An Analysis of the Scientific Status and Limitations of The Attitudinal Entropy Framework and an Initial Test of Some of Its Empirical Predictions

Pieter Van Dessel, Jan De Houwer, Sean Hughes, & Ian Hussey

Dalege, Borsboom, van Harreveld, and van der Maas (2018) describe a novel framework for the conceptualization of attitudes that draws on principles from statistical mechanics. A core idea in their framework is that systems are often characterized by randomness (i.e., entropy) and that there is both heuristic and predictive value in applying the idea of entropy to the study of attitudes and related phenomena. We applaud their initiative: the attitudinal entropy framework provides an intriguing new perspective on theoretical questions and empirical findings in social psychology. It opens up new avenues for research in many areas and is a timely contribution given the growing popularity of predictive processing theories emphasizing entropy as an important factor in human cognition (for a recent overview see Metzinger & Wiese, 2017).

Nevertheless, we believe that there is still room for improvement. In the first part of this paper, we present an epistemological analysis that clarifies the way in which the Attitudinal Entropy Framework contributes to scientific knowledge. The second section of the paper focusses on a number of limitations of the framework as it was formulated by Dalege et al. (2018), most prominently their shallow treatment of inferential processes. Finally, we present empirical data concerning two of the predictions put forward by Dalege and colleagues.

The Scientific Contribution of the Attitudinal Entropy Framework

Science is most prominently concerned with two tasks: to describe and to explain. In psychology, phenomena are typically explained at the functional level (i.e., explaining behavior in terms of elements in the environment; e.g., Skinner, 1953) or at the cognitive level (i.e., explaining the impact of environment on behavior in terms of mental mechanisms; e.g., Gardner, 1987). Dalege and colleagues (2018) suggest that the contribution of their framework is situated mainly at the cognitive level of explanation: it is assumed to deal with the nature and operation of the mental system, most prominently (changes in) the entropy of mental representations. We believe, however, that the main contribution of the framework lies at the descriptive level and at the functional level of explanation.

First, the concept of entropy is descriptive in nature. Boltzmann entropy describes the consistency between the elements of one microstate whereas Gibbs entropy describes the consistency between different microstates.

Second, the concept of entropy reduction has explanatory value at the functional level of explanation. Specifically, a State X at Time N is said to emerge because of the high entropy of State Y at Time N-1. The concept of entropy reduction as such says nothing about the (physical or mental) mechanisms via which a high entropy state gives rise to a low entropy state; it merely captures the idea that the emergence of the low entropy State X is a function of the high entropy of a preceding State Y.[1]

[1] Note that the explanatory value of entropy reduction at the functional level hinges upon the ability to manipulate directly the entropy of State Y. To the extent that the entropy of State Y can be manipulated only indirectly by manipulating elements in the environment, it makes more sense to say that State X is a function of those elements in the environment and that the entropy of State Y mediates the functional relation between the elements in the

Third, also the Causal Attitude Network on which the Attitudinal Entropy Framework is built, can be conceived of as situated at the functional level of explanation. Within the CAN model, elements are linked within a network. Whereas Dalege et al. (2018) conceive of the networks and their elements as mental entities (i.e., information represented in memory), one can also think of the elements as behaviors. In fact, if one looks at how network models are used in psychology, they are typically based on what people verbally report about their behavior, feelings, and thoughts. Rather than making the questionable assumption that verbal reports provide a direct and accurate reflection of mental representations (Schwarz, 2007), one can treat them as behaviors, much like any other movement of muscles or glands can be treated as a behavior. Within the domain of attitude research this would, for instance, imply that an inconsistency between self-reported liking of a product and buying behavior is not treated as an attitude-behavior inconsistency (which implies that self-reported liking is a proxy of the underlying mental attitude) but as a behavior-behavior inconsistency.

This perspective is compatible with the idea that attitude research deals with the study of evaluation, that is, the way in which stimuli influence evaluative responses (De Houwer, 2009; De Houwer, Gawronski, & Barnes-Holmes, 2013). It implies that both the CAN model and the Attitudinal Entropy Framework have much to contribute to attitude research. One of the big assets of network models such as the CAN model is that they provide new ways of describing relations between (evaluative) behaviors. To the extent that the relations between behaviors in a network are assumed to be directional (rather than merely correlational), networks also provide functional explanations of behavior, that is, insights into how one behavior is a function of other behaviors or states in the environment. Such functional explanations allow one to predict and influence behavior by observing and influencing other behavior or states in the environment.[2] The integration of the CAN model within an entropy framework further expands the descriptive and functional explanatory value of the CAN model by linking it with concepts such as entropy and entropy reduction. Note, however, that all of this can be achieved without invoking any reference to mental constructs such as mental representations. In fact, this conclusion is unsurprising given that both entropy frameworks and network models have been developed in areas of research such as physics and mathematics that focus on description and functional explanations.

One might argue that the Attitudinal Entropy Framework could in principle also be applied at the cognitive level of explanation by using it to describe and explain the nature of mental representations. The problem with this approach is that (the elements of) mental representations cannot be observed directly (Gardner, 1987). Hence, applying the framework at the cognitive level necessarily adds a level of uncertainty compared to when the framework is restricted to the descriptive or functional level. We therefore believe that there are advantages to applying the framework at the descriptive and functional level as compared to the cognitive level. Regardless of the level of explanation at which the framework is likely to be most successful, it would be good to always be explicit about the level of explanation at which the framework is being used. Because different questions are addressed at different levels of explanation, confounding levels can distort scientific debates. An example of such a confound can be found in the simulations that Dalege et al. (2018) present. Whereas in some simulations, nodes within the network are assumed to refer to attitude elements within the cognitive system of a single individual (an intrapersonal model at the cognitive level), in other simulations, the nodes refer to behavior of individuals in a group (an interpersonal model at the functional level). A lack of clarity around what level of explanation is being modeled (i.e., functional vs. mental level, intrapersonal vs. interpersonal) raises more questions than it answers.

Of course, our analysis does not imply that one should abandon the cognitive level of explanation in attitude research. We only argue that attitude research which focusses on description and functional explanation also has merit and that the Attitudinal Entropy Framework can contribute to attitude research at those levels. Such research can be complemented by theories about the mental mechanisms that mediate evaluation. In fact, Dalege et al. (2018) seem to be aware of this fact when they refer to the need to understand the inferences that underlie the links in networks and the motivational processes that determine the dependency within

environment and State X (see Hayes & Brownstein, 1986, for a related discussion).

[2] Note that the explanatory value of behavior-behavior relations at the functional level is limited by the fact that most if not all behaviors can be influenced only indirectly by changing the environment. Indeed, when the ability to influence a behavior is used as the criterion for the successful explanation of that behavior, a successful explanation of Behavior X in terms of Behavior Y requires the specification of those environmental variables that determine Behavior Y. This implies that functional explanations in psychology always boil down to knowledge about environment-behavior relations (Hayes & Brownstein, 1986).

networks. As we will argue below, there is indeed much merit in considering the role of motivation and inferential processes in attitude research. Although theories about mediating mental mechanisms can certainly be related to the Attitudinal Entropy Framework, much of the scientific merit of the framework itself is, in our opinion, situated at the descriptive level and functional level of explanation.

Limitations of the Attitudinal Entropy Framework

Despite its merits, the Attitudinal Entropy Framework as it was put forward by Dalege and colleagues is also limited in important ways. First, attitude elements are modeled as nodes that can only be switched on or off and are thus stripped from any (relational) content (e.g., the content of beliefs), making it difficult to see how consistency between attitude elements could be determined. The assumption that only the (momentary) valence of attitude elements (modeled as a binary variable) is compared in this process is unfeasible given that it is not specified how the valence of attitude elements (not only beliefs but also behaviors and feelings) is determined. Moreover, studies show that content-related characteristics of information about attitude objects (e.g., its diagnosticity or believability: Cone, Mann, & Ferguson, 2018) determine evaluation more than the amount of positive and negative information. For instance, Cone and Ferguson (2015) found that participants exhibited negative rather than positive implicit and explicit evaluations of a person named Bob when they learned many pieces of positive information about Bob but only one piece of negative information that was, however, more diagnostic of Bob's true character (e.g., that Bob was a child molester).

Second, as noted earlier, Dalege and colleagues refer to cognitive concepts such as inferences and motivation. However, their treatment of these concepts is rather superficial. With regard to the concept of motivation, they argue that the mental system is motivated to reduce entropy because that entropy causes distress. However, without an explanation of the motivational role of entropy, the current framework pushes the question of attitudes back from explaining attitudes to explaining entropy and distress. Note that modeling of entropy (described as consistency detection) does not solve this issue because this modeling is also merely descriptive and does not directly tie into important mental level concepts.

In the remainder of this section, we discuss in quite some detail the role of inferential processes within the Attitudinal Entropy Framework. Whereas Dalege and colleagues (2018) refer to this topic only briefly, we believe that inferential processes are vital when extending the framework to the cognitive level of explanation. In a recent paper, we described an inferential account of evaluative stimulus-action effects that focuses on the inferences that underlie evaluative learning on the basis of stimulus-based actions (e.g., repeated approach or avoidance of a stimulus) and outlines how these inferences might arise based on predictive processing principles (Van Dessel, Hughes, & De Houwer, 2018).[3] Specifically, evaluative responding is considered to result from inferences about (the value of) action outcomes. These inferences are learning-, context, and goal-dependent, and reflect the (automatic) application of inference rules to activated information on the basis of a person's belief network (which can be seen as a generative model of the world that is continuously updated on the basis of available information).

The Attitudinal Entropy framework and our inferential model share several similarities with one another. For instance, the former argues that entropy (and its reduction) may play a key role determining the structure and properties of attitudes, a claim that is certainly compatible with the inferential account given its incorporation of predictive processing theory (Friston, 2010). Second, the Attitudinal Entropy framework seems to share the position that implicit and explicit attitudes are based on a single type of mental process that involves inferential reasoning. For instance, Dalege and colleagues note that "weights between attitude elements generally arise based on inferences" (p.12). Moreover, assessing for entropy (which they conceptualize in part as consistency between attitude elements) presumably requires the mental system to be able to represent the truth value of attitude elements (and relations between these elements). This perspective is compatible with single process (propositional) models of attitudes and learning (De Houwer, 2009; De Houwer, 2014; Mitchell, De Houwer, & Lovibond, 2009) and diverges from models which distinguish between two types of attitudinal processes or systems: e.g., System 1 vs 2 (Kahneman, 2003), associative vs. rule-based processes (Smith & DeCoster, 2000), or associative and propositional processes (e.g., Gawronski & Bodenhausen, 2006). It also accords with recent recommendations to explore alternatives to dual-process theories of human cognition (e.g., Melnikoff & Bargh, 2018), a call which is especially relevant to

[3] Although our inferential model mainly focuses on evaluative stimulus-action effects it can easily be (and already has been) generalized to explain other pathways via which evaluative behavior is established or changed (for one such example in the context of evaluative conditioning see De Houwer, 2018).

attitude research where such theories remain dominant and often in the absence of clear empirical support (see Corneille & Stahl, 2018).

Importantly, however, there are two points of divergence between our inferential model and the attitudinal entropy framework. First, within the inferential model, a clear distinction is made between the functional and cognitive level of explanation (see De Houwer et al., 2013). Specifically, we model evaluations (rather than attitudes), which we define as the impact of stimuli on evaluative responses. This ensures that there is no conflation between the behaviors that need to be explained (evaluations) and the mental constructs that are used to explain these behaviors (inferences), allowing for clear, testable predictions about the moderation of evaluative responses by specific contextual variables.

Second, our model describes how inferences might arise and how they can lead to evaluative responses. To move forward, the attitudinal entropy framework might benefit from the integration of basic principles from other (e.g., inferential reasoning) models. Most importantly, the framework might integrate ideas about how evaluations are learned (e.g., on the basis of context-dependent inferences: Van Dessel et al., 2019) to allow for a more encompassing computation of attitude consistency and a model of evaluative behavior. For instance, the motivational role of attitudinal entropy might be elucidated on the basis of current theorizing on inferential reasoning. In our inferential model of evaluative stimulus-action effects, we refer to entropy as a motivational factor in the context of belief updating. We consider entropy not as a characteristic of an attitude (what would be the delineating factor of a configuration of attitude elements?) but of a more general belief system. This idea draws on predictive processing theories in which entropy reduction motivates inferences (and behavior) because it allows for the conservation of mental energy (Friston, 2010). However, we only briefly refer to entropy in the inferential theory we described. Moreover, it has been noted that the conceptualization of entropy in the predictive processing framework is implausible and requires more work (e.g., Otworowska, Van Rooij, & Kwisthout, 2018). In the spirit of the attitudinal entropy model, it might be useful to provide a more extensive description of entropy. For instance, entropy could be more clearly defined as a factor that determines the circumstances under which a person's belief system is updated. We could model entropy as the extent to which integration of information is difficult in that it requires more extensive updating of probabilities in the model. Other variables such as inferred value of information (e.g., for our survival or our self-concept) might be included in this calculation such that entropy is not the only principle that determines inferences and belief updating (which seems problematic: Otworowska et al., 2018). Such modeling that is tied to tangible mental constructs in a model that clearly separates levels of explanation might provide a clear contribution to the literature (e.g., in terms of its explanatory value).

Predictions Tested

While this commentary has primarily focused on conceptual matters, we also had the opportunity to test two of the framework's predictions that Dalege and colleagues argue flow from their model with data we already had at hand. We used data from the Attitudes 2.0 dataset (Hussey et al., 2018) to assess predictions number 1b and 3. Data to test other predictions was not at hand. This large dataset (number of experimental sessions $>$ 409,000) represents a single large study of implicit and explicit attitudes that was conducted on the Project Implicit website (https://implicit.harvard.edu). Subsets of this dataset have been used in previous research (e.g., Nosek & Hansen, 2008), and the full dataset is being curated for public release and publication (Hussey et al., 2018). Participants in the study completed one of 190 different IATs assessing attitudes within a wide range of attitude domains including politics, ideologies, popular culture figures, and everyday preferences (total N available for analysis = 155913). Self-report attitude scales also assessed multiple attitude features, such as "gut feelings" versus "actual feelings" towards the pairs of concepts used in the IAT. Relevant subsets of this data were employed to test two of the hypotheses that Dalege and colleagues put forward. Data and code for the analyses conducted below are available on the OSF (osf.io/c59y2).

Prediction 1b

"Scores on implicit measures assessing attitudes individuals regularly think about, are expected to have higher internal consistency ... than scores on implicit measures assessing attitudes individuals think only infrequently about" (p.20). We calculated internal consistency values for each type of IAT (both Cronbach's α and McDonald's ω_t). Participants were also asked how frequently they thought about the two concept categories that were used in the IAT they completed (e.g., Democrats and Republicans). For each type of IAT (k IATs = 190, mean n per IAT = 1641), mean frequency ratings were also calculated, resulting in 190 pairs of internal consistency values and mean frequency ratings. When these pairs were entered into linear regression analyses, this demonstrated that the self-reported frequency with which participants thought about the concepts

employed in the IATs was predictive of the IAT's internal consistency between domains, as predicted by Dalege and colleagues. This relationship held across both metrics of internal consistency (α: $B = 2.66$, 95% CI = [0.36, 4.95], $\beta = 0.23$, 95% CI = [0.03, 0.43], $p = .024$, $R^2 = 0.05$; ω_t: $B = 2.92$, 95% CI = [0.46, 5.38], $\beta = 0.24$, 95% CI = [0.04, 0.43], $p = .021$, $R^2 = 0.06$).

Prediction 3

"... when individuals are asked to very quickly answer attitude questions, attitudes are expected to be less polarized than when individuals are given more time to answer the questions. Note that the AE framework predicts that this would constitute a small effect." (p.24). The Attitudes 2.0 dataset also contains self-report ratings of both "[immediate] gut feelings" and "actual feelings [upon reflection]" of the 190 concept category pairs. We employed these items to assess the hypothesis that deliberative evaluations are more extreme (i.e., polarized) than gut evaluations. Self-report ratings for each evaluation type were recoded as absolute scores, so that positive scores represent deviation from neutrality/ambivalence without regard to whether those evaluations were positive or negative. A mixed-effects linear regression model that accounted for the nesting of evaluations within concept category domains (i.e., random intercept for domain and random slope for rating type) demonstrated evidence against this prediction: "deliberative" evaluations were found to be less extreme on average than "gut" evaluations ($B = -0.16$, 95% CI = [-0.18, -0.14], $\beta = -0.07$, 95% CI = [-0.08, -0.06], $p < .001$, $R^2 = 0.004$). As such, analyses using a very large existing dataset provide supportive evidence for one prediction that Dalege and colleagues put forth for the framework, however an effect in the opposite direction to that predicted was found for another prediction. Additional tests of the authors' other predictions are of course warranted.

Concluding Remarks

The Attitudinal Entropy framework interfaces concepts from statistical mechanics (entropy) and social psychology (attitudes) to offer an intriguing new perspective on the latter that has both heuristic and predictive value, as evidenced by support for one of the frameworks' predictions that we were able to test with data at hand. Unlike Dalege and colleagues (2018), we believe that the main scientific contribution of the framework, as put forward in their paper, is situated at the descriptive level and the functional level of explanation rather than the cognitive level of explanation. Nevertheless, the framework can be strengthened at the cognitive level of explanation, most prominently by incorporating more precise assumptions about the nature and role of inferential processes. Provided that researchers distinguish between the different levels of explanation to which the Attitudinal Entropy Framework contributes, the framework can provide a major step forward in attitude research.

Funding

PVD is supported by a postdoctoral fellowship of the Scientific Research Foundation, Flanders (FWO-Vlaanderen). JDH and SH are supported by BOF Grant BOF16/MET_V/002 of Ghent University to JDH. IH is supported by a postdoctoral fellowship from Ghent University.

References

Cone, J., & Ferguson, M. J. (2015). He Did What ? The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology*, *108*, 37–57.

Cone, J., Mann, T. C., & Ferguson, M. J. (2018). Can we change our implicit minds? New evidence for how, when, and why implicit impressions can be rapidly revised. *Advances in Social Psychology*.

Corneille, O., & Stahl, C. (2018). Associative attitude learning: A closer look at evidence and how it relates to attitude models. *Personality and Social Psychology Review*.

Dalege, J., Borsboom, D., van Harreveld, F., & van der Maas, H. L. J. (2018). The Attitudinal Entropy (AE) Framework as a General Theory of Individual Attitudes. *Psychological Inquiry, 29,* 175-193. doi: 10.1080/1047840X.2018.1537246

De Houwer, J. (2018). Propositional models of evaluative conditioning. *Social Psychological Bulletin, 13*(3). doi:10.5964/spb.v13i3.28046

De Houwer, J. (2009). The propositional approach to associative learning as an alternative for association formation models. *Learning & Behavior*, *37*, 1–20.

De Houwer, J. (2014). A Propositional Model of Implicit Evaluation. *Social and Personality Psychology Compass, 8*, 342-353.

De Houwer, J., Gawronski, B., & Barnes-Holmes, D. (2013). A functional-cognitive framework for attitude research. *European Review of Social Psychology, 24*, 252–287. doi:10.1080/10463283.2014.892320

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience,* 11:127–38.

Gardner, H. (1987). The mind's new science: A history of the cognitive revolution. New York: Basic Books.

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: an integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*, 692–731. doi: 10.1037/0033-2909.132.5.692

Hayes, S. C., & Brownstein, A. J. (1986). Mentalism, behavior-behavior relations, and a behavior-analytic view of the purposes of science. *The Behavior Analyst, 9(2)*, 175-190.

Hussey, I., Hughes, S., Lai, C., Ebersole, C., Axt, J., & Nosek, B. A. (2018). Attitudes 2.0: A large dataset for investigating relations among implicit and explicit attitudes and identity. https://osf.io/pcjwf

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review, 93,* 1449-147.

Melnikoff, D. E., & Bargh, J. A. (2018). The mythical number two. *Trends in Cognitive Sciences, 22*, 280-293. doi: 10.1016/j.tics.2018.02.001.

Metzinger, T., & Wiese, W. (2017) *Philosophy and predictive processing*. Frankfurt am Main: MIND Group.

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *The Behavioral and Brain Sciences*, *32*, 183–198. doi: 10.1017/s0140525x09000855

Nosek, B. A., & Hansen, J. J. (2008). The associations in our heads belong to us: Searching for attitudes and knowledge in implicit evaluation. *Cognition and Emotion*, *22*(4), 553–594. doi: 10.1080/02699930701438186

Otworowska, M., Van Rooij, I., & Kwisthout, J. (2018). Maximizing entropy of the Predictive Processing framework. doi:10.31234/osf.io/5zam7

Schwarz, N. (2007). Attitude construction: Evaluation in context. *Social Cognition*, *25,* 638-656.

Skinner, B. F. (1953). *Science and human behavior*. New York: Macmillan.

Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review, 4*, 108–131.

Van Dessel, P., Hughes, S., & De Houwer, J. (2019). How do actions influence attitudes? An inferential account of the impact of action performance on stimulus evaluation. *Personality and Social Psychology Review, 23*(3), 267-284. doi:10.1177/1088868318795730