teorema
Vol. XXXIV/3, 2015, pp. 197-219
ISSN: 0210-1602
[BIBLID 0210-1602 (2015) 34:3; pp. 197-219]

Intuition and Reason: Re-Assessing Dual-Process Theories with Representational Sub-Activation

James Trafford and Alexandros Tillas

RESUMEN

En la literatura sobre el razonamiento prevalece una distinción entre los procesos tipo-1 (rápidos, automáticos, asociativos, heurísticos e intuitivos) y los procesos tipo-2 (basados en reglas, analíticos y reflexivos). En este artículo, apoyamos la evidencia empírica reciente [De Neys (2006b); Osman (2013)] a favor de un sistema cognitivo único. En concreto, proponemos que las intuiciones (procesos tipo-1) son representaciones subactivadas, influenciadas a su vez por los pesos de las conexiones entre distintas representaciones. Asimismo, explicamos los sesgos mediante el papel de la atención en los procesos de pensamiento. Así, nuestra propuesta explica el razonamiento y el sesgo, a la vez que soluciona algunos de los problemas a los que se enfrentan las explicaciones basadas en procesos duales.

PALABRAS CLAVE: *intuiciones, razonamiento, teorías del proceso dual, sesgos cognitivos, redes representacionales, atención, corrección de sesgos*.

ABSTRACT

There is a prevalent distinction in the literature on reasoning, between Type-1 processes (fast, automatic, associative, heuristic and intuitive); and Type-2 processes (rule-based, analytical and reflective). In this paper, we follow up recent empirical evidence [De Neys (2006b); Osman (2013)] in favour of a unitary cognitive system. More specifically, we suggest that intuitions (T1-processes) are sub-activated representations, which are in turn influenced by the weightings of the connections between different representations. Furthermore, we explain biases by appealing to the role of attention in thinking processes. The suggested view explains reasoning and bias whilst dealing with extant problems facing dual-process accounts.

KEYWORDS: *Intuitions, Reasoning, Dual-Process Theories, Cognitive Biases, Representational Networks, Attention, Cognitive Debiasing.*

I. INTRODUCTION

I.1 *The Current Approach*

In the literature on reasoning, there is a prevalent distinction made between two processes, both of which underlie cognition, and often conflict with each other. Type-1 (T-1) processes are characterized as fast, automatic, associative, heuristic and intuitive; whereas Type-2 (T-2) processes are rule-based, analytical and reflective [Epstein (1994); Evans (1989); Evans and Over (1996); Sloman (1996); Stanovich (1999)].

The motivation for this distinction is, in large part, an attempt to understand the distinction between intuition and reasoning. For example, there are well-known discrepancies between the two as seen in matching biases [Wason and Evans (1975)], and biases pertaining to probabilistic reasoning [Kahneman and Tversky (1982)]. The latter are exemplified by the now infamous 'Linda problem' [Tversky and Kahneman (1982)]. In such cases, around 90% of participants routinely and systematically give a response suggesting that the probability of a conjunction (A&B) is greater than one of its conjuncts (A), (i.e. $P(A\&B) \geq P(A)$). One response to such evidence is to suggest that a 'representative heuristic' is involved in participants' judgments, which is generated from T-1 processes. This, it is suggested, is indicative of a 'belief bias' [Stanovich (1999)] across human reasoning. The bias is also suggested by experiments where subjects presented with an invalid syllogism are asked to assess its validity [Sá, West and Stanovich (1999)]. In cases where the conclusion is believable, 68% responded that it is valid. Similarly, there is evidence suggesting that subjects tend to reject valid arguments where the conclusion is unbelievable [Evans, Barston and Pollard (1983)]. Furthermore, dual process theories (DPTs) provide plausible explanations of interpersonal variance in reasoning [Stanovich (1999); Stanovich and West (1997)]; and cross-cultural variation in reasoning [Norenzayan et al. (2002)].

DPTs deal with these results by suggesting that these cases involve bringing prior beliefs to bear upon reasoning as a result of fast, automatic, associative T-1 processes.[1] These in turn, are thought to require 'override', or inhibition, by engaging slow, reflective, analytic T-2 processes [Stanovich and West (2000); Chen and Chaiken (1999)]. Many theorists have gone on to suggest that the two processes are distinct enough to be characterized as two cognitive systems [Epstein (1994); Evans (1989); Evans and Over (1996); Sloman (1996); Stanovich (1999)].

T-2 processing is generally slow and sequential, and involved in general intelligence processing such as hypothetical thinking, mental

simulation, and decision-making [Evans (2007; 2010)]. According to Evans and Stanovich [(2013)], hypothetical reasoning requires a 'cognitive decoupling', i.e. keeping real-world representations separate from representations of imaginary situations, such as planning, counterfactual reasoning and so forth. Cognitive decoupling is the central feature of T-2 processing [Stanovich (2009; 2011)]. Finally, according to DPTs working memory (WM) becomes engaged *only* in T-2 processing.

Unlike T-2, T-1 processes are largely autonomous since they do not require 'controlled attention' and have minimal demands on WM resources. Execution of T-1 processes is initiated *only* in a bottom-up manner, i.e. triggered on presence of the appropriate stimulus. Operation of T-1 is mandatory and independent from higher-level cognition [Stanovich (2004; 2009a; 2011)]. Despite the fact that all T-1 processes share an autonomous nature, they should not be construed as similar with regards to their neurophysiology and etiology.

I.2 *Issues with the Two-Systems Approach*

Whilst there has been much consensus regarding the distinction between T-1 and T-2 cognitive processes, the suggestion that there are two distinct cognitive systems has recently come under fire, and has been rejected by many of its original proponents [Evans (2006); Stanovich (2009)].[2]

Nonetheless, it remains central to this programme that biases result from intuitions that require override through reason and analysis, and only participants that can suppress their intuitive response can correct their mistakes in relation to a normative theory. This view is bolstered by experimental data suggesting that subjects that are primed by making the normative structure apparent more successfully return the normatively 'correct' response by e.g. the use of Euler circles [Sloman and Over (2003)]; or using 'foreign' terms to construct syllogisms [Sá, West and Stanovich (1999)].

However, recent work suggests that such a conception of top-down inhibition by T-2 processes is overly-simplistic: (a) recent research suggests that T-2 systems are also liable to bias [Evans (2006); Thompson and Evans (2012)]; (b) T-2 processes appear to often be engaged to post-rationalise intuitive responses, rather than to exert cognitive control [Evans and Over (1996)]; (c) the intuitive phenomenology or 'gut' feeling as well as the belief bias appears to remain after logical reasoning is engaged [Evans, Allen, Newstead and Pollard (1994)].[3]

Resultantly, we identify (at least) four significant issues in the extant literature:

A. Dual-process approaches suggest a traditionally rationalistic approach to theory of mind in which biased intuitions are 'controlled' by serialized, normative processes. This suggests that we have a standard normative theory (usually classical logic / probability theory) – but this is rather simplistic given the numerous logical systems on the market, e.g. paraconsistent logics, relevant logics, non-monotonic logic and so on. Stanovich (2004) and Oaksford et al. (1996) recognize these issues, but they are largely undeveloped.[4] Moreover, the measurement of 'deviance' from rationality in the construction of experiments is typically measured against a classical normative theory.

B. The relationship between processes of reasoning seems vastly more multifarious and mutually supportive than the dual-systems approach allows. For example, it looks like systematic reasoning processes may also distort and bias heuristic reasoning processes. Moreover, there is evidence that T-2 processes may be engaged to post-rationalise T-1 processes in relation to introspective reports on behaviour. As it stands, we have no device for determining which expressed beliefs would represent an override by engaged reasoning, and which are mere rationalizations of T-1 processes [Evans and Over (1996)].

C. The intuitive pull of heuristic reasoning processes that appears to exist long after reflection is not really dealt with by DPTs (i.e. they don't get the phenomenology right). Further work on the interrelated nature of reasoning processes is thus required – there is a problem of control here i.e. conscious reasoning does not seem to be able to exert the kind of control suggested by the dual-process approach (e.g. Kahneman and Frederick (2002) for this suggestion; Evans et al. (1994) for the suggestion that reasoning does not remove belief biases).

D. The dual-systems account may be overly internalistic, and thus not give the external scaffolding of systematic reasoning processes its due. In this respect, the role of learning, formalization, technologies (e.g. writing, math, etc.) are not adequately understood within the framework (and this is relevant to issue 1).[5] There are also rea-

soning processes that appear to begin under conscious, slow, control, but later become automated, such as learning logic and mathematics (e.g. Monsell and Driver (2000) in relation to attention and motor control). This suggests that the distinction between non-biased processes as consciously controlled, as opposed to biased processes as unconscious and automatic, may be too simplistic.

In sum, these issues suggest that it is worthwhile revisiting the explanation of biases in terms of dual processes by investigating the structure of intuition, and so, more properly understanding its relation with reasoning.

I.3 *DPTs and Empirical Evidence*

In this section we discuss recent empirical evidence in the literature against DPTs, with an eye to fleshing out a philosophical account of a unitary system.

• Accuracy and WM Loadings

De Neys (2006b) focuses on some of the necessary defining characteristics of T-1 processing such as automaticity and immunity to WM loadings, as well as accuracy of performance in tasks that are allegedly executed in virtue of T-1 processing. In this study, participants were classified in three groups (low, medium and high) with regard to their Working Memory Capacity (WMC), and were asked to evaluate a series of syllogisms. The believability of the conclusion of each syllogism was either incongruent with the logical conclusion (conflict syllogism) or congruent with it (no-conflict syllogism). While participants were evaluating the syllogisms, their WM was manipulated by using a dot memory test (high-, low-, no-loading). Namely, after evaluating a syllogism, subjects were asked to replicate a dot-array in a 3x3 grid that they saw prior to evaluating the syllogism.

• Qualitative or Quantitative differences between T-1 and T-2 processes?

Intuitively, DPTs predict that no-conflict syllogisms are processed in terms of T-1 processing. For syllogisms with cohering premises and conclusion are easily integrated and evaluated, and subjects exhibit a tendency to accept the syllogism as valid [belief bias effect; Wilkins (1928); reported in Osman (2013)]. Therefore, syllogisms of this kind should be easy-to-evaluate, without loading WM.

De Neys (2006b) shows that subjects tend to evaluate syllogisms by evaluating the believability of the conclusion. Accuracy levels, with re-

spect to validity, were near ceiling, regardless of WMC and WM loadings. In particular, no-conflict syllogisms posed no WM loadings, and accuracy levels, also with respect to validity, were near ceiling. According to Osman (2013), this suggests that T-1 processes underpin such evaluations.

On DPTs, responses in conflict tasks are generated in virtue of T-2 processing. This, presumably, is due to conflict syllogisms being hard(er) to evaluate, since when logic and belief disagree, subjects tend to base their responses on the believability of the conclusion. In evaluating conflict syllogisms, subjects integrate the premises and conclusion. This is a rather hard task since this integration is in contrast with the subjects' existing world-knowledge [Klauer, Musch, and Naumer, (2000)]. In turn, integrating premises and conclusion has an effect of WM loading [De Neys, (2006b)]. According to De Neys, regardless of the WMC of participants evaluating the syllogisms, accuracy levels were severely affected as the WM load increased. Given the defining characteristic of T-2 processing, it seems that responses were generated by T-2 [Osman (2013)].

Osman argues that the above results do not indicate a qualitative difference between the two kinds of processes but merely a quantitative distinction along a single dimension – the difficulty of the task. Similarly, participants are doing the same thing – call this belief management – in both cases of conflict and non-conflict syllogisms. Simply, in the case of valid syllogisms with believable conclusions (or invalid syllogisms with unbelievable conclusions), the 'logical response' and the 'belief management' response coincide. In turn, the issue of integrating premises that are in conflict with real-world knowledge is rather a different one, which also plays a role in the way reasoning tasks are processed.[6]

• Speed of T-1 processes and immunity to WM loadings

A further defining characteristic of T-1 processes according to DPTs is that they generate faster responses when compared to responses produced by T-2 processes. For, T-1 processes are allegedly autonomous and immune to WM loadings, which naturally slow down reasoning processes [Posner (1978); reported in Osman (2013)].

De Neys (2006a) examined the immunity of T-1 processes to WM loadings by presenting subjects with four cards alongside a conditional statement such as an indicative or a deontic task, and subjects were asked to pick out cards that would test the rule.[7] Performance accuracy levels in deontic tasks are supposedly much higher than in indicative tasks since prior knowledge about real-world infractions of regulations facilitates responses. Furthermore, deontic tasks are allegedly executed in virtue of T-

1 processing. Thus, Osman (2013) suggests that, according to DPTs, we can predict that WM loadings should have little or no effect on accuracy levels in deontic tasks since they are supposed to be generated by T-1 processes, which are immune to WM loadings. There should, however, be an effect of WM loadings on accuracy levels in indicative tasks.

De Neys (2006a) shows that performance accuracy levels in both deontic and indicative tasks were significantly compromised under WM loading condition. As a consequence, Osman argues that either reasoning in deontic tasks is not generated by T-1 processes, or that T-1 processes are not autonomous. Based on this evidence, critics of DPTs argue that there are not two qualitatively distinct kinds of processes but rather one that generates accurate responses to simpler tasks and inaccurate responses to harder tasks.

• How automatic is automatic?

With regards to the alleged 'automatic' nature of T-1 processes, De Neys (2006a) examined the response speed of subjects on deontic and indicative tasks, and recorded latencies for time spent reading the instructions and time spent making inferences. The recorded results show that under WM load conditions, participants took about 26sec. to make an inference for deontic tasks vs. 20sec. for indicative tasks. Under no WM load conditions, subjects spent approximately 22sec. before making an inference in deontic tasks, and 20sec. for indicative tasks.

Comparison of times spent making inferences between both versions of tasks, and regardless of load manipulation, showed little or no difference [De Neys (2006a); Roberts and Newton (2001); reported in Osman (2013)]. Furthermore, 26 sec. is too long for an automatic process. These results contrast with the claim that T-1 processing solves deontic tasks.

• What is automatic in T-1?

Relatedly, De Neys (2006a) also recorded judgment times of responses to the conjunction fallacy (Linda problem). Typically in the DPTs literature, T-1 processes are thought to generate erroneous responses, given that the information presented in the description of the task automatically triggers prior knowledge. In turn, prior beliefs are deployed while making judgments about the likelihood of various statements that conflict with the correct response.

De Neys reports that it took participants approximately 47sec. (excluding time spent on reading the instructions) to make the erroneous judgments, and approximately 57sec. to make the correct judgments. De-

spite the fact that generation of correct responses took, on average, longer than erroneous responses, 47sec. is again too long a process to qualify as automatic. Recall that participants spent, on average, 19sec. to generate the accurate response in deontic selections tasks. Thus T-1 processes seem to vary significantly across tasks (47sec. – 19sec.) – hardly a characteristic of an autonomic process, unless one is ready to accept that automaticity and speed of response are task-relative measurements.
Against this background, we turn to suggest a philosophical account of how a unitary system operates.

II. RSA AS A UNITARY SYSTEM

In this section, we put forth a suggestion in the form of the Representational Sub-activation Thesis (RSA), according to which intuitions are sub-activated representations, which are in turn influenced by the weightings of the connections holding between different representations. These representational structures are associatively conditioned, largely network-like, and linked with other mental states such as beliefs and emotional states. Unlike DPTs, we explain biases by appealing to the role of attention in thinking processes. The RSA framework illuminates and provides explanation of reasoning and bias, deals with the extant problems facing DPTs, and accommodates the above evidence.

II.1 *Background Commitments*

In presenting RSA, we commit to a weak version of representationalism, as well as a view of concepts as structured entities built out of perceptual representations. RSA builds upon an empirically vindicated version of Neo-Empiricism [e.g. Prinz (2002); Barsalou (1999)], and departs from traditional Empiricist accounts of concept learning. We have developed a detailed concept acquisition elsewhere [(Tillas (2010); (forthcoming 2015a)], where we explain how different kinds of concepts, including concepts with heterogeneous instances, e.g. DANGER, as well as lofty concepts like DEMOCRACY, could derive from perceptual experiences. Crucially, all concepts are linked back to experience temporarily, but they do so *indirectly*. In this sense, the causal relation between mind and world is rather sophisticated. In the light of perceptual experiences with instances of a given kind, the connection weightings of neuronal groups that ground them are adjusted. Here we appeal to the ubiquitous-

ly accepted principles of Hebbian [(1949)] learning, which is vindicated by contemporary evidence for Long Term Potentiation (LTP).[8]

Admittedly, many views hold that concepts are (or are tokened by) mental symbols; for the most part, our suggestions can be reformulated in those terms, so long as such views can deal with the idea that mental symbols are connected with adjustable weightings.

II.2 *Perceiving the World and Building Networks*

In line with what has been mentioned above, on perceiving the first instance of a given kind a representation is formed and stored in long-term memory. On encounter with a subsequent instance that does get attended and recognized as a subsequent instance of the same kind, a representation is also formed, while a scanning process becomes initiated and a match is sought for in the subject's memory. The same process was also initiated during encounter with the first instance but the scan did not yield any matching stored representations.

Once a match between the currently formed and a stored representation is found, the existing representation becomes activated.[9] The activated matching stored representations lead perception in a top-down manner to the same parts of the perceived object that have been previously attended [cf. Elman and McClelland's (1986)]. This process is enabled by representations of parts carrying information about their position in the visual field, part-whole and part-to-part relations. Furthermore, the currently formed representation get stored in the same locus [cf. Perry, (2001)] or 'closer' to the stored matching ones.[10]

On encounter with subsequent instances of a given kind, a similar process is initiated, and the appropriate mental-file becomes informationally enriched. See also [Tillas, (2014); (forthcoming 2015b)] for a detailed discussion of related issues.

Digressing slightly for a moment, given that the suggested views builds heavily upon the principles of associationism as well as top-down and bottom-up influences, it naturally converges with what is known in the literature as 'predictive coding' [e.g. Clark (2013)]. On that view, natural signals are highly redundant due to spatial and temporal uniformity. For instance, intensities of pixels tend to be correlated over time given that objects persist in time. Given these redundancies in natural signals, the suggestion is that representing a raw image directly in terms of the activity of a set of sensory receptors would be very inefficient. Thus, it is argued that early sensory processing aims at reducing such redundancies and recoding sensory inputs in more efficient ways. Predictive coding

postulates that neural networks learn the statistical regularities inherent in external stimuli, and reduce redundancy by removing components of the input that are predictable. In turn, they will transmit residual errors in prediction. In the case of visual perception, for example, cells in early visual circuits convey not the raw image intensity, but the difference between the predicted value and actual intensity.[11]

II.3. *Co-occurrence and Network Dynamics*

In shedding light upon the relations between different representations, consider Hebb's [(1949)] famous rule of learning according to which 'cells that fire together, wire together'. According to Hebb, temporal coincidence of pre- and post-synaptic activity results in a strengthening of that synapse. Thus, whenever two neurons become excited simultaneously, the connections between them are strengthened. In turn, repeated participation of neuron *A* in firing neuron *B* entails an increase of the strength of the action of *A* onto *B*. That is, on the basis of repeated simultaneous activation, the efficiency of one cell to activate the other is increased.

On the hypothesis that concepts are built out of and also become activated in virtue of activating representations, which are ultimately grounded in neuronal activations, Hebbian associations also obtain between concepts. In this sense, concepts become also associated and form networks. The more frequent concepts co-occur, the stronger the connections between them will grow. Once a given concept becomes activated, concepts strongly associated to it will ceteris paribus also become activated. Given the widely accepted view that concepts are the building blocks of thoughts, the connections between concepts determine, if only non-systematically, which thoughts will next be formed. We turn to examine this process in more detail next.

II.4. *Thinking about Thinking*

A key feature of concepts is that they are endogenously controllable or that can be activated in a top-down manner [e.g. Prinz (2002); Barsalou (1999)].[12] Endogenously controlled thinking is a form of associative thinking – current thinking caused by earlier thinking – recall that concepts are associationistic in their causal patterns. For instance, consider someone uttering the word 'Trip' and someone mistakenly hearing the word 'Grip' only to think about friction and laws of physics, rather than travelling. This thought is formed in the absence of the appropriate stimulus, seemingly in a spontaneous but actually in an associative manner.

Without committing to a connection view of the mind, evidence in support of the suggested associationistic view of thinking can be found in the work of Elman et al. (1996), who argue that artificial neural networks can be highly constrained by the network's current weight assignment. That is, the pattern of activation set by a connectionist network is determined by the weights of connections between the units. In terms of the human brain, different activation levels of synapses that connect one neuron to another place a significant constraint on what new ideas the mind can explore next. In this sense, sub-activation of certain neuronal ensembles constrains activation towards specific thoughts associated with the sub-activated links of a given conceptual network.

III. RSA & THE PROBLEMS OF DPTS

Given the current understanding of intuitions (or T-1 processes) as largely unconscious cognitive processes, which are projected into conscious cognitive processes, we argue that intuitions are precisely these sub-activated connections between different concepts and are influenced by the relative weightings obtaining between their associations – call this thesis Representational Sub-activation Thesis (RSA). The stronger the connection between sets of concepts, the more frequently certain sequences of thoughts are tokened, and the more intuitive a certain sequence of thoughts will seem to us.

As mentioned, once a given representation becomes activated, the rest of the strongly associated representations will become sub-activated. The reason why not all associated representations become activated, but rather sub-activated, is two-fold. First, different weights obtain across different sets of associated representations. In addition, the strengths of the connections determine the levels of activation. Importantly, which representations will be fully activated is also influenced by selective attention. Recall from the previous section that stored sub-activated representations drive selective attention in a top-down manner. In this way, once a stored representation is activated, the associated representations are sub-activated, and so drive selective attention to certain parts of the stimuli or to other stored representations (in the case of off-line thinking). In turn, once selectively attending to a further representation within the set in question, a similar process occurs and matching stored representations influence selective attention in a top-down manner, while further associated representations become also sub-activated. Even though

this is merely an idealised scenario of off-line thinking, which ignores for instance, contextual factors, it still illustrates how thinking occurs.

The same representational network-like structures are also available to reasoning (T-2) processes. In this sense, reasoning processes may also become activated in a top-down manner via directed endogenous control. So, thinking is constrained by the weights of different associated sets of representations, which are determined by frequencies of the co-activation of representations, whether emotional states are involved in those processes, and so forth.

Representational structures enjoy great levels of plasticity, in the sense that in light of new stimuli or reflection, new associations can be formed and, in turn, direct the system in various ways. The more frequently attention is drawn to a given new representation, the faster an association will be formed.[13] In contrast, the more rarely an existing association is activated, the weaker the connection between them will grow, and the more idle it will become in thinking. Against this background, we turn to examine the process of debiasing, and more fully explain fundamental biases such as the belief bias.

III.2 *Explaining "bias"*

Consider a classic case intended to display "belief bias" in thinkers' reasoning. Here, belief bias has to do with subjects' tendency to endorse the validity of invalid arguments that accord with extant beliefs, and reject valid arguments that do not. The bias is highlighted in experiments [Sá and Stanovich (1999)], where subjects presented with an invalid syllogism, like the following, are asked to assess its validity.

1. All living things need water.

2. Roses need water.

3. Thus, roses are living things.

According to the obtained results, 32% of subjects replied that this was not a valid argument, while 68% returned valid. The mood of the syllogism is AAA-2. A syllogism of the same mood was then presented to the subjects involving 'foreign' terms (involving an imaginary species, Wampets, and an imaginary class, Hudon), and subsequently, they were asked to evaluate the following:

1. All animals of the Hudon class are ferocious.

2. Wampets are ferocious.

3. Thus, wampets are animals of the Hudon class.

Here, 78% of the subjects gave the "logically correct" response here, that the syllogism is invalid. According to proponents of DPTs, subjects seem to bypass logical principles that would otherwise allow them to identify the syllogism as invalid. In this sense, dual-process theorists appeal to a hypothesized dichotomy between intuitions and reasoning, where reasoning is in broad accord with classical logical principles.

However, as pointed out in §I.2, we suggest that this model is overly rationalistic, and poorly reflects the role of normative reasoning strategies. In general, suggestions that involve a "biased" response to the first syllogism require a normative theoretical system to be *already in place* in the thinker's cognitive architecture, which is thought to follow the laws of classical logic or probability theory. Given the substantive disputes regarding logical systems that have emerged over the last thirty years or so (concerning paraconsistent logics, relevant logics, non-monotonic logic, and so on) this appears overly parochial. Resultantly, we prefer to employ the phrase "doxastic conservativeness" [Dutilh Novaes (2012)] to characterise the tendency of reasoners to bring to bear prior beliefs on the assessment of argument.

RSA provides a promising explanation that models doxastic conservativeness in this experiment as follows. Recall, from §2, three key elements of the RSA model that play a crucial role in explaining the obtained result from the invalid syllogism experiment.

- Thinking is influenced by connection weights between associated representations.
- RSA appeals to the role of selective attention focusing on aspects of the perceptual stimuli.
- Selective attention is driven in a top-down manner by strongly connected stored representations.

Consider the first syllogism. According to DPTs, subjects return the logically false reply because of an intuitive pull, which overrides rational processes. In contrast, we suggest that on confrontation with the syllogism in question subjects attend to certain aspects of premises 1-3. If not seen as a syllogism in which 3 is supposed to follow from 1 and 2, i.e. merely by ignoring the word 'Thus' in (3), then it is rather a list of

sentences (as a matter of fact, it is a list of truisms). Given prior beliefs, involving associations between representations of roses and living things, there is a match between the input-sentence "Thus, all roses are living things", and representations of roses being associated with living things. Given this matching, the subject feels the intuitive pull to confirm the validity of the syllogism. This, as explained, is due to the automatic representational sub-activation of the association between token representations of roses and living things.[14] Subjects cannot simply disregard the representational content of the words "rose" and the phrase "living things" because of this underlying process.[15] This explains why we display the tendency to make judgments based on extant networks of representations involving doxastic states, emotions, and other mental states.

Had it been the case that the subjects have actually treated the syllogism as a syllogism rather than a list of truisms, in the sense explained above, then the same matching would have occurred with representations or stored knowledge about certain logical rules, universal quantifiers, etc. (if the subject was trained in logic). In turn, the intuitive pull would have been against the truism that all roses are living things, and the focus would have been on "Thus" and in turn on the fact that (3) does not follow from (1) and (2).[16] In line with what has been said in the second part of the paper, once there is a matching with stored representations, attention will be driven in a top-down manner to search for existential qualifiers in the premises/steps of the syllogism. Given the way the syllogism in question is couched, the intuitive pull to resist the conclusion will be even stronger or there will be no intuitive pull towards accepting the syllogism. This is precisely what we find in the second example, where the syllogism is couched in "foreign terms" that do not activate automatic RSA processes in the same manner, because of lack of representational associations.

If the subject is not acquainted with norms of classical logic, she might still accept the invalid conclusion as true. For in this case, there will be no stored representations to allow the subject to resist the conclusion – in essence, this latter case will greatly resemble the original one. In contrast, if the subject is to a certain extent familiar with logical rules, the formalization of the syllogism, e.g. the word 'Thus', etc., will sub-activate representations associated with logical rules, priming the subject to selectively attend to the structure of the syllogism.

Note also that learning the validity of syllogisms is a slow and progressive process, since associations take time to form; influencing the network's activational pattern takes even longer. Consequently, the intui-

tive pull for accepting an invalid conclusion as true, for instance, might linger on even after it has been explained to the subjects that this syllogism contradicts the norms of classical logic. Once established though, associations trigger activation of adjacent representations in an 'automatic manner', and give rise to an 'intuitive phenomenology'.

The above suggestions also explain why the so-called 'T-2' processes are themselves susceptible to "biases", precisely because they are not distinct cognitive processes. As it happens, there is only one kind of reasoning process, which just happens to occur by deployment of different sets of representations in each of the two cases (formalised and non-formalised syllogism). In this sense, awareness of one's own reasoning processes can increase bias through post-rationalisation of the RSA process.[17]

IV. RSA and Empirical Evidence

In this section, we evaluate the RSA 'model' against the empirical evidence that we cited earlier against DPTs, and examine whether RSA can accommodate it. Given space limitations we only do that for a sample of this evidence.

De Neys (2006b) reports that performance in no-conflict tasks was near ceiling and that there was no loading on WM. DPTs predict these results by arguing that participants relied upon real-world knowledge, in evaluating these syllogisms, which is cognitively easier than implementing rules of logic. According to RSA, the knowledge upon which participants rely is captured by stronger associations. In this sense, there will be some loadings on WM but this will be less than when dealing with conflict syllogisms. Nonetheless, there will be some loading on WM regardless of the difficulty of the task (see below), which will be diminishing as these evaluating processes become automated.

DPTs predict that T-2 processes underpin evaluation of conflict tasks, since conflict syllogisms are difficult to evaluate and load WM. According to DPTs, when logic and belief disagree, subjects appeal to the believability of the conclusion; in turn integrating the premises and conclusion of a conflict syllogism is harder since this integration challenges their prior world knowledge (conflict syllogisms).[18]

For RSA in contrast, the difficulty in assessing conflict syllogisms stems from the fact that conflicting ideas do not co-occur very often, and in turn the representations that ground them are only weakly connected – if at all. Thus, in cases of conflict syllogisms participants do not

have the 'disposition' to infer the conclusion from the premises. Without this disposition in place to drive selective attention to the aspects of the syllogism that refer to principles of logic, they attend to words of the syllogisms that couch these conflicting beliefs, and bring to bear prior beliefs while assessing them (doxastic conservativeness). Consequently, participants end up evaluating in detail the associated semantics – a cognitively taxing process.

• WMC and Accuracy Levels

According to RSA, there are strong interdependencies between attention and WM.[19] Crucially, in order for attended information to contribute to higher cognitive processes, this information has to be accessible by WM. Thus, despite the top-down 'urge' to attend to certain aspects of the stimulus, if WM is loaded, the attended information will not be accessible/usable while evaluating the syllogism.

In line with De Neys (2006b), RSA predicts that high WMC subjects can exploit attended information to a greater extent when compared to low WMC subjects, since processing attended information requires WM. But WMC should not be construed in a 'use-more-if-you-can-afford-it' manner. Rather, the minimum WM resources required for a specific task are used/deployed, while the remaining WMC is standing by. Thus, RSA predicts loading of WM regardless of WMC, even though subjects with higher WMC are (ceteris paribus) expected to be more accurate either at the selection task or the dot memory test or at both.

• Why Accuracy Levels Were Higher in Deontic Tasks?

Accuracy levels in deontic tasks are higher, in comparison to indicative ones, since deontic tasks concern well-known regulations that are underpinned by strongly interconnected representations, and which trigger top-down effects in perception of such tasks.

Alternatively, Stenning and van Lambalgen (2004) predict that descriptive/indicative tasks will be highly problematical and deontic tasks rather straightforward – an empirically confirmed prediction. Their starting point is that the main interpretative problem facing subjects in reasoning tasks is assigning logical form to the task at hand or providing settings for all the involved parameters. The problem, they argue, is that psychology of reasoning has traditionally operated on an oversimplified notion of logical form. Namely, assigning logical form has been traditionally construed as translating a natural language sentence into a formal language with given semantics. Stenning and van Lambalgen expand this

notion of logical form and argue that indicative/descriptive and deontic tasks are actually different in terms of logical form. With regards to the strategy subjects follow in reasoning tasks, they argue that most likely subjects do not really know what they are doing but they certainly worry about how to set the parameters of the task. Furthermore, different subjects adopt different strategies. Against this background, DPTs interpretations that focus mainly on WM loadings seem oversimplified. In contrast, RSA is 'compatible' with this alternative view since it focuses not only on WM loadings as the main influence on performance at reasoning tasks, but also on the conceptual networks deployed in reasoning tasks. These networks might also be construed as carrying information concerning logical form [see also Tillas (2010)].

• Response-times in deontic tasks

Even though strongly connected representations are processed faster than weakly connected ones, the amount of information also influences the overall speed of its processing. Given that deontic tasks concern well-known regulations, it is likely that the volume of information associated with these regulations will be greater than this associated with rules that the subjects encounter possibly for the first time. Thus, deontic tasks processing is not necessarily faster than that of indicative ones.

• Regardless of Time Needed to Respond, Accuracy Levels in Deontic Tasks were also Influenced by WM Loadings

According to RSA, all representations deployed in reasoning need to be placed in WM. Naturally, a loaded WM will have a negative effect on the way that the relevant information becomes deployed in reasoning*.

**This paper was fully collaborative; the order of the authors' names is arbitrary.*

Department of Humanities
University for the Creative Arts
21 Ashely Rd.
Epson KT18 5BE, UK
E-mail: jtrafford2@ucreative.ac.uk

Institut für Philosophie
Heinrich-Heine-Universität
Universitätsstr. 1
40225 Düsseldorf, Germany
E-mail: atillas@phil.uni-duesseldorf.de

NOTES

[1] The suggestion is that T-1 processes do not tie-up working memory (WM), and so, they are the default mode of judgment, see Kahneman and Frederick (2005).

[2] Though Osman (2004) defends a uni-model approach.

[3] See also Gould (1992).

[4] Though see Stenning and van Lambalgen (2004), as well as the discussion in §IV.

[5] See Dutilh Novaes (2012), esp. ch. 7 for an extended discussion.

[6] We owe this suggestion to an anonymous reviewer.

[7] Deontic tasks or tasks that ask people to reason about known regulations ("If a person is drinking beer, then the person needs to be over 21 years of age") are associated with T-1 processes. Indicative tasks are tasks appearing within a context that refers to an arbitrary conditional rule, e.g. "if there is a vowel on one side, then there is an even number on the other".

[8] LTP has been found in various brain regions, e.g. piriform [Stripling et al. (1988)], prefrontal cortices [Laroche et al. (1989)].

[9] Cf. Spivey and Geng (2001); Chao, Haxby and Martin (1999), amongst others.

[10] Closer is construed here in terms of stronger positive memory effects, etc.

[11] Consider also that neurons in different areas of the visual cortex respond selectively to different stimuli, e.g. bars and edges at preferred orientations (V1), complex shapes and contours (V2 and V4), and visual motion (medial superior temporal area). These response selectivities are often understood in terms of hierarchical predictive coding of natural inputs. For example, Rao and Ballard (1999) propose a hierarchical neural network in which top-down feedback connections from higher-order visual cortical areas carry predictions of lower-level neural activities, while the bottom-up connections convey the residual errors in prediction. See also Huang and Rao (2011) for a review, and Clark (2013) for a detailed discussion of related issues.

[12] Endogenous control over a given concept is acquired by associating the set of perceptual representations that comprises the concept in question to a perceptual representation of a word or goal-directed actions over which they have endogenous control. The claim is that concepts inherit the endogenous control that we have over utterances or goal-directed action etc. In a bit more detail, human agents have the ability to manipulate external objects in relationships of agency towards them; what is argued here is that we can piggyback on that ability to manipulate and direct our own thinking. In this sense, human subjects have endogenous control over their production of linguistic items, to the extent that they are able to produce linguistic utterances at will (or silently talk to themselves). It is this executive control over linguistic utterances that gives us endogenous control over our thoughts. See Tillas (2010) for a detailed analysis.

[13] Certain kinds of mental states, like emotions for instance, allow associations to be formed after only one or in any case very few repetitions.

[14] See Meyer and Schvaneveldt (1971) for similar results at the level of semantic representations and sentences.

[15] The underlying process at hand could perhaps be further understood in terms of 'semantic activation' [e.g. d'Arcais and Schreuder (1987); Dutilh Novaes

(2012)]. It is worth clarifying though that the suggested view differs from semantic activation in that: (a) it has a broader scope to the extent that it explains how sub-activated representations become part of a representational network, while setting the background for conceptual flexibility and change (weighting calibration between concepts/representations); (b) unlike semantic activation, which is often understood lexically (or at least to concern lexical concepts), RSA does not require that and so is capable of dealing with the sub-activation of associated images and goal-directed states involving, for example, sensorimotor and proprioceptive representations; (c) resultantly, RSA is a more fine-grained approach than semantic activation.

[16] It may appear that we have sneaked-in a presupposition that participants already have the concept of "following" in the strict sense of necessary truth-preservation. However, this would be to mistake a more basic sense in which "following" indicates an inferential move or a move in a dialogue, for the formalised notion. On this point, see Dutilh Novaes (2012), and the discussion of logical form, below.

[17] This claim is in line with Mercier and Sperber (2011).

[18] It is on cases of this kind that DPTs examine and build their argument. However, there are also other cases, which do not challenge prior world knowledge. For instance, having an invalid syllogism with a believable conclusion. Furthermore, there could also be valid syllogisms with believable premises and an unbelievable conclusion. Admittedly, syllogisms of this kind are hard to come by, given the relation of necessary truth-preservation, but they are not impossible. Finally, there could be valid syllogisms with at least one unbelievable premise and an unbelievable conclusion. It is only in the second case that there will be the issue of integrating the (false) premise. In this sense, it would be interesting to see experiments dealing with these kinds of syllogisms and how DPTs interpret the results. We owe this suggestion to an anonymous reviewer.

[19] For instance, Awh et al. (1999) found that storing a location in WM led to faster detection of targets presented at that location. Also, Downing (2000) shows that the storage of an item in WM leads to automatic orienting towards similar objects in the environment.

REFERENCES

AWH, E., JONIDES, J., SMITH, E. E., BUXTON, R. B., FRANK, L. R., LOVE, T., *et al.* (1999), 'Rehearsal in Spatial Working Memory: Evidence from Neuroimaging', *Psychological Science*, 10, 5, pp. 433-7.

BARSALOU, L. W. (1999), 'Perceptual Symbol Systems', *Behavioral and Brain Sciences*, 22, pp. 577-609.

CHAO, L. L., HAXBY, J. V., and MARTIN, A. (1999), 'Attribute-Based Neural Substrates in Temporal Cortex for Perceiving and Knowing About Objects', *Nature Neuroscience*, 2, pp. 913-9.

CHEN, S. and CHAIKEN, S. (1999), 'The Heuristic-Systematic Model in its Broader Context', in Chaiken, S. and Trope Y. (eds.), *Dual-Process Theories in Social Psychology*, New York, Guilford Press, pp. 73-96.

D'ARCAIS, G.B.F. and SCHREUDER, R. (1987), 'Semantic Activation During Object Naming', *Psychological Research*, Vol. 49, Issue 2-3, pp. 153-9.

DE NEYS, W. (2006a), 'Automatic-Heuristic and Executive-Analytic Processing in Reasoning: Chronometric and Dual Task Considerations', *Quarterly Journal of Experimental Psychology*, 59, pp. 1070-100.

— (2006b), 'Dual Processing in Reasoning: Two Systems but One Reasoner', *Psychological Science*, 17, pp. 428-33.

DUTILH NOVAES, C. (2012), *Formal Languages in Logic: A Philosophical and Cognitive Analysis*, Cambridge, Cambridge University Press.

ELMAN, J. L., and MCCLELLAND, J. L. (1986), 'An Architecture for Parallel Processing in Speech Recognition: The TRACE model'. In Schroeder M. R. (ed.), *Speech Recognition*. Basel, S. Krager AG.

EPSTEIN, S. (1994), 'Integration of the Cognitive and Psychodynamic Unconscious', *American Psychologist*, 49, pp. 709-24.

EVANS J. (2008), 'Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition', *Annual Review of Psychology*, 59, pp. 255-78.

— (1989), *Bias in Human Reasoning: Causes and Consequences*, Brighton, Erlbaum.

— (2006), 'The Heuristic-Analytic Theory of Reasoning: Extension and Evaluation', *Psychonomic Bulletin and Review*, 13.3, pp. 378-85.

— (2007), 'On the Resolution of Conflict in Dual Process Theories of Reasoning', *Thinking and Reasoning,* 13, pp. 321-39.

EVANS, J. B. and OVER, D. E. (1996), *Rationality and Reasoning,* Hove, Psychology Press.

EVANS, J. B., ALLEN, J. L., NEWSTEAD, S. E. and POLLARD, P. (1994), 'Debiasing by Instruction: The Case of Belief Bias', *European Journal of Cognitive Psychology*, 6, pp. 263-85.

EVANS, J. B., BARSTON, J. L., and POLLARD, P. (1983), 'On the Conflict between Logic and Belief in Syllogistic Reasoning', *Memory and cognition,* 11, 3, pp. 295-306.

EVANS, J.B., and OVER, D. E. (1996), *Rationality and Reasoning,* Hove, England, Psychology Press.

EVANS, J.B., and STANOVICH, K.E. (2013), 'Dual-Process Theories of Higher Cognition: Advancing the Debate', *Perspectives on Psychological* Science, 8, p. 223.

GOULD, S. J. (1992), *Bully for Brontosaurus: Reflections in Natural History'*, WW Norton and Company.

HEBB, D. O. (1949), *The Organization of Behavior,* New York, John Wiley.

HOLENDER, D. (1986), 'Semantic Activation Without Conscious Identification in Dichotic Listening, Parafoveal Vision, and Visual Masking: A Survey and Appraisal, *Behavioral and Brain Sciences*, Vol. 9, 01.

HUANG, Y. and RAO, R.P.N. (2011), 'Predictive Coding', *Wiley Interdisciplinary Reviews: Cognitive Science*, Volume 2, Issue 5, pages 580-93, September/October (2011).

KAHNEMAN, D. (2011), *Thinking, Fast and Slow*. New York, Farrar, Straus and Giroux.

KAHNEMAN D., and FREDERICK, S. (2002), 'Representativeness Revisited: Attribute Substitution in Intuitive Judgement', in Gilovich, T., Griffin, D. and Kahneman, D. (eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment*, New York, Cambridge University Press, pp. 49-81.

— (2005), 'A Model of Heuristic Judgment', in Holyoak, K. J. and Morrison, R.G. (eds.), *The Cambridge Handbook of Thinking and Reasoning*. New York, Cambridge University Press, pp. 267-93.

KAHNEMAN, D. and TVERSKY, A. (1982), 'On the Study of Statistical Intuitions', *Cognition*, 11,2, pp.123-41.

KENNY, A. (1963), *Action, Emotion and Will*, London, New York, Routledge and Kegan Paul, Humanities Press.

KEREN, G. and SCHUL, Y. (2007), 'Two is not Always Better than One', A Critical Evaluation of Two-Systems Theories. Commentary Barbey and Sloman, (2007).

KLAUER, K., MUSCH, J., and NAUMER, B. (2000), 'On Belief Bias in Syllogistic Reasoning'. *Psychological Review*, 107, pp. 852-84.

LAROCHE, S., DOYERE, V., and BLOCH, V. (1989), 'Linear Relation between the Magnitude of Long-Term Potentiation in the Dentate Gyrus and Associative Learning in the Rat. A Demonstration Using Commissural Inhibition and Local Infusion of an N-Methyl-D-Aspartate Antagonist', *Neuroscience*, 28, pp. 375-86.

MERCIER, H. and SPERBER, D. (2011), 'Why do Humans Reason? Arguments for an Argumentative Theory', *Behavioral and Brain Sciences*, 34, pp. 57-111.

MEYER, D. E. and SCHVANEVELDT, R. W. (1971), 'Facilitation in Recognizing Pairs of Words: Evidence of a Dependence Between Retrieval Operations', *Journal of Experimental Psychology*, 90, 2, pp. 227–34.

MONSELL, S. and DRIVER, J. (2000), *Control of Cognitive Processes*, Cambridge, MA, MIT Press.

NORENZAYAN, A., CHOI, I. and NISBETT, R. E. (2002), 'Cultural Similarities and Differences in Social Inference: Evidence from Behavioral Predictions and Lay Theories of Behavior', *Personality and Social Psychology Bulletin*, 28, 1, pp. 109–20.

OAKSFORD, M., MORRIS, F., GRAINGER, B., and WILLIAMS, J. M. G, (1996), 'Mood, Reasoning, and Central Executive Processes', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, pp. 476-992.

OSMAN, M. (2004), 'An Evaluation of Dual Process Theories of Reasoning', *Psychonomic Bulletin and Review*, 11, pp. 998-1010.

— (2013), 'A Case Study: Dual-Process Theories of Higher Cognition – Commentary on Evans and Stanovich' (2013), *Perspectives on Psychological Science*, 8, p. 248.

PAVLOV, I. P. (1927), *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*, Translated and Edited by G. V. Anrep, London, Oxford University Press.

PERRY, J. (2001), *Knowledge, Possibility, and Consciousness*, Cambridge, MA, MIT Press.

POSNER, M. (1978), *Chronometric Exploration of Mind*, Hillsdale, NJ, Erlbaum.

PRINZ, J. (2002), *Furnishing the Mind: Concepts and Their Perceptual Basis*, Cambridge, MA, MIT Press.

— (2005), 'Are Emotions Feelings?', *Journal of Consciousness Studies*, 12, 8-10, pp. 9-25.

RAO R.P.N., and BALLARD D.H. (1999), 'Predictive Coding in the Visual Cortex: a Functional Interpretation of Some Extra-classical Receptive-Field Effects, *Nature Neuroscience*. 2, pp. 79-87.

ROBERTS, M. J., and NEWTON, E. J. (2001), 'Inspection Times, the Change Task, and the Rapid-Response Selection Task', *Quarterly Journal of Experimental Psychology*, 54A, pp. 1031-48.

SÁ, W. C., WEST, R. F., and STANOVICH, K. E. (1999), 'The Domain Specificity and Generality of Belief Bias: Searching for a Generalizable Critical Thinking Skill', *Journal of Educational Psychology*, *91*, pp. 497-510.

SHORS, T. J. and MATZEL, L.D. (1997), 'Long-term Potentiation: What's Learning got to do with it?', *Behavioral and Brain Sciences* 20, pp. 597–655.

SLOMAN, S. A. (1996), 'The Empirical Case for Two Systems of Reasoning', *Psychological Bulletin*, 119, pp. 3-22.

SOLOMON, R. C. (1984), 'Getting Angry', in Shweder R. and LeVine R. (eds.), *Culture Theory,* New York, Cambridge University Press, (ch. 9).

— (1976), *The Passions*. New York, Doubleday.

SPIVEY, M. J., and GENG, J. J. (2001), 'Oculomotor Mechanisms Activated by Imagery and Memory: Eye Movements to Absent Objects', *Psychological Research/Psychologische Forschung, 65*, 4, pp. 235-41.

STANOVICH, K. E. (1999), 'Who is Rational?', *Studies of individual differences in reasoning*, Mahwah, NJ, Lawrence Erlbaum Associates, Inc.

— (2009), 'Distinguishing the Reflective, Algorithmic, and Autonomous Minds: Is it Time for a Tri-Process theory?', in Evans, J. B. and Frankish, K. (eds.), *In two Minds: Dual Processes and Beyond*, Oxford, Oxford University Press, pp. 55-88.

— (2011), *Rationality and the Reflective Mind*, New York, NY, Oxford University Press.

STANOVICH, K. E., and WEST, R. F. (2000), 'Individual Differences in Reasoning: Implications for the Rationality Debate', *Behavioral and Brain Sciences*, 23, pp. 645-726.

STENNING, K. and M. VAN LAMBALGEN. (2004), 'A Little Logic Goes a Long Way: Basing Experiment on Semantic Theory in the Cognitive Science of Conditional Reasoning', *Cognitive Science*, 28(4), pp. 481-530.

STRIPLING, J. S., PATNEAU, D. K., and GRAMLICH, C. A. (1988), Selective Long-Term Potentiation in the Pyriform Cortex', *Brain Research*, 441, pp. 281-91.

TILLAS, A. (forthcoming 2015a), 'On the Origins of Concepts', in Kann, C., Hommen, D., and Osswald, T. (eds.), *Concepts and Categorization*. Paderborn, Mentis.

— (forthcoming 2015b), 'Internal Supervision and Clustering: A New Lesson from "Old" Findings?', in Müller V. C. (ed.), *Computing and Philosophy*, Synthese Library, Springer.

— (2014), 'How do Ideas Become General in their Signification?', in Machery, E., Prinz, J. and Skilters, J. (eds.), *The Baltic International Yearbook of Cognition, Logic and Communication*, Vol. 9, Kansas, New Prairie Press.

— (2010), *Back to Our Senses: An Empiricist on Concept Acquisition*, Doctoral Thesis, University of Bristol, UK

THOMPSON, V. and EVANS, J. B. (2012), 'Belief Bias in Informal Reasoning', *Thinking and Reasoning,* 18, 3, pp. 278-310.

WASON, P. C. and EVANS, J. B. (1975), 'Dual Processes in Reasoning?', *Cognition*, 3, 2, pp. 141-54.

WILKINS, M. C. (1928), 'The Effect of Changed Material on the Ability to Do Formal Syllogistic Reasoning'; *Archives of Psychology*, 102, pp. 1-83.