical organization that realizes this unity. In particular, it does not deductively follow that the substrate must be causally irreducible in the sense defined by the IIT metric. IIT itself acknowledges this insofar as the transition is presented not as a deduction but as an inference to a good explanation.

The second critical point therefore concerns the direction of derivation. The phenomenal axioms determine which properties the physical substrate is supposed to possess. A description of the phenomenon is thereby turned into a requirement on its realization. Such a transition is possible, but it is not justified by the phenomenal starting statement alone. An independent bridge is required between the two levels, showing why causal irreducibility, specificity, exclusion, and composition are precisely the necessary and sufficient physical counterparts of the phenomenal properties. Without such a bridge, the assignment remains theoretically underdetermined: other forms of physical organization might realize the same phenomenal property as well.

$$\text{phenomenal property} \quad \rightarrow \quad \text{physical postulate} \quad : \quad \text{bridge required}$$

The third and strongest transition is the "explanatory identity". IIT does not merely posit a correspondence between phenomenal and causal structure; it claims identity. At this point the strength of the relation must not increase unnoticed. Even a systematic one-to-one correspondence between features of experience and features of a cause-effect structure would initially be a correspondence. Identity is a stronger claim. It requires not two corresponding descriptions or properties, but one and the same state of affairs. Calling the relation an "explanatory identity" cannot itself justify that identity.

$$\text{correspondence} \quad \neq \quad \text{identity}$$

If the identity is understood as a fundamental theoretical posit, the bridge problem therefore does not disappear. It is relocated into the identity thesis. IIT then does not explain how or why a causal structure becomes experience; it presupposes that the relevant causal structure is experience. For a theory whose aim is precisely to explain why experience exists and why it feels the way it does, this is a decisive limitation of explanatory scope.

A fourth point concerns the construction of the formalism. IIT begins with properties of experience, derives physical postulates from them, and then develops mathematical quantities intended to be consistent with those postulates. IIT 4.0 itself emphasizes that its measure of intrinsic information is chosen to satisfy the postulates. The cause-effect structure constructed in this way is then invoked to explain the phenomenal properties. This is not necessarily a formally circular argument. It does, however, create a methodological feedback loop: the properties to be explained structure the requirements imposed on the formalism, and fulfillment of those requirements is then presented as an explanation of those very properties.

$$\text{phenomenon} \quad \rightarrow \quad \text{axioms} \quad \rightarrow \quad \text{physical postulates} \quad \rightarrow \quad \text{formalism} \quad \rightarrow \quad \text{explanatory identity}$$

The internal fit among axioms, postulates, and formalism can demonstrate the consistency of the theory. It does not yet provide independent evidence that the formally specified causal structure is identical with experience. Empirical predictions about conscious states or neural substrates may support parts of the theory. They do not automatically test the ontological identity of cause-effect structure and experience.

The criticism can therefore be stated precisely. IIT transforms phenomenal properties into physical postulates, retains the same guiding terms while changing their meaning, and finally strengthens the asserted relation from correspondence to identity. At each of these transitions, at least one dimension of the conceptual context changes. The required invariance is established neither by terminology alone, nor by the mathematical consistency of the formalism, nor by the label "explanatory identity".

For the present approach, this yields a clear dilemma. If IIT is understood as a model that systematically relates certain causal structures to certain conscious states, it may provide empirically informative correspondences and markers. But it then remains open why those structures are experience. If, by contrast, the strong identity thesis is maintained, that identity itself requires independent justification. As long as the bridge consists essentially of the phenomenal axioms, the postulates derived from them, and the subsequent identity stipulation, the central transition is not explained but theoretically fixed. Integration may then be a relevant system property or a marker of consciousness without thereby already being the mechanism or identity of experience.

16. Relation to Ryle, Wittgenstein, and Conceptual Engineering

The approach belongs to a recognizable tradition in philosophy of language. Wittgenstein sensitizes us to the fact that meaning should not be treated as a fixed aura accompanying a word independently of its use. Ryle shows that natural language is systematically ambiguous and that philosophical theories can go astray through type transgressions and category mistakes (Wittgenstein 1953; Ryle 1938; Ryle 1945).

The theory proposed here does not claim to replace these insights. Its additional object of analysis is the structure of the transition. It asks which dimension is preserved in a transformation, which changes, and which additional equivalence an argument requires. This also makes it possible to analyze cases in which there is no shift of meaning, such as the same computational function under different causal realizations.

There is also a clear difference from Conceptual Engineering. That field focuses on the deliberate evaluation and improvement of our conceptual or representational devices (Cappelen 2018; Koch 2021). The theory of invariance transfer, by contrast, does not primarily ask how a concept should be changed. It examines the inferential commitments that arise when expressions, functions, models, or structures have already been transferred between contexts.

17. The Analytical Procedure

The approach yields a compact procedure for analysis. For a theoretically relevant transfer, the source and target contexts must first be identified. One must then ask which dimension is actually preserved: expression, meaning, function, causal or material realization, level of description, perspective, formal structure, or theoretical status. For that dimension, the relevant equivalence criterion must be specified.

The decisive step is then to identify the bridge. Which additional invariance does the argument require? Does it follow from a definition, from empirical evidence, or from a theoretical derivation? Is that justification independent and non-circular? Does it connect precisely the two dimensions at issue in the transformation under consideration? Or is the additional equivalence merely suggested by continuity of expression, function, or mathematical structure?

The procedure is therefore neither nominalist nor anti-scientific. It prohibits no transfer. It merely separates the transfer itself from the additional conclusions drawn from it.

In compressed form:

transformation → specified preservation → additional claim → bridge assessment

Only when the bridge is missing does an unwarranted invariance transfer occur.

18. Consequences

The theoretical gain of the approach lies in making several forms of error analyzable under a common question without conflating them conceptually. Equivocation concerns a change of meaning under constancy of expression. Multiple realizability concerns constancy of function under a change of realization. Level errors concern transfer between different forms of description. Perspective errors shift attributions from the observer to the system. Model errors alter the status of a formal relation. Metaphors preserve partial structures while changing the domain of objects.

In every case, the transformation itself is initially legitimate. The error arises only when the dimension actually preserved is treated as a tacit guarantee of a further equivalence. The theory therefore shifts the focus from the question "May this concept be used here?" to the more precise question "What exactly has been preserved by this use?"

This is especially relevant to interdisciplinary theories. The further a concept travels among mathematics, physics, biology, psychology, and phenomenology, the greater the risk that continuity of notation or terminology will suggest continuity of the object. Such continuity may exist, but it must be justified wherever the argument requires it.

19. Conclusion

Conceptual transformation is not a defect but a precondition of language and science. Concepts, functions, and models must be transferable if commonalities are to be identified and new domains made accessible. Yet this very transferability makes a distinction necessary: What remains the same, and what changes?

In this paper, invariance denotes the preservation of a particular dimension relative to a particular transformation and to an equivalence criterion specified for that dimension. What is invariant may be a linguistic expression, a meaning, a function, a material or causal structure, a level of description, a perspective, or a formal relation. The invariance of one dimension, however, does not automatically entail the invariance of another.

Unwarranted invariance transfer denotes the case in which such additional preservation is presupposed or asserted for an argument without an independent, non-circular, and transformation-relevant bridge grounded in definition, evidence, or theory. This inferential form is more general than the classical category mistake. It can generate equivocations, level confusions, category mistakes, and unwarranted ontological attributions.

Philosophical conceptual analysis thereby acquires an additional question. Alongside "What does this expression mean?" and "To which category does it belong?" we must ask: "Which property has actually been preserved in this transition, and what justifies the preservation of each further property?"

References

Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A. M., Marshall, W., Mayner, W. G. P., Zaeemzadeh, A., Boly, M., Juel, B. E., Sasai, S., Fujii, K., David, I., Hendren, J., Lang, J. P., & Tononi, G. (2023). Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. PLOS Computational Biology, 19(10), e1011465. https://doi.org/10.1371/journal.pcbi.1011465

Bennett, M. R., & Hacker, P. M. S. (2003). Philosophical Foundations of Neuroscience. Blackwell.

Block, N., & Fodor, J. A. (1972). What psychological states are not. The Philosophical Review, 81(2), 159–181.

Cappelen, H. (2018). Fixing Language: An Essay on Conceptual Engineering. Oxford University Press.

Colombo, M., & Wright, C. (2021). First principles in the life sciences: the free-energy principle, organicism, and mechanism. Synthese, 198(Suppl 14), 3463–3488. https://doi.org/10.1007/s11229-018-01932-w

Friston, K. (2010). The free-energy principle: A unified brain theory? Nature Reviews Neuroscience, 11, 127–138. https://doi.org/10.1038/nrn2787

Koch, S. (2021). Engineering what? On concepts in conceptual engineering. Synthese, 198, 1955–1975. https://doi.org/10.1007/s11229-020-02868-w

Lakoff, G., & Johnson, M. (1980). Metaphors We Live By. University of Chicago Press.

Magidor, O. (2013). Category Mistakes. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199572977.001.0001

Putnam, H. (1967). Psychological predicates. In W. H. Capitan & D. D. Merrill (Eds.), Art, Mind, and Religion (pp. 37–48). University of Pittsburgh Press.

Ryle, G. (1938). Categories. Proceedings of the Aristotelian Society, 38, 189–206.

Ryle, G. (1945). Philosophical arguments. Inaugural Lecture, University of Oxford.

Shannon, C. E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27, 379–423, 623–656.

Sprevak, M., & Smith, R. (2023). An introduction to predictive processing models of perception and decision-making. Topics in Cognitive Science. Advance online publication. https://doi.org/10.1111/tops.12704

Williams, D. (2022). Is the brain an organ for free energy minimisation? Philosophical Studies, 179, 1693–1714. https://doi.org/10.1007/s11098-021-01722-0

Wittgenstein, L. (1953). Philosophical Investigations. Blackwell.