J. J. LI
R. B. H. KWOK
N. Y. FOO

# The Coherence of Theories – Dependencies and Weights

**Abstract.** One way to evaluate and compare rival but potentially incompatible theories that account for the same set of observations is coherence. In this paper we take the quantitative notion of theory coherence as proposed by [Kwok, et.al. 98] and broaden its foundations. The generalisation will give a measure of the efficacy of a sub–theory as against single theory components. This also gives rise to notions of dependencies and couplings to account for how theory components interact with each other. Secondly we wish to capture the fact that not all components within a theory are of equal importance. To do this we assign weights to theory components. This framework is applied to game theory and the performance of a coherentist player is investigated within the *iterated Prisoner's Dilemma*.

## 1. Introduction

The core of scientific theories are laws. These laws often make use of theoretical terms, linguistic entities which do not directly refer to observables. There is therefore no direct way of determining which theoretical assertions are true. This suggests that multiple theories may exist which are incompatible with one another but compatible with all possible observations. Since such theories make the same empirical claims, empirical tests cannot be used to differentiate or rank such theories. Hawking very nicely summarised this positivist approach in the philosophy of science: "A scientific theory is a mathematical model that describes and codifies the observations we make. A good theory would describe a large range of phenomena on the basis of a few postulates, and make definite predictions that can be tested" [Hawking 2001]. One property that has been suggested for evaluating rival theories is *coherence*. This was investigated qualitatively in the philosophy of science (see, e.g, summaries in [van Fraassen 80] and [Nagel 61]) until [Kwok, et.al. 98] introduced a coherence measure based on the average use of formulas in accounting for observations. Prior to this measure, the qualitative approaches considered properties of theories typified by informal notions like "tightness of coupling" of the axioms, "brevity", "predictive scope", etc. Kwok et.al. (op.cit.) took these as guides for their quantification. The idea was to identify highly coherent theories as those whose formulas are tightly coupled in accounting for observations, while low coherence theories contain many disjointed and isolated statements. It proved to

Presented by **Jacek Malinowski**; *Received* October 15, 2007

 

be quite fruitful; for instance this provided a rebuttal to Craig's method [Craig 53] for the elimination of theoretical terms by showing that the method yields theories with very low coherence.

Later work [Kwok, et.al. 03], [Kwok, et.al. 07] by the same authors generalised the approach to better mirror scientific practice. For instance, a standard way to use a theory is to design experiments with varying input and output sets. However, another way is to regard observations as inputs and explanations as outputs. The generalisation accommodates both views, and in fact permits other interpretations of input-output relations to test theories for coherence. It is also able to explain notions like theory modularisation.

It is fair to say that this approach to reifying coherence is in effect a combinatorial grounding that relies on the widely understood concept of support sets that plays an important role in artificial intelligence logic in areas as diverse as diagnoses, logic program semantics and abduction. One may question whether the hitherto qualitative notion of coherence is appropriately captured by our quantitative measure. Our response is that we propose a *plausible* way to fix the interpretation of coherence that can be tested by its efficacy in explicating some well-known examples, with the awareness that other plausible methods may emerge in future that capture variant qualitative interpretations.

In the current paper we take the above as starting points and widen the foundations of coherence as defined through support sets. Two ideas are broached, based on intuitions from scientific practice that were not considered in [Kwok, et.al. 98] and [Kwok, et.al. 03]. The first widening derives from the observation that coherence should also measure how well pairs, triples, etc. of formulas jointly account for observations or outputs. This gives rise to the quantitative notion of dependency in coherence. The second widening mirrors the practical fact that not all formulas may be considered to be equal in importance. This is already acknowledged in the works on belief revision, primarily the AGM approach [Gardenfors 88], where varying commitments to particular beliefs goes by the name of entrenchment. The possibilistic logicians' fuzzy measures aimed at capturing the same intuition have been shown to be equivalent to entrenchment. In our paper we use weights on formulas to do this. This enhanced definition of coherence reduces to the previous version when dependencies are among singletons and all weights are equal.

Numerous formal examples will illustrate the efficacy of the new definition, but we also apply it to a domain not traditionally considered in the philosophy of science which initially motivated our work. The domain is game theory, specifically forms of the (in)-famous Prisoner's Dilemma [Axelrod 81], where the one-time game is classically represented as a matrix that displays the payoffs for each of the two players depending on their choice of action (called strategy). Game theorists then assume rational decisions by each player and analyse the action choices that

must be entailed. Iterating the one-time game was then studied by a number of researchers (see, e.g. Axelrod [Axelrod 81]). It is this iterated version to which we will apply the notion of coherence. We will model a player's reasoning (using its beliefs, desires and intentions) as formulas, and the player's adaptations during the game is seen as attempts to maintain high coherence among these doxastic qualities. Computer simulations of this approach are also described and analysed.

Finally we discuss future directions that this work may profitably take. It is plausible that traditional norms of "rationality" in the evaluation of scientific theories as well as economic and social behaviour may be modulated by current discomfort with the policies that result from them. Wider notions of what it means for these theories to be coherent can contribute to modifications of the existing norms.

## 2. Internalist Coherence

This section reviews the previous contribution by [Kwok, et.al. 03], and suggests innovations in areas that were not addressed up to this date, such as the utility of a set of formulas, and the relationship between sets of formulas: how one may dominate over another, and how tightly they have coupled to account for observations. It can also be seen the other way round, as how closely they have been associated when supported by empirical evidence. For the time being, we call these nominated properties "Internalist Coherence".

### 2.1. Support Sets

The building blocks of coherence are support sets. They describe how a theory accounts for an observation from specific inputs. In this framework, a *theory*, an *input* set and an *output* set are all sets of formulas from a first–order language $\mathcal{L}$. It is appropriate to motivate the setting assumed by the next definition. We conceive of logical theories as formal models of selected aspects of the world that interest us. In science the theories of a domain such as chemistry are often painfully constructed over the course of time, and subject to much testing and revision. We do not address the revision issue here, but as we shall see the testing is implicit. A theory $T$ can be used in many ways. It may help to visualise $T$ as a blackbox into which the "input" set $I$ formulas is fed, and an "output" set of formulas $O$ is produced. The "directionality" suggested by these terms should not be taken literally. The interpretation of $I$ and $O$ depends on how $T$ is intended to be used. For instance, $O$ could be a set of observed outcomes of an experiment, in which case $I$ could describe the initial conditions of that experiment. If given certain hypotheses, we interpret $O$ as desired conclusions of $T$, $I$ could be such a set of hypotheses. Moreover, $T$ itself can have atoms which say that we are only

interested in models of $T$ that satisfy those atoms. It is a matter of modelling to decide which atoms ("facts") to place in $I$, $O$ or $T$, and different choices will yield different coherence measures. To see that this flexibility is an advantage, consider the following. Suppose someone proposes a theory $T$ that purports to account for some phenomena. If we wish to test $T$ only in settings where conditions $C$ hold, one way to do that is to consider instead the theory $T \cup \{C\}$. But if we already have a set $O$ of observations, and we wish to find conditions $C$ under which $T$ can account for $O$, then $C$ is part of $I$.

For brevity in the sequel we sometimes use the term *axiom* for an element of $T$.

DEFINITION 1 (Support Sets[Kwok, et.al. 03]). *Given input set $I$, output set $O$, a subset of the theory $T$ be $\Gamma$. $\Gamma$ is an $I$-relative support set of $O$ if*

1. $\Gamma \wedge I \models O$ *and*
2. $\Gamma$ *is minimal (wrt set inclusion).*

Let $S(T, I, O)$ denote the family of all $I$-relative support sets for $O$. As explained above different choices of input set $I$ will result in different support sets. This approach is designed to be "independent of any commitment to causality or particular use of laws" [Kwok, et.al. 98]. This definition is intended to capture the idea that $\Gamma$ alone cannot account for $O$ but it can do that with the help of $I$; moreover we want $I$ to be as small as possible, viz. no redundancy.

EXAMPLE 1 (Socrates is Mortal). *Given the input $I$:*

$$I : \{man(Socrates)\} - Socrates\ is\ a\ man$$

*output $O$:*

$$O : \{mortal(Socrates)\} - Socrates\ is\ mortal$$

*the theory $T$:*

$$T = \{\alpha_1 : \forall(x)\ man(x) \rightarrow mortal(x), \alpha_2 : \forall(x)\ deity(x) \rightarrow \neg mortal(x)\}$$

$$-all\ men\ are\ mortal, -all\ deities\ are\ not\ mortal$$

*$\{\alpha_1\}$ constitutes a support set for I and O, since it explains how O is derived from I, whereas $\{\alpha_2\}$ does not constitute a support for I and O.*

EXAMPLE 2. *Let $T$ be the theory that geniuses would only pass if they are not intoxicated; and if one is not a genius, then one would only pass after study:*

$\neg genius(x) \wedge \neg study(x) \rightarrow \neg pass(x)$

$\neg genius(x) \wedge study(x) \rightarrow pass(x)$
$genius(x) \wedge intoxicated(x) \rightarrow \neg pass(x)$
$genius(x) \wedge \neg intoxicated(x) \rightarrow pass(x)$

*Suppose we wish to explain an output set* $O = \{\neg pass(john)\}$. *Possible input sets are:*

$I_1 = \{genius(john), intoxicated(john)\}$ *and*
$I_2 = \{\neg genius(john), \neg study(john)\}$.

*Observe that we may re-interpret* $O$ *as a prediction given the input information* $I_1$ *or* $I_2$. *For this* $O$ *the second and fourth formulas in* $T$ *are not used. However, should* $O$ *be changed to* $\{\neg pass(john), pass(verana)\}$ *it can be seen that all the formulas in* $T$ *will be used to compute the input support sets.*

### 2.2. Utility of a set of formulas

Recall the informal properties of coherence, such as "tightness of coupling" and "work together", that we wish to encapsulate in our formal quantitative framework. One element missing from the previous approach [Kwok, et.al. 03] was the notion of measuring the usefulness of a group of formulas, or a sub–theory. This is an important concern as the utility of the sub–theory would reflect both the utility of the components of the sub–theory, and the tightness of the coupling between the components, and thus capture some of the desired properties of coherence in our representation.

We wish to measure the contribution of not only one formula, but several formulas in how they *together* have contributed to support observations. Building on the [Kwok, et.al. 03] definition, we now examine how a set of formulas "work together". For instance, in a theory $T$ consisting of elements $\alpha$, $\beta$ and $\gamma$; we may wish to consider not only the individual utilities of elements $\alpha$ and $\beta$, but their synergistic qualities of working together, e.g. the utility of the set $\Theta = \{\alpha, \beta\}$.

The next definition formalises this intuition. A higher level of utility for a set means that its formulas occur together often in support of observations.

DEFINITION 2 (Utility of a Set of Formulas). *Given a theory* $T$ *and a non-empty set of formulas* $A \subseteq T$, *its utility is:*

$$U(A, T, I, O) = \frac{\mid \{\Gamma : A \subseteq \Gamma \; and \; \Gamma \in S(T, I, O)\} \mid}{\mid S(T, I, O) \mid} \; if \quad S(T, I, O) \neq \emptyset$$

This formal definition provides a measure of how well all formulas in the set "work together" in supporting observations. It sees the formulas as equal, and does not discriminate one over another. Informally the idea is as follows. To measure the utility of the set $A$ we do this: first count how many times it appears within the

support sets for the given $I$ and $O$; we then express this as a fraction of the total number of those support sets — hence the more frequently $A$ so appears the higher its utility. If one formula does not work with the group, the utility for the group will be rendered as zero. The connection between the utility of individual formulas (singleton set) and the utility of sets of which it is a member is addressed in Lemma 1 below.

LEMMA 1 (Joint Utility). *Let a set of formulas A consist of two proper subsets B and* $\Delta$*, i.e.* $A = B \cup \Delta$*. The following properties hold:*
*(i)*

$$U(A, T, I, O) = U(B \cup \Delta, T, I, O)$$

*(ii) if* $S(T, I, O) \neq \emptyset$*, then*

$$U(A, T, I, O) = \frac{\mid \{\Gamma : B \subseteq \Gamma \ \wedge \ \Delta \subseteq \Gamma \ \wedge \ \Gamma \in S(T, I, O)\} \mid}{\mid S(T, I, O) \mid}$$

*since*

$$\{\Gamma : B \cup \Delta \subseteq \Gamma\} = \{\Gamma : B \subseteq \Gamma \wedge \Delta \subseteq \Gamma\}$$

*This will be useful in subsequent proofs where sets of axioms appear together.*

### 2.3. Dependencies between formulas

The "tightness of coupling" between elements of a theory can be reflected in two ways. We shall elaborate the two different senses of "tightness" over the next two sections. First, this property can be exhibited in the reliance of one set of formulas upon another. For example, to account for the observation "Socrates is mortal", the axiom "Socrates is a man" would not make sense without the other axiom "all men are mortal". However, if there are two independent explanations of Socrates' mortal nature based on he is a man, then the axiom "Socrates is a man" would be less dependent on each of the set of formulas that amounts to the respective explanations.

Formally, we wish to see how dependent a specific set of formulas is upon another. It may be that this set in isolation is not a support set, but that in combination with another set it is one; then informally the first set can be regarded as dependent on the second. More precisely, if set $\Phi$ is contained in most of the support sets that contain another set $\Theta$, then $\Theta$ would have a high dependency on $\Phi$. This dependency is generally asymmetric.

DEFINITION 3 (Dependency Coefficient).

$$D(\Theta, \Phi, T, I, O) = \frac{\mid \{\Gamma : \Gamma \in S(T, I, O) \textit{ and } \Theta \subseteq \Gamma \textit{ and } \Phi \subseteq \Gamma\} \mid}{\mid \{\Gamma : \Gamma \in S(T, I, O) \textit{ and } \Theta \subseteq \Gamma\} \mid}$$

*This defines the dependency of* $\Theta$ *on* $\Phi$.

The dependency above also reflects the importance of the set $\Phi$. Consider a formula $\alpha$ in $T$ that not only occurs in most support sets, but where other formulas are dependent on it to make a support set, this then makes $\alpha$ important in $T$. This can be captured as the *weight* of a formula which we discuss later. Section 2.3 discusses the use of dependencies.

Dependency is related to utility. Given two sub–theories $\Theta$ and $\Phi$, the dependency of $\Theta$ to $\Phi$ measures the proportion of support sets that contain both $\Theta$ and $\Phi$ against those that contain $\Theta$. The higher the dependency, the more support sets that contain $\Theta$ also contain $\Phi$.

COROLLARY 1 (Dependency-Utility Connection).

$$D(\Theta, \Phi, T, I, O) = \frac{U(\Theta \cup \Phi, T, I, O)}{U(\Theta, T, I, O)}$$

**Proof** *Recall:*

$$D(\Theta, \Phi, T, I, O) = \frac{\mid \{\Gamma_1 : \Theta \subseteq \Gamma_1 \; and \; \Phi \subseteq \Gamma_1 \; and \; \Gamma_1 \in S(T, I, O)\} \mid}{\mid \{\Gamma_2 : \Theta \subseteq \Gamma_2 \; and \; \Gamma_2 \in S(T, I, O)\} \mid}$$

*Divide numerator and denominator by* $\mid S(T, I, O) \mid$*:*

$$D(\Theta, \Phi, T, I, O) = \frac{\frac{|\{\Gamma_1 : \Theta \subseteq \Gamma_1 \; and \; \Phi \subseteq \Gamma_1 \; and \; \Gamma_1 \in S(T,I,O)\}|}{|S(T,I,O)|}}{\frac{|\{\Gamma_2 : \Theta \subseteq \Gamma_2 \; and \; \Gamma_2 \in S(T,I,O)\}|}{|S(T,I,O)|}}$$

*Then from Lemma 1, translate the numerator and denominator back to utility:*

$$D(\Theta, \Phi, T, I, O) = \frac{U(\Theta \cup \Phi, T, I, O)}{U(\Theta, T, I, O)}$$

### 2.4. Coupling of formulas

The second way to encapsulate the "tightness of coupling" property is to see how elements of a theory mutually need each other. That is, how much they "work together" in proportion to the total amount of work they do in forming I–relative support sets. The greater the ratio, the "tighter" the elements coupled together. This is different to the previous definition of dependency, as this looks at how much both sub–theories take part in accounting for observations.

We wish to formalise a notion of mutual dependency between two sub–theories. Intuitively, this will measure the degree to which the sub–theories need each other in accounting for observations. The following symmetric definition formalises this intuition.

DEFINITION 4 (Coupling Coefficient).

$$CP(\Theta, \Phi, T, I, O) = \frac{\mid \{\Gamma_1 : \Theta \subseteq \Gamma_1 \; and \; \Phi \subseteq \Gamma_1 \; and \; \Gamma_1 \in S(T, I, O)\} \mid}{\mid \{\Gamma_2 : (\Theta \subseteq \Gamma_2 \; or \; \Phi \subseteq \Gamma_2) \; and \; \Gamma_2 \in S(T, I, O)\} \mid}$$

This coupling coefficient represents how two sub–theories mutually need each other. The higher the coupling, the more they work together, reflecting the properties of coherence as stated from the informal definition proposed by [Kwok, et.al. 98].

### 2.5. Example - Socrates is wise

EXAMPLE 3 (Socrates is wise). *Consider the proposal that Socrates is wise because he had a wise student named Plato. Plato, apart from being wise, was also a prolific writer in philosophy. Therefore, we may have two possible ways of accounting for the fact that Socrates is wise, being either "The teacher of a wise man is also wise", or "The teacher of a prolific writer is wise". The theory can be formalised as:*

*I* = $\emptyset$
*T* = $\{\alpha_1$*: ∀x∀y teacher(y, x) ∧ wise(y) → wise(x),*
$\alpha_2$*: teacher(Plato, Socrates),*
$\alpha_3$*: wise(Plato),*
$\alpha_4$*: prolificWriter(Plato),*
$\alpha_5$*: ∀x∀y teacher(y, x) ∧ prolificWriter(y) ∧ philosopher(y) → wise(x),*
$\alpha_6$*: philosopher(Plato)* $\}$
*O* = $\{$*wise(Socrates)*$\}$

As the theory itself is sufficient to account for the observations, we therefore do not require inputs in this example. However, we still consider support sets to be $I$–relative as we still consider the input set together with the theory to account for observations, and in this case the input set just happens to be empty. This formalised theory enables us to investigate the utility of formulas and sub–theories, the dependencies of one component of the theory to another, and the coupling between the components. Hence we find a measure for "usefulness" of components of the theory and how they are "tightly coupled".

#### 2.5.1. Utility of sets

The two $I$–relative support sets for O are: $\{\alpha_1, \alpha_2, \alpha_3\}$, and $\{\alpha_2, \alpha_4, \alpha_5, \alpha_6\}$.

Hence the utility of formulas $\{\alpha_5, \alpha_6\}$ as a set would be $\frac{1}{2}$, since they appear together in only one of the two possible support sets; the set $\{\alpha_1, \alpha_4\}$ have the utility value of 0 since they do not work together at all; and the utility of $\{\alpha_2\}$ is 1 due to the fact that it appeared in all support sets.

### 2.5.2. Dependencies

*Case 1: High Dependency*
A formula $\alpha$ would have high dependency on a set $\Gamma$ if $\{\alpha\} \cup \Gamma$ occurs in most support sets that contain $\alpha$. So in the support sets and the theory illustrated above, the formula $\alpha_1$ has a high dependency on both $\alpha_2$ and $\alpha_3$. Because without either formula, $\alpha_1$ would not be able to account for the observation. In a theory where only one explanation is possible, the dependencies of all formulas in the support set relative to each other would be 1.

*Case 2: Moderate/Low Dependency*
A formula $\alpha$ would have a moderate/low dependency on a set $\Gamma$ if $\alpha$ occurs in multiple support sets. This way the formulas in $\Gamma$ may not always occur in support sets containing $\alpha$. In the example above, $\alpha_2$ has a moderate dependency on other formulas. This is because $\alpha_2$ is contained in two support sets, and no other formula in $T$ also occurs in the same two support sets. However, occurring in multiple support sets does not necessarily guarantee a moderate/low dependency to other formulas, for there could be another formula $\delta$ which occurs in the same support sets, thus having a high coupling coefficient. This is examined later in the section on couplings.

*Case3: Zero Dependency*
Formulas will have zero dependency if they have nothing to do with each other. In the current theoretical context it means that they do not share any support sets. Here axioms $\alpha_{1...3}$ are totally disjointed from axioms $\alpha_{4...6}$, thus any pairs selected with one from each set would yield zero dependence to each other.

### 2.5.3. Couplings

*Case 1: High Coupling*
High coupling occurs when two formulas (sets) often appear in the same support sets. In our example, $\alpha_5$ and $\alpha_6$ are required in the same support sets, since Plato needed to be both a prolific writer and a philosopher. Together, they have a coupling value of 1. However, this is different to dependency. If there were another formula $\alpha$ that also occurs across both support sets, then $\alpha$ and $\alpha_2$ would have a

high coupling value of 1 despite being spread across more than one support set.

*Case 2: Moderate/low Coupling*
Moderate/low coupling happens when two formulas (sets) appear in some support sets together, but in other support sets only one formula (set) is required. With our example, $\alpha_1$ and $\alpha_2$ have a coupling value of $\frac{1}{2}$. This value reflects the fact that both $\alpha_1$ and $\alpha_2$ appear in one support set, but only $\alpha_2$ appears in the other support set. The coupling value between sets $A$ and $B$ is greater than 0 as long as they appear together in one support set. Formally:

$$CP(A, B, T, I, O) > 0 \text{ if and only if, for some } \Gamma \in S(T, I, O), A \cup B \subseteq \Gamma$$

*Case 3: Zero Coupling*
Like dependency, two formulas (sets) have zero coupling when they have nothing to do with each other; they do not ever work together to account for an observation. In our example, $\alpha_1$ and $\alpha_5$ have zero coupling.

### 2.6. Formulas with weights

Within a theory $T$, some axioms may be considered more important than others. This quality is described in the AGM framework [Gardenfors 88]. The importance of an axiom can either be innate, judgemental or could be determined from its usage in accounting for observations (its occurrence in support sets). Although some axioms are not frequently used, they may still be essential to the integrity of the theory. The measure of utility will be generalised to take into account an innate or judgemental weighing of axioms. In AGM entrenchment a logically weaker statement entailed by a stronger one will have an entrenchment at least as high as the latter. The analog of this for utility is the following: a weaker statement would account for at least as many input-output sets as a stronger one. This property is preserved by the definitions below of observational and natural weights. However, if weights are just subjective judgements then the analog of AGM entrenchment may not hold.

DEFINITION 5 (Weight of a Formula). *Let $T$ be a finite theory $\{\alpha_1, \ldots, \alpha_n\}$, the weighing coefficient $W : T \mapsto \mathbf{R}$ is the subjective distribution of weights in $T$. $W(\alpha_i)$ reflects the innate weight of formula $\alpha_i$.*

However, it is possible to abuse the weighing process by arbitrarily adding weight to make the theory carry a high degree of coherence. Socrates may say: "My theory has half the coherence of your theory, so I just give each formula three times the weight, then mine would be more coherent!" To avoid this, and to make

different theories comparable, weighing should be normalised in order to reflect the proportionate importance of formula $\alpha_i$ to the theory $T$.

DEFINITION 6 (Normalisation Criterion). *Let $T$ be a finite theory $\{\alpha_1, \ldots, \alpha_n\}$, the normalisation criterion states that:*

$$\sum_{i=1}^{n} W(\alpha_i) = n$$

Ontologically it does not make sense to give any formula a negative weight, for at worst it plays no part in support sets. Hence we assume:

ASSUMPTION 1 (Positivity Assumption). *Let $T$ be a finite theory $\{\alpha_1, \ldots, \alpha_n\}$*

$$W(\alpha_i) > 0 \textit{ for every } i : 1 \leq i \leq n$$

### 2.6.1. Observational Weights

Thus far, utility has been defined relative to an individual input set $I$ and output set $O$. The pair $(I, O)$ can be thought of as a single experiment or application of theory $T$. However, a theory is typically applicable and testable under many situations. It is therefore natural to consider what utility might mean across a vector of experiments or applications. Consider vectors (or sequences) of input and output sets, $\mathbf{I}= (I_1, I_2, \ldots, I_m)$ and $\mathbf{O}= (O_1, O_2, \ldots, O_m)$. One may interpret this vector as a sequence of experiments, e.g., a pair $(I_k, O_k)$ being the $k$-th experiment with $I_k$ being the initial conditions and $O_k$ being the observation that results; other interpretations are of course possible, including $O_k$ being an observation and $I_k$ being the explanation. However, some observations may be considered more important than others, e.g., as in "crucial" experiments that may undermine a theory. To reflect this, experiments can be associated with a rank or weight that represents its judged significance. (In subsection 2.6.2 we propose a rather more objective assignment of weights.) The weight can then be "shared" by formulas that support this observation.

First we define a notion of *support weight* (SW). For each $O_j$ in $\mathbf{O} = (O_1, \ldots, O_m)$, we associate a payoff $P(O_j)$. Then the support weight $SW(\alpha_i, T, I_j, O_j)$ can be the "share" of the payoff for $\alpha_i$.

DEFINITION 7 (Support Weight).

$$SW(\alpha_i, T, I_j, O_j) = \frac{P(O_j)}{\mid S(T, I_j, O_j) \mid} \sum_{\Gamma \in S(T, I_j, O_j) \textit{ and } \alpha_i \in \Gamma} \frac{1}{\mid \Gamma \mid}$$

Hence from the support weight we define the observational weight that eventually reflects the importance of a formula.

DEFINITION 8 (Observational Weight). *For a theory T =* $\{\alpha_1, \ldots, \alpha_n\}$*, with input* $\mathbf{I}$ *:* $(I_1, \ldots, I_m)$ *and output* $\mathbf{O}$ *:* $(O_1, \ldots, O_m)$*, the Weight Share (WS) of axiom* $\alpha_i$ *in T is:*

$$WS(\alpha_i) = \frac{1}{m} \sum_{j=1}^{m} SW(\alpha_i, T, I_j, O_j)$$

*and the observational weight (OW) is:*

$$OW(\alpha_i) = n \frac{WS(\alpha_i)}{\sum_{j=1}^{n} WS(\alpha_j)}$$

### 2.6.2. Natural Weights

By the original definition, weighing is a subjective measure of "importance" of formulas in a theory. However, it is possible to define a scheme of weighing from the dependency coefficient as defined before, since intuitively, if a formula is more needed by others, then it is more important.

For every formula $\alpha_i$ in $T$, we can define a *dependency weight* from how each formula in $T$ depends on $\alpha_i$. This represents an implicit weight of the specific formula.

DEFINITION 9 (Dependency Weight). *For an axiom* $\alpha_i$ *in a theory* $T$*:* $\{\alpha_1, \ldots, \alpha_n\}$ *with input* $\mathbf{I}$ *:* $(I_1, \ldots, I_m)$ *and output* $\mathbf{O}$ *:* $(O_1, \ldots, O_m)$

$$DW(\alpha_i, T, \mathbf{I}, \mathbf{O}) = \sum_{j=1}^{n} \sum_{k=1}^{m} D(\alpha_j, \alpha_i, T, I_k, O_k)$$

Just as before, we could derive a measure of Natural Weight (NW) from the building block of dependency weights from the axioms.

DEFINITION 10 (Natural Weight). *So for a theory T =* $\{\alpha_1, \ldots, \alpha_n\}$ *with input* $\mathbf{I}$ *:* $(I_1, \ldots, I_m)$ *and output* $\mathbf{O}$ *:* $(O_1, \ldots, O_m)$

$$NW(\alpha_i) = n \frac{DW(\alpha_i, T, \mathbf{I}, \mathbf{O})}{\sum_{j=1}^{n} DW(\alpha_j, T, \mathbf{I}, \mathbf{O})}$$

In this way the ranking of axioms is accomplished by how other components of the theory depend on this component, and thus its weight is proportional to its importance in the theory. The advantage of this approach is that the weight of an axiom is no longer a subjective distribution given by the user, either by entrenchment or weights of observations. The natural weighing takes advantage of the natural properties of dependency, and thus weighing becomes an automated process.

### 2.7. Weighted Utility and Coherence

From this framework of weighted axioms, we can adopt a new and feature–rich definition of weighted utility. This makes utility useful in its own right, for we are able not only to compare between different theoretical systems, but components within a theoretical system. One possible application of this newly found role is in game theory, which shall be further investigated in this paper.

DEFINITION 11 (Weighted Utility of a Formula). *The* Weighted Utility *of a formula $\alpha$ in a theory $T$ with respect to an input set $I$ and an output set $O$, and a weight function $W$ is:*

$$WU(\alpha, T, I, O) = U(\alpha, T, I, O)W(\alpha)$$

Notice that we have introduced two weight functions: observational weight and natural weight. Observational weight is based on a subjective value placed on input/output $(I, O)$ pairings while natural weight is based on dependency calculations. Both weighing functions are valid instances of $W$ in the above definition.

The generalisation of coherence to weighted formulas will follow the intuition from [Kwok, et.al. 98], as the average of weighted utilities.

DEFINITION 12 (Coherence of a Weighted Theory).

$$C(T, \mathbf{I}, \mathbf{O}) = \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} WU(\alpha_i, T, I_j, O_j)$$

This culminating definition of coherence allow rival and possibly incompatible theories with weighted axioms to be evaluated and compared in a quantitative fashion. The evaluation is based on the brevity of the theory and the weighted utility of each of the theory components. It provides a perspective of how a theory can be judged based on inputs and observations, while taking into account the varying weights of different axioms in the theory.

### 2.8. Examples - Socrates is Wise 2

EXAMPLE 4 (Socrates is wise).
*T* = {$\alpha_1$: $\forall x \forall y$ *teacher(y, x)* $\wedge$ *wise(y)* $\rightarrow$ *wise(x),*
$\alpha_2$: *teacher(Plato, Socrates),*
$\alpha_3$: *wise(Plato),*
$\alpha_4$: *prolificWriter(Plato),*
$\alpha_5$: $\forall x \forall y$ *teacher(y, x)* $\wedge$ *prolificWriter(y)* $\wedge$ *philosopher(y)* $\rightarrow$ *wise(x),*
$\alpha_6$: *philosopher(Plato)* }
*O* = {*wise(Socrates)*}

Recall the above "Socrates is Wise" example. It contains two support sets for the same observation. They are $\{\alpha_1, \alpha_2, \alpha_3\}$, and $\{\alpha_2, \alpha_4, \alpha_5, \alpha_6\}$. The first support contain three axioms, where the second contained four. In this example we denote $\{\alpha_1, \alpha_2, \alpha_3\}$ as the first support set and $\{\alpha_2, \alpha_4, \alpha_5, \alpha_6\}$ as the second support set. The common element is $\alpha_2$, which is featured in both support sets. The other axioms would be called exclusive members of their support sets.

### 2.8.1. Example - Observational Weight

Suppose we assign the payoff of 100 points to the observation $wise(Socrates)$. Since both support sets adequately explain the observation, they deserve an equal share of the payoff, i.e., each support set will be apportioned 50 points. Each axiom that belong strictly to the first support set (of size 3) such as $\alpha_1$ would receive an equal share of the payoff given to that support set, i.e., $\frac{1}{3} \times 50 = 16\frac{2}{3}$. Axiom $\alpha_4$, belonging strictly to the second larger support set of size 4, would get a lesser share at $\frac{1}{4} \times 50 = 12\frac{1}{2}$.

Axiom $\alpha_2$, contained in both support sets will have the greatest support weight at $\frac{1}{3} \times 50 + \frac{1}{4} \times 50 = 29\frac{1}{6}$. Hence its observational weight would be $\frac{6}{100} \times 29\frac{1}{6} = 1\frac{3}{4}$, which is also its weighted utility, since it appears in all support sets of $O$. The utility of the other formulas would be half of their observational weight, since there is only one observation and they belong strictly to one of the two support sets. It would be $\frac{1}{2}$ for the exclusive members of the first support set and $\frac{3}{8}$ for exclusive members of the second support set. Hence the weighted coherence value would be $\frac{31}{48}$. This value reflects the degree of coherence of the given theory with respect to a set of observations with weights.

### 2.8.2. Example - Natural Weight

For the given support set, the dependency weight of the exclusive member of support sets would be $2\frac{1}{2}$ and $3\frac{1}{2}$ respectively. Since this definition values the support from other axioms, the members of the larger support set would receive more weight. The common element $\alpha_2$ would receive a dependency weight of 6.

Therefore the natural weight of $\alpha_2$ would be $\frac{72}{43}$. The exclusive members of the first support set would receive a natural weight of $\frac{30}{43}$, and the second support set $\frac{42}{43}$. This is also their utility value since there is only one observation. The weighted coherence value is $\frac{55}{86}$. This value reflects the degree of coherence with respect to the internal structure of the theory, thus the value is different to that derived from observational weights. We consider both to be valid, but different measures of coherence. The user would make the choice in selecting which measure to use depending on its applications.

Further examples of this new weighted system of coherence, particularly in observational weights, are illustrated in the following section with an application in Game Theory.

## 3. Application to Game Theory

### 3.1. Concept

Coherence, once quantified, can be used as a comparator between any two theoretical systems. In typical agent interactions, all of an agent's beliefs, desires and intentions (BDI) can be represented in formal semantics [Rao and Georgeff 91]. These enable us to assess a systemic coherence in one's belief, and the process of interaction can be seen as an effort by each agent to modify its own system in order to achieve a satisfactory outcome with respect to the other agents while maintaining a high level of its internal coherence. The intuition is that the agent will choose an action that is most coherent with its set of beliefs, desires and intentions.

### 3.2. Prisoner's Dilemma Simulation

#### 3.2.1. Background

The Prisoner's Dilemma was originally formulated by mathematician Albert W. Tucker. The iterated version of the game was proposed in [Axelrod 81]. It has since become the classic example of a "non-zero sum" game in economics, political science, evolutionary biology, and of course game theory. So that the exposition below may be independently understood, we briefly recount the set-up. In the game, two prisoners are interrogated separately in different cells. The two prisoners can either choose to cooperate (keep silent) or defect (blame the other). If they both cooperate, they receive a sentence of 2 years in prison. If one cooperates but the other betrays, the first gets 10 years in prison, and the second gets 1 year. If both betray, each will get 4 years. The payoff (years in prison) of an action is dependent on the action of the other player. It is therefore in the interests of a player to minimise this payoff. The way the payoff is set out means that whatever a player chooses to do, the other player can reduce its payoff by defecting, so in a one-time game both players will defect, resulting in 4 years for each. A better result will be for both to cooperate, suffering a sentence of only 2 years each; but they cannot communicate to negotiate, and even if they can, lack of trust may enter the picture. This "bad" solution of both defecting can intuitively be ameliorated if the game is played repeatedly, whence each player understands that if it defects now the other player can retaliate in the next iteration. Thus, in the iterated version, the players repeatedly play the game and have a memory of their previous encounters. We set

out to test the application of our coherence calculations in this scenario, and how it behaves in an iterated game with evolution of populations.

### 3.2.2. The Coherentist Agent

In coherence-based evaluative simulations in game theory, we set out to play the game repeatedly, and the histories of past games are recorded by each player. This history then forms the Belief in what had happened in the past, which can be seen as the theory $T$ in the calculation. The player's Desire (D) is to maximise its payoff (or minimise it if interpreted as a penalty). This desire can be seen as a mode of evaluating payoff as weights of each outcome. The beliefs (B), together with the criterion of selection (D), will lead to the calculation of the utility of each of the actions that the player may take. A selection of the action according to its utility will lead the player to formulate the intention to act upon this decision.

More specifically, the inputs (I) describes the rules of what the player knows about the nature of the game. This includes actions of the player, consequences of these actions, states (in a finite-state game) and payoffs associated with a particular state. The observations (O), whether a result, consequence or state, is the corresponding payoff. The desire of the player will be driven by the ranking of these payoffs. Hence the Support Set consists of the list of axioms, which together with the given input, will make a particular observation true.

In the iterated game of Prisoner's Dilemma, the prisoner evaluates the history played against the respective player to reach a rational decision. Each history element consists of the player's move at that iteration, and the returned value / payoff from that particular move. The returned value can be seen as the weight of that observation, and hence the support weight for that particular action.

The coherentist agent uses the paradigm of the observational weight as discussed in Section 2.6.1. This way the weight of a formula is reflected by the observations that it supports. The weight of an action can be evaluated from the history of payoffs for a given opponent.

The Input (I) for a particular iteration are the rules of the game, and the move of other players. Below is a summary of the rules, expressed logically. The propositions *Betray* and *otherBetray* mean respectively that a player betrays and the other also betrays; *Cooperate* and *otherCooperate* have corresponding meanings. The numbers are the payoffs for a player, depending on the move of the other player; recall that these are the years in prison, and hence a penalty to be minimised.

$$I = \left\{ \begin{array}{l} Betray \wedge otherBetray \rightarrow 4, \\ Betray \wedge otherCooperate \rightarrow 1, \\ Cooperate \wedge otherBetray \rightarrow 10, \\ Cooperate \wedge otherCooperate \rightarrow 2 \end{array} \right\}$$

This input set will remain fixed for each game of the Prisoner's Dilemma.

For each move of a player the other player has the choices *otherBetray* or *otherCooperate*. The theory, to be evaluated, are the rival options the player could adopt. viz., *Betray* or *Cooperate*, bearing in mind that in any iteration the moves of both players are to be made *simultaneously*.

$$T = \{Betray, Cooperate\}$$

For instance, consider a history (Action, Penalty) of (Cooperate, 2), (Betray, 1), (Betray, 4), (Betray, 4). This implies that at the same time the other player had made the corresponding choices of *otherCooperate*, *otherCooperate*, *otherBetray* and *otherBetray*. Hence the sequence of output observation set is:

$$\begin{aligned} \mathbf{O} = & (O_1 : \{otherCooperate \wedge 2\}, \\ & O_2 : \{otherCooperate \wedge 1\}, \\ & O_3 : \{otherBetray \wedge 4\}, \\ & O_4 : \{otherBetray \wedge 4\}) \end{aligned}$$

In section 2.7 we associated a payoff to each output set. This payoff was a measure of the importance of the output set. For our application to the Prisoner's Dilemma, we wish to measure how advantageous each output is to an agent. This would be inversely proportional to the prison sentence. In our simulation studies, we simply used the length of the prison sentence as the payoff and chose the option with the *smaller* Observational Weight. With the above example, the Weight Share of Betray is (1+4+4 )$\div$ 3 = 3, whereas the Weight Share of Cooperate is 2 $\div$ 1 = 2. Hence the Observational Weight is evaluated at $\frac{3}{4}$ for Betray, and $\frac{2}{4}$ for Cooperate. Therefore, in a system where lower weight (penalty) is favoured, Cooperate is the preferred strategy. This was the approach adopted in the experiments. However, for the analogous approach where the system favours higher weights, the payoff

can be taken as the inverse of the penalty. Therefore the agent's choice of an axiom of higher weighted utility reflects its pursuit of a higher degree of coherence.

The problem of course is that it is difficult to predict what the other player will do at any iteration. In the tournament organised by Axelrod [Axelrod 81] the system pitted many players together and simulated the iterations, looking for the best performing players. In the simulations we ran, we investigated how coherentist players performed against other kinds of players, including the best performing player in Axelrod's tournaments.

### 3.2.3. Simulation

We define five types of agents in the simulation. They are *reckless*, *cooperative*, *tit–for–tat*, *suspicious* and *trusting*. The last two types are the same coherentist agent with different initial conditions. A reckless player is one who always defects, whereas the cooperative player is one who always cooperates (does not defect). The "tit–for–tat" strategy was traditionally regarded as the best deterministic strategy developed by Anatol Rapoport, which cooperates in the first turn, and subsequently plays the opposing player's previous move. The coherentist agents are divided into two groups, one being suspicious, for its members would betray at the initial phase, whereas the other group, the trusting agents, would cooperate.

To initialise simulation, the user specifies how many of each type of agent there are in the game. The user also specifies how many iterations are to be simulated. In each iteration a player will play a round–robin tournament, playing once with every other player in the simulation. When two players meet, they have the option to either betray or cooperate. The move and the payoff will be recorded, and the agent can review this as a history element when playing this opponent in the next iteration.

After a specified number of iterations, the old players will die and a new generation of players will replace them. They will be free of the history from previous players. However, their proportions, according to agent type, will be inversely proportional to the average time the particular type of agent spent in jail. The result is then normalised to maintain the population size. Although rounding error is allowed, the overall population size will only decrease due to rounding, and increases are prohibited.

### 3.2.4. Trends and Behaviours

Initially, we set 20 iterations per generation with an equal proportion of each player type. It turns out that the coherentist player performs well compared to other agent types. As predicted the cooperative agents perish rather quickly in the simulation.

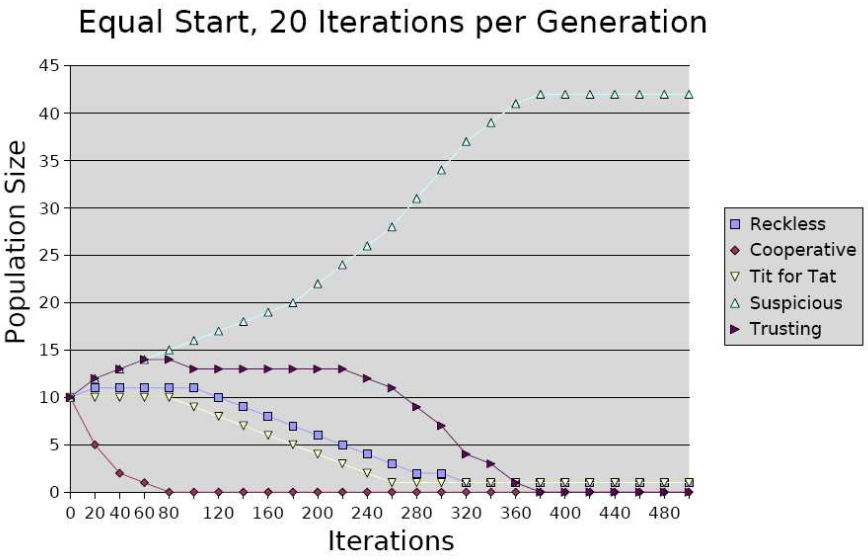


Figure 1. All five players with equal initial population, 20 iterations per generation

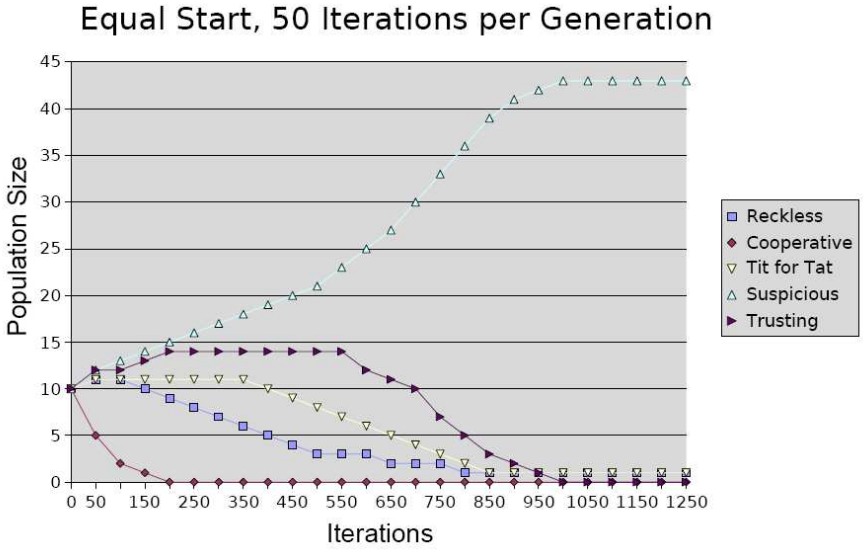


Figure 2. All five players with equal initial population, 50 iterations per generation

In the end it was the "suspicious" coherentist agents that took over the population, while others struggled to hang on. (Figure 1)

The suspicious and trusting agents only differ in their initial response when they have no previous history of playing against the other player. Yet the impact is significant as the suspicious player takes over the population after a brief initial period when both coherentist agents perform well. Both the coherentist agents gain an edge over tit–for–tat, as they exploit the cooperative agents while tit–for–tat is only nice to them.

When the number of iterations per generation is raised from 20 to 50, the results are slightly different. One feature is that the reckless agents performed much more badly, while the tit–for–tat agents played better, though not as well as the coherentist agents. (Figure 2)

In both simulations the coherentist agents came out on top. This may be associated with the coherentist agent's flexible approach of punish reckless behaviour, cooperate with rational, nice agents, and exploit the overly nice and vulnerable agents. In particular the latter characteristic is absent in the behaviour of tit–for–tat agents. However, this positive outcome may not necessarily be associated with simply a coherentist behaviour. Instead, it may be the case that the macro-environment of the game in this situation enabled the coherentist agents to be the fittest. For a different environment with different rules, coherentist agents may not perform as well as agents of the "simple faith", such as the reckless or tit–for–tat agents.

What emerges from these results is that coherence alone as a property of agents is an aid to their performance, but external factors and initial conditions (such as the first move) also matter. A way to think of the role of coherence is that it constrains agent choices in such a way that its use of its theory aligns those choices well with its observations.

## 4. Summary and Discussion

We aim to establish basic principles governing the coherence of laws within theoretical systems. Such principles provide a means for evaluating and comparing different systems. By defining a measure of how a sub–theory contributes to a theory, in terms of *Group Utility*, *Dependency* and *Coupling*, the formalism captures a number of important properties of coherence. Specifically, the formalism provides a rendering of informal characteristics of coherence, *viz.* how axioms "work together" and are "coupled tighter". The framework has also been significantly enhanced by the introduction of weights to axioms and observations. By relativising one axiom's weight, either in terms of the weight of observations or the dependency to other axioms, we derived an account of the importance and rank of axioms in a

theory.

Our proposed framework of coherence serves as a useful treatment of an old problem in the philosophy of science, namely the evaluation of rival, but possibly incompatible theories. It also provides a perspective on the development of scientific theories, where anomalies found in observations contribute to the degree of incoherence of a theory, and scientific developments to account for these anomalies can be viewed as the pursuit of a greater coherence.

This measure of coherence is not only useful for the domain of the philosophy of science, it is also useful for describing reasoning, deliberation and interaction in agents. The example of Prisoner's Dilemma illustrated how coherence can be used in game theory. When an agent chooses the option that is most coherent with its beliefs, the agent has a rational basis for reasoning and acting.

**Acknowledgements.** We would like to thank Dr. Anika Schumann, Dr. Anthony J. Flynn and the three anonymous reviewers for their proof–reading and their comments on the paper, and National ICT Australia for supporting this research under its knowledge representation and reasoning program.

## References

[Axelrod 81] R. Axelrod. *The Evolution of Cooperation*, Science, 211(4489):1390-6

[Craig 53] W. Craig. On axiomatizability within a system. In *The Journal of Symbolic Logic, 18*, pages 30–32, 1953.

[van Fraassen 80] B. van Fraassen. *The Scientific Image*, pp 14-19, Clarendon Press, Oxford, 1980.

[Gardenfors 88] P. Gardenfors. Knowledge In Flux. MIT Press, Cambridge, MA, 1988.

[Hawking 2001] S. W. Hawking. The Universe in a Nutshell. New York: Bantam Books, 2001.

[Kwok, et.al. 98] R. B. H. Kwok, A. C. Nayak, N. Foo. Coherence Measure Based on Average Use of Formulas. *Proceedings of the Fifth Pacific Rim Conference on Artificial Intelligence*, 553-564, LNCS v.1531, Springer Verlag, 1998.

[Kwok, et.al. 03] R. B. H. Kwok, N. Foo, A. C. Nayak. The Coherence of Theories. *Proceedings of the 18th Joint International Conference on Artificial Intelligence*, IJCAI03, Acapulco, Mexico, August 2003.

[Kwok, et.al. 07] R. B. H. Kwok, N. Foo, A. C. Nayak. Coherence of Laws *UNSW Computer Science and Engineering Technical Report* Number: UNSW-CSE-TR-0719. October 2007. ftp://ftp.cse.unsw.edu.au/pub/doc/papers/UNSW/0719.pdf

[Nagel 61] E. Nagel. *Structure of Science*, Harcourt 1961.

[Rao and Georgeff 91] A.S. Rao and M.P. Georgeff Modeling Rational Agents within a BDI-Architecture. *Proceedings of the Second International Conference on Principles of Knowledge Representation and Reasoning (KR'91)*, pp 473-484, 1991.

JASON JINGSHI LI
Research School of Information Science and Engineering
Australian National University
Canberra, ACT0200, Australia
jason.li@anu.edu.au

REX BING HUNG KWOK
School of Computer Science and Engineering
University of New South Wales
Sydney, NSW2032, Australia
rkwok@cse.unsw.edu.au

NORMAN Y. FOO
School of Computer Science and Engineering
University of New South Wales
Sydney, NSW2032, Australia
norman@cse.unsw.edu.au