Quantified coherence of moral beliefs as a predictive factor for moral agency

Abstract (254 words): *The notion of 'coherence' plays an important but controversial role in moral theory. Usually, the concept of coherence is discussed in a framework of justifying moral beliefs. It is then faced with objections regarding circularity, vagueness, and impracticability when analyzing coherence of large belief sets. Most of this critique states that the normative relevance of coherence is tied to its descriptive use, i.e. one has to specify what exactly it means that a system of believes "coheres" in order to argue for the normative relevance of coherence. In our contribution, we present a descriptive notion of the concept of coherence that allows us to distinguishing qualitatively different system states with respect to coherence. We compare our notion with coherence concepts in psychology and Paul Thagard's definition of coherence as constraint satisfaction. We discuss the main differences between Thagard's definition and our proposal. Finally, we outline how our concept of coherence can be applied in moral psychology as a tool for understanding how the structure of moral beliefs an individual moral agent holds may influence the behavior of the agent. In particular, we show how our approach is able to integrate different types of coherence relationships between single beliefs. In this way, our definition of coherence allows us, for example, to analyze how cognitive and affective similarities between reasons used in moral decision making may interrelate. Furthermore, it can give novel insights into phenomena like practical irrationality in decision making. We finally sketch the relevance of our descriptive notion of coherence for its normative use.*

Keywords: Coherence; moral beliefs; moral decision making; moral agency; moral psychology; ethics

1. Coherence – from an intuition to a quantified concept

The term ‘coherence’ is used in several scientific disciplines. Rigid definitions adapted to specific problems are, for example, found in quantum physics (Winter and Steinberg 2008) and signal processing (White and Boashash 1990). In social sciences, psychology and philosophy, the term usually describes the logical and/or semantic coherence of propositions representing attitudes, actions, attitudes, beliefs, judgments and the like (Thagard and Verbeurgt 1998). In particular, the notion of coherence plays an important role in truth theories (Rescher 1973) and in validating ethical arguments based on coherent moral beliefs[1] – i.e. the term has a *normative role* such that the fulfillment of the criterion ‘coherence’ serves as justification that, e.g., a structured set of beliefs is a true theory, or that an argument legitimates a specific action as moral.

Many scholars in philosophy have argued that the notion of coherence has a decisive role with respect to ethical justifications. Quine writes in his essay “On the Nature of Moral Values”:

Disagreements on moral matters can arise at home, and even within oneself. When they do, one regrets the methodological infirmity of ethics as compared with science. The empirical foothold of scientific theory is in the predicted observable event; that of a moral code is in the observable moral act. But whereas we can test a prediction against the independent course of observable nature, we can judge the morality of an act only by our moral standards themselves. Science, thanks to its links with observation, retains some title to a correspondence theory of truth; but a coherence theory is evidently the lot of ethics (Quine 1979, 477-478).

A prominent representative of the coherentist tradition in ethics, John Rawls, introduced the term when explaining “reflexive equilibrium” – probably the most influential notion of coherence within ethics:

[1] Most contemporary philosophers characterize a belief as a “propositional attitude” (see the Stanford Encyclopedia of Philosophy, http://plato.stanford.edu/entries/belief/, accessed on June 10th 2011). Whether moral beliefs are propositions in the strict sense (i.e. are truth-apt) is disputed by non-cognitivists. However, for our analysis, this controversy is not relevant. We understand beliefs as state or habit of mind in which the belief holder (the moral agent) places trust or confidence, regardless whether this state of mind refers to empirical or normative issues. Furthermore, we assume that the moral agent is (in principle) able to communicate these mental representations that reflect those issues towards others, i.e. they are accessible to empirical research. We will use the neutral term ‘moral sentences’ to indicate belief representations that are related to moral issues and that are accessible to empirical research, e.g. by using surveys.

Its [a conception of justice's] justification is a matter of the mutual support of many considerations, of everything fitting together into one coherent way (Rawls 1971, 21)

Although Rawls never defined 'coherence' in a precise way, his intuition was taken by many followers who extended his theory of justice (e.g. Daniels 1979). Within ethics, at least four theories of coherence can be distinguished that cannot be reduced to a single definition of the term (Hoffmann 2008).

However, both in truth theories and in ethics, there are well-known problems associated with the concept of coherence. A classical problem refers to the controversy between foundationalism and coherentism with respect to both justifying truth claims and finding reasons for legitimating actions as moral: A set of beliefs may be entirely coherent but still represent a wrong theory, and an argument for justifying a specific act may be coherent, although the act still is morally wrong. This critique of the normative relevance of coherence is independent of the definition of coherence (i.e. its descriptive use). It denies the normative significance of coherence even if one would have found a clear definition of the term. This point, however, is not the topic of our contribution and will not be discussed further.

We intend to overcome a second, prominent critique of coherence, referring to the vagueness of the term. This line of critique claims that the *descriptive* notion of coherence is ill-defined such that it does not make sense to attribute normative significance to coherence: Since we don't know what coherence really means, why should the fact that a system of beliefs is coherent have normative significance? Among others, Kirkham has formulated this critique as follows:

The term 'coherence' as used by coherence theories has never been very precisely defined. The most we can say by way of a general definition is that a set of two or more beliefs are said to cohere if and only if (1) each member of the set is consistent with any subset of the others and (2) each is implied (inductively if not deductively) by all of the others taken as premises or, according to some coherence theories, each is implied by each of the others individually (Kirkham 1992, 104).

The point is that even this very general definition of coherence proposed here – according to Kirkham "the most we can say" – has its drawbacks. First, the criterion of 'consistency' taken alone is too weak, as it also applies to beliefs that have nothing in common with each other. Second, the criterion of (logical) implication is hard to operationalize as soon as the set of beliefs contains more than a few propositions. Furthermore, it is doubtful whether large belief systems – and it's plausible to assume that real word agents have belief systems that include hundreds, if not thousands of beliefs – ever would fulfill the criterion of coherence with respect to inductive or deductive implication. Therefore, if the notion of coherence should remain a theoretical construct that is not only a mere intuition, but allows for fruitful applications in empirical sciences, alternative approaches are required.

Our interest lies therefore in the *practicability* of coherence as an instrument to analyze the behavior of moral agents. For example, we are interested in whether coherence can be defined in such a way that it helps to understand irrationalities in decision making and action that conflict with rational choice theories (Hastie and Dawes 2009, Gigerenzer and Gaissmaier 2011). It is well-known that the behavior of people is often inconsistent and/or conflicts with beliefs the agents hold – and when rationality is related to coherence between beliefs and/or actions (often understood as freedom from contradictions), coherence irrationality (e.g. resulting from framing effects) seems to be common place. However, this interpretation depends on the notion of coherence, i.e. what it really means when stating that an entity e_1 and e_2 "cohere" or a system of such entities is "coherent". Furthermore – as most experimental approaches to coherence irrationalities only focus on (in)coherence between only two (or very few) specified beliefs/actions in the sense of logical consistency (Jussim 2005) – it would be of interest whether a definition of coherence can be found that allows us to measure the coherence of large belief sets in a practicable way. In this way, we can analyze an important aspect of the question "What makes us moral?": the role of the structure of large belief sets moral agents hold for their actual moral behavior.

To address this issue, we organize our contribution as follows: In the next section, we briefly introduce psychological notions of coherence and we argue that no satisfying concept that goes beyond a mere intuition of coherence is used in this field. In section 3, we present the proposal of Paul Thagard, which is often called the most sophisticated philosophical notion of coherence. Our proposal is introduced in section 4, and in section 5 we compare our definition with Thagard's concept of coherence. In section 6, we outline – based on our understanding of coherence – the possible causal role of coherence for moral agency, whereas potential applications in moral research are exemplified in section 7 using the example of Thagard. In the conclusion, we briefly discuss the importance of our descriptive proposal for the normative use of coherence.

2. Coherence in Psychology

The use of 'coherence' as a theoretical term in empirical sciences does not necessarily imply that it is precisely defined and quantified. This holds true also for psychology. For example, early notions of coherence emerged in the context of Gestalt psychology, denoting the "wholeness" of specific perceptions (Silverstein & Uhlhaas 2004). This notion indeed captures an important psychological marker of the experience of coherence as something that we ultimately judge "by 'seat of the pants' feel" (Putnam 1982: 133) – but it lacks applicability in the sense that coherence of a specific percept can be computed. One may say that coherence is understood as being a Boolean variable, i.e. something is either coherent or not coherent. Later, quantified notions of coherence have been introduced in perception psychology, although the concept is merely used as a placeholder for denoting correlations between elements (Rodwan 1965) or predictability in temporal sequences of elements (Trumbo et al. 1968).

Since the 1950s – although the term 'coherence' was not used – being "consistent" (with respect to beliefs, beliefs and actions, etc.) became a major topic within psychology. Here, it is not possible to review fully the conceptual and empirical contributions of the various different theories of cognitive consistency (see Abelson et al 1968, Abelson 1983). However, both in conceptual explanations (e.g. the Balance theory of Heider 1958)[2] and in experimental tests (like the "forbidden-toy paradigm", a classic dissonance paradigm)[3] consistency was discussed involving in most cases only few beliefs or cognitive elements. Later, in personality psychology, the term 'personality coherence' has been coined to grasp three related phenomena: The first is coherence in social behavior and experiences. Across different circumstances, people's experiences and actions are often meaningfully interconnected. People respond con-

[2] The classic condition in Balance theory involves the relations among three cognitive elements constituting a triangle pattern that can be in balance or imbalance.

[3] In this paradigm, children were forbidden to play with a very attractive toy and had to resist the temptation to play with it while the experimenter was out of the room – creating a dissonance between the children knowing they very much wanted to play with this toy and knowing they were not playing with it even thought, if they did, the experimenter would be a bit (or strongly) annoyed (Aronson & Carlsmith 1963).

sistently across some contexts and display distinctive patterns of variation across others. The second phenomenon is the organization among multiple psychological processes. Personality variables do not function as independent mechanisms but as coherent, integrated systems. The third aspect of coherence is phenomenological. People generally achieve a coherent sense of self. They have a stable sense of their attributes and develop a coherent life story (Cervone and Shoda 1999). Again, this notion of coherence captures an important intuition – but it remains unclear how these quite distinct phenomena can be merged into one defined measure of coherence. However, as soon as the set of entities that are the subject of coherence attributions was refined, more precise notions emerged. An example is the central coherence theory which defines autism as an inability to integrate sources of information to establish meaning, where words and sentences are the objects of coherence measurements (Jolliffe and Baron-Cohen 1999; see also Silverstein & Uhlhaas 2004).

This short overview outlines the difficulties when trying to define coherence in a precise way. Conceptually, there is the problem of "holism": Virtually all elements of a system whose coherence is assessed may be connected to all others. Therefore, the problem emerges: How to define a subclass of elements whose coherence should be assessed in order to make a coherence computation feasible? This holism problem may be surmountable by heuristics that prune out links to other elements that are below a certain level of density; but such heuristics are difficult to implement in practice and are highly sensitive to context. Furthermore, if coherence should be made useful for empirical sciences (e.g. for explaining behavior), one must analyze whether and to what extend people are actually sensible for the coherence of their beliefs. For example in decision making, coherence may be most powerful when all relevant elements are explicitly considered at the same time. However, given the limits of human working memory, the number of such elements actually considered is likely to be quite small (Keil 2006). But it might be that the interrelation between potentially accessible concepts (whose number is high) frames which elements become explicit in a decision task – i.e. an underlying "coherence" of those concepts may be a decisive element in perception and decision making. This view is supported by research on semantic priming demonstrating that the exposure to a concept influences the response to a later presented, semantically similar concept (e.g. Friederici et al. 1999).

Therefore, quantifying coherence not only involves the task of properly defining the term but also to showing whether this definition is actually computable for real-world purposes and whether it can be linked to a specified problem that one wants to solve. Certainly, these tasks are interrelated: a feasible and operational definition is required in order to test the importance of coherence. And finally, demonstrating a role of coherence in real world (moral) agency is an important aspect for discussing its normative significance. For example, the fear of coherence critics, that there may be coherent belief sets (of practical relevance) that represent a wrong theory, could turn out to be practically impossible given a more precise definition of coherence.

3. The suggestion of Paul Thagard

Probably the most prominent proposal to define coherence in a precise way and to operationalize it for various empirical questions was published by Paul Thagard (Thagard and Verbeurgt 1998, Thagard 2000). We both share the same aim – i.e. turning coherence into a scientifically useful concept – as well as (to some degree) the conceptual framework in which coherence is quantified – i.e. a connectionism or network-based approach. Therefore, we briefly present the main points in Thagard's theory. In Section 5, we will outline the major differences between his and our proposal, which is presented in the next section.

Thagard frames the coherence problem as the maximization of satisfaction of a set of positive and negative constraints that exist between elements $\{e_1, \dots e_n\}$ of a system, whose coherence should be assessed (Thagard and Verbeurgt 1998). The elements are understood as representations (of beliefs, actions, etc.) that "cohere" (fit together) or "incohere" (resist fitting together). The relationships between the elements are included either as positive or negative constraints, which can be weighted to denote the strength of the constraint. For denoting the type of a constraint, Thagard made several proposals: deductive, explanatory, analogical and deliberative coherence (Thagard 1998). The coherence problem then consists in dividing the set of all elements of the system in two subsets *A* (for accepted) and *R* (for rejected) such that most of the constraints are satisfied. Satisfying a constraint means that when there is a positive constraint between e_i and e_j, they should be in the same set, whereas when there is a negative constraint between e_i and e_j, they should be in different sets. This partition of the system is associated with a number called coherence, which is the sum of the weights of the constraints which are satisfied. When this number is maximized the partition is optimal in the sense that subset A contains those elements that "best fit together" by excluding the elements that do not fit together with the A-elements. This definition is finally related to five algorithms that could be used to calculate the coherence of a specific set of representations. Among them, Thagard favors the connectionist algorithm. This method maps the elements and their constraints onto a neural network. Each element of the set is related to a node of the network, a positive constraint between two elements results in an excitatory connection (with specified weight) and a negative one in an inhibitory link. After each node is assigned the same initial activation value, the network is updated in parallel until a stable state is reached (i.e. the activations of each node don't change any more). Nodes with activations above a certain threshold are then assigned to the set *A*.

Thagard exemplified his notion of coherence by a decision making task in the murder case of Paul Bernardo, who was convicted in 1995 of the prolonged sexual torture and murder of two young women. In Canada, this case led to a discussion whether capital punishment would be appropriate for Bernardo's crimes, although the death penalty was abolished in the country. For Thagard, the case is an exemplar that a person may have different, even contradicting beliefs with respect to that question. These beliefs form a constrained network that becomes the object of a coherence analysis. His example network includes different types of positive and negative constraints; he calls them deductive, explanatory, analogical and deliberative coherence. For example, the statements "Capital punishment is sometimes justified" and "Paul Bernardo should be executed" are (positively) connected by a deduction, the statements "Killing a defenseless victim is wrong" and "Capital punishment is wrong" are (positively) connected by analogy. Statements[4] with respect to the question whether capital punishment is deterrent or not are connected by explanatory relations and the statements "Paul Bernardo should be executed" and "Reduce prison expenses" are (positively) connected by deliberative coherence, referring to intrinsic goals an agent has for biological or social reasons. Figure 1 shows the network (adapted by us to exemplify the different types of constraints). As Thagard stated, this network only shows an extract of a (possibly) much larger network.

[4] The fact that Thagard includes the statement "empirical evidence" in his example is somehow misleading, as the phrase actually is a placeholder for the actual evidence (i.e. either supporting "Capital punishment is not a deterrent" or "Capital punishment helps to prevent serious crimes"). It would make more sense to connect former sentence with "Capital Punishment is wrong" and to understand the empirical evidence as a factor that defines the weight of either connection.

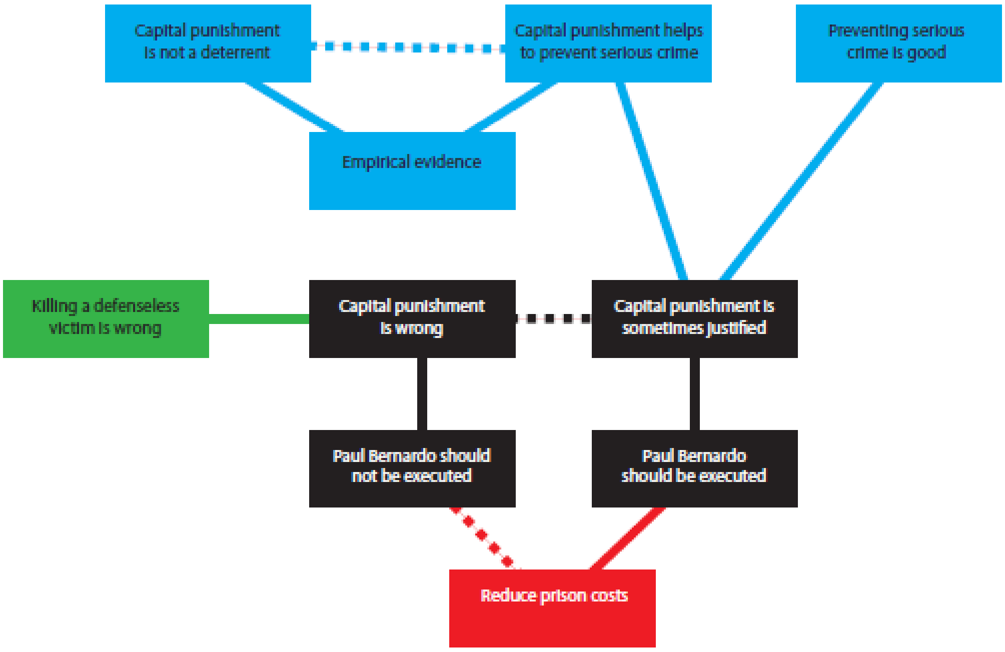
Capital punishment is not a deterrent
Capital punishment helps to prevent serious crime
Preventing serious crime is good
Empirical evidence
Killing a defenseless victim is wrong
Capital punishment is wrong
Capital punishment is sometimes justified
Paul Bernardo should not be executed
Paul Bernardo should be executed
Reduce prison costs

Figure 1: The belief network with respect to the question, whether Paul Bernardo should be executed or not, adapted from Thagard (2000, 143). Solid lines are positive constraints, dashed lines are negative constraints, and the colors indicate explanatory (blue), deductive (black), analogical (green) and deliberative (red) coherence.

The example exemplifies the problem of holding many beliefs that are related in very different ways – probably a more realistic account to understand moral decision making and moral agency compared to pure deductive reasoning. However, he does not “finish the game”, i.e. the example only shows the problem, not the solution. Although he wrote that his computing algorithms were able to provide a solution (i.e. the partition of the belief set that maximizes constraint satisfaction), he did not say how the partition looks (Thagard 2000: 144). This probably results from two problems. First, he needs to quantify the weights of the constraints; a point for which he does not offer a solution: “I do not have an algorithm for establishing the weights on people’s constrains” (Thagard 1998, 418). The reason for this may rely in the second problem: the kinds of interrelations between the beliefs he has introduced are hard to operationalize – and sometimes the classification of a specific interrelation between two beliefs as, e.g., either being “deliberative” or “explanatory” is ambiguous.

To sum up: Thagard’s definition of coherence is precise and includes proposals for actually computing the coherence of a set of representations. He also offers a solution for the problem of circularity (Thagard and Verbeurgt 1998, Section 6) – i.e. the notion also addresses the critique with respect to the normative significance of coherence. His proposal is therefore able to address central issues for turning coherence into a useful scientific concept.

But the definition also has some drawbacks. From a theoretical point of view, the coherence measure is (in respect of computational complexity) NP-hard, as the number of all possible partitions of a set of size *n* scales according to 2^n. In other words, the “best” solution cannot be computed when the set contains more than a few elements. Although the connectionist algorithm is able to find a good estimate, there is no guarantee that the global maximum has been reached. Furthermore, there is no guarantee that the neural network reaches a (quasi-)stable state. Thus, these formal analyses are difficult to apply to everyday data sets. From a practical point of view, we note – as mentioned above – that the problem of finding the weights of the connections remains unsolved. In section 5, we briefly describe other problems when comparing his definition with our proposal. The practicability of Thagard’s proposal has also been criticized by others (e.g. Keil 2006). Therefore, we will now propose an alternative definition of coherence that takes these theoretical and practical issues into account.

4. Our definition of coherence

We begin our considerations by outlining the basic idea of coherence. In its most general form, the term ‘coherent’ is a property of a set of entities that are interrelated in a specific way. Therefore, both the entities as well as the kind of interrelations have to be defined – when using the network-terminology the entities refer to the nodes and the interrelations refer to the edges of the network. The entities we are interested in are beliefs (moral sentences) that are related to the evaluation of actions and states of affairs. They form a belief system, i.e. – by following Converse’s classical definition (1964, 207) – a “configuration of ideas and attitudes in which the elements are bound together by some form of constraint or functional interdependence”. These beliefs can take different characteristics. Daniels distinguished three types of beliefs, namely “moral judgments”, “moral principles” and “relevant background theories” (Daniels 1979, 258) – unquestionably very different kinds of beliefs, as it is not really clear how complex a “background theory” (that probably consists of several sentences) is, compared to a mere judgment.

In the following – as we have an interest in defining coherence such that it can be used for empirical applications – we will refer to beliefs that are stated in one or only a few sentences (or even single concepts) such that they can be understood as distinct moral schemas (Jordan 2009; Lapsely and Narvaez 2004). Furthermore, we include not only purely normative beliefs (e.g.: “abortion is wrong”), but also beliefs on (disputed) matters of facts (e.g.: “abortion destroys a human being”), as long as they are normatively loaded (i.e. include ‘thick’ terms, Williams 1985) or serve as important reasons for justifying moral beliefs (i.e. belong to the framework of “explanatory coherence” using Thagards’ terminology). The type of question under consideration may lead to new specifications of what counts as a single element of the system whose coherence is measured.[5]

The crucial point in defining coherence is, however, the definition of the relationship between the beliefs. Within truth theories, these relationships are defined in logical terms (implicative or deductive relationships between beliefs) and every belief that represents, for example, a wrong deduction makes the system incoherent and is, consequently, excluded from the system. In this understanding, coherence is a Boolean variable, i.e. a system is either coherent or not coherent[6] (an understanding that is probably close to the interpretation of coherence in Gestalt psychology). In formal sciences, e.g. mathematics, Boolean coherence is useful. For example, an incoherent axiom system involves contradictions that cannot be accepted.

If one allows weighting the relationships between beliefs, as Thagard has proposed, this is equivalent to defining a similarity relation (or a distance function) between them. This means that for each pair of elements e_i and e_j a number (usually between 0 and 1, if the similarity relation can be normalized) is given that reflects the similarity of these two elements.[7] Doing this for all beliefs of the system sets up a matrix that describes the pairwise similarity of all system elements. Depending on the kind of similarity relation, this matrix is symmetric or not

[5] For example, if one would be interested to assess the coherence of beliefs a whole society holds, this may require defining the agents that hold the beliefs as single entities of the system. In that way one may assess, whether there are mutually strong but diverse sub-cultures with respect to morality that may be a danger for the cohesion of a society. However, we will not elaborate on this point in this contribution. See Christen et al. (submitted) for an example of this approach, where we showed how degrees of coherence of political beliefs party member hold served as a predictor of party splits.

[6] Also in this framework one may introduce a gradual measure of coherence by counting the number of beliefs that are not coherent with the system. However, this misses the point, as the function of a coherence measure in a logical setting is to identify the incoherent beliefs in order to exclude them from the system.

[7] As the similarity of two elements is equivalent to the notion of the distance between two elements, 0 usually denotes maximal similarity (= zero distance), whereas (if the distance function is normalized) 1 denotes no similarity (= maximal distance).

symmetric.[8] The choice of the similarity relation depends on the question under consideration, and the values of the similarity matrix are evaluated in empirical investigations. This allows the understanding of coherence as a continuous and multidimensional variable that can be related to qualitatively distinct kinds of coherence of the system. These qualitative different system states with respect to coherence can then be correlated to different types of behavior. We consider this to be an adequate understanding of coherence when the concept is applied to empirical questions, as it allows for a more fruitful analysis. In moral psychology, for example, the predictive value of Boolean coherence is probably zero, as it can be assumed that no real world moral agent has a coherent moral belief system in a strictly logical sense.

The quantitative notion of coherence we have in mind builds on the following two properties, which a belief system that is understood as a network of beliefs (i.e. entities/nodes with specified interrelations/edges) usually holds: First, we assume the network to be inhomogeneous. It will probably display sub-structures that can be understood as clusters of beliefs with stronger mutual relationships compared to beliefs from other clusters. This allows the introduction of some quantification of the diversity of the system. Second, these structures may display some property of stability that depends on the strength of the mutual relationships of beliefs. In that sense, the coherence of a belief system can be related to its diversity and stability.

Take the example of Thagard as a simple illustration of these two aspects. Obviously, the belief system has a sub-structure formed by at least two sets of beliefs which are either close to the decision “Paul Bernardo should be executed” or “Paul Bernardo should not be executed”. Within each set, further sub-structures may be present that reflect the stability of the cluster. For example, someone may consider the analogy between “Killing a defenseless victim is wrong” and “Capital punishment is wrong” as rather weak, which would lower the stability of the “not execution” cluster.

To operationalize these two dimensions “coherence diversity” and “coherence stability”, we suggest the adaption of the concept of superparamagnetic clustering (Blatt et al. 1996; Ott et al. 2005) to define coherence. Generally, superparamagnetic clustering is a nonparametric method suitable for detecting and characterizing group structures in data without imposing a

[8] For example, if the matrix represents belief connections as deductions, then the similarity (distance) between element e_i and e_j is 0 (when e_j is deduced from e_i), but the similarity (distance) between e_j and e_i is 1, i.e. the matrix is not symmetric.

prior bias. The algorithm is inspired by a self-organization phenomenon in magnetic spin systems. In physics, this is described as follows: In an inhomogeneous spin system, clusters of correlated (synchronized) spins can emerge, corresponding to groups of spins with strong couplings. Upon an increase in temperature, i.e. an increase in stress on the system, these clusters decay into smaller units in a cascade of (pseudo-)phase transitions. Hence, the physical properties ('coherence') of the spin system are contingent on two factors: stability under stress (captured by the notion of a temperature) and diversity of the clusters. A formal introduction is provided in the appendix.

A translation of this picture into the world of moral science yields the following correspondence: the spin system is the belief system, the single spins are the beliefs, the spin couplings reflect the similarity of the beliefs, and the clusters stand for the internal structure of the belief system and allow us to define a continuous notion of coherence along the two dimensions of diversity and stability. This allows for the identification of four idealtypic states of a belief system (Fig. 1). Those can be described in terms of their role of the system for powering decisions of the moral agents as follows (see also section 6): (1) coherence stability and coherence diversity may be low. Such a state probably does not induce a clear direction towards a decision problem for which the beliefs serve as guiding principles or motivational force. Such a structure may be typical for a decision problem in which the agent has no specific interest in and where the beliefs do not have any decisive role. One may model this type of decision problem as random decision making (with respect to moral beliefs, i.e. moral beliefs do not matter). (2) If coherence diversity is high and coherence stability is low, the system exhibits a plurality of sub-groups and lacks a strong and stable core. In this situation, the belief system may support several options in the decision process, although no single option is clearly distinguished. The moral agent holds reasons for several options, although he or she does not understand the problem as conflicting. (3) Low coherence diversity and high coherence stability may indicate a belief system with a high degree of unity. Such a system offers a clear direction within a specific decision problem and represents the 'classic' understanding of coherence (i.e. a set of mutually supporting beliefs giving rise to clear reasons for moral action). (4) Of particular interest is the combination of high coherence diversity and stability, as several (at least two) strong sub-groups of more or less equal size exist that are inherently stable and mutually incohesive. Such a system may be representative for a dilemmatic decision situation (see sections 6 and 7 for further explanations).

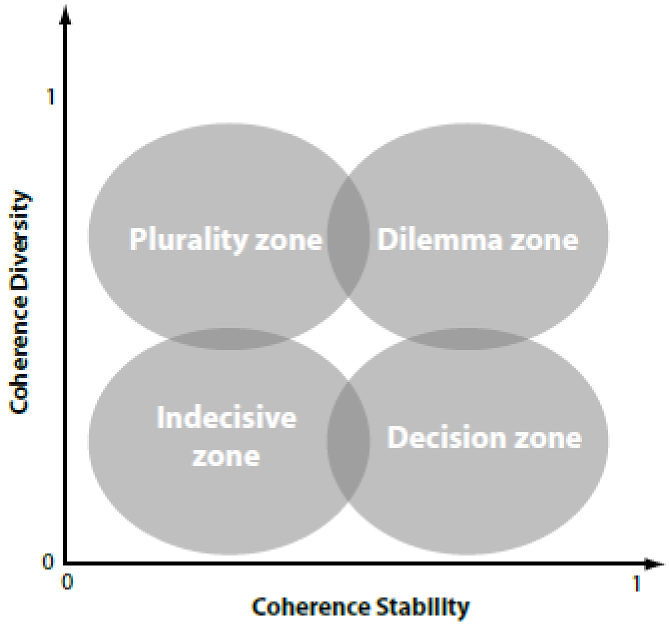

Figure 2: A formal outline of four idealtypic kinds of coherence along the dimensions coherence stability and coherence diversity.

This understanding of coherence requires two things: a distinction between single elements and a conceptualization of the interrelations between these elements. Both aspects include controversially discussed problems and may, in an empirical setting, require predefinitions. For example, one may have to define whether a specific moral sentence is understood as a single belief or as consisting of more than one belief. Also for defining the type of relationship and for quantifying it (i.e. the distance metrics) more than one option is possible. Whenever a practical problem has to be solved using this approach, one has to take these problems into account and, e.g., perform a sensitivity analysis in order to evaluate to which extent different distance metrics affect the qualitative result.

5. Comparison to the proposal of Thagard

In the following, we briefly outline some distinctions between our and Thagard's definition of coherence. First, we note that Thagard uses the term 'cohere' (or 'coherent') both to describe a system property as well as the interrelation between two elements of the system. This, however, mixes two different levels of the system. We use 'coherent' solely as a descriptor of the system, whereas the pairwise interrelations of the elements of the systems are denoted by the term 'similarity'.

A second, important point refers to the issue of weighting the connections in the network. In the original proposal of Thagard and Verbeurgt (1998) the role of these weights has not been discussed further, but in other contributions (Thagard 1998, 2000), more emphasis has been put on that aspect, although he did not offer an explicit solution of how to generate these weights (see our comment in section 3). We suspect that the underestimation of this aspect by Thagard is grounded in the fact that he started his considerations by discussing coherence relations (in our terminology: similarity) of a Boolean type, i.e. deductive or explanatory coherence (Thagard 1998), in which it is a simple yes-no-issue whether entity e_1 and entity e_2 fit together or not. By later distinguishing between types of coherence relations and allowing for weights at the same time creates some questions that are not discussed in detail by Thagard. In deductive coherence, for example, where a general principle e_1 and a particular moral judgment deduced from that principle e_2 are linked together, the role and assessment of the weight remains unspecified. Basically, one would assume that the (normalized) weight representing the distance should be 0, as e_2 is deduced from e_1. However, Thagard writes that the constraints "are typically soft rather than hard. A soft constraint produces a tendency to accept two positively constrained elements together, but this constraint can be overruled if overall coherence maximization suggests that one of the elements be accepted and the other rejected" (Thagard 1998: 408). In this description, it remains unclear how the fact, that one element may be rejected and the other not (which results from the algorithmic procedure that maximizes coherence), is related to the weight of the connection between these two elements. In our proposal, the type of similarity relation defines the weights (expressed by the similarity matrix) of all edges of the network. With respect to the deduction example just mentioned, the weight could reflect the relevance of the deduction between e_1 and e_2 for the agent, i.e. refers to a psychological property that is linked to a specified measurement method. In this way we avoid the problematic entanglement of different coherence relations in the sense of Thagard by introducing a general notion of similarity that is linked to a method to assess the similarity, rather than to distinguish between different types of coherence relations.[9]

[9] The general notion of similarity also avoids another potential source of confusion in the proposal of Thagard, the distinction between positive and negative constraints. This distinction is introduced as being of a categorical type – but it is actually blurred when allowing weights. When using the connectionist algorithm to compute coherence, negative weights are included in the network (the weight of the inhibitory links) and contribute to the activations of the neurons in the quasi-stable state – i.e. they are considered in a similar way as positive weights. In our proposal, we do not distinguish between positive and negative constraints but we reflect the degree of similarity in a distance function with endpoints "maximal similarity" and "no similarity".

A third distinction refers to the algorithmic procedure that Thagard proposes. His goal is to subdivide the set of beliefs in two subsets – one of which would be the “optimal set” with respect to coherence. This procedure probably reflects the foundation of Thagard’s concept of coherence in truth theories – i.e. the aim is to find one single set of elements that cohere in an optimal way. However, by constructing set *A* consisting of elements with maximal mutual positive constraints, Thagard does not discuss the possibility that *W* may also consist of elements that share some degree of mutual similarity (although less than those in *A*) – and it remains unclear whether this fact has any explanatory role when relating the coherence of a belief set an agent holds and the behavior of the agent. In our concept of coherence, we allow partitioning the set into various subsets (the diversity dimension) and we are able to distinguish between several qualitative states of coherence of the system (see Fig. 2). This has advantages with respect to relating the result of a coherence computation to agent behavior, as we will outline in the next section.

Finally, our approach is also able to deal with the theoretical problem of Thagard’s proposal (the issue of convergence of the computation, see section 3), although both proposals are similar insofar as they are based on a network perspective that allows for an interpretation in terms of statistical physics (Thagard and Verbeurgt 1998: 10f, Ott et al. 2005). In both approaches, coherence is related to the network’s ability to synchronize the ‘coherent’ elements. However, Thagard’s approach defines the problem as an optimization problem for a zero-temperature spin glass. As a consequence, convergence to the global optimum, or even to a local optimum, cannot be guaranteed. In our approach we take advantage of a solution to this problem that had been presented in connection with data clustering (Blatt et al. 1996). It proposes a ferromagnetic spin system, rather than a spin glass, in order to avoid the (computationally expensive) problem of spin frustrations. Furthermore, it introduces a statistical description which is not affected by problems of convergence. Consequently, in the ferromagnetic interpretation, incoherence is understood as the inability to synchronize, whereas in the spin glass interpretation, incoherence is related to frustration.

In summary, our proposal has several advantages compared to Thagard’s notion of coherence, as we avoid several open problems and questions that his proposal raised. In the following, we will now sketch a practical application of our proposal. This requires in a first step an outline, why (descriptive) coherence could be a factor that may explain the behavior of agents.

6. Outlining the (possible) causal role of coherence

In the following, we will focus on the use of our concept of coherence in moral psychology and moral philosophy. We outline a possible causal role of coherence along the four ideal types of coherence degrees. We repeat that our goal is to make use of this concept under consideration of the fact that real word moral agents may hold many beliefs, from which only a subset of considerable size is related to a specific decision problem and can be accessed by the agent given limited cognitive resources (e.g. regarding memory). These beliefs can be of different kinds and are (in terms of cognitive psychology) accessible to different degrees (Higgins, 1996). The challenge is, therefore, to gain an understanding of the coherence of *large* belief systems, as one can expect a connection between the similarity of beliefs and their accessibility (whether such a connection exists would be a topic of empirical research using our definition). This is why it's important to fulfill the practicability criterion – i.e. the definition of coherence should allow computing the coherence of large systems as well as identifying the mutual similarities between the elements with reasonable effort.

In this contribution, we will not enter into a discussion of whether beliefs (that may serve as reasons in specific decision situations) have *any* causal role in actions of moral agents (psychologists like Haidt (2001) raised doubts upon that point) or whether moral beliefs require motivational force by necessity in order to be called *moral* beliefs (the internalism-externalism debate, e.g. Brink 1997). We assume that (1) moral agents have beliefs of various types (regarding both factual and normative issues, whereas it will not be possible in all cases to draw a clear distinction between them), (2) some of these beliefs are recruited in specific decision situations, and (3) there exists at least one type of similarity between these beliefs that is relevant for the specific decision situation. We then claim that the structure of this belief-subset, in terms of coherence, is a decisive factor in understanding the actions of moral agents with respect to the specific decision problem. This claim requires (a) to find correlations between different degrees of coherence and specific behavior patterns and (b) to show some causal relation between belief coherence and behavior. If the claim turns out to be true, it would give coherence a predictive value for understanding moral agency. Although this describes basically an empirical project, it is based on a new understanding of the concept of coherence, which makes this claim also interesting from a philosophical perspective.

We acknowledge that our framework has to be related to a model of moral agency that includes the current knowledge of moral science (moral psychology etc.) on the processes moral agents use or rely upon when executing moral agency. Such a model will be necessary in order to investigate a possible causal relation between degrees of belief coherence and observed behavior (e.g. in terms of decisions a moral agent makes). We call this model that refers to the agent's capacity to process and manage moral problems "Moral Intelligence"[10] (Tanner and Christen, submitted). Explaining this model in detail goes beyond the scope of this contribution. However, we briefly outline that moral intelligence both includes (psychological) capacities that are structured along a process model of moral reasoning (moral commitment, moral sensitivity, moral problem solving, moral assertiveness; the model can be seen as a variant of the four component model of Rest 1986) and a reference system containing one's (either existing or newly formulated) moral standards, values or convictions which provide the basis for moral evaluation and regulation (moral compass). The elements of the reference system are conceptualized as 'schemas', a standard notion to represent mental representations of beliefs an agent holds within cognitive psychology. In terms of content types, the moral compass is multifaceted. Moral values, moral convictions, ethical principles, religious beliefs, personal goals, self-related beliefs as well as behavioral scripts, etc., form such ingredients. Depending on the agent and the decision problem, it may be possible that only a single element guides decision making in a specific case. An example of this are protected values that refer to non-instrumental values involving strong moral convictions about the impermissibility of trading of specific values in exchange for other good, in particular monetary benefits (Tanner et al. 2009). However, in the majority of cases we expect that a set of beliefs is involved in moral decision making – and the question is whether the structure of these beliefs tells us something about how the beliefs influence decision making. To answer this question we believe that the notion of coherence could play a crucial role.

[10] As far as we know, Lennick and Kiel (2005) were the first who introduced this term in the context of business ethics.

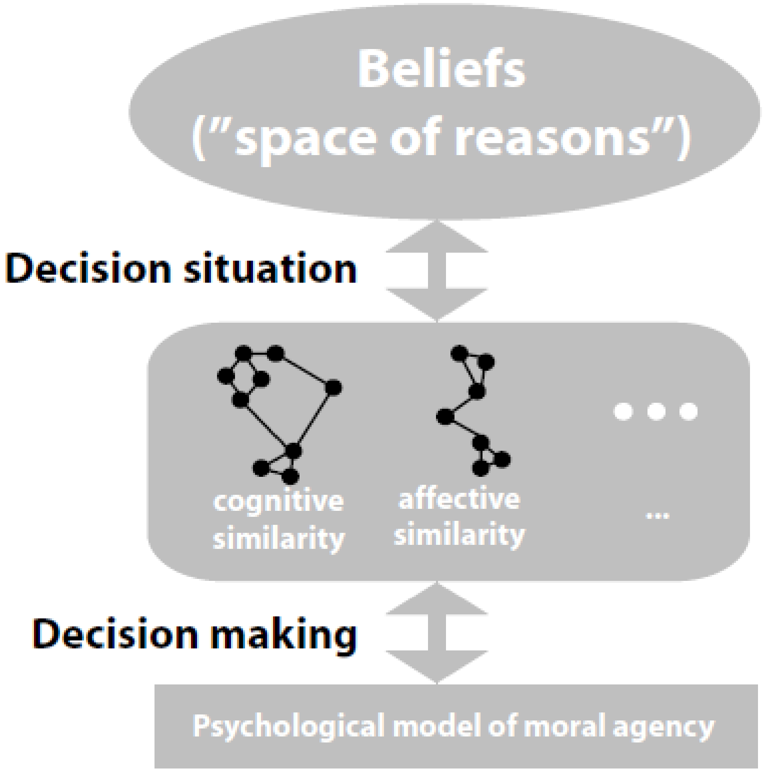

Figure 3: Illustrating the framework for applying a coherence measure in moral psychology: Moral agents possess a large number of moral beliefs accessible to different degrees that are recruited during decision situations. Depending on the kind of similarity between beliefs the belief network (both the single beliefs as well as their interrelations) whose coherence is assessed will differ. Using different similarity measures assessed through empirical research, one may relate the coherence type with the actual decision performed, whereas a psychological model of moral agency serves as an instrument to analyze causalities between coherence type and behavior.

We conceptualize the framework in which the notion of coherence is applied as follows (see Fig. 3 for an illustration): A moral agent possesses probably many thousands of beliefs about the world and evaluations of matters of fact. These beliefs – that can be understood as forming a "space of reasons" using Sellars' (1956) terminology – are accessible to different degrees to the agent and serve as potential reasons in a decision making process upon moral issues. A specific decision problem recruits a subset of those beliefs that may be activated both through fast and intuitive processes and through deliberation. This subset forms the (potential) reasons the agent has in order to decide upon the various options the decision problem poses – it forms the moral compass of the agent. In real world decision making we can expect that the number of elements within this set is not fixed and may change during the decision process (e.g. because the agent realizes that a specific problem involves additional aspects). In empirical research settings, however, one may get a better framed set of beliefs, e.g. by defining a survey that includes specific questions and thus activates in this way the beliefs the agent has regarding a specified decision problem.

After having generated a set of beliefs dedicated to a specific decision problem the question emerges as to how to model the interrelations between these beliefs. This problem both includes a qualitative (what type of interrelation?) and a quantitative (which distance metrics?) aspect. In terms of the qualitative aspect, we can assume that all beliefs recruited in the first step share some semantic similarity in respect of the decision problem for which they have been recruited for – the semantic similarity in relation to the decision problem is actually the reason why they form the belief set in question. However, the decision problem consists of (at least) two options that can be taken – and therefore, it will be possible to define a similarity metric of some beliefs in respect of these options. In empirical applications, this can be operationalized e.g. by a survey that asks, whether a specific belief supports a specific option. This is one way to create a similarity measure within the belief set.[11] It maps the "cognitive structure" of the belief set in question, i.e. if a specific option is able to recruit many beliefs of the set with strong internal similarity, the moral agent has many reasons for that specific option.

An alternative similarity metric relies on the motivational force of specific reasons, as one may expect that not only quantitative issues (i.e. how many reasons does the agent have to do option X?), but also qualitative issues (do the "important" reasons support option X?) play a role in decision making. One way to model this could be to quantify affective responses in relation to the sympathy or aversion that pairs of specific beliefs may induce. For example, an agent may have strong emotional objections that a belief like "doing X is fair towards the employees of the company" and "doing X maximizes profits of the company" are in the same cluster of supportive beliefs for doing X[12] – and a similarity measure that takes this into ac-

[11] This point involves a practical challenge as the number of possible interrelations between *n* elements scales by n^2. Although (depending on the problem) not all possible interrelations may have to be evaluated in empirical research, statistical reasons may require multiple tests of the same pairwise interrelation, i.e. the number of tests that has to be performed can be high. However, a general statement about this issue is not possible, as the number of tests depends on the kind of problems one wants to solve.

[12] This may be the case irrespective to the cognitive similarity these two beliefs may have in a specific decision context. I.e. although it may be the case that option X is both fair to the employees and maximizes profits, the emotional aversion of seeing a similarity between these two beliefs may encode the past experience of the agent that although these two beliefs are often suggested to support the same claim they turned out to be mutually exclusive.

count could impose a sub-structure within a seemingly coherent set of beliefs. In empirical applications, the “affective distance” between beliefs may be measured using both survey techniques as well as physiological measurements to assess unconscious responses of the agent (e.g. skin conductance).

Although we do not claim that those two similarity measures (“cognitive” and “affective” distance) are the only ones one could use, we suggest that they are plausible candidates for a coherence analysis. Technically, both measures can be combined and weighted individually in order to assess those different aspects of similarity and their weight towards the coherence of a decision-specific belief system. Suggestion of how to operationalize these similarity measures in empirical settings will be discussed the next section.[13]

7. Coherence types of moral belief systems

A theory of moral agency should explain how entities act with reference to right and wrong. Such a generalized theory of moral agency involves the clarification and explanation of various aspects: Agency (e.g. individual and collective agency), the ontogenesis of moral agency (moral development), moral cultural history and the phylogenies of moral agency (the evolution of morality). As the essence of human morality is not only the ability to follow moral norms and principles, but also the ability to question an existing moral framework based on new justifications, one cannot understand moral agency without taking beliefs into account that serve as reasons in the justification process. Not only compliance to moral norms, but also the way we justify the application of specific norms – e.g. in dilemmatic situations – will play a role when moral agency is assessed. Our understanding of coherence intends to give new insights into this justification process by offering a way to empirically assess the belief systems of moral agents. In this section, we briefly sketch the application of our concept of coherence to understand specific aspects in moral agency and exemplify this by discussing the example of Thagard (the Bernardo Case).

[13] We remind the reader that the similarity measure alone (i.e. how to quantify a single belief in respect of semantic or affective aspects) is not the only point to consider in practical applications. The distance metrics (i.e. how to quantify the distance between beliefs) may play a role, too. This goes along with the obligation to evaluate several distance metrics in practical applications in order to find a metric that is suited to the problem. Usually, standard distance metrics like the Euclidean or Manhattan distance serve the desired purpose.

A well-studied topic both in moral psychology and moral philosophy are dilemmas, i.e. decision situations that force individuals to make trade-offs between moral values that have a similar status for the agent, or between conflicting consequences of a single value. These „tragic trade-offs conflicts" that, e.g., pit two protected values against each other (such as human life vs. another human life) are particularly stressful. Such situations are not only eliciting high levels of negative feelings, they are also perceived as highly difficult to solve, as the decision-maker is forced to violate one of the values, if no other new solution can be found that allows upholding both values. In our framework of quantified coherence, we predict that such dilemmatic situations would be described as a combination of high-coherence diversity and stability, as several strong sub-groups exist that are inherently stable and mutually incohesive. This prediction per se is not very surprising – however, the application of different types of similarities (cognitive and affective) may allow distinguishing different types of dilemmatic settings. For example, one setting using cognitive similarity could show a paradigmatic case-4 situation (see Figure 2), i.e. a situation in which two contradicting options are both supported by two belief sets of similar size and stability – whereas the application of an affective similarity measure does not provide this picture but reveals a case-3 situation. This may indicate a dilemmatic situation in terms of fulfilling a certain rationality standard (i.e. the situation is dilemmatic because two contradicting options are both supported by many good reasons) – but not in terms of emotional involvement of the agent (as the "important" reasons are not in mutually exclusive belief clusters). In this way, different kinds of dilemmatic situations could be distinguished and their effect on different decision settings (e.g. personal decision making versus group decision making) could be investigated.

The Bernardo case serves as an example of such a situation, as it illustrates the dilemmatic situation that – confronted with a brutish murder case – an agent may reconsider his initial rejection of capital punishment. Just to exemplify our method, we have applied it to this (small) belief set[14] using three similarity relations: First, we translated the network of Thagard in a similarity matrix. Whenever two beliefs are positively connected, the entry in the matrix is "0" (otherwise "1" for indicating maximal dissimilarity or distance). Second, we performed a simple experiment to evaluate the "cognitive" similarity. In this experiment, one subject weighted the probabilities for each pair of beliefs that these beliefs can be part of a common argument for or against the execution of Paul Bernardo. This probability was translated into a

[14] The node „empirical evidence" in Thagard's initial network has been excluded, as it represents a different kind of belief; see footnote 4 for further explanations.

distance (i.e. a distance close to 0 represents a high probability that there is a common argument, whereas a number close to 1 represents low probability). Third, we repeated the experiment to evaluate the “affective” similarity. In this experiment, the subject had to judge his feelings with respect to the concurrent use of two beliefs in an argumentation pro/contra capital punishment (i.e. a distance close to 0 represents a positive feeling with respect to the concurrent use, whereas a number close to 1 represents a negative feeling).

The cognitive and the affective similarity of two beliefs can be different. For example, one may put a high probability on the concurrent use of the beliefs “reduce prison costs” and “capital punishment is justified” (low cognitive distance), whereas one may have emotional difficulties to jointly use these beliefs, because they seem to indicate that a person is executed to save money (high affective distance). After performing the coherence analysis, we find the following result (Fig. 4): Although all three measures put the belief set in the “dilemma zone” (which is not surprising given the setup of the belief system), there are differences with respect to diversity and stability. In the clustering-based coherence analysis. the “Thagard-similarity” produced three clusters resulting in a higher diversity – which is not surprising given the fact that the similarity matrix is sparse and does not take into account other potential relations between the beliefs. More interesting are the differences between the cognitive and the affective distance. The cognitive distance represented the dilemma in its classical form: the most stable cluster includes the statements “Capital punishment helps to prevent serious crimes”, “Preventing serious crime is good”, “Capital punishment is sometimes justified”, “reduce prison costs” and, finally, “execute Paul Bernardo”. Using the affective distance, however, changed the picture. Here, the most stable cluster included the statements “Capital punishment is not a deterrent”, “Preventing serious crime is good”, “killing a defenseless victim is wrong”, “capital punishment is wrong”, and, surprisingly, “execute Paul Bernardo”. The “emotional cluster” is less stable than the “cognitive cluster” – and it obviously represents the conflicting situation the Bernardo Case induced: namely that a person who is generally against capital punishment still thinks that in the case of Bernardo, the person should be executed. Thus, a decision that seems to reflect practical irrationality is reframed as resulting from a decision procedure in which affective, and not cognitive, similarities between the beliefs involved may have played the decisive role. This also has normative consequences, as it is not so clear which of the two kinds of similarities should count as “more rational” given the important role of emotions in establishing what is important for humans (Rolls 2005).

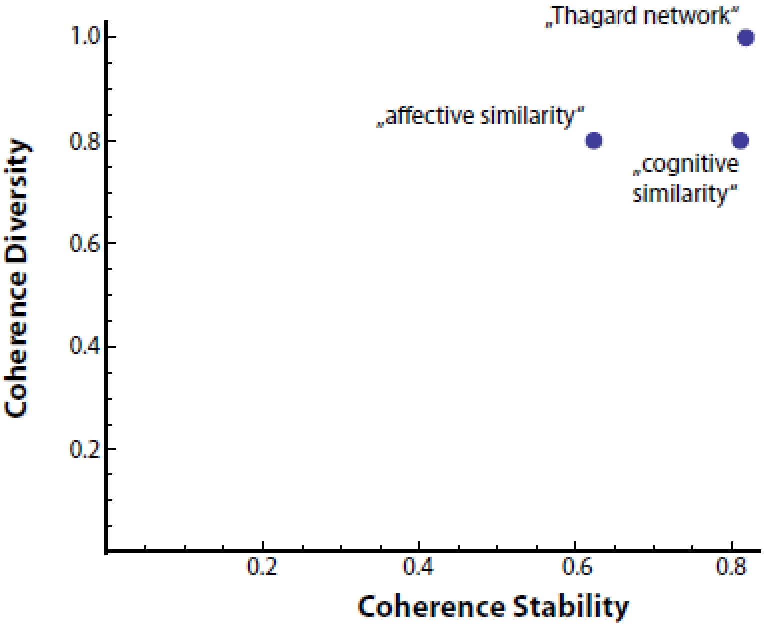

Figure 4: Calculating the coherence of Thagard’s belief set for the “Bernardo murder case” example using our measure with three different similarity measures: a direct mapping of Thagard’s network (see Fig. 1) into a distance matrix, a “cognitive similarity” and an ”affective similarity”.

We remind the shortcomings of our analysis whose role is purely to exemplify our method: First, the number of the beliefs is small and does probably not include all relevant beliefs a person may have to judge this case. One would have to collect these beliefs in a first round of the experiment in order to make a more valid statement. Second, one would have to perform this experiment with various persons in order to validate the differences that emerge using the cognitive and affective similarity. However, the example shows that allowing for degrees of coherence along two dimensions gives a more detailed picture of the decision situation and allows analyzing specific facets of the problem.

8. Conclusion

In moral philosophy, coherence is an often-mentioned, but rarely defined concept in order to state whether moral beliefs are justified guides of actions. In our contribution, we presented an alternative understanding of the coherence of belief systems that is quantified (i.e. precisely defined), that allows dealing with large belief sets and that can be adapted to investigate empirical issues about moral agency. We based our argument on the assumption that a moral agent has a large number of (moral) beliefs that are turned into reasons by the agent in specific decision situations. These beliefs are related to each other not by various kinds of similarities that presupposes distance metrics. Our concept allows quantification of the degree of coherence of this set of moral beliefs along two dimensions: the diversity of sub-groups of beliefs and the stability of a set of beliefs. In this way, four ideal types of coherence of belief-sets that are associated with moral decisions are identified, that are predicted to have different implications for the behavior of the moral agent. We compared our proposal with the definition of Thagard and showed how we are able to overcome some shortcomings of latter definition.

The question remains, to what extend clarifications with respect to the descriptive use of coherence are of relevance for its normative importance. We belief, that there are several answers to this question. The first one refers to the very basic critique that seems to be independent of any definition of coherence – namely the objection that the coherence of a set of belief representing, for example, a theory does not imply that the theory itself is true. However, although it seems to be possible to construct simple sets of coherent beliefs that are obviously false, it remains an open question whether it is really possible to create larger sets that indeed refer to real word problems (or theories about them) that are coherent, but wrong. A gradual understanding of coherence furthermore may ease the critique: one may indeed find systems of (lower) coherence that are wrong and are outcompeted by systems of higher coherence. The second answer refers to the (psychological well studied, see section 2) fact, that people seem to "like" coherence – an important factor that motivates the normative significance of coherence –, although they are sometimes (or maybe: often) unable to maintain it. Our methodology allows the assessment of this gap in more detail and could find reasons for it, which may support the motivational foundation of the normative use of coherence. Finally, making coherence an observable that can be assigned to qualitative different system states could also serve as a tool for self-understanding and thus could become an instrument that allows training the justificatory use of coherence.

Appendix: Exposition of the Measure and Operationalization

Our measure of coherence is based on superparamagnetic clustering (SPC, Blatt et al. 1996) and sequential superparamagnetic clustering (SSC, Ott et al. 2005). Based on these algorithms, we have defined a measure of coherence that captures both the stability and the diversity component of coherence (Christen et al. 2009). The stability component of coherence

$C_{stability}$ is calculated in the SPC framework. It is evaluated with respect to the disintegration of the largest cluster $\bar{c}$ for increasing temperature T until the system's order completely breaks apart, where T is the parameter that models the stress on the system. This involves the assumption that the largest cluster represents the 'core' of the belief system that disintegrates under stress.

Let $CS(t)$ be the size of the largest cluster for $T = t$. We assume that $CS(0) = n$, where n stands for the total number of data points; i.e., without stress, all beliefs are in the same cluster. Upon an increase in stress, $CS(t)$ decreases until $CS(t) = 1$ for some $t = T_{end}$. The average decay curve serves as a measure of coherence stability.

$$C_{stability} = \frac{1}{T_{end}} \int_{0}^{T_{end}} \frac{CS(t)}{n} dt$$

The measure is normalized to the interval [0,1]. $C_{stability}$ is close to 1 if the largest cluster remains intact for a long time and then disintegrates rapidly for high T, whereas $C_{stability}$ is close to 0 if the largest cluster disintegrates rapidly and only a small core is stable over a longer interval.

In the actual analysis, $CS(t)$ is calculated in $l+1$ discrete steps $t = 0, \Delta T, 2\Delta T, ..., T_{end} = l\Delta T$. For the approximate calculation of the integral, the trapezoidal rule, known from basic calculus, is used.

$$C_{stability} = \sum_{i=0}^{l-1} \frac{CS(i\Delta T) + CS((i+1)\Delta T)}{2nl}$$

Coherence diversity $C_{diversity}$ is calculated using SSC, yielding a binary tree in which the size of each of the k sub-clusters is evaluated. Again, we consider the largest cluster $\bar{c}$ as the 'core' of the system. $C_{diversity}$ is calculated as the sum of the distance of each cluster c_i from the largest cluster in the tree diagram weighted with its size $|c_i|$. The 'tree distance' $\bar{d}_i$ is the number of bifurcation points in the tree between $\bar{c}$ and c_i. Both the maximal tree distance $\bar{d}_{\max}$ and the size of the largest cluster serve as calibration factors, leading to the definition:

$$C_{diversity} = \sum_{i=1}^{k} \frac{\bar{d}_i}{\bar{d}_{\max}} \cdot \frac{|c_i|}{|\bar{c}|}$$

$C_{diversity}$ is not normalized to 1 according to the current definition. Its value is 0 if SSC does not reveal any sub-clusters, and it is close to 0 if only small clusters emerge. However, many large clusters that have a large tree distance from the largest cluster, or fewer clusters with similar size, lead to an increase in $C_{diversity}$. Since $C_{diversity}$ is typically far below the maximally possible value, the normalization was skipped to simplify the calculation.

In this way, the measure consisting of the two components $C_{stability}$ and $C_{diversity}$ is able to capture the intuition of coherence outlined in Figure 2. The concept was tested extensively and approved on the basis of toy data (Christen et al. 2009).

Acknowledgments

We thank Clare Estelle Jackson and Mark Alfano for English corrections of the original draft. This work has been supported by the Cogito Foundation (Wollerau, Switzerland, grant number R-143/08) and by the Hasler-Stiftung (Berne, Switzerland).

References

Abelson RP (1983) Whatever became of consistency theory? Personality and Social Psychology Bulletin 9: 37-54

Abelson RP, Aronson E, McGuire WJ, Newcomb TM, Rosenberg MJ, Tannenbaum PH (eds.) (1968) Theories of cognitive consistency: A sourcebook. Rand McNally, Chicago

Aronson E, Carlsmith JM (1963) The effect of the severity of threat on the evaluation of forbidden behavior. Journal of Abnormal and Social Psychology 66: 584-588

Blatt M, Wiseman S, Domany E (1996) Superparamagnetic clustering of data. Physical Review Letters 76: 3251-3254

Brink D (1997) Moral Motivation. Ethics 108: 4-32

Cervone D, Shoda Y (1999): Beyond traits in the study of personality coherence. Current Directions in Psychological Science 8(1): 27-32

Christen M, Starostina T, Schwarz D, Ott T (2009) A spin-based measure of the coherence of belief systems. Proceedings of NDES 2009, 21-23 June 2009. Rapperswil

Converse PE (1964) The Nature of Belief Systems in Mass Publics. In: Apter DE (ed) Ideology and Discontent. Free Press, New York, pp 206-261

Daniels N (1979) Wide Reflective Equilibrium and Theory Acceptance in Ethics. Journal of Philosophy 76 (5): 256-282

Friederici AD, Steinhauer K, Frisch S (1999) Lexical integration: Sequential effects of syntactic and semantic information. Memory & Cognition 27 (3): 438-453

Gigerenzer G, Gaissmaier W (2011) Heuristic Decision Making. Annual Reviews in Psychology 62: 451-482

Haidt J (2001) The emotional dog and its rational tail: A social intuitionist approach to moral judgment. Psychological Review 108: 814–834

Hastie RK, Dawes RM (2009) Rational Choice in an Uncertain World: The Psychology of Judgment and Decision Making, 2nd ed. Sage Thousand Oaks, CA

Higgins ET (1996) Knowledge activation: Accessibility, applicability, and salience. In: Higgins ET, Kruglanski AW (eds) Social psychology: Handbook of basic principles. Guilford Press, New York, pp 133-168

Hoffmann M (2008) Kohärenzbegriffe in der Ethik. Walter de Gruyter, Berlin

Jolliffe T, Baron-Cohen S (1999) A test of central coherence theory: linguistic processing in high-functioning adults with autism or Asperger syndrome: is local coherence impaired? Cognition 71: 149-185

Jordan J (2009) A social cognition framework for examining moral awareness in managers and academics. Journal of Business Ethics 84: 237-258

Jussim L (2005) Accuracy in Social Perception: Criticisms, Controversies, Criteria, Components, and Cognitive Processes. Advances in Experimental Social Psychology 37: 1-93

Keil FC (2006) Explanation and understanding. Annual Review of Psychology 57: 227-254

Kirkham RL (1992) Theories of Truth: A Critical Introduction. MIT Press, Cambridge MA

Lapsley DK, Narvaez D (2004) A Social-cognitive approach to the moral personality. In: Lapsley DK, Narvaez D (eds) Moral development, self and identity. Erlbaum, Mahwah NJ, pp 189-212

Lennick D, Kiel F (2005) Moral Intelligence: Enhancing Business Performance and Leadership Success. Wharton School Publishing, New Jersey

Ott T, Kern A, Steeb W-H, Stoop R (2005) Sequential clustering: tracking down the most natural clusters. Journal of Statistical Mechanics, P11014

Putnam H (1982): Reason, Truth and History. Cambridge University Press, Cambridge

Quine WV (1979) On the Nature of Moral Values. Critical Inquiry 5(3): 471-480

Rawls J (1971) A Theory of Justice. Harvard University Press, Cambridge

Rescher N (1973) The Coherence Theory of Truth. Oxford University Press, Oxford

Rest JR (1986) Moral development: Advances in research and theory. Praeger, New York

Rodwan AS (1965): A coherence-criterion in perception. The American Journal of Psychology 78(4): 529-544

Rolls ET (2005) Emotions Explained. Oxford University Press, Oxford

Sellars W (1956) Empiricism and the Philosophy of Mind. Harvard University Press, Cambridge, MA

Silverstein SM, Uhlhaas PJ (2004) Gestalt Psychology: The Forgotten Paradigm in Abnormal Psychology. The American Journal of Psychology 117(2): 259-277

Thagard P, Verbeurgt K (1998) Coherence as constraint satisfaction. Cognitive Science 22: 1-24

Thagard P (1999) Ethical coherence. Philosophical Psychology 11: 405-422

Thagard P (2000) Coherence in thought and action. MIT Press, Cambridge, MA

Trumbo D, Noble M, Fowler F, Porterfield J (1968): Motor performance on temporal tasks as a function of sequence length and coherence. Journal of Experimental Psychology 77(3): 397-406

White LB, Boashash B (1990) Cross Spectral Analysis of Non-Stationary Processes. IEEE Transactions on Information Theory 36(4): 830-835

Williams B (1985): Ethics and the Limits of Philosophy. Fontana, London

Winter RG, Steinberg AM (2008) Coherence. AccessScience, McGraw-Hill Companies. Available through: http://www.accessscience.com. (Accessed on September 28th 2011)