CITY
UNIVERSITY OF LONDON
EST 1894

Dissertations on Decision-Making: Similarity, Constructive Judgments, Morality and Social Dilemmas.

Albert Barqué-Duran

Submitted to City University London for the degree of Doctor of Philosophy.

Department of Psychology

July 2016

Contents

Theme 1: A Quantum Cognitive Approach on Similarity Judgments

Theme 2: A Quantum Cognitive Approach on Constructive Judgments

Theme 3: On Moral Judgments

List of figures and tables

Theme 1

Chapter 1

Chapter 2

Chapter 3:

Chapter 4:

Theme 2

Chapter 5:

Theme 3

Chapter 6:

Chapter 7:

Theme 4

Chapter 8:

Abstract

The present thesis (mostly) concerns the application of alternative mathematical methods to understand patterns in human cognition and to model them. The different chapters presented in this thesis show research that concerns the application of quantum probability (QP) theory in the modeling of human decision-making. Quantum probability (QP) theory is a theory for how to assign probabilities to events. QP theory can be thought of as the probability rules from quantum mechanics, without any of the physics. This work is not about the application of quantum physics to brain physiology. Rather, we are interested in QP theory as a mathematical framework for cognitive modelling. This theory is potentially relevant in any behavioural situation that involves uncertainty. QP theory is analogous to classical probability theory, though QP theory and classical probability (CP) theory are founded from different sets of axioms (the Kolmogorov and Dirac/von Neumann axioms respectively) and so are subject to alternative constraints. In this thesis we show that especially over the last decade, there has been a growing interest in decision-making and cognitive models using a quantum probabilistic (QP) framework. We see how this development encompasses publications in major journals (see Pothos and Busemeyer, 2013; Wang et al., 2014; and Yearsley and Pothos, 2014; among others), special issues, and dedicated workshops, as well as several comprehensive books (Busemeyer and Bruza, 2012; Khrennikov, 2010; and Haven and Khrennikov, 2010).

However, uncertainty itself is neither ethical nor unethical – yet it is inherent to most situations in which, for instance, moral judgments and decisions have to be made. For a descriptive understanding of judgment and decisions in moral situations, it is an important lesson to acknowledge both the cognitive side (bounded rationality) and the environment (ecological rationality) – and thus the uncertainty of the world and how the mind deals with it. This thesis also shows significant interest in moral and social psychology. Specifically, we consider present technologies that suggest a need for evaluating alternative contexts for ethical decision-making. How the research on human-machine interaction feeds back into humans' understanding of themselves as moral agents? This key question ultimately relates to the nature of ethical theory itself.

Overall, this dissertation presents and addresses not only standard aspects of decision- making processes, such as similarity judgments (*Chapters 1 to 4*) or the constructive role of articulating impressions (*Chapter 5*), but also standard aspects of social psychology, such as moral judgments (*Chapters 6 and 7)* and game theory (*Chapter 8).* As stated in the *Declarations* section, the present thesis is a combination of a standard and a publication- based dissertation.

Acknowledgements

I would like to thank the Air Force Office of Scientific Research, Air Force Material Command, USAF, who, through their grant FA 8655-13-1-3044 made the work in this thesis possible. I would also like to thank the Psychology Department at City University London.

I am grateful to my family and those special people for their support throughout all my life.

I would like to express the deepest gratitude to my supervisors, Prof. Emmanuel Pothos and Prof. James Hampton, for all the opportunities, guidance, encouragement and inspiration they provided me with throughout my PhD.

Declarations

This thesis is submitted to City University London in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree.

The present thesis is a combination of a standard and a publication-based dissertation. On each *Theme*, a *Statement of Contribution* is presented which specifies the work carried out by the author for each of the projects (development of the study concepts, design of the experiments, data collection and analysis and interpretation of results).

The present thesis also presents some of the artwork that the author produced during the time of the PhD. Some of the artwork was exhibited in different countries and won an award from the *Cambridge Neuroscience Society (University of Cambridge)*. After each *Theme*, a *Special Artwork Chapter* is presented with some of the selected artworks[1] inspired by the research on this thesis.

List of publications including under review papers:

Theme 1: A Quantum Cognitive Approach on Similarity Judgments

Barque-Duran, A., Pothos, E. M., Yearsley, J., Hampton, A., Busemeyer, J. R. & Trueblood, J. S. (2015). Similarity Judgments: From Classical to Complex Vector Psychological Spaces. In, E. Dzhafarov, R. Zhang, S. Joardan, and V. Cervantes (Eds.) *Contextuality from Quantum Physics to Psychology*. World Scientific.

Pothos, E., Barque-Duran, A., Yearsley, J., Trueblood, J., Busemeyer, J., Hampton, J. (2015). Progress and current challenges with the Quantum Similarity Model. *Frontiers in Psychology*. 6, 205.

[1] © Albert Barqué-Duran. 2016. All rights reserved.

Yearsley, J. M., Pothos, E. M., Barque-Duran, A. & Hampton, J. A. (2015) Diagnosticity: Some Theoretical and Empirical Progress. *Proceedings of the 37th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society. 2739-2744.

Yearsley, J. M., Pothos, E. M., Hampton, J. A., & Barque-Duran, A. (2015). Towards a Quantum probability theory of similarity judgments. *Lecture Notes in Computer Science.* 8951, 132.

Yearsley, J., Barque-Duran, A., Pothos, E., Hampton, J., Scerrati, E. The Triangle Inequality Constraint in Similarity Judgments (*submitted*).

Theme 2: A Quantum Cognitive Approach on Constructive Judgments

White, L., Barque-Duran, A., Pothos, E. (2015) An investigation of a quantum probability model for the constructive effect of affective evaluation. *Philosophical Transactions of the Royal Society A.* 374, 20150142.

Theme 3: On Moral Judgments

Barque-Duran, A., Pothos, E., Yearsley, J., Hampton, J. Contemporary Morality: Moral Judgments in Digital Contexts. (*under review*)

Barque-Duran, A., Pothos, E., Yearsley, J., Hampton, J. (2015). Patterns and Evolution of Moral Behavior: Moral Dynamics in Everyday Life. *Thinking and Reasoning*. 22, 31-56.

Barque-Duran, A., Pothos, E., Yearsley, J., Hampton, J. (2015). Moral Dynamics in Everyday Life: How morality evolves in time? *Proceedings of the 37th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society. 154-159.

Theme 4: A Quantum Cognitive Approach on Game Theory

Denolf, J., Martinez-Martinez, I., Barque-Duran, A. (*forthcoming*) A quantum-like model for complementarity of preferences and beliefs in dilemma games. *Journal of Mathematical Psychology.*

Martínez-Martínez, I., Denolf, J., Barque-Duran, A. (2016). Do Preferences and Beliefs in

Dilemma Games exhibit Complementarity? *Lecture Notes in Computer Science*. 9535, 142-153.

List of exhibitions (artwork produced and inspired by this thesis):

- **Creative Reactions – Pint of Science Festival**. London, UK. (May 2016)
  "Albert vs. Machine". On Artificial Intelligence and Computational Creativity.
- **Cambridge Neuroscience Society** (University of Cambridge). Cambridge, UK. (July 2015)
  "Untangling Mind and Brain".
- **SciArt Center**. New York, USA. (June 2015)
  "The New Unconscious".
- **UCL Neuroscience Society** (University College London). London, UK. (May 2015)
  "The Brain & Mental Health".
- **TEDxYouth@Barcelona**. Barcelona, Spain. (March 2015).
  "Painting Contemporary Morality".
- **Department of Psychology** (City University London), UK. (March 2015)
  "Drawing Neurons,Writing Brains, Painting Minds".

Prologue

Probability theory is of central importance in the modeling of human decision-making. Fundamentally, humans have to make their decisions on the basis of uncertain information, whether this uncertainty relates to ambiguity or lack of knowledge in the information provided or to the implications of a decision. Understanding the computational principles which guide human decision-making is of obvious significance. Such an understanding would allow us to predict consistent patterns in judgment, anticipate perhaps problematic decisions, and attempt to predict reactions to particular decision-making problems. The particular characteristics of human decision-making are a fundamental aspect of what it means to be human.

The present thesis (mostly) concerns the application of alternative mathematical methods to understand patterns in human cognition and to model them. The different chapters presented in this thesis show research that concerns the application of quantum probability (QP) theory in the modeling of human decision-making. Quantum probability (QP) theory is a theory for how to assign probabilities to events. QP theory can be thought of as the probability rules from quantum mechanics, without any of the physics. This work is not about the application of quantum physics to brain physiology. Rather, we are interested in QP theory as a mathematical framework for cognitive modelling. This theory is potentially relevant in any behavioural situation that involves uncertainty. QP theory is analogous to classical probability theory, though QP theory and classical probability (CP) theory are founded from different sets of axioms (the Kolmogorov and Dirac/von Neumann axioms respectively) and so are subject to alternative constraints.

CP theory has been the dominant approach in the modeling of human decision-making. It is widely assumed that human decisions conform to CP principles. This is not to say that naïve observers are familiar with the abstract mathematical principles of CP theory. Rather, the assumption is that humans make such decisions as would be predicted by an application of CP principles and processes. But, when can we say that a person is behaving rationally? When irrationally? CP theory provides a standard for rational decision-making, against which human behavior is typically assessed. That is, CP theory prescribes certain decisions as rational or irrational, correct or incorrect, on the basis of consistency with CP principles. Take as an example The Dutch Book Theorem (DBT; e.g., Howson and Urbach, 1993), which shows that if one assigns probabilities to events in a way inconsistent with the axioms of CP theory, then it is possible to identify a combination of stakes (money to be won or lost, depending on whether the events occur or not), which guarantees a loss (or gain, depending on the sign of the stakes). That is, according to the DBT, when failing to follow the rules of CP theory, you may be vulnerable to a sure loss. Extensive evidence on the so-called probabilistic fallacies shows that naïve observers routinely behave in a way that superficially diverges from CPT principles, thus inviting the conclusion that humans are irrational. Quantum probability theory (QPT) is a probabilistic framework, alternative to CPT, that has been employed to model behavior for some of these fallacies, but the rational status of QPT is under examination. We could argue that *Homo Economicus* is no longer a reality (where here we imply that traditional economic rationality concerns consistency with the CP principles; this is clearly an approximation to a complex debate); but is *Homo Heuristicus* our best approach (where we allude to the alternative tradition in explaining human decision-making, based on heuristics)? While the latter has been successful in explaining judgment and decision-making in many domains (Gigerenzer et al., 1999) this thesis explores the extent to which an alternative system for probability, QP theory, could provide a descriptive model for human decision-making, as well as a normative model for human rationality.

In this thesis we show that especially over the last decade, there has been a growing interest in decision-making and cognitive models using a quantum probabilistic (QP) framework. We see how this development encompasses publications in major journals (see Pothos and Busemeyer, 2013; Wang et al., 2014; and Yearsley and Pothos, 2014; among others), special issues, and dedicated workshops, as well as several comprehensive books (Busemeyer and Bruza, 2012; Khrennikov, 2010; and Haven and Khrennikov, 2010).

However, uncertainty itself is neither ethical nor unethical – yet it is inherent to most situations in which, for instance, moral judgments and decisions have to be made. For a descriptive understanding of judgment and decisions in moral situations, it is an important lesson to acknowledge both the cognitive side (bounded rationality) and the environment (ecological rationality) – and thus the uncertainty of the world and how the mind deals with it. This thesis also shows significant interest in moral and social psychology. Specifically, we consider present technologies that suggest a need for evaluating alternative contexts for ethical decision-making. How the research on human-machine interaction feeds back into humans' understanding of themselves as moral agents? This key question ultimately relates to the nature of ethical theory itself.

Overall, this dissertation presents and addresses not only standard aspects of decision-making processes, such as similarity judgments (*Chapters 1 to 4*) or the constructive role of articulating impressions (*Chapter 5*), but also standard aspects of social psychology, such as moral judgments (*Chapters 6 and 7)* and game theory (*Chapter 8).* As stated in the *Declarations* section, the present thesis is a combination of a standard and a publication-based dissertation.

Theme 1

A Quantum Cognitive Approach on Similarity Judgments

Statement of Contribution

Theme 1 is a collaborative work, mainly with Emmanuel Pothos, James Hampton and James Yearsley but also with Jerome Busemeyer and Jennifer Trueblood. This *Theme 1* consists of several challenging projects, involving complex methods and complex models; the author focused and led on the empirical parts while most of the modelling/ mathematical part was led by others. Specifically, the author contributed to the development of the study concepts, designed the experiments and collected, analysed and interpreted the data for the studies presented, with input from all other authors, mainly Emmanuel Pothos, James Hampton and James Yearsley. The author contributed to the writing of the manuscripts published.

List of publications for *Theme 1*:

Barque-Duran, A., Pothos, E. M., Yearsley, J., Hampton, A., Busemeyer, J. R. & Trueblood, J. S. (2015). Similarity Judgments: From Classical to Complex Vector Psychological Spaces. In, E. Dzhafarov, R. Zhang, S. Joardan, and V. Cervantes (Eds.) *Contextuality from Quantum Physics to Psychology*. World Scientific.

Pothos, E., Barque-Duran, A., Yearsley, J., Trueblood, J., Busemeyer, J., Hampton, J. (2015). Progress and current challenges with the Quantum Similarity Model. *Frontiers in Psychology*. 6, 205.

Yearsley, J. M., Pothos, E. M., Barque-Duran, A. & Hampton, J. A. (2015) Diagnosticity: Some Theoretical and Empirical Progress. *Proceedings of the 37th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society. 2739-2744.

Yearsley, J. M., Pothos, E. M., Hampton, J. A., & Barque-Duran, A. (2015). Towards a Quantum probability theory of similarity judgments. *Lecture Notes in Computer*

Science. 8951, 132.

Yearsley, J., Barque-Duran, A., Pothos, E., Hampton, J., Scerrati, E. The Triangle Inequality Constraint in Similarity Judgments (*submitted*).

Chapter 1

Similarity judgments: from classical to complex vector psychological spaces.

Abstract

This chapter reviews progress with applications of quantum theory in understanding human similarity judgments. We first motivate and subsequently describe the quantum similarity model (QSM), which was proposed by Pothos, Busemeyer and Trueblood (2013), primarily as a way to cover the empirical findings reported in Tversky (1977). We then show how the QSM encompasses Tversky's (1977) results, specifically in relation to violations of symmetry, violations of the triangle inequality and the diagnosticity effect. We next consider a list of challenges of the QSM and open issues for further research.

1. Background and motivations for a new model

Similarity judgments play a central role in many areas of psychology (e.g. Goldstone, 1994; Medin, Goldstone & Gentner, 1993; Pothos, 2005, Sloman & Rips, 1998). Consequently, they have received much attention (e.g. Goodman, 1972), especially in relation to Tversky's (1977) findings, which have been a major focus of subsequent theoretical work on similarity judgments.

One traditional way to understand similarity uses a geometric approach, whereby similarity is assumed to correspond to a function of the distance between concepts in a psychological space. According to this approach, stimuli or concepts are represented as points in a multidimensional psychological space, with similarity being a decreasing function of distance in that space. The origin of the debate, criticism and the several attempts to empirically refute this approach all relate to the fact that similarity measures based on distance must obey various properties, called the metric axioms, that all distances (and simple related measures) are subject to. The most famous demonstration that human similarity judgments are inconsistent with these properties is due to Tversky (1977). The importance and the impact of Tversky's paper come from the fact that his findings questioned the fundamental properties of any model of similarity based on distance in psychological space.

Specifically, Tversky's approach was to provide empirical tests of the metric axioms, regardless of the specifics of the similarity approach. Showing, as he did, that the metric axioms are inconsistent with human similarity judgments, he concluded that human similarity judgments cannot be modeled with any distance-based approach.

Specifically, Tversky (1977) reported violations of (1) minimality: identical objects are not always judged to be maximally similar; (2) symmetry: the similarity of A to B can be different from that of B to A; (3) the triangle inequality: the distance between two points cannot exceed the sum of their distances to any third point; (4) a diagnosticity effect: the similarity between the same two objects can be affected by which other objects are present. In the next four sections we elaborate on all these findings and we consider some notable previous theoretical efforts to account for Tversky's (1977) challenges. Note we do not consider minimality, since most models (including the QSM) can become consistent with violations of minimality through some process of noise in how representations are specified and compared.

1.1. Asymmetries

If similarity is determined by distance, then how could it be the case that the similarity between two objects depends on the order in which the objects are considered? Directionality can arise from the fact that the relevant stimuli are not (always) simultaneously presented. For example, the temporal ordering of the stimuli can impose directionality structure in the similarity comparison. Alternatively, directionality can be conveyed in a syntactical way, e.g., if an observer is asked to evaluate sentences like "A is similar to B". Whenever this happens, there is a potential for asymmetry. This can be readily seen in the kind of task Tversky (1977) employed to explore putative violations of symmetry. He asked participants to choose which they preferred between these two statements: "North Korea is similar to Red China" and "Red China is similar to North Korea" (for simplicity we will use only Korea and China). Most participants preferred the former to the latter statement (this demonstration involved several other pairs of counties and was generalized to other kinds of stimuli). This result implied that the similarity of Korea to China (expressed as sim(Korea, China)) is higher that of China to Korea (expressed as sim(China, Korea)), and thus revealed a violation of symmetry in similarity. Tversky's interpretation about why such asymmetries arise related to differences between the two stimuli in the extent of featural knowledge combined with differential weight given to the features specific to each concept (the parameters α and β in his model, see below). But, asymmetries in similarity judgments can also arise in other ways:

Polk et al. (2002) proposed that they can also be the result of differences in the frequency of occurrence of one of the compared stimuli (a higher similarity was observed when comparing a low frequency stimulus with a high frequency one). Even before Tversky's (1977) work, Rosch (1975) had proposed similarity asymmetries can arise when a less prototypical stimulus is compared to a more prototypical one.

Asymmetries are difficult to reconcile with the idea of similarity-as-distance. Some kind of mechanism that can produce asymmetries, in some circumstances, in a more natural way is clearly desirable. We will see below that a quantum approach provides such a mechanism.

1.2. The triangle inequality

Tversky (1977) also considered how similarity judgments can lead to violations of the triangle inequality, another one of the metric axioms. In his paper, he states that (p.329) "the perceived distance of Jamaica to Russia exceeds the perceived distance of Jamaica to Cuba, plus that of Cuba to Russia – contrary to the triangle inequality." We can assume that perceived distance is either the same or approximately the same as dissimilarity, so that consistency with the triangle inequality requires $Dissimilarity(Jamaica, Russia) < Dissimilarity(Jamaica, Cuba) + Dissmilarity(Cuba, Russia)$. Regarding the implications from this statement for similarity, we need a function that takes us from dissimilarity to similarity (or at least some indication of its properties) and Tversky does not provide this. Instead, he says "…the triangle inequality implies that if a is quite similar to b, and b is quite similar to c, then a and c cannot be very dissimilar from each other. Thus, it sets a lower limit to the similarity between a and c in terms of the similarities between a and b and between b and c." But, this expression is too vague to lead to a quantitative constraint. If one assumed that similarity is just the negative of dissimilarity, then one could write $Similarity\ (a, c) > Similarity\ (a, b) + Similarity\ (b, c)$, but such an expression leaves us with some problems (e.g., we would need another function to take a negative, unbounded similarity measure to something that corresponds to e.g. similarity ratings; assuming the latter are closer to psychological similarity, in itself another assumption). No doubt some readers will find it unsatisfactory that a discussion, which is overall about similarity, actually is restricted to claims only about dissimilarity. But, for our purposes it is not necessary to resolve these issues, since we can easily formulate our discussion in terms of the inequalities based on dissimilarities above.

With these points in mind, Tversky's example was as follows. Consider A=Russia and B=Jamaica; Dissimilarity(Russia, Jamaica) is high. Consider also C=Cuba. But Dissimilarity (Russia, Cuba) is low (these countries are similar because of political affiliation) and Dissimilarity(Cuba, Jamaica) is also low (these countries are similar because of geographical proximity). Thus, Tversky's example suggests that Dissimilarity(Russia, Jamaica) > Dissimilarity(Russia, Cuba) + Dissimilarity(Cuba, Jamaica), which suggests a violation of the triangle inequality. Interestingly, more elaborate theories of similarity, specifically developed to address Tversky's (1977) findings, do not always deal with violations of the triangle inequality straightforwardly (we will consider Krumhansl's, 1978. theory shortly, in Section 1.4.4).

1.3. Diagnosticity

The diagnosticity effect, a particular type of context effect, is another major finding from Tversky (1977). Participants were asked to identify the country most similar to Austria, from a set of alternatives including Hungary, Poland, and Sweden. Participants typically selected Sweden. However, when the alternatives were Hungary, Sweden, and Norway, participants typically selected Hungary. Thus, the same similarity relation (e.g., the similarity between Sweden and Austria or the similarity between Hungary and Austria) appears to depend on which other stimuli are immediately relevant, showing that the process of establishing a similarity judgment may depend on the presence of other stimuli, not directly involved in the judgment. Tversky's (1977) explanation was that the diagnosticity effect arises from the grouping of some of the options. For example, when Hungary and Poland were both included, their high similarity made participants spontaneously code them with their obvious common feature (both were Communist bloc countries at the time), which, in turn, increased the similarity of the other two options, (Austria and Sweden) which were both Western democracies.

1.4. Previous theoretical formalisms

In the following sections we consider some significant previous theoretical efforts to account for Tversky's (1977) challenges. Such efforts have the same objective, but can vary widely in their assumptions, implementation, and structure, thus sometimes making it hard to identify their key distinguishing characteristics. Consideration of these previous theoretical approaches motivates our own proposal for a new approach, based on quantum theory.

1.4.1. Extensions of the Geometric Model

Let us first repeat the point that simple extensions of geometric models of similarity are unsatisfactory. In standard models (e.g. Shepard, 1980), the similarity between two entities A and B is given by $Sim(A, B) = e^{-c \cdot distance(A,B)}$, where c is a constant. Clearly, such a function of similarity obeys symmetry. This basic definition could lead to an asymmetric similarity measure with the introduction of a directionality parameter, p_{AB}, indexed in a way to indicate that it may have a different value depending on whether we are considering the similarity of A to B or the similarity of B to A (see Nosofsky, 1991, for these ideas). However, without a scheme for motivating particular values of the directionality parameter, this proposal cannot be said to explain asymmetry in similarity judgments a priori (even if it can post hoc reproduce the empirical results).

The basic geometric scheme also fails in the case of diagnosticity, since there is no mechanism by which to augment the computation of similarity for two entities by information for other, assumed relevant, objects. One could augment a basic similarity scheme with attentional weights, which could vary depending on contextual influences (cf. Nosofsky, 1984). However, an approach like this would be incomplete without a precise understanding of how attentional weights can change, across different contexts. Overall, this simple extension of geometric models is a straw man and it is unsurprising that it fails. We will shortly see Krumhansl's (1978; see also Nosofsky, 1991) proposal.

1.4.2. Tversky's proposal

Tversky's (1977) Contrast Model proposed that

$$Similarity(A, B) = \theta f(A \cap B) - af(A - B) - \beta f(B - A)$$

where θ, a, β are constant parameters, $A \cap B$ denotes the common features between A and B, A-B the features of A which B does not have and B-A the features of B which A does not have (see also Bush & Mosteller, 1951; Eisler & Ekman, 1959). Such a scheme can predict violations of symmetry if A has more features than B and the parameters a, β are different to each other and suitably set (e.g., $\theta > 0, a = 1, \beta = 0$ allows the emergence of asymmetries from Tversky's contrast model, in the predicted direction). For example, regarding the Korea-China example, Tversky assumed that China has more features than Korea, because the average observer will know more about China than Korea. First we must assume that $\alpha > \beta$ in a directional judgment of similarity, so that distinctive features of the subject are more relevant than the distinctive features of the referent. Then, the similarity of Korea to China would be fairly high (a minimal negative contribution from $af(A-B)$, since Korea has very few features which China does not have). However, in comparing China to Korea, there is now a larger contribution from $af(A-B)$, which lowers the overall similarity result. Thus, according to Tversky's similarity model, China is predicted to be less similar to Korea than Korea is to China. Tversky's model of similarity is appealing, but still involves two independent parameters, which must have appropriate values to account for violations of symmetry. For example, if instead of assuming $a > \beta$, we assume the reverse, then the model fails to predict the right direction for symmetry violation in the Korea-China example. There are some similarities between the a, β parameters in Tversky's similarity model and the directionality ones above (Nosofsky, 1991), in the sense that both kind of parameters are about defining a 'preferred' direction in similarity comparisons (that is, a direction that leads to higher similarities).

So, why did Tversky (1977) set the contrast model parameters one way, as opposed to another? Tversky's (1977) assumption was that when assessing $Sim(A,B)$ then A is the subject and B is the referent and so "…the features of the subject are weighted more heavily than the features of the referent" (p. 333, Tversky, 1977). This allows one to set $a > \beta$, which enables possible violations of symmetry, as long as the compared objects differ in the number of distinctive features. While this assumption seems reasonable, it is also one which does not follow naturally from the rest of Tversky's (1977) model. Moreover, it is hard to logically exclude the alternative assumption (which leads to the exact opposite prediction), i.e., that it is the referent's features which are more heavily weighted. It is this assumption which basically allows the prediction of the asymmetry in a specific direction, so the extent to which it can be justified a priori goes hand in hand with our perception of whether the contrast model can explain the China, Korea asymmetry in an a priori way.

Tversky's (1977) contrast model provides the same elegant account for both the triangle inequality and the diagnosticity effect, in terms of how different contexts lead to the emergence of different diagnostic features (but see Krumhansl, 1978, for some criticisms, relating to how weights are assigned to features, with varying contexts; her theory is considered in Section 1.4.4 below). His explanation for these empirical results is theoretically appealing, but some concerns can be expressed regarding the number and precise form of the emerging diagnostic features. In closing the discussion for Tversky's (1977) model, it is perhaps worth remarking that this detailed scrutiny of his work, so as to motivate the need for a new model (the quantum model), should not detract from the fact that his theory has had a profound and lasting influence on the similarity literature, probably more so than any other similarity theory.

1.4.3. Classical Probability Theory

In this section we consider whether classical (Bayesian) probability theory can provide an account of Tversky's (1977) challenges. It has to be said that classical probability theory is not obviously relevant to human similarity judgments. Nevertheless, cognitive models based on classical probability theory have been extremely successful in recent years (e.g., Griffits et al., 2010; Oaksford & Chater, 2009; Tenenbaum et al., 2011) so it is worth exploring possible extensions in relation to similarity. The similarity between two instances could be modeled as the joint probability of both instances together, $Prob(A \wedge B)$. Such a joint probability could be understood in terms of statements corresponding to both instances being concurrently true or in terms of the ease of having both thoughts together. We stress that our aim here is not to develop an operational model of similarity based on classical probability theory! Rather, we look at the Quantum Similarity Model and consider which operations are analogous in classical probability theory. The Quantum Similarity Model basically models similarities as conjunctive probabilities (the ease of having a thought about the first between two compared concepts and then the second). So, without worrying too much about operational details, we consider whether a similar approach might work with classical probability theory.

However, the joint probability operator is symmetric in classical probability, so that $Prob(A \wedge B) = Prob(B \wedge A)$, and so this scheme fails to account for violations of symmetry in similarity. Note that one could say that $Prob(A \wedge B|\ Order\ 1) \neq Prob(B \wedge A|\ Order\ 2)$, but such a scheme offers a trivial solution to the problem of asymmetry (it is equivalent to the directionality parameter one above). Alternatively, one could model the similarity between two instances in terms of a conditional probability function, which can be asymmetric. In other words, one could postulate that $Sim(A, B) = Prob(A|B) \neq Prob(B|A) = Sim(B, A)$. However, such a scheme does not work. Consider the paradigmatic Korea-China example again, from Tversky (1977), and assume that $Sim(Korea, China) = Prob(China|Korea)$, so that the similarity process involves assessing the probability of the second predicate given knowledge of the first (note, something like $Prob(China|Korea)$ could be interpreted as the conditional probability of thinking about China, given that we have been thinking about Korea). Then, $Sim(Korea, \sim China) = Prob(\sim China|Korea) = 1 -$

$Prob(China|Korea) = 1 - Sim(Korea, China)$. Since $Sim(Korea, China)$ is assumed to be high, it follows that $Sim(Korea, {\sim}China)$ has to be low. The latter conclusion seems reasonable, as all the predicates which satisfy ${\sim}China$ would, on average, have a low similarity to Korea (there are more countries which are dissimilar to Korea than ones which are similar). However, this approach can also lead to paradoxical predictions. Consider $Sim(Alaska, {\sim}China) = Prob({\sim}China|Alaska) = 1 - Prob(China|Alaska) = 1 - Sim(Alaska, China)$. Therefore, following this set of equalities in the reverse order, as $Sim(Alaska, China)$ is very low, it must be the case that $Sim(Alaska, {\sim}China)$ is very high. However, such a prediction seems counterintuitive.

Of course, classical probability theory is a sophisticated computational framework and it is possible that a satisfactory account of symmetry violations (and the rest of Tversky's, 1977, challenges) can emerge. Our purpose here was to assess whether a basic classical probability model is consistent with violations of symmetry. This appears not to be the case. Moreover, it is not clear how this basic classical probability model could be extended in the case of the other relevant empirical results. A critic might note that (classical) probability theory has nothing to do with similarity judgments and this entire section is misguided. Nevertheless, the Quantum Similarity Model does exactly this: it provides a formalism in which probabilities (corresponding to the ease of having sequences of thoughts) lead to similarity judgments. Indeed, we think that approaching similarity judgments as probabilities (defined in a suitable way) is a worthwhile endeavor, insofar that this provides a framework for exploring commonalities between similarity and probabilistic inference (Shafir et al., 1990; Tversky & Kahneman, 1983).

1.4.4. Krumhansl's distance-density model

Krumhansl's (1978, 1988) distance-density model provides a principled extension to the basic geometric model of similarity. Her proposal rests on the assumption that alternatives lying within dense subregions of psychological space are subject to finer discrimination than alternatives lying in less dense subregions. With regards to similarity judgments, this implies that a pair of points a given distance apart in a dense region would have a lower similarity (greater psychological distance) as compared to an identical pair of points in a less dense region. More specifically, the distance between two points A and B in psychological space should be affected by the local density around each point, $D(A)$ and $D(B)$. The local density around a point reflects the number of other points within a certain radius. Thus, $d'(A,B) = d(A,B) + aD(A) + bD(B)$, where $d(A,B)$ is the standard geometric distance, a and b are parameters that reflect the weight given to each density, and $d'(A,B)$ is the modified distance measure, as affected by local densities. As with Tversky's (1977) similarity model, it is immediately clear that if $a = b$ then $d'(A,B) = d(A,B) + a\big(D(A) + D(B)\big) = d'(B,A)$; that is, unless $a \neq b$ no violations of symmetry are predicted. But, also as with Tversky's model, setting the a, b parameters in different ways (e.g., $a, b > 0$, as in the original formulation of the model, versus $a, b < 0$) predicts asymmetries in different directions. In order to account for asymmetries, Krumhansl (1978) adopted an assumption equivalent to that of Tversky (1977), that is, that the density of one object influences the comparison more than the density of the other.

In the particular case of the Korea-China example, for a violation of symmetry to occur, one would need to assume that the local density around China is different from the local density around Korea. Krumhansl (1978) suggested that prominent objects are likely to have many features and so these objects are likely to share features with a greater number of other objects, as compared to objects with fewer features. Therefore, prominent objects are more likely to exist in denser regions of psychological space. Krumhansl's (1978) logic is perhaps intuitive, but it does raise some questions. For example, why should prominent objects (with more features) *share* a greater number of features with other objects? These additional features could be distinctive, as indeed is implied in Tversky's (1977) analysis.

Krumhansl's (1978) explanation for the triangle inequality is based on the idea that similarity judgments emphasize dimensions and features that objects have in common. As a result, stimuli which are far apart in the overall psychological space may be close to each other in a low dimensionality subspace, corresponding to the common dimensions between the stimuli. For example, Russia and Cuba are similar in the subspace of Communism, which corresponds to their common dimension. Krumhansl (1978, p.12) notes "Subspaces defined by obvious stimulus dimensions would seem to be likelier projections than subspaces not corresponding to such dimensions" and goes on to observe that such a scheme may be able to account for similarity relations inconsistent with the triangle inequality. However, this explanation involves some ad hoc assumptions. For example, why is similarity assessed in a subspace (such an assumption does not follow from her density model, nor is employed elsewhere; cf. Ashby & Perrin, 1988), why is the appropriate subspace determined in such a way etc. (these issues are similar to the corresponding criticism for Tversky's, 1977, account).

The density model is easily consistent with the diagnosticity effect (Krumhansl, 1978, 1988). Recall, the distance between two concepts increases as the density of one object increases. For example, $Sim(Hungary, Austria)$ would change depending on the $D(Hungary)$ term, which is the density around Hungary. If we add Poland to the choice set, $D(Hungary)$ increases, the effective distance between Hungary and Austria also increases, and so the similarity between the two countries decreases. A perhaps unsatisfactory aspect of Krumhansl's (1978) account for the diagnosticity effect is that it fails to capture the (reasonable) intuition that different comparisons do evoke different relevant features (in Tversky's, 1977, approach) or perspectives (in the quantum approach).

1.4.5. Ashby & Perrin's general recognition theory

Ashby and Perrin's (1988) general recognition theory is an established probabilistic approach to similarity for perceptual stimuli, also based on representations in a psychological space. In brief, the theory can readily account for violations of symmetry in similarity judgments of perceptual stimuli. Each time a stimulus is perceived it can correspond to a different point in psychological space, according to a particular probability distribution. Psychological space is divided into response regions, such that within each response region it is optimal to make a particular response. Thus, similarity between two stimuli depends on the extent to which the distribution of perceptual effects for the first stimulus overlaps with the optimal response region for the second stimulus. Formally, for a pair of two-dimensional stimuli A and B, $Sim(A,B) = \iint_{\mathbf{R_B}} f_A(x,y)dxdy$, where $\mathbf{R_B}$ is the region in the x,y perceptual plane associated with response R_B and $f_A(x,y)$ is the probability density function for the distribution of perceptual effects of stimulus A (note that similarity is actually defined as a function of the above integral, but this is not relevant here). As Ashby and Perrin (1988) note, such a scheme can lead to violations of symmetry in a number of ways. For example, if stimulus B is associated with a greater response region than A then, in general, $Sim(A,B) > Sim(B,A)$ and if the perceptual effects distribution for A has a greater variability than X, then it is also the case that $Sim(A,X) > Sim(X,A)$.

As Ashby and Perrin (1988) observed, these intuitions can be related to the Korea-China example. First, because for many observers Korea will be a 'more vague and poorly defined concept' (p.133), the representation of Korea in psychological space will have a greater variability. Second, they argued that the response region for Korea would be smaller than that of China, because Korea is very similar to many other countries. According to general recognition theory, both these factors predict that $Sim(Korea, China) > Sim(China, Korea)$. But there are some problems with this account.

Whether Korea or China is more similar to other countries is unclear. Ashby and Perrin (1988, p.133) note that "…for many people North Korea is very similar to several other countries." But, recall, Krumhansl (1978, p.454) made the exact opposite assumption, "If prominent countries … are those stimuli having relatively many features, then these objects have features in common with a larger number of different objects….". In other words, Krumhansl (1978) assumed that it is China, not Korea, which is similar to a greater number of other countries. Thus, Ashby and Perrin (1988) and Krumhansl (1978) make the exact opposite assumption, regarding whether it is Korea or China which is similar to a greater number of other countries. This shows the fickle nature of this assumption and how it can be (fairly easily) made one way or another, so that the corresponding models can describe an asymmetry in similarity judgments in the predicted direction and, equally easily, in the opposite direction as well.

Regarding the triangle inequality, Ashby and Perrin (1988) show how one can manipulate the perceptual effects distributions, so that two stimuli can be dissimilar to each other and yet both similar to a third stimulus, hence violating the triangle inequality. Such a situation can clearly be mapped to Tversky's (1977) Russia-Cuba-Jamaica example. One weakness of Ashby and Perrin's (1988) demonstration is that it appears to assume (see their Figure 4, p.133; one distribution circular on a plane, the other two elliptical, if one considers a suitable cross-section of the distributions) asymmetric and inequivalent perceptual effects distributions for the three stimuli. In the case of simple perceptual stimuli, presumably their form can be manipulated to produce arbitrary perceptual effects distributions. However, it is unclear whether such an assumption is reasonable in the case of, for example, comparisons between Russia, Cuba, and Jamaica. Why would the distributions for such countries have a different shape?

Regarding the diagnosticity effect, Ashby and Perrin's (1988) model can account for context effects on similarity judgments, in terms of how the presence of an additional stimulus C can modify the response region relevant in computing the similarity between two other stimuli, A and B. Specifically, the similarity between two stimuli A and C is $s(A, C) = \iint_{\mathbf{R_C}} f_A(x, y)dxdy$. Suppose that a third stimulus B is introduced near stimulus C. This means that the response region $\mathrm{R_C}$ is decreased, since a part of what used to be $\mathrm{R_C}$ is now the response region for $\mathrm{R_B}$. Therefore, the integral $s(A,C) = \iint_{\mathbf{New\ R_C}} f_A(x,y)dxdy$ sums probability weight over a smaller area and so $s(A,C)$ is reduced. But, such a reduction of similarity between A and C is predicted regardless of where exactly a stimulus intermediate to A and C is introduced, as long as it is in between (in psychological space) A and C, and so leads to a reduction in the response region for C. In other words, with such a scheme, a 'diagnosticity' effect can emerge for stimuli without a corresponding natural grouping of some stimuli in psychological space, in contrast with the intuition in Tversky's (1977) empirical demonstration.

A more general issue with general recognition theory is that it is not a theory best suited for dealing with conceptual stimuli (a limitation which Ashby & Perrin, 1988, themselves acknowledged). For example, the argument for asymmetry in the Korea-China example or the diagnosticity effect also assumes that the decision boundary between response regions is optimal. Perhaps such an assumption is valid for perceptual stimuli studied across multiple repetitions but it is questionable as to whether it applies for one-shot similarity judgments between previously unencountered object pairs (Ashby & Perrin, 1988, say that in one-shot cases additional assumptions can be made regarding the form of perceptual effects distribution). Moreover, the notion of confusability itself does not apply to most conceptual stimuli. Ashby and Perrin (1988) recognized this and provided a generalization to their similarity function, so that the overlap integral also includes a weighting term. Crucially, their generalized similarity function still implies that similarity depends on the proximity of the perceptual effects distributions, and this proximity is likely to be low for many conceptual object pairs (since it is rarely the case that we can confuse one real-world object for another, even for objects which are quite similar, such as apples and pears). Thus, the general recognition theory guides us to a prediction of universally low similarity in the case of pairs of conceptual objects.

Overall, the key strength of general recognition theory is that a researcher can produce predictions regarding the classification of simple, perceptual stimuli, on the basis of precise manipulations of the perceptual effects distribution for each stimulus. In such cases, the general recognition theory is probably the best of the available theories. However, applying this approach to the case of conceptual stimuli, such as the ones in Tversky's (1977) challenges, leads to difficulties.

We complete our short review by a (fairly obvious, we hope) qualification: our review was extremely selective, focusing primarily on the formal models, which have emerged as major candidates for explaining the key findings from Tversky (1977). It is important to bear in mind that there have been other, influential theoretical perspectives for these results, not based on theory specified in mathematical terms (especially in relation to asymmetries, e.g., Bowdle & Gentner, 1997; Bowdle & Medin, 2001; see also Gleitman et al., 1996). Moreover, there has been extensive work on various relevant methodological aspects of how asymmetries in similarity and the other relevant effects are demonstrated (e.g., Aguilar & Medin, 1999). Note, however, most researchers currently do accept the reality of most of these effects.

2. The Quantum Similarity Model (QSM)

Next, we present an alternative model for similarity judgments based on Quantum Probability (QP) theory (Pothos et al., 2013; Pothos & Trueblood, 2015). QP theory is a theory for how to assign probabilities to events (Hughes 1989; Isham 1989), alternative from classical probability theory. We call QP the rules for how to assign probabilities to events from quantum mechanics, without any of the physics. QP has the potential to be relevant in any area of science, where there is a need to formalize uncertainty. Regarding psychology, clearly, a major aspect of cognitive function is the encoding of uncertainty and therefore QP is potentially applicable in cognitive modeling. QP theory and classical probability theory are founded from different sets of axioms and so are subject to alternative constraints. The use of QP for modeling cognitive processes follows on from a number of recent attempts to describe various phenomena in psychology, and the social sciences more generally, using non-classical models of probability. Certain types of cognitive processing, in situations where it appears there may be incompatibility between the available options (Busemeyer et al, 2011), may be better modeled using QP theory.

The QSM follows the recent interest in the application of QP theory to cognitive modeling. Applications of QP theory have been presented in decision-making (White et al., 2014; Busemeyer, Wang, & Townsend, 2006; Busemeyer et al., 2011; Bordley, 1998; Lambert, Mogiliansky, Zamir, & Zwirn, 2009; Pothos & Busemeyer, 2009; Trueblood & Busemeyer, 2011; Yukalov & Sornette, 2010); conceptual combination (Aerts, 2009; Aerts & Gabora, 2005; Blutner, 2008); memory (Bruza, 2010; Bruza et al., 2009), and perception (Atmanspacher, Filk, & Romer, 2004). For a detailed study on the potential of using quantum modeling in cognition see Busemeyer and Bruza (2011) and Pothos and Busemeyer (2013).

A unique feature of the QSM is that, whereas previous models would equate objects with individual points or distributions of points, in the quantum model objects are entire subspaces of potentially very high dimensionality. This is an important generalization of geometric models of similarity, as it leads to a naturally asymmetric similarity measure. We first present an outline of the QSM and its main features. Subsequently, we consider again the violations of symmetry, triangle inequalities and the diagnosticity effect, from Tversky (1977), and how the QSM helps provide relevant explanations.

2.1. A new psychological space

Representations in QP theory are based on a multidimensional space. These representations are geometric ones, but such that the represented entities (stimuli, concepts, etc.) are not just single points in a geometric space, but rather entire subspaces. This provides a very natural approach to the problem of capturing differences in knowledge: the more you know about something (stimuli, concepts, etc.) the greater the dimensionality of the subspace for that entity. Thus, QT theory provides a unique, novel way to approach representation, that extends previous efforts both in psychology (Shepard, 1987) and generally (cf. Kintsch, 2014). Furthermore, the idea of an overlap between vectors and subspaces as a measure of similarity has a long history in psychology (Sloman, 1993); QP theory provides a more principled approach to this idea.

The QSM is based on a Hilbert space, which is a complex vector space (with some additional properties), that represents the space of possible thoughts. The overall space can be divided into (vector) subspaces representing particular concepts. Imagine a concept A. The subspace corresponding to this concept is associated with a projection operator P_A. Note that, in general, suitable spaces for modeling similarity judgments would be of very high dimensionality. However, in specific experimental situations, low dimensionality spaces usually provide adequate approximations.

In quantum models, in general, the current state of the system is given by a density operator ρ on H or, where simplifying conditions apply, a state vector. In psychological applications, including in the QSM, the state of the system corresponds to whatever a person is thinking at a particular time. More specifically, in the QSM, the relevant state is the mental state of a participant, just prior to a similarity judgment. Note, the state vector will often be at an angle to the various subspaces in the Hilbert space and it is determined by, for example, the experimental instructions; in other cases, the state vector may represent the expected degree of knowledge of participants. By projecting this current state onto the different subspaces of the relevant Hilbert space and then computing the squared length of the projected vector, we have a measure of the consistency between the state vector and the other entities represented in our quantum space. Below, (Figure 1) we will see a graphical illustration for how to compute these operations.

The QSM is a departure from classical geometric representation schemes. It offers a rigorous framework for associating concepts with subspaces and it provides us with representational flexibility, in that there are no constraints in the number of features one can employ for representing different concepts (within the same application, one can have subspaces varying greatly in dimensionality). Note, in a classical representational approach based on psychological spaces, each object must be represented with the same number of dimensions (all the available ones).

Consider next how to compute the similarity between two concepts in the QSM. The similarity between two concepts A and B is computed as $\mathrm{Sim}(\mathrm{A}, \mathrm{B}) = \mathrm{Tr}(\mathrm{P_B P_A \rho P_A})$, where ρ is the mixed knowledge state of the system; if the knowledge state is pure, the expression for similarity reduces to $\mathrm{Sim}(\mathrm{A}, \mathrm{B}) = |\mathrm{P_B P_A}|\psi\rangle|^2$. One of the important parts of the model is how to specify the current state of system or the knowledge state vector, as we called it before. We discuss this shortly.

We are going to follow the China-Korea example from Tversky(1977) to explain how the subspaces of the Hilbert space should be specified in our model. China would correspond to a subspace of the relevant knowledge space and Korea would correspond to another subspace. A subspace could be a ray spanned by a single vector, or a plane spanned by a pair of vectors, or a three dimensional space spanned by three vectors, etc. In this example, we represent China as a subspace spanned by two orthonormal vectors, $|\mathrm{v}_1\rangle$ and $|\mathrm{v}_2\rangle$, that is, the China subspace is two-dimensional and $|\mathrm{v}_1\rangle$ and $|\mathrm{v}_2\rangle$ are *basis* vectors for the China subspace. All the vectors of the form $\mathrm{a}|\mathrm{v}_1\rangle + \mathrm{b}|\mathrm{v}_2\rangle$, where $|\mathrm{a}|^2 + |\mathrm{b}|^2 = 1$ (as is required for a state vector in quantum theory) represent the concept of China. The concept of China itself is about lots of things. For example, when we think about China we think about culture, food, language, etc. To represent China as a subspace means that all these thoughts and properties, of the form $\mathrm{a}|\mathrm{v}_1\rangle + \mathrm{b}|\mathrm{v}_2\rangle$, are consistent with this concept and are contained within the China subspace. Here, we can see a key feature of the QSM, and at the same time, some commonalities with other models of psychological similarity, that is, that concepts correspond to regions of psychological spaces (Ashby & Perrin, 1988; Gärdenfors, 2000; Nosofsky, 1984). Further, imagine that we want to represent the idea that we have a greater knowledge for China than for Korea. We would represent China as a two dimensional space (we have a greater range of thoughts/properties/statements) and Korea as a one dimensional space.

Let us note that a thought of the form $|\psi\rangle = a|v_1\rangle + b|v_2\rangle$ is neither about $|v_1\rangle$ nor $|v_2\rangle$, but rather reflects the potentiality that the person will end up definitely thinking about $|v_1\rangle$ or $|v_2\rangle$. In QP theory we cannot assign definite meaning to superposition states such as $a|v_1\rangle + b|v_2\rangle$. This is a result of the Kochen-Specker theorem. If $|a| > |b|$, this means that the person has a greater potential to think of $|v_1\rangle$ than $|v_2\rangle$. The mathematical expression for the concept of China would be a projector denoted as $P_{China} = |v_1\rangle \langle v_1| + |v_2\rangle\langle v_2|$. Therefore, the mathematical expression of the collection of thoughts about China $|\psi\rangle$ is that $P_{China}|Thought\rangle = |Thought\rangle$; so, the collection of these vectors represents, in the QSM, the range of thoughts consistent or part of the concept. For example, if we think about Chinese food, then $|\psi\rangle = |ChineseFood\rangle$, and $P_{China}|ChineseFood\rangle = |ChineseFood\rangle$, showing that this is a thought included in the China concept. How are we to determine the set of appropriate vectors, properties, or dimensions, especially given that different subsets of properties of a particular concept are likely to correlate with each other? This is an issue common to all geometric approaches to similarity. Recent work, especially by Storms and collaborators (e.g., De Deyne et al., 2008), shows that this challenge can be overcome, for example, through the collection of similarity information across several concepts or feature elicitation. Then, the relatedness of the properties will determine the overall dimensionality of the concept.

In the next section we discuss how to compute similarity in our QSM.

2.2. Computing similarity

In QP theory, to examine the degree to which the state vector is consistent with the subspace we need to employ a projector. We need (1) a particular subspace, which is China in our case and (2) a suitable knowledge state vector (or, more generally, a density matrix). A projector can be represented by a matrix, which takes a vector and projects it (lays it down) onto a particular subspace. In other words, (2) has to be projected into (1). Let us illustrate this in Figure 1, where we can see how we project vector B onto vector A; note, both vectors are unit length. We represent in red the projection, which would be another vector that corresponds to the part of B which is contained in A.

Mathematically, this is denoted by $|A\rangle\langle A|B\rangle$, noting that $P_A = |A\rangle\langle A|$ is the projector onto the A ray. Indeed, the notation $|A\rangle\langle A|B\rangle$ indicates a multiplication between a vector and an inner product. But, from elementary geometry we have that the inner product between two real vectors is $\langle A|B\rangle = |A| \cdot |B| \cdot \cos\theta$, where θ is the angle between the two vectors (see also Sloman, 1933). If the two vectors are normalized, then $\langle A|B\rangle = \cos\theta$.

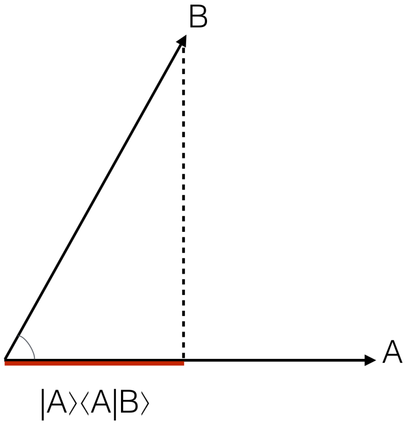


Figure 1: Illustration of the projection of vector B onto vector A.

Let us follow the same procedure following the China example above. The projector onto the China subspace is denoted by P_{China}. Then, to compute the part of the vector $|\psi\rangle$ that is contained in the China subspace we need to compute the projection $P_{\mathrm{China}}|\psi\rangle$. To compute the probability that the state vector is consistent with the corresponding subspace, we need to compute the length of the projection squared. The probability that a thought is consistent with the China concept equals $|P_{\mathrm{China}}|\psi\rangle|^2 = \langle\psi|\ P_{\mathrm{China}}|\psi\rangle$. If the state vector is orthogonal to a subspace, then the probability is 0. This can also be written as $\mathrm{p(China)} = \langle P_{\mathrm{China}}\rangle_\rho =$ $\mathrm{Tr}(P_{\mathrm{China}}\rho)$, if the initial state is a density matrix ρ, instead of a pure state $|\psi\rangle$. Thus, the probability that the initial knowledge state is consistent with the concept China is given as a measure of the overlap between the knowledge state and the subspace.

The QSM proposes that the similarity between two concepts is determined by the sequential projection from the subspace corresponding to the first concept to the one for the second concept. In other words, making a similarity judgment or comparison is a process of thinking about the first of the compared concepts, followed by the second. The similarity between Korea and China may therefore be written as, $\mathrm{Sim(Korea, China)} =$ $\mathrm{Tr}(P_{\mathrm{China}}P_{\mathrm{Korea}}\rho P_{\mathrm{Korea}})$ or $\mathrm{Sim(Korea, China)} = |P_{\mathrm{China}}P_{\mathrm{Korea}}|\psi\rangle|^2$, depending on whether the initial state is a density matrix ρ or a pure state $|\psi\rangle$.

2.3. Reproducing Asymmetries

As just noted, in the QSM, similarity between two concepts A and B is defined as $sim(A, B) = |\mathrm{P_B P_A}\psi|^2$, that is, a process of thinking about concept A first and concept B second. Critically, the term $|\mathrm{P_B P_A}\psi|^2$ depends on four factors. First, how the initial state is set. In the case of comparing two concepts, we think the most plausible assumption is that ψ is set so that it is neutral/unbiased, between A and B. Second, similarity judgments are often formulated in a directional way (Tversky, 1977). When this is the case, we suggest that the directionality of the similarity judgment determines the directionality of the sequential projection, i.e., the syntax of the similarity judgment matches the syntax of the quantum computation. Thus, there is a mechanism which potentially allows asymmetries in similarity judgments, when the projectors corresponding to the compared concepts do not commute (this will be the case, in general). Third, of course it depends on the angle between the subspaces. Finally, it depends on the relative dimensionality of the subspaces for concepts A and B (recall, greater dimensionality means greater knowledge).

In this section we are interested in how asymmetries can emerge from the QSM. We compare $|\mathrm{P_{Korea}P_{China}}\psi|^2$ and $|\mathrm{P_{China}P_{Korea}}\psi|^2$, noting that in both cases, the state vector is set so that it is neutral between the concepts compared, China and Korea. Note that $|\mathrm{P_{Korea}P_{China}}\psi|^2 = |\mathrm{P_{Korea}}\psi_{China}|^2|\mathrm{P_{China}}\psi|^2$ and $|\mathrm{P_{China}P_{Korea}}\psi|^2 = |\mathrm{P_{China}}\psi_{Korea}|^2|\mathrm{P_{Korea}}\psi|^2$, where $\psi_{China} = \mathrm{P_{China}}\psi/|\mathrm{P_{China}}\psi|$ and ψ_{Korea} are normalized (length=1) vectors in the corresponding subspaces. Note also that, by assumption, $|\mathrm{P_{Korea}}\psi|^2 = |\mathrm{P_{China}}\psi|^2$, which is the condition that the mental state vector is unbiased between the two concepts. Then, the similarity between Korea and China vs. China and Korea reduces to comparing $|\mathrm{P_{China}}\psi_{Korea}|^2$ (similarity of Korea to China) and $|\mathrm{P_{Korea}}\psi_{China}|^2$ (similarity of China to Korea). But, in the former case, we project a vector to a higher dimensionality subspace, than in the latter. Thus, in the former case, there is more opportunity, so to say, to preserve the vector's amplitude. Thus, in the former case, the projection will (on average) have greater length.

In Figure 2, we illustrate the relevant subspaces and projections. The green line corresponds to a one-dimensional subspace(Korea), the blue plane to a two-dimensional subspace(China), and the black line to the state vector (set in such a way that it is neutral between the two subspaces, as postulated by the QSM). The length of the first projection corresponds to a solid red line and, by assumption, is the same regardless of whether we project to the ray or onto the plane. But, the length of the second projection, illustrated as a yellow line, differs depending on whether it is to a ray or to a plane, so that when this second projection is onto the plane, it is longer. Panel (a) shows a process of thinking about Korea first and then China, that is, $\mathrm{P}_{\mathrm{China}}\mathrm{P}_{\mathrm{Korea}}$; $sim(Korea, China) = |\mathrm{P}_{\mathrm{China}}\mathrm{P}_{\mathrm{Korea}}\psi|^2$, which is the squared length of the yellow line. Analogously for panel (b). This illustrates how $|\mathrm{P}_{\mathrm{China}}\mathrm{P}_{\mathrm{Korea}}\psi|^2 > |\mathrm{P}_{\mathrm{Korea}}\mathrm{P}_{\mathrm{China}}\psi|^2$, that is, the square of the yellow line in panel (a) is greater than the square of the yellow line in panel (b).

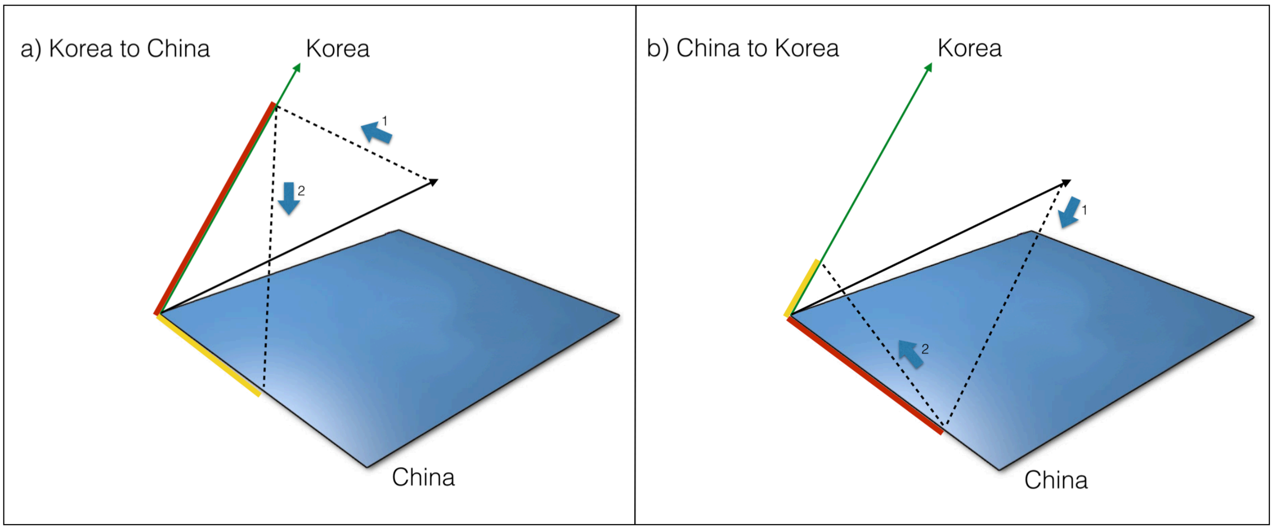


Figure 2: Illustration for how to compute Sim(Korea, China) and Sim(China, Korea) using the QSM.

As extensively discussed in Pothos et al. (2013), the QSM thus allows a prediction of asymmetry in the case of the Korea, China example (and, obviously, all cases where there is a difference in degree of knowledge) to emerge naturally. As we noted above, in order to generate these asymmetries we need some principle for fixing the initial state. Usually we will (partly) fix the initial knowledge state by demanding that it is unbiased, that is, that there is equal prior probability that the initial state is consistent with either, say, Korea or China. Such an assumption is analogous to that of a uniform prior in a Bayesian model. Then, it is straightforward to show that $\mathrm{Sim}(\mathrm{Korea}, \mathrm{China}) \sim |\mathrm{P}_{\mathrm{China}}\ |\psi_{\mathrm{Korea}}\rangle|^2$, whereby the vector $|\psi_{\mathrm{Korea}}\rangle$ is a normalized vector contained in the Korea subspace. Therefore, the quantity $|\mathrm{P}_{\mathrm{China}}\ |\psi_{\mathrm{Korea}}\rangle|^2$ depends on only two factors, the geometric relation between the China and the Korea subspaces and the relative dimensionality of the subspaces.

2.4. Reproducing violations of the triangle inequality

The QSM leads to violations of the triangle inequality in a way similar to how Tversky (1977) suggested such effects arise. As our representations are subspaces, different regions in the overall space end up reflecting the features characteristic of the corresponding concepts. We will follow another of Tversky's experiments as an example (Figure 3). We have a Hilbert space with Russia(in blue), Cuba(in red) and Jamaica(in green). All of Russia, Cuba, and Jamaica are represented as one dimensional subspaces, for simplicity. The region between Russia and Cuba will overall reflect the property of communism, noting that both countries are consistent with this property. Next, we can imagine a different region to the communist one containing Cuba and Jamaica. The shared characteristic of Cuba and Jamaica is their geographical proximity (they are both in the Caribbean), so this second region will likewise correspond to this property. It should be hopefully straightforward to then see how, if Cuba is on the boundary of the communism and Caribbean regions in psychological space, we can have Cuba highly similar to Russia (represented as (1) in dashed lines in Figure 3), Cuba highly similar to Jamaica(represented as (2)), but Russia and Jamaica dissimilar from each other(represented as (3)), thus violating the triangle inequality, i.e., producing $Dissimilarity(Russia, Cuba) + Dissimilarity(Cuba, Jamaica) < Dissimilarity(Russia, Jamaica)$.

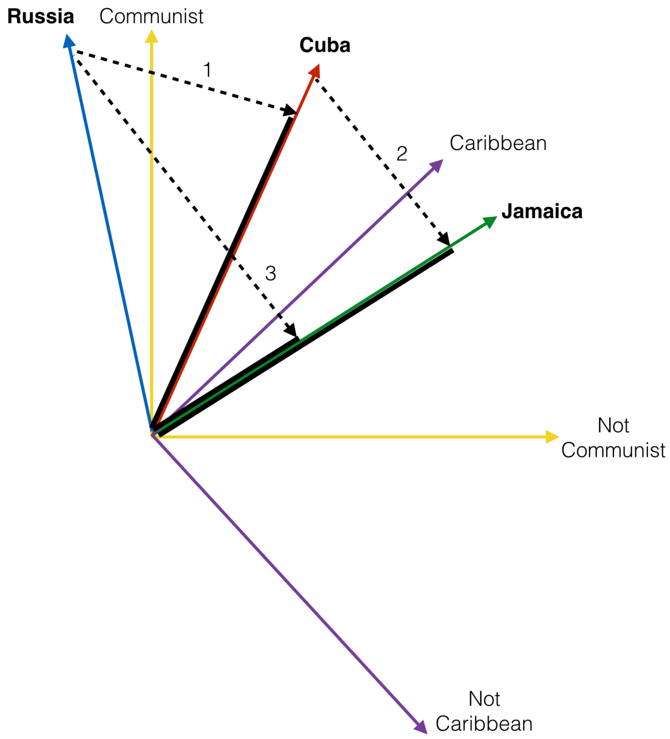


Figure 3: An illustration of how the QSM can accommodate Tversky's(1977) finding, which is often interpreted as a violation of the triangle inequality.

2.5. Reproducing Diagnosticity

The diagnosticity effect is central in the debate on whether distance-based similarity models are adequate or not and in this section we will show how the QSM can accommodate context when computing similarity judgments.

Sometimes what we think just prior to a comparison may be relevant to the comparison itself. Therefore, when computing the similarity of A and B we have to take into account the influence of some contextual information, C. As in all other computational examples we have seen, in the QSM C has to be represented by a subspace. Following Tversky's (1977) diagnosticity effect experiment, this information C could correspond to the alternatives in the task he employed. The similarity between A and B should then be computed as, $\mathrm{Sim}(A, B) = |P_B P_A|\psi'\rangle|^2 = |P_B|\psi'_A\rangle|^2 |P_A|\psi'\rangle|^2$, where $|\psi'\rangle = |\psi_C\rangle = P_C|\psi\rangle / |P_C|\psi\rangle|$ is no longer a state vector neutral between A and B, but rather one which reflects the influence of information C. If we minimally assume that the nature of this contextual influence is to think of C, prior to comparing A and B, then $\mathrm{Sim}(A, B) = |P_B P_A|\psi'\rangle|^2 = |P_B P_A(P_C|\psi\rangle)/(|P_C|\psi\rangle|)|^2 = |P_B P_A P_C|\psi\rangle|^2 / |P_C|\psi\rangle|^2$. In other words, if we first think about A and then about B when making a similarity comparison between A and B, then in the context of some other information C should involve an additional first step of first thinking about C. Computationally, we prefer to employ $|P_B P_A P_C|\psi\rangle|^2$, since $|P_B P_A P_C|\psi\rangle|^2 = |P_B P_A|\psi_C\rangle|^2 |P_C|\psi\rangle|^2 = |P_B|\psi_{AC}\rangle|^2 |P_A|\psi_C\rangle|^2 |P_C|\psi\rangle|^2$, where $|\psi_C\rangle = (P_C|\psi\rangle)/(|P_C|\psi\rangle|)$ and $|\psi_{AC}\rangle = (P_A|\psi_C\rangle)/(|P_A|\psi_C\rangle|)$. Therefore, the similarity comparison between A and B is now computed in relation to a vector which is no longer neutral, but contained within the C subspace. Depending on the relation between subspace C and subspaces A and B, contextual information can have a profound impact on a similarity judgment.

As we have done in the previous sections, we present an illustration (Figure 4) of how the diagnosticity effect arises from the QSM, using Tversky's (1977) example. As we explained in a previous section, in his experiment participants had to identify the country most similar to a particular target, from a set of alternatives, and the empirical results showed that pairwise comparisons were influenced by the available alternatives. Specifically, participants were asked to decide which country was most similar to Austria, amongst a set of candidate choices. When the alternatives were Sweden, Poland, and Hungary, most participants selected Sweden (49%), so implying that Sim(Sweden, Austria) was the highest (panel (a) in Figure 4) . When the alternatives where Sweden, Norway, and Hungary, Hungary was selected most frequently (60%), not Sweden (14%). Thus, changing the range of available alternatives can apparently radically change the similarity between the same two alternatives. Tversky's (1977) explanation for this result was that the range of alternatives led to the emergence of different diagnostic features (either 'Eastern European' countries or 'Scandinavian' countries), which in turn impacted on the similarity judgment. Analogous demonstrations were provided with schematic stimuli. Figure 4 shows a plausible QSM arrangement for Austria, Sweden, Poland, Norway and Hungary and the corresponding projections that lead to the diagnosticity effect.

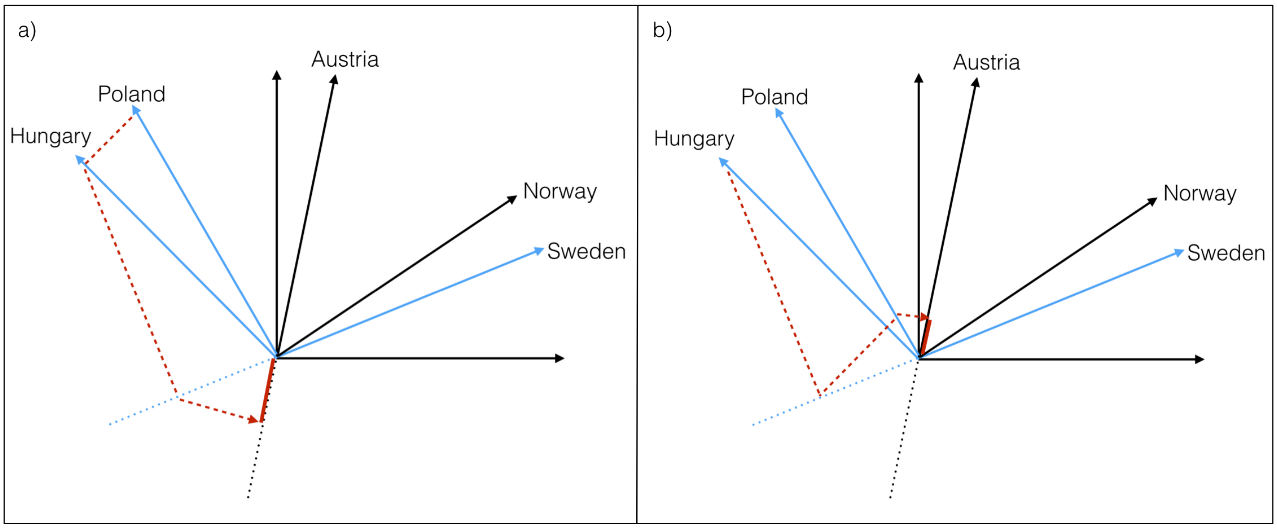


Figure 4: An illustration of how QSM can account for the results from Tversky's(1977) diagnosticity experiment. The order of projection on panel (a) is such that we start from the context elements, Poland (arbitrarily chosen first in the illustration), then Hungary, then Sweden, then Austria. If we were computing a similarity without context, we would just have a projection from Sweden to Austria, to correspond to the similarity of Sweden (first projection) to Austria. Analogously for panel (b). The key aspect of this illustration is that, when the 'grouped' elements (Poland, Hungary) are the context elements, the initial projections are such that not much amplitude is lost across successive projections (of the context elements; panel a). When the context elements are e.g. Hungary and Sweden, so that the similarity comparison concerns Poland and Austria (in the context of Hungary and Sweden), then the initial projections lead to a massive loss of amplitude, with the resulting similarity judgment being lower.

The QSM is able to reproduce the main empirical findings from Tversky's (1977) diagnosticity effect experiment and this approach also leads to qualitative predictions about when the effect is likely to be present or absent, based on the geometric relationships between the stimuli in psychological space. Nevertheless, regarding the emergence of the diagnosticity effect, the QSM involves a number of assumptions worth evaluating in detail. These assumptions concern mainly the way the context items influence the similarity judgment and the role of the initial knowledge state.

Let us first consider how the diagnosticity effect emerges in the QSM. As discussed above, context corresponds to successive projections between the context elements. When the context elements are grouped together (as for Hungary, Poland), projecting across them leads to little loss of amplitude of the state vector, so that the similarity judgment ends up being higher. When there is no grouping across any of the possible contexts, then the effect of context is simply to uniformly scale the similarity judgments. Thus, context can make the same similarity comparison appear higher or lower, depending exactly on the grouping of the context elements (Pothos et al., 2013). The intuition for how the quantum model produces the diagnosticity effect is thus not much different from that of Tversky's (1977). But, in Tversky's (1977) model, it has to be assumed that diagnostic features are 'invoked', as a result of grouping, while in the quantum model, the diagnosticity effect emerges directly from the presence of a grouping. In the next section, we address some challenges regarding how the QSM reproduces diagnosticity effects, the kind of novel predictions that the model can produce and alternative motivations for some of the QSM assumptions.

3. Conclusions, challenges and further directions

The objective of this chapter was to present the QSM and consider how it can account for Tversky's (1977) key challenges. The QSM generalizes the notion of geometric representations, but the emergent similarity metric is not distance-based, thus avoiding many of the criticisms Tversky (1977) made against distance-based similarity models. The QSM can be seen as an example of a new way of thinking about cognitive modeling, that may also be applied to constructive judgments (i.e. White et al., 2014), belief updating and many other analogous areas of research.

The QSM was developed to associate knowledge with subspaces. This idea of representations as subspaces allows us to capture the intuition that a concept is the span of all the thoughts produced by combinations of the basic features that form the basis for the concept. The QSM also helped us to cover some key empirical results: basic violations of symmetry, violations of the triangle inequality and the diagnosticity effect, all from Tversky (1977).

Nevertheless, we offer below a list of challenges for the QSM and open issues for further research (some of which we are in the process of addressing). It is important to establish whether the QSM model makes any novel predictions about similarity judgments in particular cases. These could either take the form of new qualitative effects or of quantitatively accurate predictions for similarity judgments. Our overall conclusion is that further work is clearly needed with the QSM, though the new results are encouraging for the overall potential of the approach.

3.1. Fixing the initial state

One problem with the QSM, as presented, is that it relies on a particular choice of initial state in order to reproduce the asymmetry/diagnosticity effects. Even in set-ups where one can partially fix the initial state by demanding it to be unbiased, this typically leaves some degrees of freedom unfixed (that is, this requirement does not always produce a unique state vector; there are equivalent neutral state vectors and it is unclear why one would prefer one option, as opposed to another). Further research is needed in terms of determining in a reliable way how to set the knowledge state vector for a participant or a group of participants. Moreover, we noted that the state vector could be affected by information relevant to the similarity judgment. In the diagnosticity effect example, there is a specific procedure for incorporating relevant effects, but we would like a more general scheme for how relevant prior information impacts on the state vector.

3.2. Interpreting the subspaces

More work is needed concerning the interpretation of the dimensions of the subspaces, which represent each concept (or stimulus, etc.). The dimensions of each subspace may correspond to the independent feature/characteristics, which collectively capture our knowledge of a concept. As an example, consider the standard Cartesian xyz coordinate system. There are many vectors which are in between the xyz coordinates. However, we can represent all this information, in terms of coordinates just along the three main axes (xyz) of the overall space. Likewise, when considering the subspace representing e.g. China, there are going to be many characteristics which highly correlate with each other. For example, our knowledge of Chinese art and culture relates to our knowledge of Chinese language etc. So, for a particular concept, we may have a greater or smaller number of individual features, but the extent to which the dimensionality of the corresponding subspace will be greater or smaller depends on the relatedness of the features. Regarding the emergence of asymmetries in similarity judgments, this is the main difference between Tversky's (1977) thinking and the QSM: the former predicts asymmetries in terms of the number of features, the latter as some function of the number of independent features. It is clearly desirable to empirically examine this difference in prediction.

3.3. Modeling context effects

One important challenge in developing the QSM is further formalizing the way in which contextual influences are taken into account. The idea of incorporating context as prior projections works well, but can the QSM be extended such that these prior projections can be motivated in a more rigorous way?

A great focus for further work with the QSM concerns the diagnosticity effect. This effect has proved difficult to replicate (e.g., see Evers & Lakens, 2014) and it would be interesting to see whether the QSM could generate any new predictions, regarding the emergence or suppression of the diagnosticity effect. We are interested in exploring whether the QSM model can provide insight into why the diagnosticity effect has proved elusive in its replicability.

The diagnosticity effect is also significant because the quantum formalism, overall, is often said to embody strong contextual influences. So, perhaps, quantum theory would be particularly suitable for modeling context effects in similarity judgments? The diagnosticity effect does emerge fairly naturally from the QSM, but the mechanisms that allow this are not the traditional contextual mechanisms in quantum theory (e.g., relating to entanglement or incompatibility). The difficulty lies in the fact that contextual influences in similarity appear to arise depending on the degree of grouping of some of the options in the relevant choice set. The QSM is sensitive to the grouping of the context elements, but there is still a challenge to embed the contextual mechanism in the QSM within a more rigorous, formal framework.

3.4. Dealing with frequency and prototypicality

An important gap in the QSM concerns how to deal with asymmetries arising from differences in the frequency of presentation of stimuli (Polk et al., 2002) or from differences in prototypicality (Rosch, 1975). This failure is interesting when we note that there appears to be an obvious way to include such effects. Presumably what distinguishes a prototypical stimulus from a non-prototypical one, or a stimulus presented many times from one presented only infrequently, is the increased potentiality for a participant to think about this stimulus. It would be interesting to see how the QSM could account for how differences in frequency/prototypicality can lead to asymmetries in similarity.

3.5. Analogical similarity judgments

Another important focus concerns so-called analogical similarity judgments (e.g., Gentner. 1983; Goldstone, 1994; Larkey & Love, 2003). Analogical similarity is a vast topic and here we focus on one aspect of it, namely the idea that, for example, when comparing two people, Jim and Jack, if they both have black hair, this will increase their similarity, but if Jim has black hair and Jack has black shoes (and blond hair), this will have less impact on their similarity. That is, work on analogical similarity recognizes that objects often consist of separate components. Commonalities on matching components (e.g., black hair) increase similarity more so than commonalities on mismatching components (e.g., black hair and black shoes). It is currently unclear whether there is a genuine distinction between cognitive processing corresponding to basic similarity tasks (as in Tversky, 1977) and analogical similarity ones (some researchers have suggested that different cognitive systems may mediate the two types of judgments; Casale et al., 2012). Nevertheless, there have been largely separate corresponding literatures for these two kinds of similarity judgments, with different objectives. We think that the QSM can be extended to incorporate analogical similarity, because quantum theory already has extensive machinery in place for combining individual components into a whole (cf. Smolensky, 1990). Indeed, we have been pursuing an approach based on tensor products (Pothos & Trueblood, 2015).

Chapter 2

Asymmetries: Theoretical and Empirical Progress

Abstract

Tversky's (1977) seminal demonstration of asymmetries in similarity judgments has been a consistent focus of theoretical work in similarity. His explanation for such asymmetries was that degree of knowledge, formalized as numbers of features, drove salience, which in turn drove asymmetries in similarity judgments. We reviewed in *Chapter 1* (see also Pothos et al. 2013) a novel model of similarity, based on the mathematics of quantum theory. This similarity model predicts that asymmetries do not arise from the number of features, but rather from the number of 'independent dimensions', which represent a concept. Using the similarity database of de Deyne et al. (2009) and with another set of stimuli (countries) we provide an experimental demonstration of this idea (5 experiments). The results are inconclusive regarding the quantum similarity model (QSM) but, moreover, cast doubt on the validity of either the asymmetry effect or the De Deyne et al. stimuli, at least as we utilized them in the present experiments.

Introduction

Similarity judgments, as we pointed out in *Chapter 1*, play a central role in many areas of psychology (Goldstone, 1994; Pothos, 2005, Sloman & Rips, 1998) and it is hardly surprising that intense effort has been directed towards elucidating the relevant formal principles. A focal point in this effort is Tversky's (1977) highly influential work, in which he reported several general properties of similarity judgments.

One of Tversky's (1977) objectives was to evaluate (and eventually criticize) the dominant, distance-based approaches to similarity, according to which similarity is a (simple) function of distance in a psychological space. If such a conceptualization of similarity were correct, we would expect similarity judgments to be consistent with the metric axioms (general properties that all distances must obey). Instead, Tversky (1977) reported violations of all metric axioms. He reported violations of minimality (when identical objects are not always judged to be maximally similar); of symmetry (the similarity between A to B can be different from that of B to A); and the triangle inequality (the distance between two points cannot exceed the sum of their distances to any third point). Tversky (1977) also identified a context effect in similarity judgments, called the diagnosticity effect, such that the similarity between the same two stimuli could be affected by the properties of whichever other stimuli were present (and broadly relevant to) the similarity comparison.

It is worth briefly reviewing that researchers have attempted to reconcile distance-based similarity models with Tversky's (1977) challenges (this point is also discussed in *Chapter 1*). Asymmetries in similarity could be accommodated if one modifies the similarity function from $\mathrm{sim}(\mathrm{A},\mathrm{B}) = \mathrm{f}(\mathrm{dist}(\mathrm{A},\mathrm{B}))$ to $\mathrm{sim}(\mathrm{A},\mathrm{B}) = \mathrm{p}_{\mathrm{AB}}\mathrm{f}(\mathrm{dist}(\mathrm{A},\mathrm{B}))$, where f is a function transforming distance to similarity, dist is distance, sim is similarity, and p_{AB} is a 'directionality' parameter, such that it can be different, depending on whether the comparison involves stimulus A with B or B with A (Nosofsky, 1991). Such efforts (especially regarding asymmetries) are not universally considered satisfactory and researchers have sought approaches to similarity from which Tversky's (1977) main findings can emerge (more) naturally (Ashby & Perrin, 1988; Krumhansl, 1997).

An important challenge in such theoretical efforts is not only the coverage of Tversky's results (and extensions, e.g., Aguilar & Medin, 1999; Polk et al., 2002) but also their generative value, in terms of extending the scope of empirical prediction. In this vein, the focus of the present chapter concerns whether our own proposal to account for Tversky's (1977) findings, the quantum similarity model (QSM), can motivate novel empirical directions, specifically in relation to violations of symmetry. Violations of symmetry are particularly significant in similarity research, because they run against the grain of basic intuition. How can it be the case that, for example, the similarity between a dog and a cat can be different from that between a cat and a dog? Yet, Tversky (1977) demonstrated exactly this, when using countries as stimuli, such as (North) Korea and (Red) China. Retrospectively, naïve observers agree that there is something more natural about the statement 'Korea is like China' compared to 'China is like Korea' (forced choice between such statements was Tversky's, 1977, task).

Even though we have already briefly covered Tversky's proposal in *Chapter 1*, we will first consider again Tversky's (1977) explanation for how violations of symmetries arise, not least because this explanation is still arguably current (and so as to make chapters reasonably self-contained). Then, we can extract the key relevant implication from his model and contrast it with a corresponding one from the QSM.

Tversky's (1977) Account and the QSM

Tversky's (1977) contrast model of similarity is that

$$\text{Similarity}(A,B) = \theta f(A \cap B)\text{-}af\big(A\text{-}B\big)\text{-}\beta f(B\text{-}A),$$

where θ, a, β are parameters, $A \cap B$ denotes the common features between A and B, A – B the features of A which B does not have, and B – A the features of B which A does not have (see also Bush & Mosteller, 1951; Eisler & Ekman, 1959). Such a scheme can predict violations of symmetry if A and B have a different number of features and the parameters a, β are different from each other and suitably set (e.g., $\theta > 0, a = 1, \beta = 0$). For example, regarding the Korea-China example, Tversky assumed that China has more features than Korea, because his average participant would know more about China than Korea. Therefore, the similarity between Korea and China would be fairly high (no contribution from $af(A - B)$, since Korea has no, or very few, features which China does not have). However, in comparing China and Korea, there is now a contribution from $af(A - B)$, which lowers the overall similarity result. Thus, according to Tversky's similarity model, the similarity of China to Korea is predicted to be less than that of Korea to China. Tversky's model of similarity is appealing, but still involves two parameters, which must have appropriate values before accounting for violations of symmetry. If instead of setting $a = 1, \beta = 0$, we e.g. set $a = 0$ and $\beta = 1$, the model fails to predict the right direction for symmetry violation in the example above.

Note that Tversky (1977) did consider this issue of parameterization carefully and provided a rationale for setting the parameters in the appropriate way (relating to which of the two constituents in a similarity statement correspond to the subject and the referent). It is beyond the scope of this work to consider whether Tversky's (1977) parameterization scheme is justifiable or not. Rather, we scrutinize his key assumption for how asymmetries arise. According to Tversky (1977), higher similarities will be observed when less salient stimuli are compared to more salient ones. His analysis (e.g., see p. 332 in his paper) concludes that the number of features determines salience – the more the features, the greater the degree of salience of the corresponding object. Note that linking asymmetries to differences in numbers of features ties in well with interpretations of the origin of asymmetries as serving communication principles. For example, arguably 'Korea is like China' makes more sense than 'China is like Korea', because in the former case we can understand the concept we know less about (Korea) in terms of the features of the concept we know more about (China; cf. Bowdle & Gentner, 1997).

The explanation for asymmetries from the QSM is similar to that of Tversky (1977). Our development of the QSM was, at least originally, intended as a refinement/ formalization of Tversky's (1977) key ideas. But, there is an important difference (regarding asymmetries), which leads to an interesting prediction. The crucial aspect of the present work concerns the interpretation of the dimensions of the subspaces, which represent each concept (or stimulus etc.). The dimensions of each subspace have to correspond to the independent feature/ characteristics, which collectively capture our knowledge of a concept. As an example, consider the standard Cartesian xyz coordinate system. There are many vectors which are in between the xyz coordinates. However, we can represent all these vectors, in terms of coordinates just along the three main axes (xyz) of the overall space. Likewise, when considering the subspace representing e.g. China from the Korea-China example provided in *Chapter 1*, there are going to be many characteristics which highly correlate with each other. For example, our knowledge of Chinese art and culture relates to our knowledge of Chinese language etc. So, for a particular concept, we may have a greater or smaller number of individual features, but the extent to which the dimensionality of the corresponding subspace will be greater or smaller depends on the relatedness of the features. Regarding the emergence of asymmetries in similarity judgments, this is the main difference between Tversky's (1977) model and the QSM: the former predicts asymmetries in terms of the number of features, the latter as some function of the number of independent features.

The empirical challenge is to disentangle these two predictions. We needed a database for representation information, in terms of features, of a large number of objects. De Deyne et al. (2008) carried out the largest exercise of this sort to date, to our knowledge. Presently relevant is the fact that they considered several categories of objects (e.g., clothes) and within each category several objects (they called the results Type I Exemplar Feature Matrices). Then, four judges were asked to consider for each object the applicability of a very large number of features (these were determined after a feature generation task in an experiment where 1003 participants were asked to write down, preferably, 10 different features for 6 up to 10 different stimulus words, without a time limit). For example, for clothes, there were overall 258 possible features, so that a rating between 0 and 4 indicated the applicability of each feature to each object. For clothes, there were 29 exemplars considered (e.g., bra, blouse). Thus, for the category of clothes, we have information for 29 exemplars, so that each exemplar is represented by a feature vector consisting of 258 features.

So, the number of features relevant to the category of clothes is 258 (cf. Tversky's, 1977, account for how asymmetries emerge). But, through some measure of feature relatedness or data reduction, we can compute the number of independent features, for the category of clothes (cf. the QSM prediction for asymmetries). These considerations were the basis for our empirical test in experiments 1, 2, 3 and 4 (more details about the stimuli will be provided in the following sections).

Pilot Study

We intended to employ de Deyne et al.'s (2008) general categories as stimuli, but, we were concerned that, in some cases, similarity ratings may be subject to floor effects (e.g., what is the similarity between fish and professions?). Therefore, the objective of the Pilot Study was to identify pairs of stimuli, such that their similarity would not be too low. An additional objective was to include at least some category pairs where the stimuli differed by a large amount on some measure of feature relatedness. At the time of running the pilot, we were fairly agnostic regarding the most suitable measure of feature relatedness. So, we opted for computing the average of the correlations of all possible, unique exemplar pairs, within each category (each exemplar was represented as a vector, with values along all the possible features). This average correlation can be taken to be a (fairly simplistic, but nonetheless approximately valid) measure of feature relatedness, within each category. For example, in the clothes category, there are 29 exemplars and so 406 unique exemplar pairs. We computed 406 correlations and averaged them, to obtain a measure of feature relatedness for the clothes category. This process was repeated for each of the 15 stimuli in the Type I Exemplar Matrices.

Participants

Thirteen experimentally naïve students at City University London received course credit for taking part in the study.

Materials, Procedure, and Results

The experiment was computer-based and implemented in Excel; it lasted 15 minutes. First, the 15 stimuli in de Deyne et al.'s (2008) Type I Exemplar Feature Matrices were initially presented for a generality task. For each category, we tested the perceived generality, that is, how general or specific each category was on a scale from 1 to 9, where 1 represents a concept as specific as possible and 9 a concept as general as possible. We considered generality as another factor which may impact on asymmetries, and this aspect of the pilot allowed some preliminary impression of the differences in generality, amongst the stimuli employed. Then, participants were asked to provide similarity ratings for 105 pairs of stimuli, using a rating scale, with anchors 1 (not similar) and 9 (very similar).

We selected all pairs of stimuli, for which the average similarity rating was above 4.5. Then, we ordered these pairs in terms of average feature correlation difference, and selected for the main experiments the top 10 pairs with the highest difference (this was done in a purely exploratory way: since feature difference was, approximately speaking, the basis for predicting asymmetries from the quantum similarity model, we wanted to develop a design that was broadly sensitive to this prediction; note, in other experiments in this series, the design considerations of the quantum model were more carefully balanced against those from alternative models). From these, we had to further cull two pairs involving Professions, as this category was over-represented in the set and there is some indication that absolute frequency may impact on similarity ratings. Finally, we selected filler category pairs, to ensure that the frequency of each category across the main experiments would be the same (cf. Polk et al., 2002). These were Clothes-Tools, Clothes-Insects, Fish-Reptiles, Insects-Fish, Musical Instruments-Fish, Musical Instruments-Reptiles, Tools-Weapons, Tools-Musical Instruments, Weapons-Insects, Weapons-Reptiles.

Category 1	**Category 2**	**Av. Sim**	**Abs. Diff. in feat correlation**
Clothes	Mammals	5.1	|0.69-0.57|=0.12
Mammals	Reptiles	5.2	|0.57-0.71|=0.14
Fish	Mammals	5.9	|0.73-0.57|=0.16
Musical Instruments	Professions	7.1	|0.63-0.45|=0.18
Insects	Mammals	4.8	|0.75-0.57|=0.18
Professions	Weapons	5.0	|0.45-0.65|=0.20
Professions	Tools	5.9	|0.45-0.68|=0.23
Clothes	Professions	6.2	|0.69-0.45|=0.24

Table 1: List of categories used as experimental stimuli.

Experiment 1

The objective of both Experiments 1 and 2 was to collect data, using the stimulus pairs identified in the pilot, so as to assess QSM's main prediction regarding asymmetries. The stimulus pairs differed in number of features and absolute difference in feature correlation, so providing a dissociation between the prediction regarding the origin of asymmetries according to Tversky (1977) (number of features) and the prediction from the QSM (number of independent features, broadly conceptualized in terms of these averaged correlations).

Participants

A total of 29 experimental naïve participants at City University London were recruited and received £3 for doing the task.

Materials, Procedure and Results

The experiment, designed in SuperLab, lasted about 15 minutes. Participants saw a single screen of instructions, informing them that they would have to see pairs of 'categories of everyday objects' and that they would have to rate their similarity, on a 1 to 9 scale. Then, each of 18 trials in a first block of trials involved presenting on a computer screen both stimuli simultaneously and the rating scale (in a way that reminded participants of the anchors), with a prompt for participants to respond. Participants could view the category pair for as long as they needed to respond and, once a response was provided, they proceeded onto the next trial. The second block of 18 trials was identical to the first, but for the fact that the order of stimuli in each pair was reversed. For example, if in the first block participants were asked to rate the similarity between A and B, in the second one they would be asked to rate the similarity between B and A. Two versions of this experiment were run, counterbalancing the order of blocks. Trial order within each block was randomized.

First, we ran a reliability test concerning any findings of asymmetry in the similarity ratings. We randomly divided participants into two groups and computed the effect size (for asymmetry in a given direction, for each category pair). Then, we correlated these effects sizes across the two subsets in the sample, r(8) = .78, p= .01 (recall, there were eight critical pairs of stimuli). Given this result, a more details reliability analysis was not perceived necessary.

The main dependent variable in the study concerned differences in similarity between A and B and between B and A, for the eight category pairs (we call this variable asymmetry difference or just asymmetry, in what follows). We considered the following independent variables as predictors of asymmetry across category pairs. First, we considered differences in the generality of each category in a pair, that is, when considering category pair A, B, the difference in the average generality ratings for category A with that of category B. Second, we computed the difference in numbers of features for each category pair. This variable was based on the number of features listed for each category by de Deyne et al. (2008). For example, concerning the clothes - professions category pair, the feature difference variable was -112, indicating that there were 112 more features in the professions category, compared to the clothes one. Third, we used the average correlations across all exemplar pairs in each category to compute a corresponding average correlation difference variable. This variable was meant to approximately correspond to the QSM prediction, in that it concerned differences in the relatedness amongst features for the stimuli. Fourth, for each category, we ran a Principal Components Analysis, to identify the number of principal components which capture 90% of the variance in feature ratings, across the category exemplars. The principal components most closely correspond to subspace dimensions, implicated in the QSM. We then computed a difference in number of principal components, for the stimuli in each pair. Note, considering instead 80% of the variance leads to a nearly identical variable (r=.99). Note also that, to carry out the Principal Components Analysis, we had to reject features which were identically represented across all exemplars in a category (only a handful of features were rejected in this way, for each category). Table 2 shows all the variables.

Category pair	Asym Diff Exp.1	Propor. Exp.2	Gen Diff	Featu. Diff	Corr Diff	PCA Diff
Clothes Professions	.31	.52	-1.3	-112	.24	-3
Clothes Mammals	-.24	.72	-.31	-30	.12	-2
Fish Mammals	.31	.76	-1.5	-132	.16	-9
Insects Mammals	-.79	.79	-.55	-74	.18	-4
Mammals Reptiles	-1.00	.21	1.1	109	-.14	12
Musical Instr. Professions	.31	.59	-1.8	-152	.18	-6
Professions Tools	.17	.64	.00	85	-.23	2
Professions Weapons	.55	.48	1.14	189	-.20	9
Correlation of IV with the DV, Exp. 1			-.34, ns	-.10, ns	.02, ns	-.30, ns
Correlation of IV with the DV, Exp. 2			-.57, ns	-.53, ns	.49, ns	-.80, sig.

Table 2: Empirical results and predictors, in Experiments 1, 2; the significance level was .05.

None of the independent variables predicted similarity asymmetries. Note, we carried out an items-based analysis. But, this is a potential issue only for the predictor relating to category generality (since we collected category generality ratings from each participant); all the other predictors, are identical (for each pair) across participants. In any case, running a within participants regression analysis (Lorch and Meyers, 1990) did not push the generality predictor to significance.

Experiment 2

Similarity ratings are commonly employed in psychology, but, in the case of studying similarity asymmetries, they may not constitute the most appropriate procedure. The directionality of a similarity comparison may not be so obvious, with similarity ratings. Moreover, given that any asymmetries are likely not to be large, a problem with similarity ratings are scaling issues (both between and within participants, since even the same participant may adjust his/her ratings, between start and end of the task). Experiment 2 is identical to Experiment 1, but for the use of a forced choice similarity task, analogous to that of Tversky (1977).

Participants

We recruited 30 experimentally naïve participants, all at City University London, who received £3 for their time.

Materials, Procedure, and Results

The experiment, designed in SuperLab, lasted about 15 minutes. Participants were simply told that they would see pairs of statements about categories of everyday objects and that they would have to decide which one they prefer. Each of 18 trials involved participants choosing between two simultaneously presented statements, with the general structure "Category 1 is similar to Category 2" vs. "Category 2 is similar to Category 1".

An index of similarity asymmetry is readily provided by computing the average proportion of selecting a particular statement (Table 1). As in Experiment 1, we first ran a reliability analysis on the proportion selected variable, which revealed the variable to be (marginally) reliable, r(8) = .59, p= .059. Regarding the problem of predicting asymmetries, in this case, it can be seen that the variable corresponding to the difference in principal components correlated significantly with the dependent variable (note, as in Experiment 1, employing a within participants logit regression, did not salvage the generality predictor). It is worth ascertaining that the obtained result is in a direction consistent with the prediction of the QSM. The dependent variable was the proportion of times the statement "A is similar to B" was preferred to the reverse statement and it correlated negatively with the variable difference in principal components for item A minus item B. This means that the higher the preference for the "A is similar to B" statement, the more the principal components for category B, than A. Put differently, when category B has more principal components, than category A, then the statement "A is similar to B" is preferred. This is indeed the prediction of the QSM.

Experiment 3

The objective of this experiment was twofold. First, we wanted to test whether when PCA is controlled for, no asymmetries in similarities arise as a result of differences in features (Hypothesis 1). And second, we wanted to test whether when features are controlled for, asymmetries in similarities do arise, as a result of differences in PCA (Hypothesis 2).

Participants

We recruited 60 experimentally naïve participants, all at City University London, who received course credits for their time.

Materials, Procedure, and Results

We designed a paper-based experiment to run in a lecture class; the experiment lasted about 15 minutes. Participants were simply told that they would see pairs of statements about categories of everyday objects and that they would have to decide which one they prefer. Participants had to choose between two simultaneously presented statements, with the structure of a forced choice task: "Category 1 is similar to Category 2" vs. "Category 2 is similar to Category 1". They had to circle the left arrow if they preferred the first statement or circle the arrow on the right if they preferred the second one.

For Hypothesis 1, we wanted two sets of category pairs. For both sets, category pairs would not differ in terms of PCA. That is, the category pairs in the first set would (on average) have the exactly the same overall PCA difference as the category pairs in the second set (so, there is no opportunity for asymmetries to arise as a result of differences in PCA). For SET ZERO, we would have pairs that, in addition, do not differ in terms of numbers of features. For SET HIGH, by contrast, we would have pairs that differ in terms of numbers of features.

For Hypothesis 2, again we have two sets of category pairs. For both sets, category pairs would not differ in terms of features. So, the average feature difference amongst stimuli in each pair, in the first set would be the same as the average feature difference amongst stimuli in each pair, in the second set. In addition, in SET ZERO category pairs would not differ in terms of PCA as well. But, in SET HIGH category pairs would differ in terms of PCA.

In practice, the way this was achieved was as follows. First, we identified the category pairs for Hypothesis 1. We ordered all category pairs in terms of PCA difference (ensuring that as many differences as possible were positive, by switching the order of the stimuli in each pair).

For Hypothesis 2, we followed an analogous procedure. We ordered all category pairs in terms of features difference. Here, it was rarely the case that there were category pairs with exactly the same features difference. But, there were category pairs with similar feature differences. We therefore aimed to identify category pairs for SET ZERO and SET HIGH, such that they had similar values for feature differences, but one would have a low PCA difference value (SET ZERO), while the other would have a high PCA difference value (SET HIGH).

In this way, for Hypothesis 1 we identified 24 category pairs for each of SET ZERO and SET HIGH and, for Hypothesis 2, 18 category pairs for each of SET ZERO and SET HIGH. Note, some of these pairs were ones that were identified as leading to very low similarity ratings in the pilot study. However, as we planned to use a forced choice similarity task in this iteration of the experiment, we thought this would not cause any problem.

For the final selection of stimuli, we used 15 members in each set, for each hypothesis, that is 30 category pairs for Hypothesis 1 and 30 category pairs for Hypothesis 2.

For Hypothesis 1, we ended up with category pairs for SET ZERO and SET HIGH, such that the average difference in feature differences was 115 features (the range was from 94 to 163). That is, the category pairs in SET HIGH differed in terms of features by 115, more so than the category pairs in SET ZERO. The category pairs in SET ZERO and SET HIGH were exactly matched in terms of PCA differences.

For Hypothesis 2, we ended up with category pairs for SET ZERO and SET HIGH, such that they differed in terms of feature differences by 2.2 features (the range was from 4 to 10). That is, the category pairs in SET HIGH had, on average, a difference of 2.2 features, more so than category pairs in SET ZERO. This shows that the category pairs were well balanced across the two sets, in terms of feature differences. Importantly, the category pairs in SET HIGH differed, on average, by 7.2 dimensions, compared to category pairs in SET ZERO.

We tested the material set for Hypothesis 1 in half of our sample (30 participants) and the material for Hypothesis 2 in the other half. We also introduced five filler statements and a control statement to test that participants were paying attention to the task. This control statement was meant to have a very establish effect in similarity (i.e. “Birds are similar to robins” vs. “Robins are similar to birds”).

First, our results showed that in our first group (where we tested Hypothesis 1), the percentage of responses going in the predicted direction was 49.59% ± 4.4% of confidence interval. Second, they showed that in our second group (where we tested Hypothesis 2), the percentage going in the predicted direction was 49.54% ± 4.52% of confidence interval. That is, there was no evidence in support of either hypothesis. In other words, we could not conclude that when PCA is controlled for, asymmetries in similarities arise as a result of differences in features; and that when features are controlled for, asymmetries in similarities arise, as a result of differences in PCA.

Experiment 4

Given that the design in Experiment 3 did not produce useful findings, we sought an alternative way to dissociate the two possible predictors of asymmetries, difference in features and difference in PCA. First, we ran a simple simulation randomly selecting subsets of 15 pairs of stimuli, with the view to identify the 15 pairs with the lowest correlation between the two predictors possible. Then, we ran an experiment with the aim to test the QSM prediction, that is, that the statement “Category 1 is similar to Category 2” should be more likely to be preferred, if Category 1 had more PCA dimensions than Category 2 (we employed a forced-choice task to test for similarity asymmetries). Specifically, the aim of this experiment was to test for the presence of a negative correlation between the proportion of statements selected by participants in a forced-choice task and the variable PCA difference. Note, the QSM hypothesis is that when PCA_Category1 - PCA_Category2 is low (or negative), the statement “Category 1 is similar to Category 2” should be selected more often. The experiment was also designed to test for whether similarity asymmetries can emerge from differences in just features, as one would predict following Tversky (1977). The features hypothesis would be supported with a negative correlation between Features_Category1 – Features_Category2 and the proportion of preference for the statement “Category 1 is similar to Category 2”.

Participants

A total of 300 participants, all of them US residents, were recruited online and received $0.90 for doing the task.

Materials, Procedure, and Results

The study designed in Qualtrics and run on Amazon Mechanical Turk, lasted about 15 minutes. The list of pairs we presented in the task is shown in Table 3. We picked 15 appropriate pairs after identifying the lowest correlation (0.24) for PCA vs. Features difference (we designed a Matlab script to accomplish that[2]). Since we had a small number of items, we included 5 items that were catch questions (in order to exclude participants who were not paying attention to the task). The catch questions were statements that would fit in the rest of the task, but were such that one statement was obviously correct (i.e. New York is a city in Europe vs. New York is a city in North America).

Category Pairs	
Birds	Weapon
Insects	Music
Insects	Reptile
Kitchen Utensils	Mammal
Kitchen Utensils	Vehicle
Mammal	Vehicle
Mammal	Vegetable
Mammal	Tool
Music	Weapon
Profession	Sport
Profession	Vehicle
Reptile	Weapons

[2] The function was based on a simple, undirected search for a pair selection, such that the correlation between the two predictors was the lowest possible. It worked by randomly identifying a candidate selection of pairs, noting the correlation between the predictors, and comparing to the lowest correlation identified by that point in the programme function.

Sport	Vegetable
Sport	Tool
Tool	Vehicle

Table 3: Pairs of stimuli used in experiment 4.

The procedure was based on a forced-choice task, with a double presentation of the 15 different pairs; we counterbalanced the presentation order of the two statements in each trial (i.e. in the trial “Birds are similar to Weapons” vs “Weapons are similar to Birds”, the first comparison was presented on the left side of the screen in one trial and the same comparison was presented on the right in another one). Trials corresponding to each of the 15 pairs were presented in a randomized order. Participants were simply told that they would see pairs of statements about categories of everyday objects and that they would have to decide which one they prefer. Participants had to choose between two simultaneously presented statements, with the general structure “Category 1 is similar to Category 2” vs. “Category 2 is similar to Category 1”.

In the analysis of the results, we first eliminated participants who did not answer correctly 2 or more of the catch questions presented (7 participants in total). We then, computed the PCA difference for each of the pairs of stimuli (i.e. PCABirds - PCAWeapon) and Features difference (i.e. FeaturesBrids - FeaturesWeapon). Finally, we computed the proportion of times each statement (i.e. “Birds is similar to Weapon”) was preferred over the reverse. A negative correlation between the proportion of statements selected and the variable PCA, would support the QSM hypothesis (since when the difference PCA_Category1 - PCA_Category2 is low or negative, the statement “Category 1 is similar to Category 2” should be selected more often, and analogously for the features hypothesis).

We conducted an items-based analysis (N=15) and overall, we got a positive relationship between PCA difference and asymmetry (the proportion of times “Category 1 is similar to Category 2” was preferred to “Category 2 is similar to Category 1”). That is, the statement ‘Category 1 is similar to Category 2’ was preferred more, if Category 1 had more distinct dimensions (see Figure 1). Note, the correlation was significant only if we employed a non-parametric correlation coefficient, such as Kendall’s tau, $r_\tau = .044$, p<.05 for PCA difference and $r_\tau = -.029$, p<.05 for Features difference. A non-parametric correlation is potentially justifiable if one accepts that differences in PCA do not map to differences in similarity in a straightforward way (which is plausible). Nevertheless, it is readily acknowledged that the use of a non-parametric correlation procedure here is somewhat post hoc.

Our results showed that the correlation between asymmetries and PCA difference was the exact opposite of what we expected. In other words, we did not observe the predicted effect and we could not conclude that the statement "Category 1 is similar to Category 2" was more likely to be preferred, if Category 1 had more PCA dimensions than Category 2.

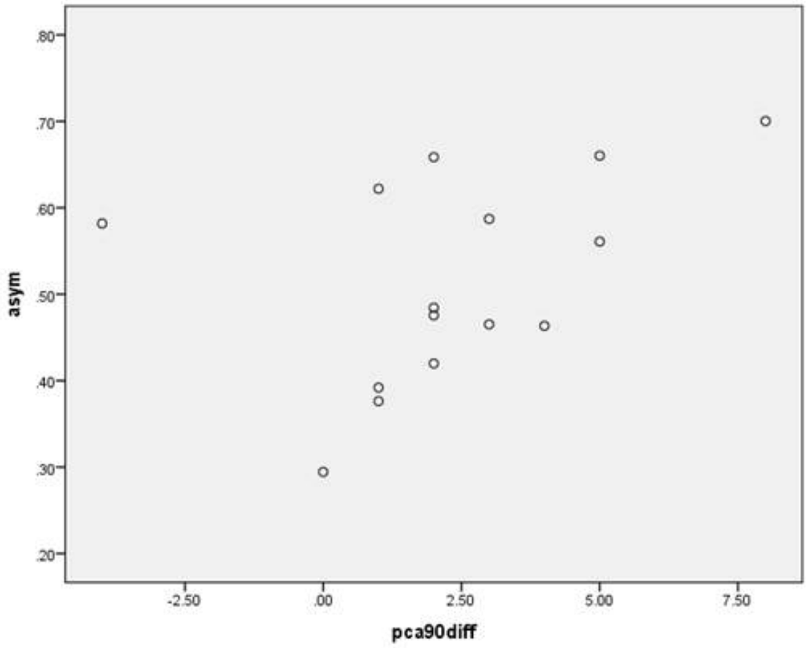


Figure 1: Relationship between PCA difference and Asymmetry in experiment 4.

Experiment 5

Given the mixed results of the previous two experiments regarding both the hypothesis that it is feature differences which drive asymmetries (we consider this hypothesis the most straightforward implication from Tversky's, 1977, model) and the hypothesis that it is PCA differences (which is the implication from the QSM), in this final experiment we sought to take a step back and examine the basic evidence for asymmetries in relation to features. We abandoned the de Deyne et al. (2008) stimuli and instead opted for straightforward countries stimuli (e.g., similarity (Greece, Albania) etc.). In this occasion, we measured directly a variable which would correlate with the degree of knowledge participants had for each of the items (countries) presented and tested the hypothesis that the number of preferred statements of the form “Country 1 is similar to Country 2” ought to be higher, where the difference in knowledge Country 1- Country 2 is more negative.

Participants

A total of 300 participants, all of them US residents, were recruited online and received $0.90 for doing the task.

Materials, Procedure, and Results

The study designed in Qualtrics and ran on Amazon Mechanical Turk, lasted about 15 minutes. We specified 20 pairs of countries (40 countries, all unique), and we presented a forced-choice similarity task, with 5 different pairs (10 countries) in a randomized order. Participants were simply told that they would see pairs of statements about countries and that they would have to decide which one they prefer. Participants had to choose between two simultaneously presented statements, with the general structure “Country 1 is similar to Country” vs. “Country 2 is similar to Country 1”. Then, after the similarity ratings (counterbalanced), participants spent a minute per country (a timer appeared on the screen and after 60 seconds the experiment jumped into the following screen) to list as many facts about the countries they could think of. We specified in the instructions that “It does not matter if you know much or little about a country; it is really important for our study to get an idea of how much you can think about a country spontaneously (i.e., without e.g. using Google!). We just want you to write the facts about each country that pop into mind.” We used the data as a measure to account differences in degree of knowledge. We also included 2 items as catch questions (in order to exclude participants who were just responding randomly).

In the stimuli used in the experiment, each pair of countries had an intermediate similarity rating (average similarity for each pair was 5 from a scale 1-9). The rationale for this selection was that if the similarity ratings between the two items/countries were too high, then there would be a risk that any effect from asymmetries might be obscured by ceiling effects (and likewise for low similarities and floor effects).

In the analysis of the results, we first computed the degree of knowledge by counting the number of facts that each participant wrote for each of the two countries in a specific pair and then we normalized it to take into account the relative degree of knowledge of each participant[3]. We also computed a score (ranging from 0 to1) to capture the preference for different statements. In other words, this was a score to measure the proportion of statements selected. For example, in the statement "USA is similar to Canada" vs. "Canada is similar to USA", a score of 0 would mean that the first statement was never preferred compared to the second. A summary of the results is shown in Table 4.

	Score Similarity	**Degree of Knowledge A**	**Degree of Knowledge B**
USA_Canada	0.21	0.75	0.71
India_Pakistan	0.29	0.65	0.45
Italia_Greece	0.40	0.69	0.63
Poland_Germany	0.71	0.39	0.70
Slovakia_Latvia	0.49	0.17	0.16
Brasil_Portugal	0.58	0.65	0.35
UK_Australia	0.23	0.71	0.74
France_Spain	0.46	0.73	0.45
China_NorthKorea	0.17	0.69	0.53
Ghana_Nigeria	0.69	0.28	0.43
Turkey_Cyprus	0.22	0.33	0.22
DominicanRep_CaymanIslands	0.38	0.40	0.39
Malaysia_Singapoore	0.52	0.33	0.37
Hungary_Austria	0.49	0.26	0.41
Madagascar_Mozambique	0.44	0.43	0.16
Mongolia_Nepal	0.46	0.35	0.38
Colombia_Panama	0.38	0.51	0.44
Serbia_Croatia	0.48	0.24	0.16
Belarus_Lithuania	0.68	0.16	0.18
Iceland_Norway	0.60	0.48	0.40

Table 4. Empirical results for the pairs of countries presented in Experiment 5.

Then, we computed the correlation between the similarity scores and the difference in degree of knowledge. Our results showed that there was a negative correlation but it was not significant, r(20) = -.392, p > .05. In other words, the results appeared in the predicted

[3] We used the following equation to normalise the data: $\frac{a+(x-A)\cdot(b-a)}{(B-A)}$, whereby a=0, b=1, x=the number we wanted to normalise, A= the maximum degree of knowledge that a participant stated and B= the minimum degree of knowledge that a participant stated. 'Amount' of knowledge was simply quantified in terms of the stated facts for different countries.

direction but we could not conclude that when the difference in knowledge Country 1 - Country 2 is more negative, the proportion of preferring "Country 1 is similar to Country 2," as opposed to the converse statement, is higher. This result motivated the transformation of the knowledge index variable to a binary one; it is possible that fine differences in the knowledge variable simply reflect noise. When we computed the same correlation after expressing the variable Difference in Knowledge in a binary form, then the results showed a significant negative correlation, r(20) = -.445, p < .05. Specifically, we transformed the variable Difference in Knowledge according to the following rule: if a participant stated that Country 1 had more degrees of knowledge than Country 2, then the variable was coded as 1. Otherwise, the variable was coded as 0. When the knowledge index was expressed as a binary variable, we could find an association between degree of knowledge and similarity asymmetry, in the way we anticipated.

Finally, we wanted to test the within-subjects predictor of degrees of knowledge (in its continuous form) against the appropriate error term. We adopted the individual regression equation method from Lorch and Myers (1990). A separate simultaneous regression was run for each participant with Diff_Knowledge (difference in the degree of knowledge between Country 1 and Country 2) and Sim(score of the similarity of the 1st to the 2nd country) as the dependent variable. This analysis provided one equation for each of the 300 participants in the experiment. The mean regression coefficient was calculated across subjects for the predictor variable and a one-sample t test was then used to assess whether the predictor variable differed reliably from zero. The predictor was not significant (M=-0.005, SD=0.519), t(292)=-0.165, *p*=0.869.

2.3 Conclusion

Throughout this chapter, we have presented a series of empirical tests (five experiments) with the aim to predict asymmetries in similarity judgments using two approaches. One approach for such asymmetries is that degree of knowledge, formalized as number of features, drives salience, which in turn drives asymmetries in similarity judgments (Tversky, 1977). The other approach was the one proposed by Pothos et al. (2013), where a novel model of similarity, based on the mathematics of quantum theory, predicts that asymmetries do not arise from the number of features, but rather from the number of independent features or dimensions, which are needed to represent a concept.

First, in Experiment 1 and 2 we used the stimulus pairs identified in a pilot study (using the similarity database of de Deyne et al., 2008) with the objective to assess QSM's main prediction regarding asymmetries. The stimulus pairs differed in number of features and absolute difference in feature correlation, so providing a dissociation between the prediction regarding the origin of asymmetries according to Tversky (1977) (number of features) and the prediction from the QSM (number of independent features). In Experiment 2, where we presented a two-alternative forced choice task instead of a similarity rating (Experiment 1), the obtained results were in the direction consistent with the prediction of the QSM, Specifically, we showed that when category B has more principal components, than category A, then the statement "A is similar to B" was preferred.

Second, in Experiment 3, still using the similarity database of the Deyne et al. (2008), we could not find evidence in support of our hypothesis: we could not conclude that when PCA is controlled for, asymmetries in similarities arise as a result of differences in features; and that when features are controlled for, asymmetries in similarities arise, as a result of differences in PCA.

Third, in Experiment 4, where we sought an alternative way to dissociate the two possible predictors of asymmetries, difference in features and difference in PCA, we did not observe the predicted effect and we could not conclude that the statement of the form "Category 1 is similar to Category 2" was more likely to be preferred, if Category 1 had more PCA dimensions (or features) than Category 2.

Finally, in Experiment 5 we sought to take a step back and examine the basic evidence for asymmetries in relation to features. We noted that de Deyne et al.'s (2008) database was compiled in Dutch with Belgian participants, but the participants in the present experiments were predominantly British or North American. Therefore, we abandoned the de Deyne et al. (2008) stimuli and instead opted for straightforward countries stimuli (e.g., Greece, Albania etc.). With some additional assumptions regarding the predictor variables (notably that the knowledge index variable was more appropriately expressed as a binary variable), we could find an association between degree of knowledge and similarity asymmetry, in the predicted direction. It has to be noted, however, that this was not the case when the degree of knowledge variable was employed in its continuous form. It is known that sometimes discretizing a variable can lead to confounding results, nevertheless, for reasons outlined above, we think this procedure was appropriate in the present case. In any case, with a conservative stance, we can conclude that the results in Experiment 5 are promising (regarding the prediction of similarity asymmetries), but inconclusive.

Overall, the work in this chapter showed unsatisfying results corresponding directly to both Tversky’s (1977) (based on absolute number of features) and QSM’s (based on principal components) accounts on predicting asymmetries. Nevertheless, plenty of promising directions for future research remain. We considered empirical implications from the QSM only regarding asymmetries, but what about the other metric axioms or the diagnosticity effect? In *Chapters 3, 4* and with future work we hope to make progress regarding these issues.

Chapter 3

Diagnosticity: Theoretical and Empirical Progress

Introduction

As we pointed out in *Chapters 1* and *2*, our capacity to perceive degrees of similarity appears essential for cognition. In previous chapters we stated that a standard approach in psychology has been to represent objects as points in multidimensional psychological spaces and similarity between these points as some decreasing function of distance. We argue that because of some of the requirements and constraints of the distance-based similarity models, a corresponding abstract representational hypothesis can be explored as a psychological theory.

In the seminal study we presented in previous chapters from Tversky (1977), context effects in similarity judgments were reported, such that the same similarity relation between the same two objects can be affected by the presence or not of other available options. This context effect, called the diagnosticity effect, is also completely beyond simple distance-based models of similarity.

The diagnosticity effect is a fascinating demonstration for why similarity cannot be understood as a pairwise relation. However, it is really difficult to find any replications of this effect (see Evers and Lakens, 2014, and Medin et al., 1995, for notable attempts). By contrast, there have been numerous replications in relation to the other key findings in Tversky's (1977) paper (e.g., regarding violations of symmetry, see Bowdle & Gentner, 1997; Catrambone et al., 1996; Op de Beeck, Wagemans, & Vogels, 2003; Rosch, 1975). So, how can we reproduce the diagnostic effect? And what is the most appropriate way to incorporate context in similarity judgments?

The diagnosticity effect is fundamental in the discussion of whether distance-based models are adequate or not. This is because, when it comes to violations of the metric axioms, standard distance-based similarity models can be easily salvaged. In *Chapters 1* and *2* we have already noted how different types of violations arise naturally in distance-based similarity models, but when considering the diagnosticity effect, there are no simple ways to modify standard distance-based similarity models (cf. Goldstone & Son, 2005).

As we noted in *Chapter 1 (section 1.3),* Tversky's (1977) diagnosticity experiment asked participants to identify the country most similar to Austria (denoted as the target stimulus), from a set of three alternatives, e.g. (in one condition) Hungary, Poland, and Sweden. Participants typically selected Sweden (49%), so implying that $\mathrm{Sim}(\mathrm{Sweden}, \mathrm{Austria})$ was the highest. However, when the alternatives were Hungary, Sweden, and Norway, participants typically selected Hungary (60%). Thus, the same similarity relation (e.g., the similarity between Sweden and Austria or the similarity between Hungary and Austria) appears to depend on which other stimuli are immediately relevant, showing that the process of establishing a similarity judgment may depend on the presence of other stimuli, not directly involved in the judgment. Tversky's (1977) explanation was that the diagnosticity effect arises from the grouping of some of the options. In other words, that the range of alternatives led to the emergence of different diagnostic features, which in turn impacted on the similarity judgment. For example, when Hungary and Poland were both included, their high similarity made participants spontaneously code them with their obvious common feature (both were Communist bloc countries at the time), which, in turn, increased the similarity of the other two options, (Austria and Sweden) which were both Western democracies. Tversky (1977) employed 20 pairs of four countries and one further demonstration of the diagnosticity principle, based on schematic faces.

So while the idea behind Tversky's (1977) diagnosticity principle is that the similarity between some of the options leads to grouping, which increases the salience of diagnostic features, which in turn alters the similarity between the target and different options, it is still unclear how similar the two options need to be, before grouping takes place, and how much grouping is needed, before diagnostic features become salient or emerge.

Although we mentioned before that the diagnosticity effect has been difficult to replicate, if we look at the decision-making literature we find extensive replications of the so-called "similarity" effect (e.g., Pothos, Busemeyer, & Trueblood, 2013; Shafir, Smith, & Osherson, 1990; Sloman, 1993; Tversky & Kahneman, 1983). The similarity effect is entirely analogous to the diagnosticity one, but concerns the grouping of options in a decision task. So, there are some differences between the task revealing the similarity effect and the one relevant to the diagnosticity effect (see shortly), but, in essence, the key idea is identical: grouping of some options enhances preference for the isolated one. The literature on the similarity effect provides relevant empirical evidence, which, in some sense, can support our perception of the reality of the diagnosticity effect as well.

Specifically, the kind of task which has been employed to demonstrate the similarity effect involves asking participants to choose the option they prefer, between two alternatives, *A* or *B*. It is well established that when introducing a third option *C*, similar to *B*, then preference for *A* increases. Instead, in the diagnosticity effect case, participants have to choose an option most similar to a target.

More interestingly, the decision-making literature also reports an attraction and a compromise effect apart from the similarity one. To illustrate the attraction, compromise, and similarity effects, suppose an individual is choosing among different cars. Available cars are described in terms of the two attributes, quality and economy, where Car *A* is better on the quality dimension but Car *B* is better on the economy dimension. The attraction effect is produced by adding Car *D* to the choice of Cars *A* and *B*. Car *D* is inferior to Car *A* in both quality and economy dimensions and should thus be discarded but, after adding this decoy, Car *A* becomes more likely chosen and Car *B* becomes less likely chosen (Huber, Payne, & Puto, 1982). Adding Car *C* to a choice between Cars *A* and *B* produces the compromise effect. Car *C* has extremely good quality but poor economy. Importantly, Car *C* makes Car *A* a compromise between the other cars, and with Car *C*'s presence, Car *A* becomes more likely to be chosen than Car *B* (Simonson, 1989). The similarity effect is produced by adding Car *S* instead. Car *S* is similar to Car *B*, and Car *S*'s introduction results in the higher probability of Car *A* being chosen than Car B (Tversky, 1972). So clearly, there is a "competition" between the attraction effect and the similarity one. And perhaps the interplay between the similarity/diagnosticity effect and the attraction effects is what explains the fragility of the diagnosticity effect.

The main purpose of this chapter is to provide a collection of empirical results examining the key interplay between the diagnosticity and attraction effects, with a set of novel experimental paradigms and different stimuli (single feature stimuli: spirals; simple schematic stimuli with more features: triangles). In the first case, we continuously varied the similarity structure between three options, where the positions of the targets are fixed, but the positions of the intermediate alternative, *C*, and the 'extreme' elements *A* and *B* are variable (Experiments 1). In further experiments (Experiments 2, 3 and 4) we attempted a more direct replication of Tversky's (1977) original diagnosticity paradigm. In an additional experiment, we varied the similarity structure between only two options (Experiment 5). This chapter focuses on the empirical aspect of this work, but we note that a detailed mathematical framework based on the QSM (*Chapters 1* and *2*) is being developed and pursued as a separate research direction. This framework enables predictions for both the diagnosticity and the attraction effect in similarity judgments. The empirical results presented here (and the relevant modelling, which goes beyond the scope of the thesis) reveal that the conditions for obtaining a diagnosticity effect are indeed highly constrained.

Diagnosticity using Spirals

We first report an initial set of experimental results examining the interplay between the diagnosticity and attraction effects. In these experiments we used 17 spirals of different sizes, with spiral 1 the smallest one, spiral 17 the largest one and spiral 9 the target one. The size of each spiral was given by the formula,

$$S_n = S_0(1.1)^n,$$

where $S_0$was the size of the initial target spiral (the available choices would be created by having n>0 or n<0, with n=0, as noted, the target). The reason for this choice was that according to Weber's law, participants should rate the similarity between spirals as a function only of the difference in values of n, so that,

$$sim(S_n, S_m) = sim(S_{n+k}, S_{m+k}).$$

This means the expected perceived similarity between neighbouring spirals should be constant, which simplifies the analysis (of course, this is an assumption, but one that is supported by pilot results). We chose the size of S_0 to be 7cm, where physical sizes refer to the appearance of the stimuli on the computer screen on which they were designed. The exact sizes of the spirals as they appear to participants depend on the resolution of the screen on which they take the experiment, but given that such differences would correspond to a constant scaling, this is an issue we need not be concerned with. That is, regardless of differences in the exact size of the stimuli across different screens, their *relative* sizes are resolution independent, and this is all that is important for the present experiments.

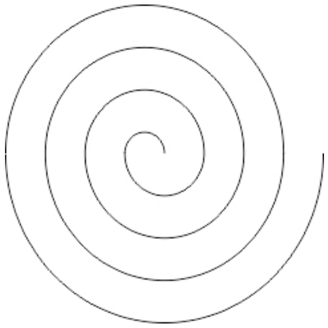


Figure 1. One of the spirals used in the Pilot study and Experiment 1. The sizes of the spirals also depend on the screen on which the experiment is taken. However the relative sizes do not.

Pilot Study

The aim of the pilot study was threefold; firstly to check that the perceived similarity decreases with increasing difference in the size of the second spiral, secondly to check that the perceived similarities are sensitive to the value of the geometric factor, so that participants are not simply ranking the spirals in order of similarity, and thirdly to see whether the perceived similarities obey Shepard's (1987) law of generalization, which fixes the exact relationship between percentage size difference and perceived similarity (this last objective was an aside though an interesting one, given the data was available).

Participants

100 experimentally naïve US residents were recruited via Amazon Turk, and were paid $0.50 for their time.

Procedure, Materials and Results

Participants were asked to rate the similarity between the target spiral and another spiral whose size was given by the formula stated before, $S_n = S_0(1.1)^n$, for $n > 0$ and by a similar expression but with a geometric factor of 10/9 rather than 1.1 for $n < 0$ (note, this difference in the Weber exponents was unintentional). Each participant was shown every pair of the target spiral plus one of the 17 possible spirals (so that participants also saw the pair (target, target)). The order of presentation of the pairs was randomized. The pairs were shown simultaneously on the screen, side by side, and we also randomized the position (left or right) of the target. Participants were asked to rate the similarity on a scale from 1 (very dissimilar) to 9 (very similar). We applied a linear transformation to the ratings produced by each participant to ensure all had a maximum value of 9 and a minimum value of 1.

The results of the pilot confirmed our expectations; although the error bars are large, the (log) data for both smaller and larger spirals show a monotonic decrease in perceived similarity as the distance (in terms of the exponent n) from the target increases, which supports the assumption that similarities are governed by $S_n = S_0(1.1)^n$. The decrease in similarity for the smaller spirals is slightly more rapid, as expected. Finally both sets of data are good fits ($R^2 > 0.98$) to a linear curve (when plotting the logarithm, of the similarity ratings), providing good evidence for Shepard's (1987) law of generalization, for these stimuli (Figure 2).

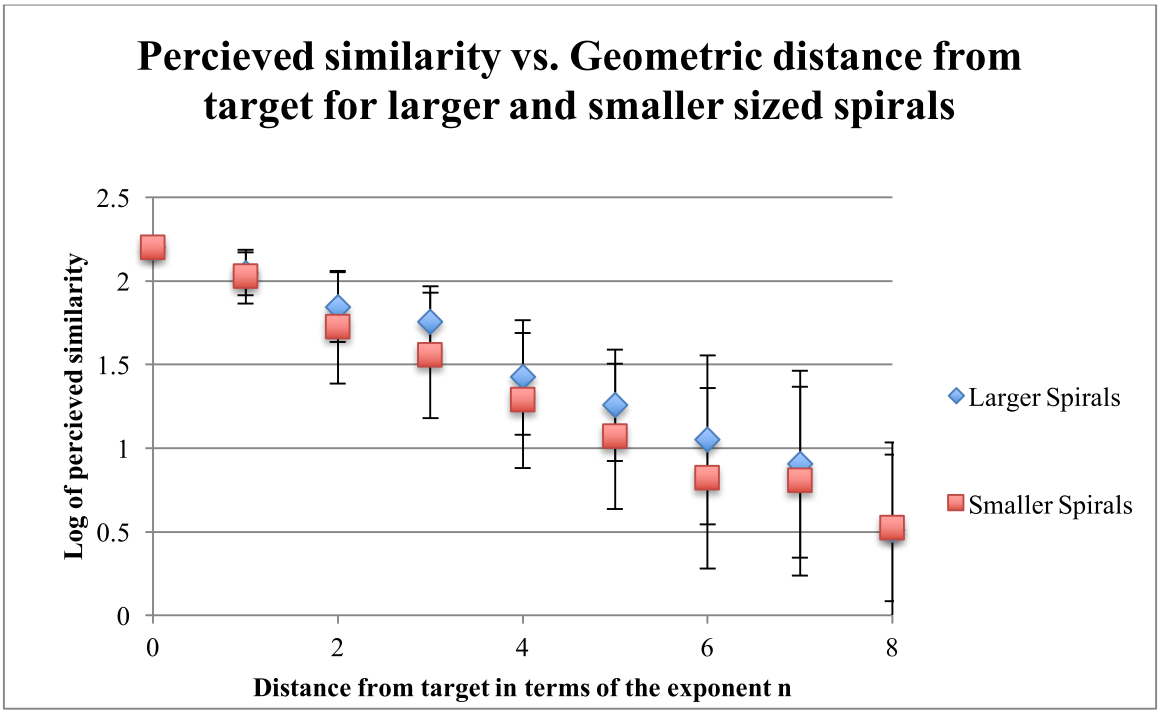


Figure 2. Perceived similarity vs. Geometric distance from target stimuli for larger and smaller spirals in the Pilot study.

Experiment 1

The aim of Experiment 1 was to report an initial set of experimental results examining the key interplay between the diagnosticity and attraction effects using single-feature spirals as stimuli.

Participants

200 experimentally naïve US residents were recruited via Amazon Turk, and were paid \$0.50 for their time.

Procedure, Materials and Results

Participants were given a series of trials where they were shown a target spiral *T* with size S_0 and below three other spirals *A*, *B*, *C*, of sizes S_A, S_{-A}, S_C, and were asked to indicate which of *A*, *B*, *C* they judged most similar to *T*.

All stimuli can be represented in a notional scale of -8 to 8. The experimental trials were designed so that *A* took values from 1 to 8, and *C* took values from -8 to 8, but excluding 0 (note, *T* corresponded to the value of 0). Participants were split into two groups, where one group saw values of *C* from 1 to 8, and the other from -1 to -8. The presentation of the spirals *A*, *B*, *C* on the screen in each trial was partly randomized, where the spirals were presented horizontally across the screen, with *C* in the centre and either *A* on the right and *B* on the left or vice versa. The order of presentation of trials within each group was randomized.

In this section, we restrict the report of the data to the similarity judgments corresponding to *A* = 4. Additional conditions were relevant only with respect to the modelling, which is developed in a research programme exceeding the scope of this thesis. For the present purpose, the objective is purely empirical and is focused on demonstrating an interplay between diagnosticity and the hypothesized attraction effect. Choosing the A=4 condition was motivated because this means we have a target of intermediate size, relative to the range of possible sizes and, importantly, relative to the alternative choices (other choices, like A=3 or A=5, would be nearly as good, but their individual consideration does not add anything for the present purposes). Having A=4 means we have a target spiral *T* of size 7cm, spirals *A* and *B* of sizes 10.2cm and 4.8cm respectively, and then a spiral *C* whose size varied from 3.3cm to 15.0cm. So, *C* is a variable context element, which impacts on choice probabilities for all alternatives, while *A, B* are fixed choices, *equidistant* from the target. We organize the data according to which other spiral is closer, and which one is further away from *C*. Diagnosticity means participants should prefer the spiral furthest from *C*, while attraction means participants should prefer the spiral closer to *C*, provided it is more similar to the target than *C*. The data is presented in Figure 3.

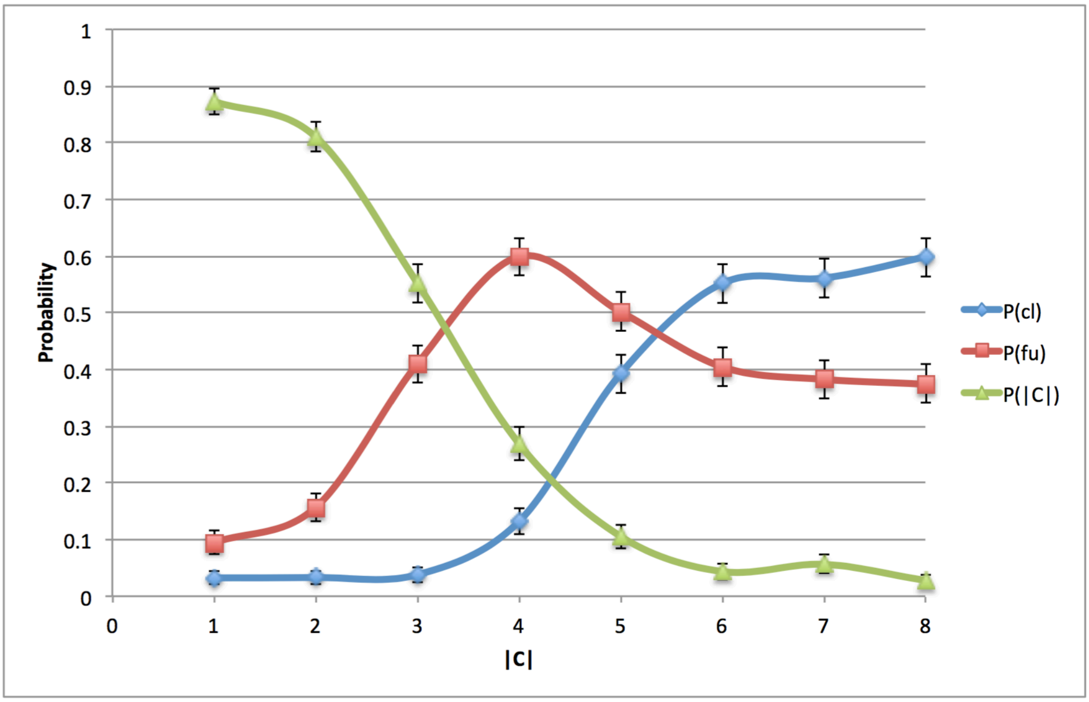


Figure 3: Probabilities for choosing spiral closest to *C* (cl), spiral furthest from *C* (fu) or *C* itself (C), as most similar to *T* as function of |*C*|. *A* = 4 in all cases. Error bars show standard errors. (Source: Yearsley, J. M., Pothos, E. M., Barque-Duran, A. & Hampton, J. A. (2015) Diagnosticity: Some Theoretical and Empirical Progress. Proceedings of the 37th Annual Conference of the Cognitive Science Society. Pasadena, California.)

The data showed both a diagnosticity (|*C*| = 4,5) and an attraction effect (|*C*| ≥ 6). In addition there was a preference for the ungrouped over the grouped stimuli even for |*C*| ≤ 3 which, though not strictly a diagnosticity effect (since *C* is preferred over *A* and *B*), still represents a context effect. Note the key point that, because the *cl* and *fu* stimuli are equidistant from the target, if there were no contextual influences on similarity, then we would expect the red line to coincide with the blue line. Note also that, because of the large number of observations per data point, for all positions of |*C*| any differences in choice probabilities can be assumed to be reliable.

Overall, our data clearly demonstrate both the existence of a diagnosticity effect for these single-feature stimuli and that this effect can break down/ compete with an attraction effect. As noted, the QSM can be extended to account for the interplay between attraction and diagnosticity, though this is a mathematically involved exercise and is pursued elsewhere (an introduction of these ideas is in Yearsley et al., 2015).

Diagnosticity using Triangles

In this section we report a second set of experimental results potentially examining the interplay between the diagnosticity and attraction effects. In these experiments, instead of using spirals as stimuli, we used simple schematic triangles with one or more features on the edges. The motivation for using such stimuli is that they correspond most closely to one of Tversky's (1977) diagnosticity conditions. There were 8 triangles in total (see Figure 4). For simplicity we first present Experiments 2, 3 and 4 together, which all followed a similar experimental paradigm but with different sets of stimuli. We finally present Experiment 5, which explores an extension to the basic paradigm and addresses the challenging results from our previous studies in this section.

Experiments 2, 3 and 4

The aim of Experiments 2, 3 and 4 was to further report experimental results examining the key interplay between the diagnosticity and attraction effects using variations of simple, schematic triangles with different features.

Participants

300 experimentally naïve US residents were recruited via Amazon Turk, and were paid $0.50 for their time (100 participants for Experiment 1, 100 for experiment 2 and 100 for Experiment 3).

Procedure, Materials and Results

We used eight simple, schematic triangles with three different features. In Figure 4 we display the set of stimuli used in Experiment 2 and how the stimuli varied across Experiment 2, 3 and 4. In Experiment 2 we used the set of stimuli based on triangle (x), where the three different features on the sides were a circle, a triangle and a rectangle. In Experiment 3 we used the set of stimuli based on triangle (y), where the three different features were a circle, diamond and a rectangle. The difference between (x) and (y) sets is the triangle-diamond feature. This difference was motivated by a consideration that using a little triangle as one of the features on the side of the larger triangle may have unbalanced the relative salience of the three features. Finally in Experiment 4 we used the set of stimuli based in triangle (z) where the three different features were coloured circles. We used primary colours for all three features (yellow, magenta and cyan) in the hope of avoiding certain colours being more salient than others.

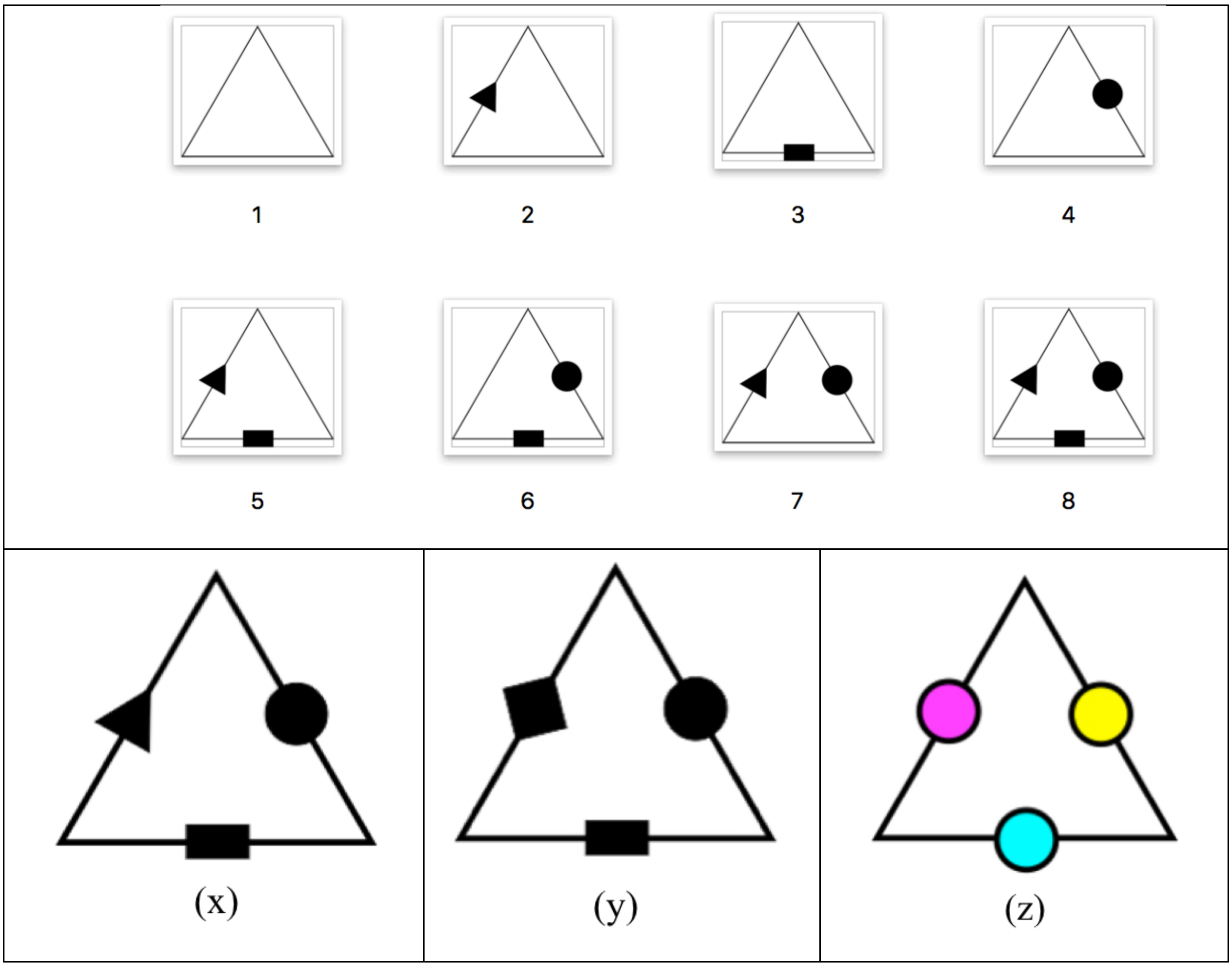


Figure 4. Top: The eight simple, schematic triangles with the three different features used in Experiment 2. Bottom: Three versions of the stimuli, that were used in Experiments 2, 3 and 4. We use the label (x) for the Experiment 2 version (also referred to as Triangle version), (y) for the Experiment 3 version (also referred to as Diamond version) and (z) for the Experiment 4 one (also referred to as Coloured version).

We designed two *Sets* of four triangles, which differed in only one of their elements (*C* and *C'*, in Figure 5), in a way inspired by the studies of the diagnosticity principle in Tversky (1977; Figure 5).

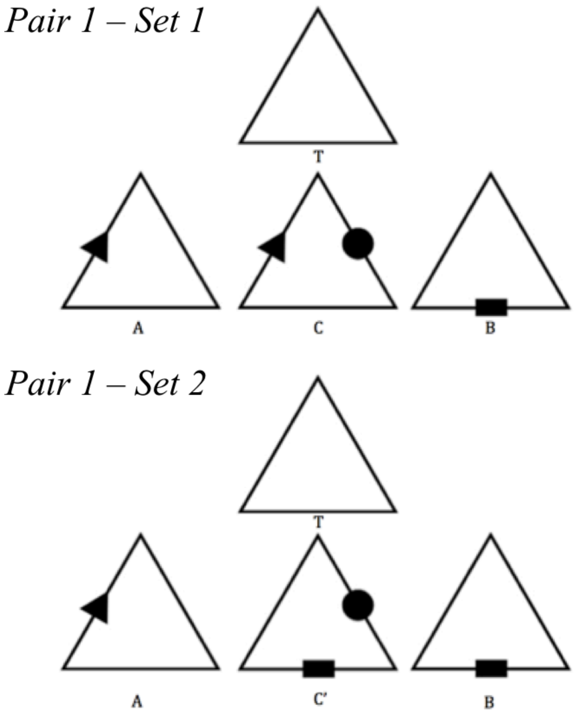


Figure 5. An example of one of the three possible *Pairs* for *Set 1* and *Set 2* using the Simple triangle as a target *T*. Note that this example corresponds to the Triangle version of stimuli (x).

Participants in *Set 1* were given a series of trials, which we call *Pairs*, where they were shown the target triangle *T* and three other triangles *A, B, C,* and were asked to indicate which of the three triangles they thought was most similar to the target. The triangles in the *Pairs* on *Set 1* were designed to have *A* and *C* (similar in one feature) versus *T* and *B* (different features). To be more specific, *A* and *B* were each different from *T* in one feature. That is, they were 'equidistant' from *T* (cf. Experiment 1 above). *C* was designed to be close to *A* but far from *B* and *T*. That is, *C* shared one feature and had one feature away from *A*, two away from *T* and three away from *B*. So, *C* is the context element that is grouped with one of the options that are equally similar to the target (these equally similar options are *A*, *B*).

Participants in *Set 2* were also given a series of trials (*Pairs*) where they were shown the target triangle *T* and three other triangles *A, B, C',* and were asked to indicate which of the three triangles they thought was most similar to the target. The triangles in the *Pairs* on *Set 2* were designed to have *B* and *C'* (similar in one feature) versus *T* and *A* (different features). To be more specific, *A* and *B* were different from *T* in one feature, as in *Set 1*. But *C'* was designed to be close to *B* but far from *A* and *T*. That is, *C'* shared one feature and had one feature away from *B*, two away from *T* and three away from *A*.

By the diagnosticity hypothesis, choice behaviour should follow the grouping. That is, the similarity of *T* to *B* should be greater in *Pairs* on *Set 1*, where the other choices (*A* and *C*) are grouped together, than in *Pairs* on *Set 2*, where the choices *A* and *C'* are not. Likewise, the similarity of *T* to *A* should be greater in *Pairs* on *Set 2*, where the other choices (*C'* and *B*) are grouped together, than in *Pairs* on *Set 1*, where the choices *C* and *B* are not.

To test this prediction, we also created two different targets: Simple Triangle (a triangle with no features, Figure 5) or Complete Triangle (a triangle with a feature on each edge, Figure 6).

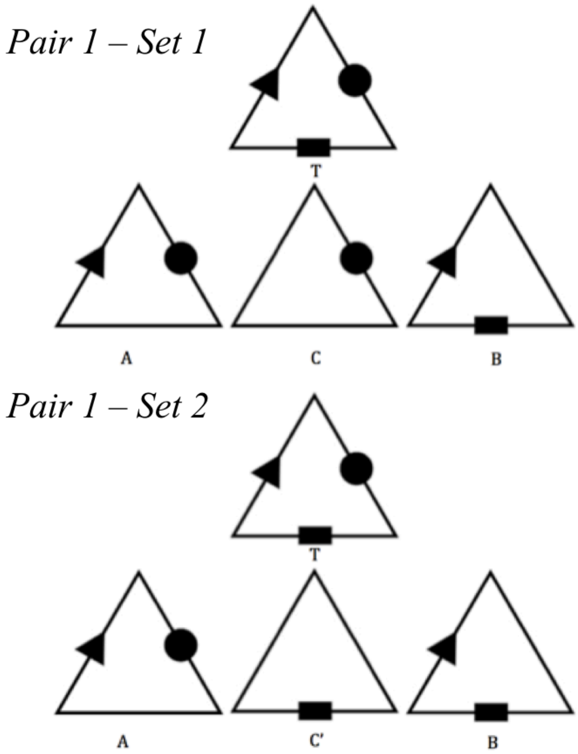


Figure 6. An example of one of the three possible *Pairs* for *Set 1* and *Set 2* using the Complete triangle as a target *T*. Note that this example corresponds to the Triangle version of stimuli (x).

In total, there were six different possible arrangements or *Pairs*: three for the Simple target and three for the Complete target. We presented all of six *Pairs* (in the form displayed in Figure 5 and 6) to each participant and they repeated the task three times. One group of 50 subjects (in each experiment) responded to three different *Pairs* from *Set 1* for the Simple target and three different *Pairs* from *Set 1* for the Complete target. The other group of 50 subjects (in each experiment) responded to three different *Pairs* from *Set 2* for the Simple target and three different *Pairs* from *Set 2* for the Complete target. The order of presentation of triangles within each trial was partly randomized, the order of *A* and *B* was randomized, but *C* or *C’*, the context item, was always presented in the middle. The order of presentation of *Pairs* within each group (*Set 1* or *Set 2*) was also randomized.

We report the data focusing on the diagnosticity hypothesis, where choice behaviour should follow the grouping. That is, in Table 8 we compare the percentage of selected responses when options are grouped or ungrouped. In all three experiments (Experiment 2: Triangle, Experiment 3: Diamond and Experiment 4: Coloured), we found a similar pattern of results. For example, if we focus on the results from Experiment 2, although in *Set 1* participants preferred the grouped option in *Pair 2* (65.73%) and in *Pair 3* (64.41%), participants also preferred the ungrouped option in *Pair 1*. In other words, two out of three *Pairs* followed the diagnosticity hypothesis (similarity of *T* to *B* should be greater in *Pairs* on *Set 1*, where the other choices are grouped together, than in *Pairs* on *Set 2*). Similar results appeared in *Set 2*, where participants preferred the ungrouped option in *Pair 2* (75.9%) and in *Pair 3* (72.41%), but preferred the grouped option in *Pair 1*. Again, two out of three *Pairs* followed the diagnosticity hypothesis (similarity of *T* to *A* should be greater in *Pairs* on *Set 2*, where the other choices are grouped together, than in *Pairs* on *Set 1*). Following Experiment 2, we introduced new versions of stimuli for Experiments 3 and 4 as we noticed that *Pair 1* in Experiment 2 seemed to be problematic. As stated before, the purpose was to avoid the similarity from the small triangle (feature) to the big triangle (background) when designing Experiment 3; also, when designing Experiment 4, we used primary colours (yellow, magenta and cyan) to avoid saliency effects. A similar pattern of data appeared in all three experiments.

		Triangle (Exp. 2)		**Diamond (Exp. 3)**		**Coloured (Exp. 4)**	
	Pair	**%Grouped**	**%Ungrouped**	**%Grouped**	**%Ungrouped**	**%Grouped**	**%Ungrouped**
Set 1	**P1**	35.80	61.93	40.63	56.25	47.87	49.47
	P2	65.73	33.15	64.06	30.73	60.64	36.17
	P3	64.41	29.38	70.83	26.04	65.96	31.91
Set 2	**P1**	46.12	48.71	56.25	41.35	48.56	49.52
	P2	20.17	75.97	34.13	62.98	37.98	56.25
	P3	25.00	72.41	29.81	66.35	24.04	69.71

Table 7. Percentage of grouped and ungrouped responses for each *Set* and *Pair* and for each of the three versions of the stimuli used in Experiments 2, 3 and 4.

The data from all three attempted replications of Tversky's (1977) diagnosticity paradigm (Experiments 2, 3 and 4) pointed to the same conclusion: rather than having a consistent diagnosticity effect, we appear to have a combination of diagnosticity and attraction. Moreover, we could argue that there are no context effects at all and the results arise simply because of differences in feature salience. We support this idea by showing in Table 7 the percentage of selected responses when stimuli that have a single feature (singe rectangle, a single triangle or a single circle) were presented grouped or ungrouped in a trial. A Chi-square test of independence was performed comparing the frequency of grouped and ungrouped endorsement responses, when these stimuli having a single feature were presented. Indeed, no significant differences were found χ^2 (2) = 0.49, p=.078. Essentially, this result shows that what may look like a context effect in a choice paradigm, very similar to the one employed by Tversky (1977), does not uniquely provide evidence for contextual influences.

	Blank Target/Single Feature	
	% when Grouped	**%when Ungrouped**
Rectangle	47.88	52.11
Triangle	45.86	54.13
Circle	45.76	54.23

Table 8. Percentage of grouped and ungrouped responses when stimuli with a single feature (rectangle, triangle and circle) were presented in Experiment 2.

Experiment 5

The aim of Experiments 5 was also to examine a variation of the basic paradigm employed in Experiments 2 to 4, with a view to provide more direct evidence for (putative, at this point) contextual influences on choice behaviour. In Experiment 5, we employ a choice set of two items instead of three as in our previous experiments or as in the studies of the diagnosticity principle from Tversky (1977).

Participants

400 experimentally naïve US residents were recruited via Amazon Turk, and were paid $0.50 for their time.

Procedure, Materials and Results

We used the same eight simple and schematic triangles as in Experiment 2 (see Figure 4, Top), where the three different features on the sides of each triangle were a circle, a triangle and a rectangle. First, participants were presented with a pairwise similarity rating task, which had 28 trials in order to measure the similarities between each possible combination between a choice and a target. Participants were asked to rate the similarity on a scale from 1 (extremely different) to 9 (near identical). Then, participants were presented with a forced choice task where they had to choose the triangle most similar to a target of either triangle 1 or triangle 8 from a pair of options (see Figure 5 for target stimuli; see Figure 9 for trial structure). We designed the task with a choice set of two items, instead of three as in our previous experiments. Participants saw a total of 12 randomised trials (six using triangle 1 as the target and six more using triangle 8 as the target). We also introduced two filler trials to test that participants were paying attention to the task. These control trials were meant to have a very obvious answer (i.e. one of the options was identical to the target stimulus). In Figure 9 we present one experimental trial as an example, where participants were asked which item, between 2 or 3, was most similar to 1 (see Figure 9).

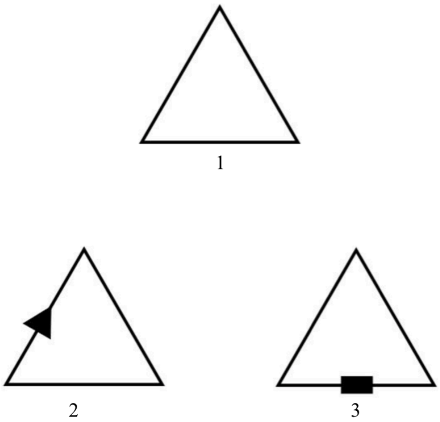
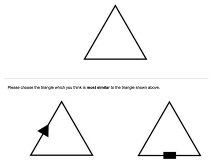
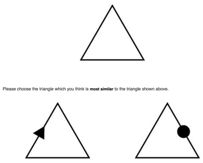
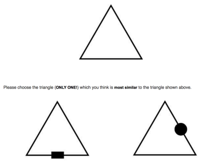
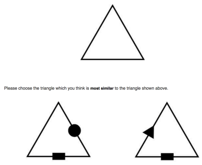
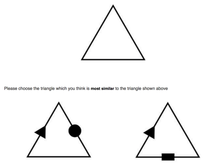
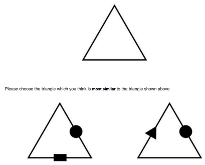


Figure 9. An example of one of the trials using triangle 1 as the target stimulus in Experiment 5.

Here, we report the data for all the different trials presented and we focus on the percentage of selected answers for each item (see Figure 10). However, before analysing the results, we applied some data cleaning. Following the example of the trial presented above, we ignored the responses from any participant that rated the pairwise similarities between item 1 and 2 and item 1 and 3 as different by more than one point on the 1-9 scale we used. In other words, we controlled for participants who did not report fairly similar pairwise similarities (i.e. similarity between item 1(target) and item 2 should be similar to the similarity between item 1(target) and item 3, because it is convenient to restrict data in a way that the assumption that the two features in the choice candidates were equally salient). Critically, this means that the similarities between the target and each of the possible choices individually were (fairly) identical.

Trial (Target 1)	 Please choose the triangle which you think is most similar to the triangle shown above.		 Please choose the triangle which you think is most similar to the triangle shown above.		 Please choose the triangle (ONLY ONE!) which you think is most similar to the triangle shown above.		 Please choose the triangle which you think is most similar to the triangle shown above.		 Please choose the triangle which you think is most similar to the triangle shown above.		 Please choose the triangle which you think is most similar to the triangle shown above.	
Item	2	3	2	4	3	4	6	5	7	5	6	7
%Selected	33.5	66.5	76.4	23.6	82.9	17.1	35	65	48.4	51.6	36.2	63.8
χ^2 (1), *p*<.05	29.45		748.78		112.02		27.81		0.31		24.85	

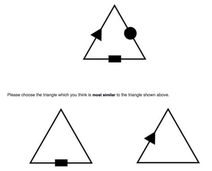
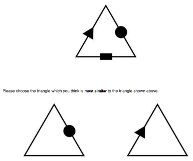
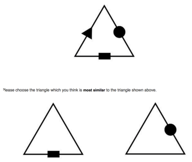
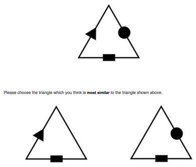
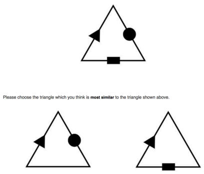
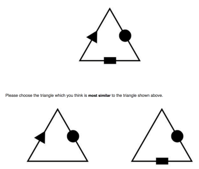
Trial (Target 8)	 Please choose the triangle which you think is most similar to the triangle shown above.		 Please choose the triangle which you think is most similar to the triangle shown above.		 Please choose the triangle which you think is most similar to the triangle shown above.		 Please choose the triangle which you think is most similar to the triangle shown above.		 Please choose the triangle which you think is most similar to the triangle shown above.		 Please choose the triangle which you think is most similar to the triangle shown above.	
Item	2	3	4	2	3	4	5	6	7	5	7	6
%Selected	37.7	62.3	31.3	68.7	51.3	48.7	69.2	30.8	59.8	40.2	78.1	18
χ^2 (1), *p*<.05	18.63		44.39		0.20		45.21		10.97		99.46	

Figure 10. Percentage of selected answers for each item and trial in Experiment 5.

The results showed some strong preferences for particular choices in the forced choice task. First, we focused on the results for the trials using item 1 as the target stimulus. For example, when participants were asked to choose which item, between 3 or 4, was most similar to 1, 82.9% selected item 3. This is a contextual influence, since the first phase of the experiment and subsequent data selection ensured that the similarity between the target and each of the choices should be (fairly) identical. Chi-square tests of goodness-of-fit were performed for each of the trials to determine whether the two items were equally preferred (see Figure 10). Preferences were not equally distributed in all trials (i.e. significant differences), except for the trial with items 7-5. More interesting is the following: if the target is 1, participants preferred item 3 to 2 (66.5% vs. 33.5%). However for the same target, participants preferred item 7 to 6 (63.8% vs. 36.2%). That is, adding the circle on the side of each of the choice stimuli, reversed the effect of the triangle, rectangle features on choice probabilities. Nevertheless, the other two sets when the target is 1 did not show this effect.

Second, we focused on the results for the trials using item 8 as the target stimulus. For example, when the target is 8, participants preferred item 7 to 5 (59.8% vs. 40.2%) but they marginally preferred item 3 to 4 (51.3% vs. 48.7%). Nevertheless, the other two sets when the target is 8 did not show this effect. Again, Chi-square tests of goodness-of-fit were performed for each of the trials to determine whether the two items were equally preferred (see Figure 10). Preferences were not equally distributed in all trials, except for the trial with items 3-4.

We think there is a reasonable post hoc explanation for some of the results provided above. For example, stimulus 3 and 7 display a vertical symmetry that for instance stimulus 2 and 6 (their respectively pairs) do not show. One could argue that the reason for participants choosing options 3 and 7 is that a forced choice task is first guided by default features or dimensions (in this case the presence or absence of triangles, circles or rectangles) and when theses default features fail to discriminate, then it is either the case that other features emerge or the distribution of attention changes. These explanations do not deter, however, from appreciating that this paradigm did lead to reliable contextual influences on choice behaviour. Whether these contextual influences are theoretically interesting or not ultimately depends on the kind of model developed to account for them – this latter objective is, however, beyond the scope of this thesis.

In fact, following the description from Tversky’s (1977) diagnosticity experiment provided above, one can think of an analogous situation. If we assume that the pairwise similarities between Austria, Sweden and Poland are identical, then in a forced choice task with Austria as a target, the default similarity features (e.g., geography) fail to discriminate between Sweden and Poland, so other features guide the choice (e.g., whether a country is considered in Western Europe or not). This explanation is motivated by the take-the-best heuristic, whereby when the most typically useful cues fail to discriminate, other cues are recruited (Gigerenzer & Goldstein, 1996).

Overall, even though the pairwise similarities of the two choice items were fairly identical to the target, when presented together revealed an effect of contextual influence. That is, when pairwise similarities are equal, one still gets a preference for a particular stimulus. Experiment 5, together with Experiments 2, 3 and 4 are significant in that they show both why people have mostly failed to replicate the original Tversky (1977) diagnosticity results, with shapes, and because we identified an extension of Tversky's (1977) with potential to reveal more surprising and interesting contextual influences.

Conclusion

In this chapter we provided an extensive collection of empirical results examining contextual influences in choice in a similarity task, with a set of novel experimental paradigms and different stimuli with different features. In Experiments 2, 3 and 4, motivated by Tversky's (1977) diagnosticity experiment, we varied the similarity structure between three options (where the positions of the targets were fixed and varied the position of the intermediate alternative, *C*, and the 'extreme' elements *A* and *B)*. Then, in Experiment 5 we varied the similarity structure between only two options.

Overall, the results from Experiment 1 pointed to the conclusion that rather than having a consistent diagnosticity effect, we appear to have a combination of diagnosticity and attraction. Based on Experiments 2 to 4, one could argue that there are no context effects and the corresponding results simply illustrate differences in feature salience. Moreover, it seems that feature salience is malleable and is susceptible of attention changes, as the results of Experiment 5 appear to indicate. In other words, other features can emerge as salient, when the most typically useful cues fail to discriminate. So the question is still whether a quantum model can predict all these patterns simultaneously. If we assume that attention weights are stable across the tasks, then it seems clear that any distance-based model of similarity or any feature-based model of similarity (with no interactions between feature weights) cannot account for these (especially Experiment 5) findings. Can a quantum model account for these effects computationally? This is an issue for future work (for an outline of some of the relevant ideas see Yearsley et al., 2015).

Chapter 4

Distinguishability: Asymmetries in Similarity Judgments.

Introduction

As stated in *Chapter 1*, an important gap in the quantum similarity model (QSM) concerns how to deal with asymmetries arising from differences in the frequency of presentation of stimuli (Polk et al., 2002) or from differences in prototypicality, for stimuli for which prototypicality does not have to do with more extensive knowledge (e.g., it is unlikely that we 'know' more about a prototypical red, compared to a non-prototypical red; Rosch, 1975). In this chapter we focus on the later. Presumably what distinguishes a prototypical stimulus from a non-prototypical one, or a stimulus presented many times from one presented only infrequently, is the increased potentiality for a participant to think about this stimulus. Can we use this idea to motivate or outline a prediction from the QSM, that will enable a test of whether the QSM can accommodate such asymmetries? (Or be extended in a way that can accommodate such asymmetries?).

Similarity asymmetries in the quantum model can arise in two ways. The first way concerns asymmetries which have to do with differences in the degree of knowledge between the compared entities. So, in the classic example from Tversky (1977), because we know more about China than Korea, the representations of China and Korea involve subspaces of different dimensionalities. If this is the case, then as shown by Pothos et al. (2013), the similarity between Korea, China can be predicted to be higher than the similarity between China, Korea. The second way can arise from differences in the distinguishability between the compared stimuli; note, such differences could relate to relative figural goodness or frequency of the stimuli (Polk et al., 2002; Rosch, 1975). Then, in the QSM, one could consider similarity comparisons between, for example, a more focal and a less focal red, in the context of other stimuli from the relevant category (here, other variations of red). With pilot modelling work, under such circumstances, it appears that the QSM predicts an emergence of asymmetries of the form

$$Sim\ (NP, P) > Sim(P, NP),$$

where *NP* denotes a stimulus that is not prototypical and *P* one that is prototypical. However, it also appears that such asymmetries would disappear when the stimuli become more discriminable. These predictions, coming from mathematical derivations of the QSM discussed in the previous chapters, are preliminary and we do not wish to fully develop them here. Rather, the objective of this chapter is to collect some empirical data, broadly conforming to the above ideas, and so inform any further application of the QSM.

We employed meaningless, schematic stimuli (e.g., Gibson & Gibson, 1955, scribbles), which can be constructed to have a roughly prototypical structure (e.g., using the classic distortion procedure of Homa, 1978). Presumably, changes in discriminability for such stimuli are meaningful (while this is clearly not the case for conceptual stimuli, like China or Korea). Specifically, we employed one dimensional stimuli arranged along two clusters (see below). It is worth mentioning the significant work from Rosch (1978) on the psychological principles of categorisation (i.e. Subordinate: kitchen chair, living-room chair; Basic: Chair; Superordinate: Furniture). Rosch distinguishes between vertical and horizontal levels of categorization. The vertical dimension concerns the level of inclusiveness of the category - the dimension along which the terms collie, dog, mammal, animal vary. The horizontal dimension concerns the segmentation of categories at the same level of inclusiveness - the dimension along which the terms car, dog, chair etc. vary. Rosch argues that the use of prototypes, which contain the most representative attributes inside the category, would increase the flexibility and distinctiveness of categories along the horizontal dimension. Nevertheless, as we employed meaningless, schematic stimuli, we are not going to strictly follow the taxonomy that Rosch uses, as we tried to develop our own methods to measure prototypicality.

Then, in the main experiment we performed a series of categorization and similarity tasks with the aim to create stimuli which would vary in their degree of prototypicality, but also vary in other ways that can impact on their salience (notably, some stimuli were diagnostic, some ideals). We also included some manipulations with a view to vary the discriminability of the stimuli.

Pilot Study 1, 2 and 3

In this series of pilot studies we tested the level of contrast at which the distinguishability between simple perceptual stimuli starts to break down and also aimed to identify a suitable percentage size difference for constructing a target category of stimuli, that is, the percentage by which immediately adjacent stimuli differ from any stimulus (i.e. a percentage size difference of 5% means that the adjacent stimuli from any stimulus have a ±5% difference in its size (cm)). We used three different percentage size difference values to construct stimuli and we manipulated the contrast between the stimuli and the background against which they would appear (using an RGB scale). The results obtained helped us to design the main similarity study, where we tested the idea that similarity asymmetries might be reduced or eliminated when increasing the distinguishability of the compared stimuli. The difference between Pilot 1, 2 and 3 related to the mask presented between trials. Here, we only report the results from Pilot 3. Further details and a justification for choosing the procedure from Pilot 3 are reported in Appendix 1: The effect of masks in similarity judgments.

Participants

Sixteen experimentally naive students at City University London received course credit for participating in Pilot 3.

Procedure

The experiment, designed in Superlab, lasted approximately 20 minutes. We used a 2-alternative forced choice to test objective distinguishability. On each trial, we presented two stimuli (see specific details in Materials section) sequentially in time and asked participants to say which was the biggest one. The outline of a trial was as follows. A fixation cross appeared in the middle of the screen, indicating that all was ready for the subject to initiate the trial. When the space bar key was pressed, the first spiral of the pair of stimulus figures was flashed upon the screen (three seconds) and immediately followed by a mask (composed of random curve segments of similar curvature to the spirals). Then the second spiral was flashed immediately followed by a mask, which was the same as before. The subjects' task was to respond ‘1’ if they thought that the first item was bigger or ‘2’ if the second item was bigger. Corrective feedback was not provided.

Materials

We used spirals, which varied in overall diameter, for our stimuli, as they had the basic desired properties (meaningless and schematic). The overall distribution of the eighteen spirals in a psychological space was such that there were two distinct groups of nine items each (See Figure 1).

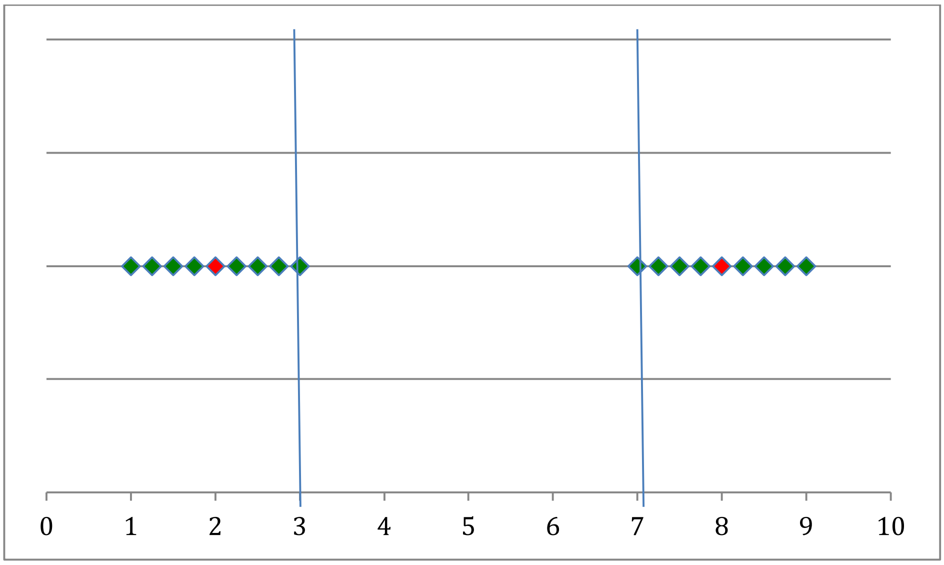


Figure 1: Overall distribution of the stimuli in psychological space. Red items are the assumed *P* (prototypical) items. Green items are the *NP* (non prototypical) items, that is, away from the centre. Note, that in this psychological space the dimension just corresponds to experimenter-defined values, for representing the stimuli.

In this pilot we wanted to sample from the stimuli that would be relevant in the main similarity experiment (i.e., the stimuli in the two clusters). So from these eighteen stimuli, we used a small subset of adjacent pairs. We used the two adjacent smallest stimuli and the two adjacent largest stimuli for the cluster in the left (See Figure 1 and Table 1) and the two adjacent smallest and the two adjacent largest stimuli for the cluster on the right. So, regarding this distinguishability experiment, we had eight items (four pairs).

Then, we created different sets of stimuli based on different values for percentage size difference. We employed 5%, 7% and 9%, which, recall, refer to the percentage increase in size, between any immediately adjacent stimuli (see Table 1). Note that we employed the percentage size difference values not assuming that they correspond to just discriminable differences between adjacent stimuli. Still, based on Weber's law, we expect the psychological impact of differences between any adjacent (in psychological space) stimuli to be the same, regardless of size.

		Percentage size difference		
Stim id	**Psych Space Dimension**	**5%**	**7%**	**9%**
1	1	1.50cm	1.5cm	1.5cm

2	1,25	1.58cm	1.61cm	1.64cm
3	1,5	1.65cm	1.72cm	1.78cm
4	1,75	1.74cm	1.84cm	1.94cm
5	2	1.82cm	1.97cm	2.12cm
6	2,25	1.91cm	2.10cm	2.31cm
7	2,5	2.01cm	2.25cm	2.52cm
8	2,75	2.11cm	2.41cm	2.74cm
9	3	2.22cm	2.58cm	2.99cm
10	7	4.84cm	7.61cm	11.87cm
11	7,25	5.08cm	8.14cm	12.93cm
12	7,5	5.33cm	8.71cm	14.10cm
13	7,75	5.60cm	9.32cm	15.37cm
14	8	5.88cm	9.97cm	16.75cm
15	8,25	6.17cm	10.67cm	18.26cm
16	8,5	6.48cm	11.42cm	19.90cm
17	8,75	6.81cm	12.22cm	21.69cm
18	9	7.15cm	13.07cm	23.64cm

Table 1. Dimensions of all the stimuli depending on their percentage size difference condition.

For each percentage size difference, we further manipulated the contrast between the stimulus colour and the background against which the stimuli appear. Cortical neurons tend to be sensitive to contrast, which we can define as the luminance difference divided by the mean luminance (Doubling of contrast). So, we designed three conditions where contrast has been specified to have an approximately equivalent perceptual effect, assuming what was important was the proportional increase in luminance difference. Specifically, the background for each stimulus was set to 255 (in RGB scale this is white colour) and the colours of the stimuli for each of the three conditions of spirals were: 192, 210 and 219; note, 219 corresponds to least contrast (light grey stimuli appearing against a white background). Note also, the difference in contrast between the first two conditions is 18 points (in RGB scale) and the difference to the following contrast condition is 9 points, that is, we multiplied the difference in contrast by a constant factor in order to increase the internal effect by a constant additive amount.

In sum, we manipulated the percentage size differences (three conditions) and the contrast between stimuli and background (three conditions), so that overall there were nine sets of stimuli. The task for all sets of stimuli was identical; it was a 2-alternative forced choice task (asking participants which item was the biggest); recall, each set of stimuli involved eight unique stimuli, that is, four pairs. The pairs in each set were presented three times, for a total of 36 trials. In total we tested 108 trials (half of them in the opposite direction).

Results

On percentage size differences effects:

First, a one-way ANOVA was used to test if there were any differences in performance (percentage of correct responses) depending on percentage size difference (P1=5%, P2=7% and P3=9%). The analysis showed that there were significant differences, F(2, 177) = 7.397, p=.001; the bigger the percentage size difference, the better the performance, as indeed expected.

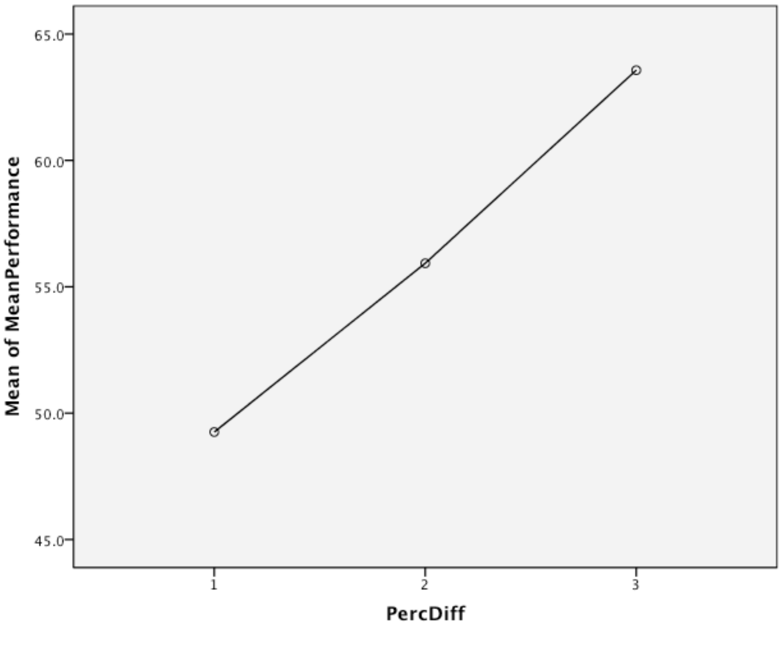


Figure 2. Mean performance (percentage of correct responses) depending on percentage size difference (P1=5%, P2=7%, P3=9%).

We decided to eliminate the P1 condition (Percentage Difference 1=5%) because performance in that condition was the lowest (in terms of accuracy); indeed it was so low so as to be effectively indistinguishable from chance.

Second, we performed a two-way ANOVA with two within participant factors, percentage size difference (2 levels: P2 and P3) and pairs (4 levels: Pair 1 to 4). The results indicated that there was a main effect of percentage size difference in accuracy (percentage of correct responses), F(1, 112) = 5.301, p=.023, but not a significant main effect of pairs, F(3,112) = 1.680, p=.175. The results also indicated that there was a non-significant interaction between percentage size differences and pairs, F(3,112) = 1.680, p=.175. The mean values of correct responses for each pair and percentage size difference are shown in Table 2:

Percentage Size Difference	**Pair**	**Mean**	**Std. Error**
2	1	54.073	4.590
	2	57.053	3.579
	3	48.140	6.295
	4	64.460	5.629
3	1	61.493	4.587
	2	57.973	5.551
	3	62.233	5.511
	4	72.587	7.021

Table 2. Mean performance (% of correct responses) for each pair and percentage size difference.

Even though there was a non-significant interaction between percentage size differences and pairs in the previous analysis, we still considered any performance differences between pairs in the two conditions (note, these analyses were exploratory, with a view to identify the optimal form of stimuli, rather than inferential). We ran a separate ANOVA to test if the differences in accuracy (percentage of correct responses) across pairs varied for each of the percentage size conditions. The analysis showed that there was no main effect in P2 (Percentage Difference 2=7%), F(3,42) = 1.382, p=.261. However, there were significant effects in P3 (Percentage Difference 3=9%), F(3,42) = 3.022, p=.040. Recalling that we performed this analysis to explore any interesting differences across pairs with the aim to identify the optimal form of stimuli for our Main Experiment, the lack of interaction reassures us regarding the appropriateness of the design.

Finally, we also assessed whether the distribution of scores was normal within each of the relevant categories of data points. We used a non-parametric statistic test of normality (Kolmogorov-Smirnov) for the data of each percentage size difference. For both P2 and P3 we accepted the null hypothesis that the data was normal (p>0,05). Therefore, normality of the data was assumed.

In conclusion, and in terms of percentage size difference effects, we decided to use both P2 and P3 conditions for the main similarity study, which were used to manipulate higher and lower discriminability between stimuli. We decided to use both conditions because on the one hand the results indicated that there was a main effect of percentage size difference in accuracy (percentage of correct responses) and on the other hand because there was no evidence for any difference between percentage size differences and pairs since the interaction was absent.

On contrast effects:

First, we performed an ANOVA to test if there were any differences in performance (percentage of correct responses), depending on contrast (C1= 192RGB; C2= 210RGB and C3= 219RGB) and percentage size difference (P1=5%, P2=7% and P3=9%). The analysis revealed two significant effects. Contrast had a significant effect on performance, F(2,28) = 4.109, p=.027 and, consistently with what we observed in the analyses above, percentage size difference had a significant effect on performance too, F(2,28) = 13.170, p<.05. Nevertheless, the interaction between contrast and percentage size difference was not significant, F(4,56) = 1.974, p=.111. A graph with the mean values of correct responses obtained depending on contrast are shown in the figure below (Figure 3).

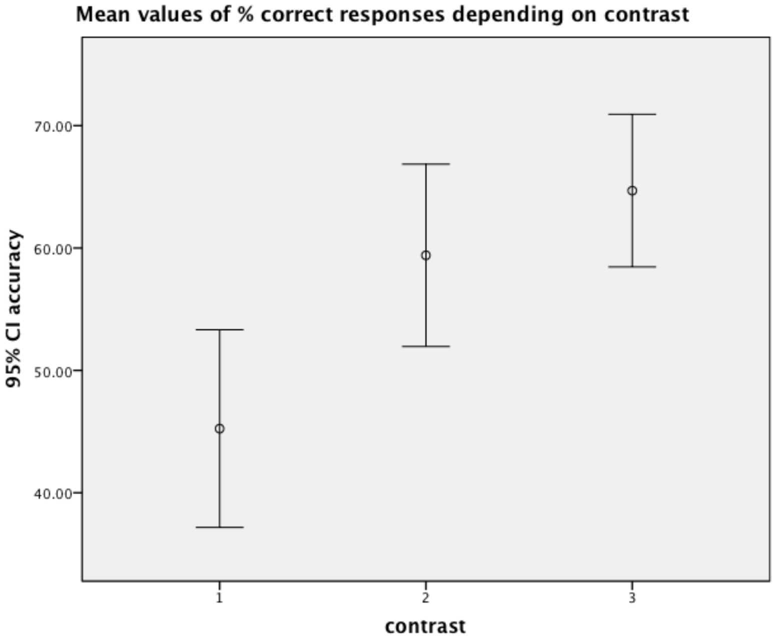


Figure 3. Mean performance (percentage of correct responses) depending on contrast levels (C1= 192RGB; C2= 210RGB and C3= 219RGB).

Second, we performed an ANOVA to test if there were any differences in response time, depending on contrast and percentage size difference. That is, we wanted to know if reaction time can change due to any of the relevant perceptual characteristics of our stimuli. The ANOVA showed that there were no significant main effects for contrast (F(2,28) = .619, p=.546) and for percentage size difference (F(2,28) = .060, p=.942). Also, the interaction between contrast and percentage size difference was not significant, F(4,56)=1.034, p=.398. The response time values obtained are shown in the table below (Table 3).

Contrast	Percentage Difference	Mean	Std. Error
1	1	982.400	128.430
	2	946.331	122.460
	3	961.061	107.458
2	1	965.689	125.077
	2	1086.107	169.778
	3	954.302	139.842
3	1	936.111	98.645

	2	905.520	108.153
	3	992.412	135.893

Table 3. Mean response time depending on contrast levels and percentage size difference.

Overall, the C1 condition, which corresponded to having an RGB value of 255 as a background on the screen and an RGB value of 192 for the stimuli, appeared to be the condition in which the percentage of correct responses was the lowest. C2 condition, with the same background contrast as C1 and a spiral with an RGB value of 210 showed a better percentage of correct responses. Finally, C3 condition, with a spiral with an RGB value of 219 seemed to give the best percentage of correct responses. The results were somewhat surprising because they indicated that the lower the contrast (i.e., the harder it is to differentiate between the stimuli), the better the performance. This finding is counterintuitive. It might be the case that when a question feels easy, people deliberate on it less and make more errors than when the same question appears more difficult (cf. Alter, Oppenheimer, Epley, and Eyre, 2007). Whether such an explanation for the present results is valid or not is beyond the scope of this work (and indeed not entirely relevant to the present stimulus design considerations). Therefore, we decided to eliminate contrast as a condition in the main experiment and use the highest RGB level possible (84 RGB), as the pilot results indicated that participants found challenging even the highest contrast we employed in this study, that is the contrast in the C3 condition (255 RGB value for the background and 219 RGB value for the spiral).

Discussion

Two main conclusions are derived from these analyses. On the one hand, in terms of percentage size difference effects, we have seen that there is evidence for differences in accuracy between P2 and P3, but not between pairs. On the other hand, in terms of contrast difference effects, we discovered that the lower the contrast (i.e., the harder it is to differentiate between the stimuli), the better the performance; a finding that is counterintuitive. All these results suggested to us to eliminate P1 (Percentage Difference of 5%), because of the near chance results in performance in this condition, and to eliminate contrast as a condition in the main experiment, because of the counterintuitive results from this pilot study. Therefore, in order to manipulate higher and lower discriminability between stimuli we decided to use P2 (Percentage Difference of 7%) and P3 (% size difference 9) and use the highest RGB level possible (84 RGB) for the stimuli, which is completely black. It is on this basis that we designed the stimuli for the main study, where we aimed to explore the possibility that similarity asymmetries might depend on the distinguishability of the compared stimuli.

Main Experiment

In this experiment we performed a series of categorization tasks, a forced-choice similarity task and slider tasks, with the aim to explore the hypothesis that asymmetries of the form $Sim\ (NP, P) > Sim(P, NP)$ disappear when the stimuli are more discriminable (*NP* denotes a non-prototypical stimulus and P a prototypical one). The forced-choice similarity task was the task testing for similarity asymmetries; it was based on Tversky's (1977) classic manipulation. In our task participants were asked whether they preferred a statement along the lines 'stimulus 1 is similar to stimulus 2' vs. 'stimulus 2 is similar to stimulus 1' (Tversky employed a forced choice task, with a very similar structure, i.e., participants were asked 'is Korea like China' vs. is 'China like Korea'). We employed two categorization tasks, which were meant to teach to participants a simple two-cluster category structure. The purpose of the categorization tasks was to make certain some stimuli could be considered *P*, others *NP*; equally, the category structures were such that other stimuli would acquire significance that might be relevant to similarity asymmetries (e.g., some stimuli would be ideals, others diagnostic). Finally, the slider tasks were based on a simple procedure, which were used to test participants' knowledge of various important stimuli in the task (e.g., averages).

We used the results from our pilot study to design two sets of schematic stimuli of nine spirals each and meaningless labels to indicate the intended categories (*Chomps* and *Blibs*).

Participants

Fifty experimentally naive students at City University London received course credit for participating in the study.

Procedure

The experiment, designed in Superlab, lasted approximately 30 minutes. We split participants in two conditions: for half the participants all the stimuli used in the tasks involved stimuli corresponding to percentage size difference P2 (7%) and for the other half stimuli corresponding to percentage size difference P3 (9%).

Participants first had to carry out a supervised learning task. A first learning block presented the stimuli to participants, so that each stimulus was shown with its label (there were two categories, each having nine stimuli; the labels that were used to indicate the two categories were meaningless linguistic labels, Chomps or Blibs); note, stimulus order was randomized. Prior to each stimulus, a fixation cross was shown for 500ms. Two keyboard keys were labeled 'Chomps' and 'Blibs' and participants had to click on the correct category label button for a stimulus, for the trial to end and the next trial to start. This first learning block consisted of 36 trials (there were 18 unique stimuli, presented twice). There was then a test block, with eighteen randomized trials (one for each stimulus), where participants were asked if the stimulus on each trial was a Chomp or a Blib. Participants had unlimited time to answer, by clicking on the Chomp or Blib key. Visual and auditory feedback was provided for 1000ms on each trial and this test block was repeated until participants made no mistake (if participants made even a single mistake, then this test block was simply run again).

Once participants completed this supervised categorization task with no mistakes, they were presented with yet another supervised categorization task, comprised of 18 trials. Participants followed the same procedure as before but during this new test block we asked them to respond as fast as possible, if the stimulus shown was a Chomp or a Blib. Response time was recorded from appearance of the stimulus until the response key was pressed. The test block was repeated three times and feedback was not provided. Note, there is certainly a sense of over-learning for what was an extremely simple category structure (this will be shortly discussed, in the next section), comprising of two well-separated clusters (one with stimuli having smaller diameters, another with stimuli having larger diameters). However, our intention was to have a procedure that would lead to as entrenched a representation of the two intended categories as possible, since, otherwise, it would be arguably meaningless to talk about prototypes or ideals.

We then presented the forced-choice similarity task. This involved three test blocks (with a break of a few seconds between them), with thirty trials each (presentation order was randomized). In each trial, participants were asked to rate the similarity between two stimuli by preferring one of two statements, "1st item is similar to 2nd item" or "2nd item is similar to 1st item". They had to indicate their responses by clicking on the relevant key (1 or 2).

Finally, participants were presented with a simple task, involving a moving slider to check their knowledge of salient category members. Specifically, we asked them: “Can you move this slider to show me what is an average Chomp?”. Then we asked: “Can you move this slider to show me the Chomp that best characterizes the category? We finally asked: “Can you move this slider to show me the Chomp that is most easily distinguished from other Chomps?” We repeated the same procedure for the Blibs category. We assigned arbitrary points to the slider in order to have a rating scale for the responses (Condition P2: 161 points was the value for the prototypical Chomp and 202 points was the value for the prototypical Blib; Condition P3: 162 points for the prototypical Chomp and 204 points for the prototypical Blib).

Materials

We used the stimuli designed in the pilot study, that is, spirals which varied in overall diameter. We used the highest RGB level possible (84 RGB) to design the spirals. The overall distribution of the eighteen spirals in psychological space was such that there were two distinct groups of nine items each (see Figure 1), which corresponded to the Chomps or to the Blibs category. Details of the pairs of stimuli are shown in Table 4 (each pair was presented in both directions, that is, both AB and BA):

stim id	Psych space dim	
1	1	
2	1.25	A_NP_C
3	1.5	
4	1.75	
5	2	P_C
6	2.25	
7	2.5	
8	2.75	D_NP_C
9	3	
10	7	
11	7.25	D_NP_B
12	7.5	
13	7.75	
14	8	P_B
15	8.25	
16	8.5	
17	8.75	A_NP_B
18	9	

Critical Pairs	Filler Pairs
A_NP_C, P_C	1_3
D_NP_C, P_C	4_6
A_NP_C, D_NP_C	7_9
D_NP_B, P_B	10_12
A_NP_B, P_B	13_15
D_NP_B, A_NP_B	16_18
	1_12
	4_15
	7_18

Table 4: Details of the stimuli presented in the main experiment. *A* stands for away, that is, away from the other stimuli in the corresponding category. These stimuli can be considered ideal category members. *D* stands for diagnostic, that is, a stimulus diagnostic with respect to the other stimuli in the corresponding category. So, diagnostic stimuli would be right at the boundary with the other category. *P* stands for prototypical stimuli. In these tables we also indicated whether a stimulus is assumed to be *P* or *NP*. Finally, *C* and *B* index the Chomps or Blibs category.

Results

Regarding the similarity results, we first performed a mixed-design ANOVA with percentage size difference (P2 vs. P3) as a between-subjects factor and type of asymmetry (6 levels, for all possible pairs/asymmetries) as a within-subjects factor. All main effects and interactions were non-significant, p>.05. Even though none of the main effects or the interaction were significant, for purely exploratory purposes we persisted with the intended post hoc single sample t-tests.

So, we considered the two conditions (P2, P3) separately. For each of the six critical pairs of stimuli (see Table 5), we coded preference for the statement "1st item is similar to 2nd item" with a 1 and preference for the statement "2nd item is similar to 1st item" with a 2. Therefore, if there is no asymmetry at all in choice behaviour, we would expect the average score for each statement to be 1.5. Then, we compared scores for each pair with single-sample t-tests (against 1.5), but without a multiple comparisons correction (as it turns out this does not impact on the conclusions).

Regarding P2 (7%), there was no evidence at all for directionality in choice behaviour. For the P3 (9%) condition, there was a single significant t-test, for the pair A_NP_B, P_B (that is, the trial with the Away stimuli vs. the Prototypical one for the Blibs category), t(24) = 2.286, p = .031 (see Table 5).

Note, we averaged choice behaviour for the Chomps pairs and the Blibs pairs. That is, for example, we averaged choice behaviour for the A_NP pair, for Chomps and for Blibs. This was done simply with a counterbalancing motivation: clearly, we are not interested in whether asymmetries may be evidenced in just Chomps, but not Blibs.

Critical Pairs	**Mean Similarity Scores**	
	P2	**P3**
A_NP_C, P_C	M = 1.47, SD = 0.24	M = 1.56, SD = .18
D_NP_C, P_C	M = 1.47, SD = 0.19	M = 1.54, SD = 0.23
A_NP_C, D_NP_C	M = 1.48, SD = 0.22	M = 1.54, SD = 0.27
D_NP_B, P_B	M = 1.43, SD = 0.22	M = 1.54, SD = 0.2
A_NP_B, P_B	M = 1.44, SD = 0.19	M = 1.58, SD = 0.17*
D_NP_B, A_NP_B	M = 1.45, SD = 0.27	M = 1.51, SD = 0.27

Table 5: Mean similarity scores for P2 and P3 conditions. For P2, all results were lower than the similarity score of 1.5, but there were no statistically significant mean differences, p > .05. For P3, all results were higher than the similarity score of 1.5, but there were no statistically significant mean differences, p > .05. The * indicates significance at the 0.05 level.

We finally explored the results regarding the slider task. The results are shown in Figure 4.

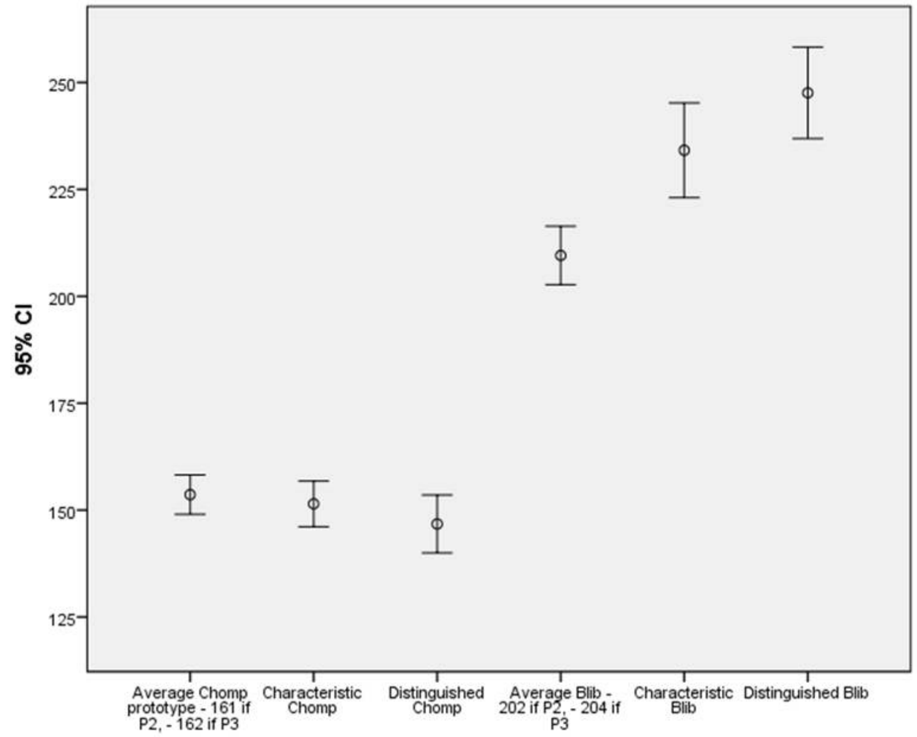


Figure 4: Scatterplot of the values from the rating scale for the three questions asked (for each category) in the slider task.

We correlated the results for Chomps and Blibs separately (ignoring the distinction between P2 (7%) and P3(9%), since this should not impact on these results). The results for both Chomps and Blibs are very similar, indicating high correlations all round, with Average correlating most highly with Characteristic (see Table 6). These correlations show that perhaps it is less meaningful to expect differences between e.g. the most distinguishing Chomp from the average Chomp, in the context of this task.

Variables *(N=50)*	1	2	3	4	5	6
1. Average Chomp	-					
2. Characteristic Chomp	.71**	-				
3. Distinguishing Chomp	.69**	.43**	-			
4. Average Blib				-		
5. Characteristic Blib				.73**	-	
6. Distinguishing Blib				.67**	.56**	-

Table 6: Correlations for the three questions asked (for each category) in the slider task.

**. Correlation is significant at the 0.01 level (2-tailed).

For both the Chomps and the Blibs categories, the estimates participants provided for the average Chomp and the average Blib were very close to the real values. Note, however, that the estimates for the average Chomp and Blib were actually significantly different from the true value. This was established with single sample t-tests against the true values, which were significant for both Chomps (t(49) = -3.23, p < .05; M = 209.54, SD = 24.06) and Blibs (t(49) = 2.21, p < .05; M = 209.54, SD = 24.06). These results suggest that perhaps the lack of asymmetry effects may be due to participants not learning the intended categories quite with the degree of fluency that we were anticipating. However, rejecting participants with estimates for the average Chomp and average Blib two standard deviations or more away from the corresponding means, and rerunning the analyses, did not qualitatively alter any of the above results.

Conclusion

In this chapter we pursued an exploratory direction regarding similarity asymmetries and possible extensions for the QSM. Our starting points were similarity asymmetries of the form $Sim\ (NP, P) > Sim(P, NP)$, where the stimuli are simple perceptual ones, so that no differences in degrees of knowledge are expected and with a manipulation to alter distinguishability of the stimuli.

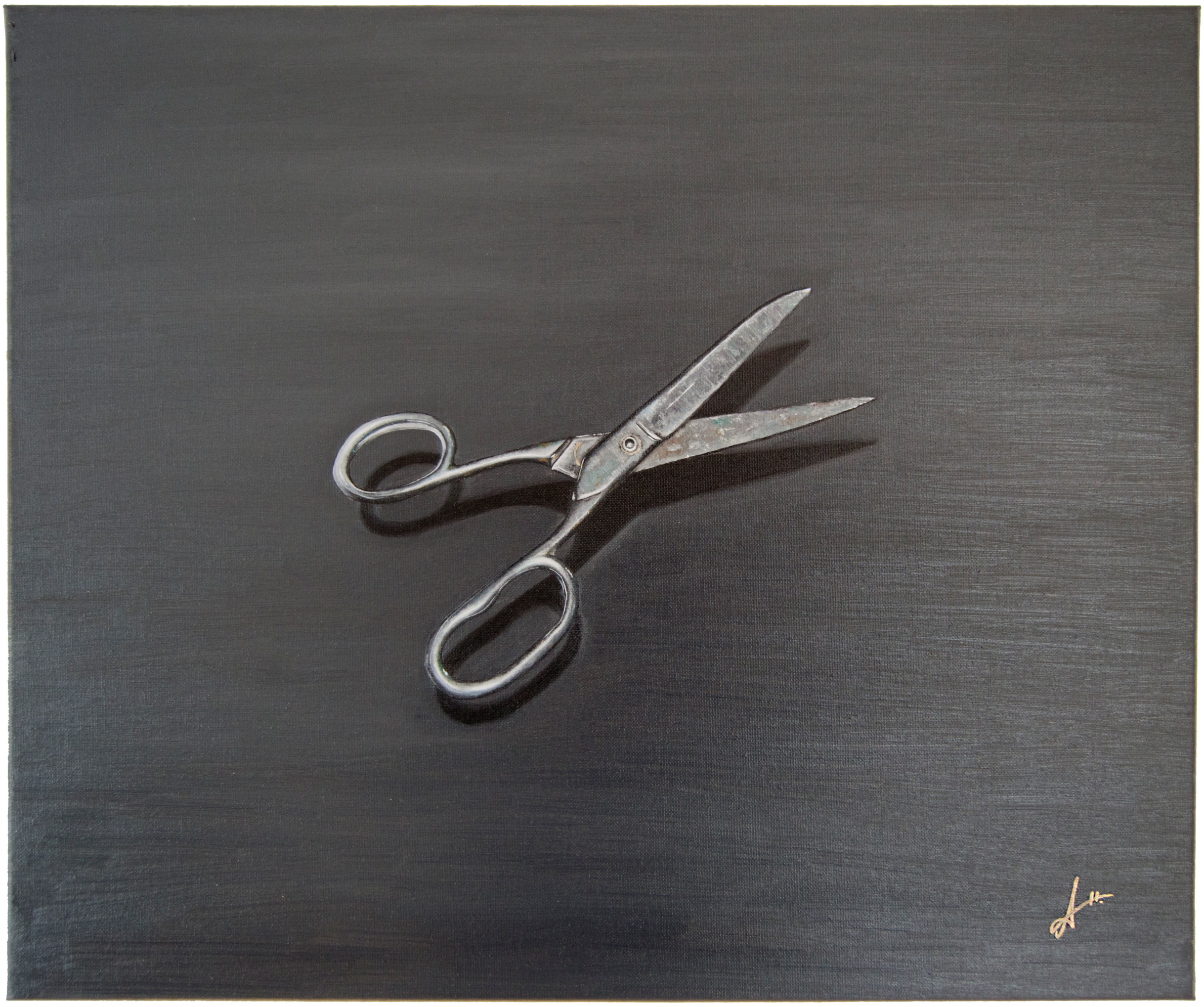
Stimulus design was guided by results from three pilot studies. For the main experimental study, stimuli were generated to conform to simple, two-cluster category structure. The two clusters had an easily identifiable category boundary, which we believed would make it more likely that prototypical structure would emerge. Then, we explored our hypothesis in the Main Experiment, where we performed a series of categorization tasks, a forced-choice similarity task and a slider task.

Overall, the results from our Main Experiment pointed to the conclusion that for both percentage size difference conditions there were no significant preferences for statements in one direction to statements in another direction, that is, that there were no asymmetries in similarity judgments. The results showed an asymmetry only in one case, where the preference was towards the prototypical stimuli. Specifically, the statement *Away stimulus* to *Prototypical stimulus* was preferred than the converse. However, on the basis of only one result for the stimuli from only one of the two clusters, it is not possible to really draw any conclusions. Our results really do make unlikely any evidence for asymmetries, at least given the present procedure. A sceptic might argue that perhaps participants failed to learn the intended categories to a degree sufficient for stimuli to emerge as prototypes (or ideals, etc.), but given the extensive learning procedure, this seems somewhat unlikely. Another related concern is that maybe participants did learn the categories, but rather than develop prototype-based representations for the categories, they represented them in an exemplar or rule-based way. These possibilities cannot be easily dismissed and would require far more extensive methodologies to fully address.

Special Artwork Chapter

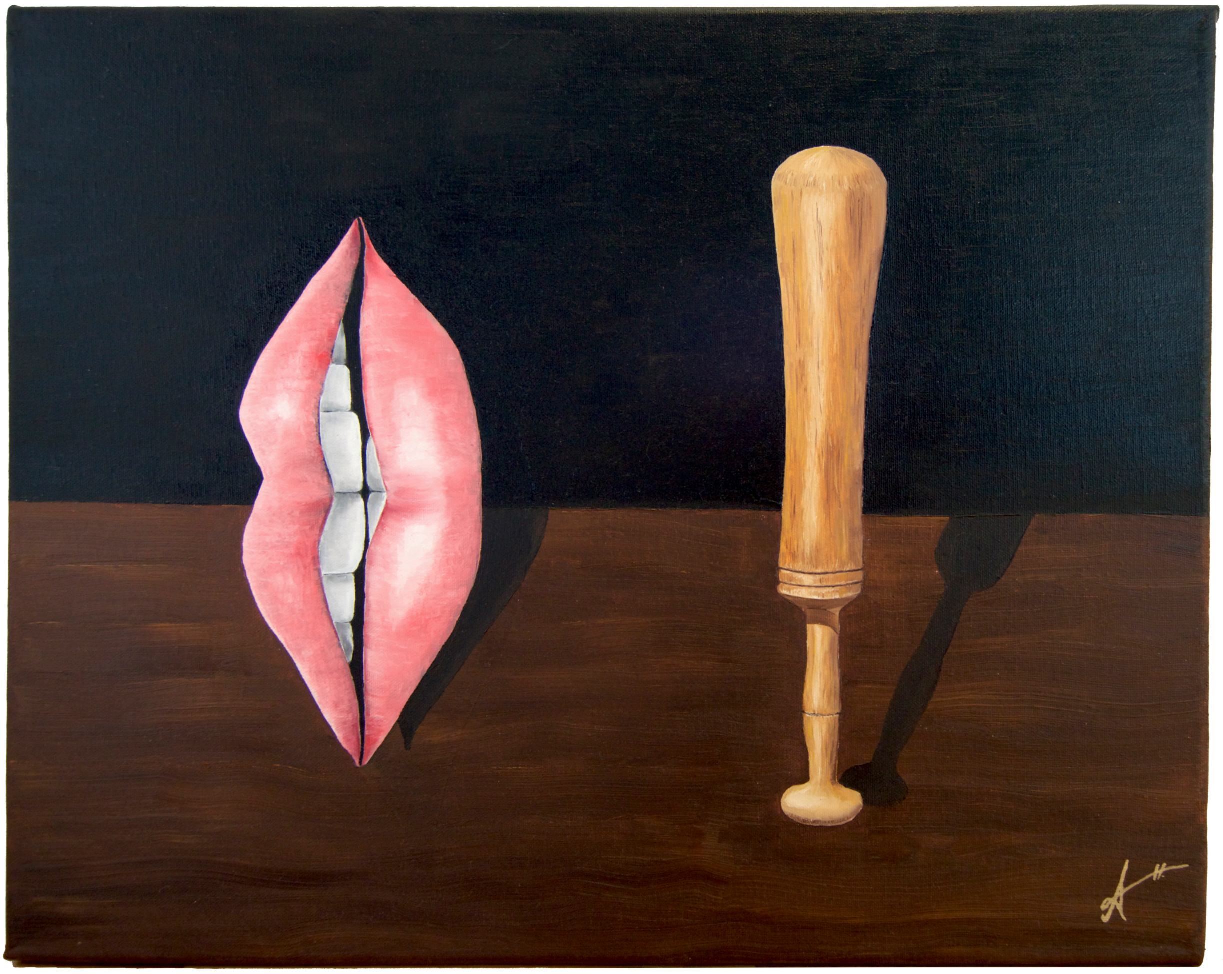
Artist Statement:

What do Cognitive Science and Surrealism have in common? My work proposes a reinterpretation-actualisation of the surrealist movement through the contemporary knowledge about the human mind. My paintings are inspired by my scientific research at City University London. These works discuss the conceptual excesses, the melancholy, the wonder, the reflection and the sensitive violence. This work arises from the pursuit of scientific objectivity, but expressed figuratively through experiential subjectivity.



***Summer Institute on Bounded Rationality* 2015 – Tribute Painting**
Max Planck Institute for Human Development, Berlin (Germany)
Oil on canvas (60 x 49.5cm)
Private Collection

A pair of metallic scissors suspended in the air. Why?
There is a concept in our field of research that we call bounded rationality, which is the idea that when individuals make decisions, their rationality is limited by the available information, the tractability of the decision problem, the cognitive limitations of their minds and the time available to make the decision. Decision-makers in this view act as satisficers, seeking a satisfactory solution rather than an optimal one. Herbert A. Simon, a leading academic in the field, used the analogy/metaphor of a pair of scissors, where one blade represents “cognitive limitations” of actual humans and the other the “structures of the environment”, illustrating how minds compensate for limited resources by exploiting known structural regularity in the environment.



The Origin of Species (2015)
Oil on canvas (50.8 x 40.6cm)
Private Collection

Theme 2

A Quantum Cognitive Approach on Constructive Judgments

Statement of Contribution

Theme 2 is a collaborative work with Emmanuel Pothos and Lee White. The author contributed to the design of the studies, the analyses of the results, and the writing of the manuscripts; he collected all the data of the studies from the manuscripts published.

List of publications for *Theme 2*:

White, L., Barque-Duran, A., Pothos, E. (2015) An investigation of a quantum probability model for the constructive effect of affective evaluation. *Philosophical Transactions of the Royal Society A.* 374, 20150142.

Chapter 5

The constructive effect of affective evaluation

Abstract

In this chapter we first explore the work from White et al. (2014) on the constructive role of articulating an impression, for a presented visual stimulus. Second, we review the Quantum Probability (QP) cognitive model that formalizes such constructive processes, in that work. Finally, as we conclude with some outstanding methodological questions in relation to this previous research, this chapter reports the results of three experiments designed to resolve these questions. Experiment 1, using a binary response format, provides partial support for the interaction predicted by the QP model and Experiment 2, which controls for the length of time participants have to respond, fully supports the QP model. Finally, Experiment 3 sought to determine whether the key effect can generalize beyond affective judgments about visual stimuli. Using judgments about the trustworthiness of well-known people, the predictions of the quantum probability model were confirmed. Together these three experiments provide further support for the quantum probability model of the constructive effect of simple evaluations.

Introduction

One of the main themes that has emerged from behavioural decision research is the view that people's preferences are often constructed in the process of elicitation. Sometimes it seems that the process of choosing one alternative over another alters their relative qualities. For example, selecting a particular alternative appears to increase our preference for this option (e.g. Ariely & Norton, 2008; Kahneman & Snell, 1992; Payne et al., 1993; Sharot et al., 2010; Sherman, 1980; Slovic 1995). This phenomenon, whereby the process of choosing actually influences the subsequent decision, is known as the constructive effect of choice. This concept is derived in part from studies demonstrating that normatively equivalent methods of elicitation often give rise to systematically different responses. These "preference reversals" violate the principle of procedure invariance that is fundamental to theories of rational choice and raise difficult questions about the nature of human values. If different elicitation procedures produce different orderings of options, how can preferences be defined and in what sense do they exist? Describing and explaining such failures of invariance will require choice models of far greater complexity than the traditional models. This chapter will aim at this.

Take as a convincing example the studies provided by Sharot et al. (2010). These authors had participants select between two holiday destinations. After first rating how happy they would be at various destinations, participants then made a blind choice between destinations (they were told that the study concerned subliminal decision-making). Subsequently they were informed which destination they had chosen, before participants again rated the destinations. The results showed a choice-induced change in preference and furthermore no such effect was observed when participants were given a choice from a computer.

The idea that judgments can be constructive is not novel, but there has been controversy regarding the particular origins of the effect. In this chapter we start by reviewing the work from White et al. (2014), who introduced a novel approach to this issue, namely that there is a fundamental limitation in how uncertain information is cognitively represented (e.g., our preference for alternatives in relation to a particular choice). Then, a choice or judgment can be constructive, simply because of how potentialities regarding different options translate into a certainty for a particular option. These ideas are formalized with a Quantum Probability (QP) model, first introduced by White et al. (2014), but which is employed and tested in the research reported in this chapter as well. One of the important predictions from the QP model is that constructive effects may be present just for simple affective evaluations, so that simply articulating how one feels about a positively or negatively valenced stimulus also leads to constructive effects. All this work, while promising, raised some key questions, notably regarding the robustness of the finding and the extent to which it generalized to other kinds of stimuli. For this reason, in this chapter we also describe some methodological restrictions to the original paradigm and we present three new experiments we ran as an attempt to resolve those questions.

Past experiments and paradigms

In White et al. (2014)'s experiments, fictitious adverts for insurance and mobile phones were created which had positive or negative content. The valence of images was either confirmed in a pilot study or images were selected from the Geneva Affective Picture Database (GAPED; Dan-Glauser and Scherer KR. 2011), a database which contains images whose valence has been externally validated.

We will briefly describe the details of the three experiments White et al. (2014) ran, but to summarize, they employed a 2x2 within subjects design, where participants were asked to view two images which were displayed sequentially in either a positive and then negative order (PN condition) or vice versa (NP condition; see Figure 1). In the double rating condition participants were asked to give a simple affective rating for the first image in the sequence and were then again asked for a rating for the second image. In the single rating condition, they saw the first image but provided no rating, instead moving on to view and rate the second image.

Experiment 1: The influence of an intermediate evaluation on mixed adverts.

With Experiment 1, their aim was to establish the effect of interest: does the act of articulating an impression for the first image impact on the rating for the second image? As stated in White et al. (2014), they chose the first stimulus as a single image advert and the second as that image augmented with another image of opposite affect, to create a mixed advert with the aim to examine the impact of an intermediate measurement, in identical stimulus presentation orders (see Figure 1). For each advert, when asked, participants should answer the question ‘how does this advert make you feel?’, responding on a nine-point scale, with anchors “1: very unhappy to 9: very happy”.

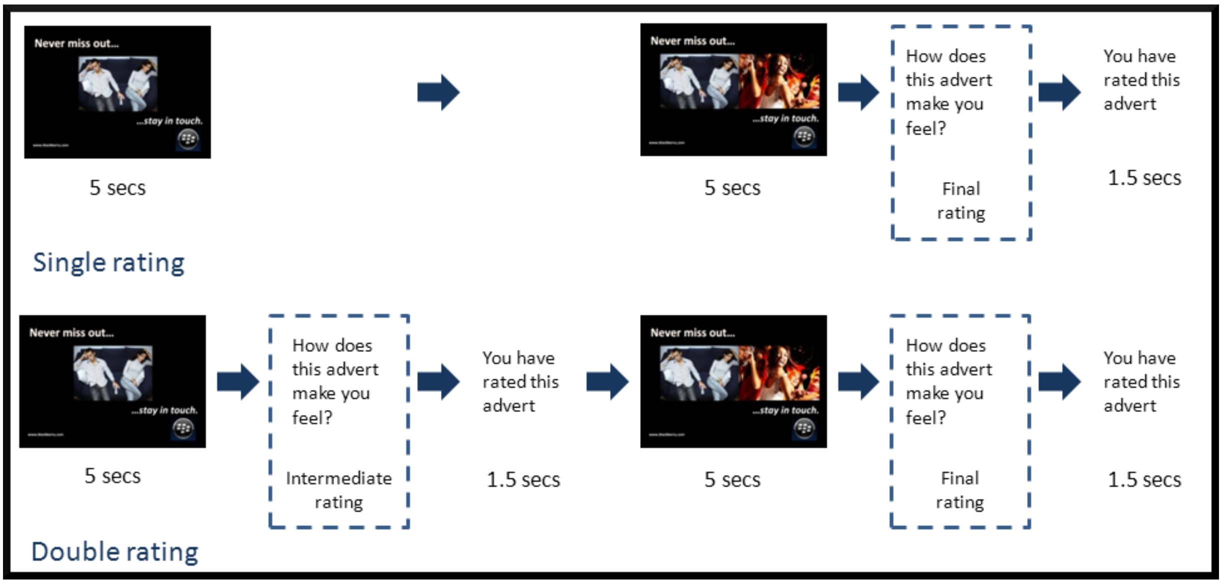


Figure 1: Sample adverts used in the NP condition and the procedure for presentation of single and double rated adverts used in Experiment 1. (Source: White et al. (2014).

Sometimes it does hurt to ask: the constructive role of articulating impressions. Cognition, 133, 48-64. Adapted with permission from the authors.)

Experiment 2: The influence of an intermediate evaluation on single image adverts.
The procedure of this experiment was as in Experiment 1, except that all adverts included single images instead of mixed adverts. They used positive and negative images from Experiment 1 but they incorporated new images, previously piloted to create realistic-looking adverts. Participants were distributed following the same conditions as before and they answered in the same way, responding on a nine-point scale.

Experiment 3:
The aim in Experiment 3 was to replicate the main result of Experiments 1 and 2 with different materials and slightly different procedures. Specifically, regarding the between-subjects control manipulation, after the presentation of the first advert, some participants were shown a rating of an allegedly random participant and asked to confirm the rating by pressing the appropriate key. Some other participants were simply told that the computer had rated the advert but were not told the rating. So, all participants were tested with one of the control manipulations (random participant rating or computer rating), as well as the main experimental manipulations.

Overall, in all three experiments reported by White et al. (2014), they obtained the same main result, which showed that, when two stimuli are presented in identical orders, the presence of an intermediate affective judgment can impact on the last judgment. In other words, whether or not someone articulated an affective evaluation for the first image influenced how participants rated the second image. Specifically, when participants saw images in the PN condition, the ratings for the second negative image in the single rating condition were significantly more positive than the ratings of the same image in the double rating condition. In the NP condition, the ratings of the second positive image in the single rating condition were significantly less positive than ratings of the same positive image in the double rating condition. Thus, in both conditions, it appeared that the intermediate rating increased the affective contrast between the two images (see a summary of results in Figure 2).

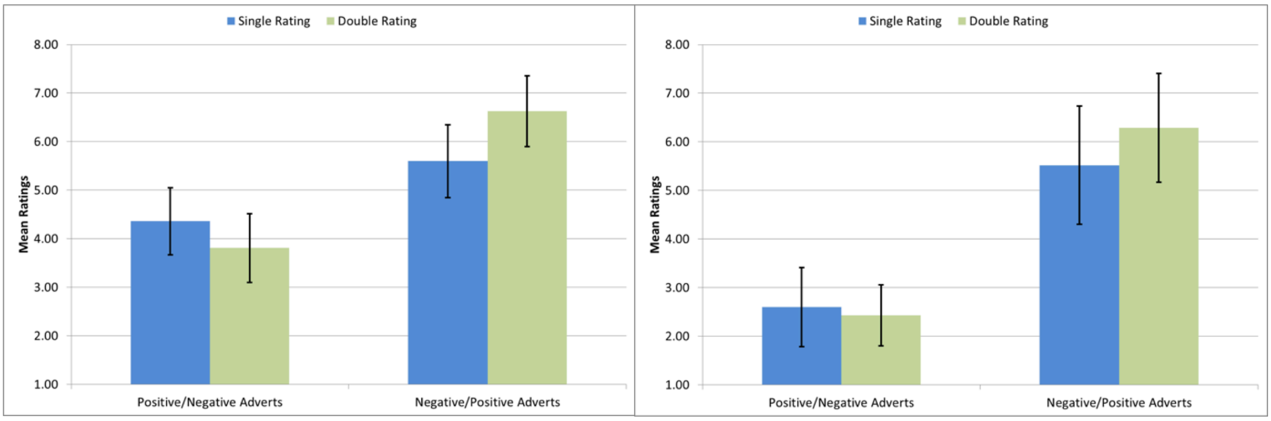


Figure 2: Experiment 2 and 3 results, respectively. Mean participant ratings of single and double rated PN and NP adverts (error bars represent standard deviations). (Source: White et al. (2014). Sometimes it does hurt to ask: the constructive role of articulating impressions. Cognition, 133, 48-64. Adapted with permission from the authors.)

Furthermore, White et al. (2014) argued that this result could not be explained by other approaches, such as order effects, anchoring or Hogarth and Einhorn's (1992) belief-adjustment model, which are in principle applicable in situations concerning the impact of intermediate judgments. Instead, the authors argued that the cognitive model based on QP, which we will present in the next section, could predict the empirical results they observed.

A QP model for the constructive effects of affective evaluation.

Throughout this thesis, we have presented several examples on the use of QP in cognitive modelling and its applications ranging across decision-making (Busemeyer et al., 2011; Trueblood and Busemeyer, 2011; Want et al., 2014), similarity (Pothos et al., 2013; Goldstone, 1994; Medin, Goldstone & Gentner, 1993;), memory (Bruza et al., 2009), concept combination (Aerts, 1995) and other areas (Atmanspacher, 2010).

We have previously argued that QP has some unique features, with no equivalents in CP, such as incompatibility, entanglement and superposition. In this section and in the whole chapter we will focus on the latter, as it offers a natural and straightforward way to model constructive processes in judgment and decision-making. Classical models in decision theory naturally assume that, as the cognitive state changes from moment to moment, at any specific moment it is considered to be in a definite state (even if this state is unknown). Alternatively, QP models allow the cognitive system, at each moment, to be in a superposition state regarding a question (or a stimulus), which reflects ontic indefiniteness for the question outcomes (or feelings about the stimulus) – that is, the question outcomes do not exist, prior to a measurement. Superposition is a technical term in QP and indicates a special kind of uncertainty, such that the cognitive system has the potential for any of the possible decisions at each moment, but which one is selected cannot be determined until the system is measured (in our case, when a judgment or affective evaluation is made). According to models based on standard CP theory, the measurement taken of a system reflects the state of the system immediately prior to the measurement. However, in QP theory, taking a measurement of a system can create, rather than record, a property of the system (Peres, 1998), which means that the subsequent state of the cognitive system is constructed from the interaction between the superposition state and the measurement taken (Bohr, 1958).

Here, we briefly review the model devised by White et al. (2014) and we provide a detailed version of the model in Appendix 2, where we provide a simple illustration of the model that shows how the key prediction of the QP model emerges.

The QP model predicts a difference in the ratings of a positively or negatively valenced visual stimulus, depending on whether a previous unrelated oppositely valenced stimulus was rated or not. The model leads us to the following insight into the psychological processes that underpin the observed effect. The participant's initial cognitive state is set by the first image in the sequence. Following the intermediate affective evaluation of the first image, the cognitive state is changed to being one corresponding to either positive or negative affect. This change is represented in the model by a collapse of the state vector onto either a positive or negative affect ray, which represents a positive or negative affect. This collapse can be explained as an abstraction process, whereby some of the information about the first image is forgotten and attention is focused on information related to its affective properties. It is also the critical constructive step in the model: the intermediate rating changes the mental state in a certain way. This means that having made the intermediate rating, when the second oppositely valenced image is presented, it is evaluated from the perspective of a different cognitive state, than it would have been without an intermediate rating. As the second image is opposite in valence to the first, when the cognitive state is a pure affective one, there is a greater contrast in the impression made by the second image. Without the intermediate rating, the differences between the images concern aspects of their affective quality, but also differences between the images that are not related to affect, so the affective contrast between the first and second image is less pronounced. It is in this way that the QP model prediction arises, that the intermediate rating increases the affective contrast between the two stimuli.

Methodological developments to the present paradigm.

We identified three methodological questions with respect to the experiments and findings we summarized from White et al. (2014). In this chapter, we attempted to resolve them by running three different experiments (this work is reported in White, Barque-Duran, & Pothos, 2015):

Experiment 1: The influence of an intermediate evaluation using a binary judgment.

The first problem concerned the specification of the QP model. Given that the experiments employed a nine point rating scale with anchors "1: very unhappy to 9: very happy", ideally judgments should be represented by a nine-dimensional vector space in the QP model, rather than the simplified two-dimensional vector space, used in White et al. (2014). A demonstration of the same result, when participants are required to make a simple binary choice between being either happy or unhappy in response to the images, would provide further support for the model and this was the focus of Experiment 1 (see details below). We predicted that, for example, in the NP condition, there would be participants who indicate that the second advert makes them feel happy in the double rating condition, and also indicate that the same advert makes them feel unhappy in the single rating condition.

Experiment 2: Controlling for the amount of time to process the first stimulus.

A second methodological question concerned the amount of time that participants had to process the images in the different conditions in the three experiments reported in White et al. (2014). In the double rating condition, they saw the first image for five seconds and then had no limit on the amount of time they could take before providing their response. But in the single rating condition, they just saw the first image for five seconds, before being presented with the second image. So a difference in ratings might arise from the fact that people process the first image in the double rating condition for longer, perhaps increasing the likelihood of more deliberative or strategic processing. Such additional processing possibly implies that the image would leave a stronger impression and have greater saliency or become more accessible, as a point of reference, when considering the second image. In other words, it is possible that that the rating of the second image in our experiments may be affected by a process of affective priming of the first image, more so than in the single rating condition, because of the extra time the image is processed in the double rating condition. But the research on affective priming, which is the finding that the processing of an affective stimulus can be faster and more accurate when preceded by stimulus of the same valence as opposed to an oppositely valenced prior stimulus, suggests us the following: the influence that the first image, in our experiments, has on the second image, is actually not dependent on whether the first image is processed for a long time (Barh et al., 1992; Damasio, 1994; Duckworth et al., 2002; Fazio et al., 1986; Greenwald et al., 1989; LeDoux et al., 1996; Zajonc, 1980). Instead, affective priming research indicates that the affective content can be processed relatively quickly and, in spite of the actual speed in which it is processed, can still have an influence on subsequent judgments. In our experiment, it seemed reasonable to suppose that, in either condition, initial exposure would lead participants to rapidly form an affective impression of the image. But it was also possible that, as participants had longer than 500 milliseconds to view the first image, affective priming was not relevant, as the longer time scale provided them with ample time to process the image more deliberately.

It should be clear that the current literature provides a somewhat unclear picture of how affective priming could or could not impact on the second rating, depending on whether a first rating was provided or not. It should also be clear that it is desirable to rule out the length of time that participants had to process the first image as an explanation for the key effect. This was the purpose of Experiment 2 (see details below), which controlled the amount of time people had to process the first image and make their ratings. The predictions were the same as those in Experiment's 1 – 3 in White et al. (2014); in the PN condition, the second image would be more likely to be rated negatively in the double rating condition than in the single rating condition and vice versa for the NP condition.

Experiment 3: The influence of an intermediate evaluation on judgments of celebrity trustworthiness.

Our third critical methodological concern was related to the generalization of the QP predictions and whether we could use some other kind of stimuli and judgments, which would in turn inform about the boundary conditions in the applicability of the model. In Experiment 3 (see details below) we decided to use judgments of the trustworthiness of celebrities and well-known people. The stimuli used were facial images and studies have found that, even when faces are unfamiliar to participants, there is a large degree of consensus between participants about the trustworthiness of those faces (Engell et al. 2007), even for strangers (Rule et al., 2013; Porter et al., 2008). This research suggested that people should easily be able to make judgments about the trustworthiness of celebrities[4], given that making such a judgment is a basic human ability. We expected that whether or not someone provided an intermediate rating of trustworthiness, regarding the first celebrity, would influence their rating of the second celebrity. So, for the PN condition, where a more trustworthy celebrity was seen first, before viewing a less trustworthy celebrity, an intermediate rating of trustworthiness for the first celebrity would result in a less trustworthy rating for the second celebrity, than if the first celebrity was not rated. Note, we use 'P' for 'trustworthy' and 'N' for not trustworthy, by analogy to the other experiments in this chapter. The prediction for the NP condition, was reversed, in that the intermediate judgment would make the second celebrity appear more trustworthy. In both cases, the intermediate evaluation was predicted to increase the difference in the perception of trustworthiness for the two celebrities.

Experiment 1

The aim of Experiment 1 was to change participants' response measure from a nine point rating scale with anchors "1: very unhappy to 9: very happy" (used in White et al. 2014), to a simple binary choice between being either happy or unhappy in response to the images presented. We predicted that for the NP condition, the probability that second adverts in the double rating condition elicit a happy response would be greater than in the single rating condition. Analogously, for the PN condition, we predicted that the probability that second adverts in the double rating condition elicit a happy response would be lower than in the single rating condition.

Participants and design

[4] The use of celebrities as stimuli is not novel in QP cognitive research. A focus of QP modelling has been Moore's result (Moore, 2002). Moore, using Gallup poll data, found that the American Vice President Al Gore was rated as being less honest and trustworthy, if the previous question was about the honesty and trustworthiness of President Bill Clinton. This order effect can be described using a QP model (Wang et al., 2012; Wang & Busemeyer, 2013). The QP explanation involves the idea that the first stimulus and the participant's response both provide a context, against which the judgment about the second stimulus is made.

Forty City University London students participated in the experiment for course credit (31 women, average age 22.03 years). We employed a within-subjects design with two independent variables: advert order (PN, NP, neutral) and rating (single, double). The inclusion of a neutral condition for advert order (to mean one neutral stimulus was presented after another neutral one) was the only difference between this experiment and White et al. (2014)'s original experiments. It was thought that these stimuli might serve to accentuate the positivity or negativity of the other stimuli.

Stimuli

The same positive and negative images from White et al. (2014)'s Experiment 2 were used but rather than using the same filler adverts as in the previous experiment, we created a new set of adverts for a camera which involved neutral images. These neutral images were drawn from GAPED. The neutral stimuli were evaluated in the same way as the experimental stimuli (i.e. single and double rated). Stimulus materials were presented using Superlab.

Procedure

The procedure was the same as that employed in White et al. (2014) (see Figure 1) with only the kind of rating of the adverts being different. Participants were told that they would see several adverts and that for each advert, when asked, they should answer the question 'how does this advert make you feel?', by pressing the appropriate key to indicate one of two possible choices, "Z: Happy or M: Unhappy" (keys were appropriately labelled). Trials were organized into two blocks. One block contained the six single rating PN smartphone adverts, six double rating PN insurance adverts, six single rating NP insurance adverts, six double rating NP smartphone adverts, six single rating neutral camera adverts and six double rating neutral camera adverts. The other block contained the same adverts, but switching the requirement for single vs. double rating. Block order was counterbalanced between participants and trial order within blocks was randomized.

Results

As for White et al. (2014)'s previous experiments, as the valence of the images had been established in the pilot study, we excluded four participants whose ratings for the first rated images in the double rating condition were over one standard deviation below the mean for positive adverts (*M*=0.87, *SD*=0.23) and one standard deviation above the mean for negative adverts (*M*=0.13, *SD*=0.21).

Happy responses were coded “1” and unhappy responses were coded “0”. We conducted a three (advert order: PN, NP, neutral) × two (rating: single, double) repeated measures ANOVA on participant ratings for the second adverts. There was a main effect of advert order (*F*(2,70)=93.23, *p*<.001), but not of rating (*F*(1,35)=0.13, n.s.). Importantly, the advert order × rating interaction was significant (*F*(2,70)=4.74, *p*=.012). Paired samples t-tests showed that in the NP condition, the positive advert was more likely to be rated positively, when there was an intermediate rating (*M*=0.93, *SD*=0.10), than without an intermediate rating (*M*=0.86, *SD*=0.25; *t*(35)=-2.18, *p*=.035; *d*=0.37). For the PN condition, the second negative advert was more likely to be rated negatively, when there was an intermediate rating (*M*=0.18, *SD*=0.22), than when there was no intermediate rating (*M*=0.23, *SD*=0.25) but not significantly so (*t*(35)=1.41, *p*=.17; *d*=0.22). In the neutral condition, there was no significant difference between single rated (*M*=0.37, *SD*=0.25) and double rated (*M*=0.35, *SD*=0.24) second neutral adverts (*t*(35)=1.19, n.s.). With the exception of the non-significant trend (but in the right direction) for the PN order, these results replicate White et al. (2014).

Experiment 2

The aim in Experiment 2 was to control the amount of time people had to process the first image and made their ratings. The predictions were the same as those in Experiment’s 1 – 3 in White et al. (2014); in the PN condition, the second image would be rated more negatively in the double rating condition than in the single rating condition and vice versa for the NP condition.

Participants, design and stimuli

Twenty-five City University London mostly undergraduate students participated in the experiment for course credit (15 women, average age 24.84 years). We employed a within-subjects design with two independent variables: advert order (PN, NP) and rating (single, double). The same stimuli as used in White et al. (2014)’s Experiment 2 were used in this experiment.

Procedure

The timings for the presentation and rating of all adverts were controlled (see Figure 3). Based on an analysis of the reaction times for rating adverts in White et al. (2014)’s Experiment 2 (*M*=3259 milliseconds, *SD*=2412 milliseconds), in the current experiment, participants were given 5000 milliseconds to view the first image in the double rating condition, followed by 3300 milliseconds to rate it. If participants took longer than 3300 milliseconds to rate the image, they were presented with a message informing them that they had been too slow and they proceeded to the next image, without rating the first. In the single rating condition, they were given a total of 8300 milliseconds to view the first image. The same timings were used when participants rated the second image in both single and double rating conditions. In all other respects, the procedure, including ordering of trials, block order and counter balancing was identical to that used in in White et al. (2014)'s Experiment 2.

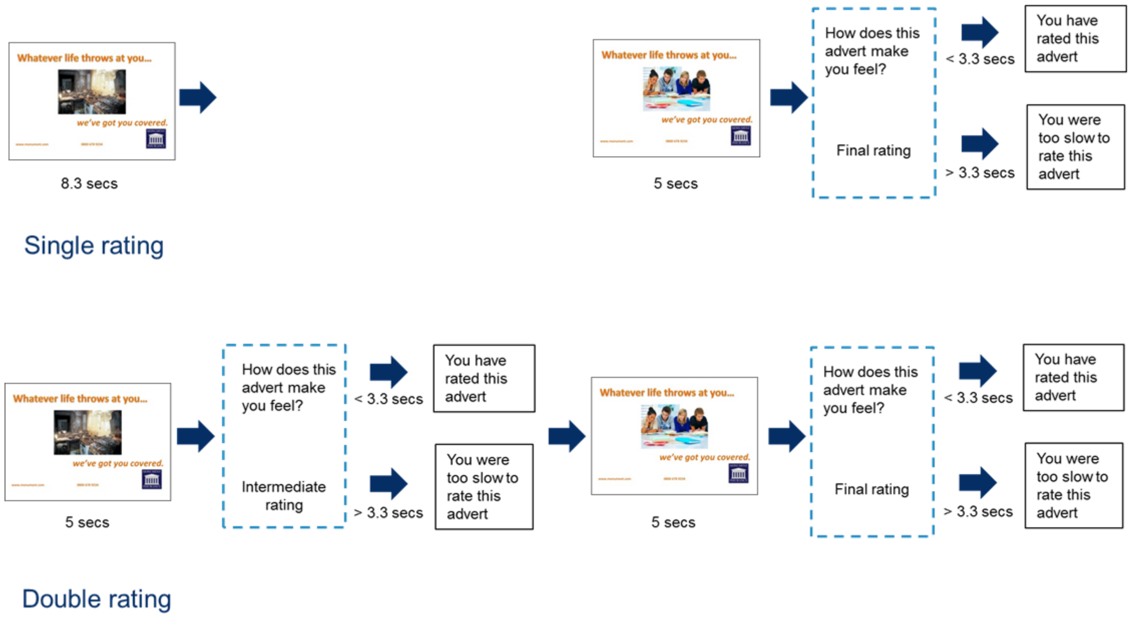

Whatever life throws at you...
we've got you covered.
8.3 secs
Single rating
Whatever life throws at you...
we've got you covered.
5 secs
How does this advert make you feel?
Final rating
< 3.3 secs
You have rated this advert
> 3.3 secs
You were too slow to rate this advert
Whatever life throws at you...
we've got you covered.
5 secs
How does this advert make you feel?
Intermediate rating
< 3.3 secs
You have rated this advert
> 3.3 secs
You were too slow to rate this advert
Whatever life throws at you...
we've got you covered.
5 secs
How does this advert make you feel?
Final rating
< 3.3 secs
You have rated this advert
> 3.3 secs
You were too slow to rate this advert
Double rating

Figure 3: Procedure for Experiment 2: sample advert used in NP condition and procedure for presentation of single and double rated adverts. (Source: White et al. (2015). An investigation of a quantum probability model for the constructive effect of affective evaluation. Phil. Trans. R. Soc. A.)

Results

Trials in the double rating condition in which participants failed to respond in time to the first image were eliminated from analysis. One participant was too slow on 17 out of 72 trials (23.5%), and so the calculation of average ratings for PN and NP single rated adverts was not possible. This participant was not included in further analyses. As for previous experiments, as the valence of the images had been established in a pilot study, we excluded one participant whose ratings for the first rated images in the double rating condition were over one standard deviation below the mean for positive adverts (*M*=6.48, *SD*=1.20) and one standard deviation above the mean for negative adverts (*M*=3.36, *SD*=1.19).

We conducted a two (advert order: PN, NP) × two (rating: single, double) repeated measures ANOVA on the ratings for the second adverts. There was a main effect of advert order (*F*(1,22)=69.51, *p*<.001), but not of rating (*F*(1,22)=3.22, *n.s*). The advert order × rating interaction was significant (*F*(1,22)=12.51, *p*=.002). Paired samples t-tests showed that in the NP condition, the positive advert was rated more positively when there was an intermediate rating (*M*=6.76, *SD*=1.25) than without an intermediate rating (*M*=6.07, *SD*=1.46; *t*(22)=3.77, *p*=.001; *d*=0.79). For the PN condition, the second negative advert was rated more negatively when there was an intermediate rating (*M*=3.34, *SD*=1.25) than when there was no intermediate rating (*M*=3.77, *SD*=1.16; *t*(22)=-2.59, *p*=.017; *d*=0.55). These results exactly replicate White et al. (2014), showing that length of exposure or processing time is not a viable explanation for the key QP prediction.

Pilot Study Experiment 3

The stimuli used in Experiment 3 were the images and names of pairs of celebrities drawn from various areas of public life e.g. music, politics and sport. The celebrities were selected on the basis that they were sufficiently related, so that one would expect the trustworthiness of one to change our perspective for the trustworthiness of the other and so that one celebrity would be regarded as more trustworthy than the other. The purpose of this pilot study was to collect data on the trustworthiness of the selected celebrities, which is essential in order to realize the necessary design.

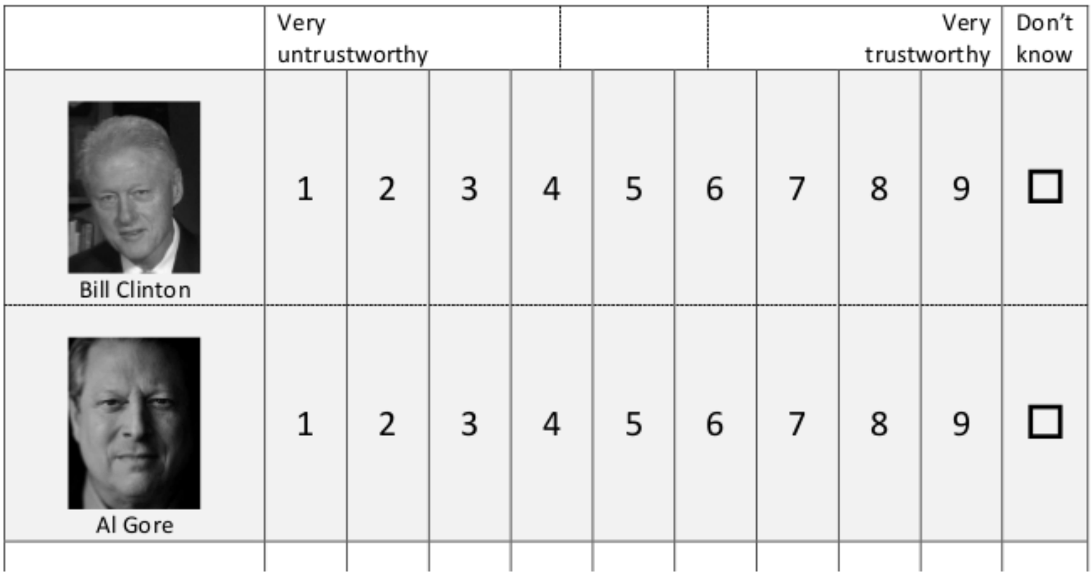
Participants and Design

Seventeen Swansea University students participated for course credit (16 women, average age 19.7 years).

Stimuli and procedure

Twenty seven pairs of celebrities were collected from various internet sources as being likely, in the experimenter’s estimation, to show differential levels of trustworthiness. Images were selected that showed the celebrity looking directly at the camera and with a neutral, non-emotional expression (e.g., not smiling). The images, as in other experiments (e.g., Brehm & Miron, 2006; Rule & Ambady, 2006), were cropped to the celebrity’s head and shoulders, converted to grayscale and scaled to the same size.

Using these images a questionnaire was constructed (see Figure 4), asking people to rate the trustworthiness of each celebrity on a 9 point scale (1 is “Very untrustworthy” and 9 is “Very trustworthy”). Images of celebrities were presented in their intended pairings. There was also an option to say “don’t know”. Participants completed the questionnaire after taking part in other experiments conducted by the experimenter.



	Very untrustworthy								Very trustworthy	Don’t know
Bill Clinton	1	2	3	4	5	6	7	8	9	□
Al Gore	1	2	3	4	5	6	7	8	9	□

Figure 4. Sample page from famous faces pilot questionnaire. (Source: White et al. (2015). An investigation of a quantum probability model for the constructive effect of affective evaluation. Phil. Trans. R. Soc. A.)

Results and discussion

The mean rating of trustworthiness for each celebrity and the difference in mean ratings for pairs was calculated. We also calculated how well recognised a celebrity was by summing the responses to the “don’t know” question (see Table 1). The results indicated that five celebrity pairs in particular were not very well-recognised, as compared with the other celebrity pairs, since the corresponding number of don’t knows was over 1 standard deviations above the mean (M=2, SD=3). These pairs (Condolezza Rice & George Bush, Yoko Ono & John Lennon, Ed Balls & Gordon Brown, Bill Clinton & Al Gore, and Ed Milliband & David Milliband) were eliminated. In Experiment 3 we were not interested in whether or not participants could recognise a celebrity, only in how trustworthy they judged the celebrities to be, based on whatever it was they know about them or just their impression of their faces (cf. Rule et al., 2013; Porter et al., 2008). We simply eliminated celebrities that were unfamiliar, because we wanted to ensure a degree of uniformity amongst the stimuli, regarding familiarity. We decided to retain the remainder of celebrity pairs, in spite of the small degree of difference in trustworthiness between some pairs, in order to ensure that we had sufficient numbers of stimuli for the study.

Experiment 3

The aim of this experiment was to establish whether the QP prediction could generalize to at least some other kinds of stimuli and judgments, which would in turn inform the boundary conditions in the applicability of the model. In Experiment 3, we tested whether the QP model applies to judgments of the trustworthiness of celebrities and well-known people using stimuli based on facial images. We first ran the pilot study presented before, with a view to set up comparisons that would provide the low and high trustworthiness contrast that we needed, to examine the prediction of the QP model. This prediction was entirely analogous to that in Experiments 1 and 2.

Participants and design

Given the novelty of the task and the uncertainty about consistency in ratings of trustworthiness of celebrities we recruited more participants than in previous experiments. Eighty-one mostly undergraduate students from Swansea University and City University London participated in the experiment for course credit (69 women, average age 20.15 years). We employed a within-subjects design with two independent variables: order of celebrity trustworthiness (PN, NP) and rating (single, double). We use the same notational convention as in previous experiments to represent high and low levels of trustworthiness. So P represents higher trustworthiness and N represents lower trustworthiness.

Stimuli

To ensure that celebrity pairs would be familiar to participants and that they were differentiated in terms of their perceived trustworthiness, we conducted a pilot study to evaluate each celebrity's trustworthiness. Further details on the pilot can be found in the section *Pilot Study Experiment 3* and the results are shown in Table 1.

High Trustworthy Celebrities		Low Trustworthy Celebrities			
Name	M	Name	M	DK	Difference
Al Gore*	4.25	Bill Clinton*	4.24	9	0.01

Bill Gates	6.71	Steve Jobs	6.67	2	0.04
Ed Milliband*	4.27	David Milliband*	4.21	6	0.05
William Hague	3.14	George Osborne	3.07	4	0.07
Brad Pitt	5.82	Angelina Jolie	5.65	0	0.18
Victoria Beckham	5.53	David Beckham	5.35	0	0.18
Gordon Brown*	3.50	Ed Balls*	3.22	8	0.28
Catherine Zeta Jones	6.06	Michael Douglas	5.59	0	0.47
Zara Phillips	6.00	Mike Tindall	5.29	0	0.71
John Lennon*	5.75	Yoko Ono*	5.00	8	0.75
Beyonce	7.29	Jay Z	6.53	0	0.76
Stephen Merchant	6.43	Ricky Gervais	5.50	3	0.93
Katie Holmes	5.81	Tom Cruise	4.81	1	1.00
Vince Cable	3.92	Nick Clegg	2.88	5	1.03
Condolezza Rice*	4.11	George Bush*	2.94	8	1.17
Prince Charles	5.47	Camilla Parker Bowles	4.29	0	1.18
Dawn French	7.00	Lenny Henry	5.65	0	1.35
Tess Daley	6.56	Vernon Kay	5.13	1	1.44
Gary Barlow	6.82	Robbie Williams	5.12	0	1.71
Paul McCartney	5.59	Heather Mills	3.87	2	1.72
Charlotte Church	5.41	Gavin Henson	3.53	0	1.88
Barack Obama	6.76	Hilary Clinton	4.65	0	2.12
Boris Johnson	5.41	David Cameron	3.24	0	2.18
Coleen Rooney	5.65	Wayne Rooney	3.35	0	2.29
Katy Perry	6.12	Russell Brand	3.41	0	2.71
Peter Andre	5.94	Katie Price	2.88	0	3.06
Cheryl Cole	6.06	Ashley Cole	2.29	0	3.76

Notes: M=Mean rating on a 9 point scale (1 is "Very untrustworthy" and 9 is "Very trustworthy"). DK=number of times that someone responded 'don't know' to one or both of a pair. Difference =difference in Means. *Indicates celebrity pair that was not used in the main experiment. Celebrity pairs are matched by row, so the high and low trustworthiness classification is to be interpreted only within individual rows.

Table 1. Pilot study celebrity trustworthiness ratings.

Within each pair there was a celebrity perceived to be less trustworthy (N) than the other celebrity (P). To mitigate variability in participants' responses, we broadly matched images on colour, clothing or background, and emotional expression. Moreover, following the procedure in related experiments (Brehm & Miron, 2006; Rule & Ambady, 2006), the images were standardised by cropping to the celebrity's head and shoulders, converting to grayscale and scaling to the same size. We also included the name of the celebrity, under their image, to aid recognition.

We constructed a second set of stimuli, which was identical to the first, except that the order of presentation of celebrity pairs was switched. For example, in one set of stimuli, Angelina Jolie was shown first followed by Brad Pitt, as a celebrity pair in the NP condition

(Angelina Jolie was rated as being less trustworthy than Brad Pitt in the pilot). In the second set of stimuli, Brad Pitt was shown first followed by Angelina Jolie, as a celebrity pair in the PN condition. Stimulus materials were presented using Superlab.

Procedure

Participants were randomly assigned to view one of the two sets of stimuli (the sets only differed in the order of faces in each pair), as in other experiments.

Participants were then told that they would be shown various well-known people and that they would be asked to evaluate their trustworthiness. They rated each celebrity's trustworthiness on a 9 point scale, with anchors "1: very untrustworthy to 9: very trustworthy". They were also given the option of pressing "D" (corresponding to "don't know") if they did not know the celebrity at all.

Each trial involved the presentation of a celebrity followed by a request for rating (double rating condition) or not (single rating condition), followed by the second celebrity image and a final request for a rating. Trials were organized into two blocks (within participants). One block contained five single rating PN celebrity pairs, six double rating PN celebrity pairs, six single rating NP pairs and five double rating NP pairs. The other block contained the same pairs, but switching the requirement for single vs. double rating (i.e., participants rated pairs twice, once in the single rating condition, once in the double rating one). Trial order within blocks was randomized.

Results

Of the eighty-one participants who took part in the experiment, seven answered "don't know" to more than 50% of the trials, which meant there was insufficient data to analyse their responses. These seven were eliminated from further analysis[5].

As in previous studies, we checked the ratings to ensure they were in line with the ratings for trustworthiness that had been established in the pilot study. Two celebrity pairs, Angelina Jolie & Brad Pitt and David Beckham & Victoria Beckham were not rated as they had been in the pilot. Angelina Jolie should have been rated less trustworthy than Brad Pitt, but the reverse was observed in the main experiment. Similarly for David Beckham, who should have been rated less trustworthy than Victoria Beckham but was rated as more

[5] There were a number of foreign students taking part in the study, which might explain why some did not know the particularly UK-centric set of celebrities.

trustworthy[6]. As these pairs had been explicitly chosen because they were perceived in a way suitable for the condition they were in, they were eliminated from further analysis.

We then followed the same procedure as was used previously to check the ratings of the first celebrities in the double rating condition. For the first set of stimuli, 11 participants showed ratings that were either over one standard deviation below the mean for trustworthy celebrities (*M*=6.10, *SD*=1.21) or above the mean for less trustworthy celebrities (*M*=4.10, *SD*=1.02). For the second set of stimuli, another 11 participants showed ratings that were either over one standard deviation below the mean for trustworthy celebrities (*M*=5.5, *SD*=1.22) or above the mean for less trustworthy celebrities (*M*=4.40, *SD*=1.25). These 22 participants were eliminated from the analysis, leaving 52 participants.

We conducted a two (order of celebrity trustworthiness: PN, NP) × two (rating: single, double) repeated measures ANOVA on the ratings for the second celebrities. There was a main effect of order (*F*(1,51)=81.19, *p*<.001), but not of rating (*F*(1,51)=0.03, n.s.). The order × rating interaction was significant (*F*(1,51)=7.11, *p*=.01). Paired samples t-tests showed that, with the intermediate rating, the second celebrity was rated less trustworthy in the PN condition, compared to without the intermediate rating (*M*=4.36, *SD*=0.98 vs. *M*=4.54, *SD*=0.94; *t*(51)=-2.23, *p*=.029; *d*=0.3) and the trustworthy celebrity was rated more trustworthy in the NP condition (*M*=6.02, *SD*=0.90 vs. *M*=5.85, *SD*=1.05; *t*(51)=2.23, *p*=.029; *d*=0.3). In other words, the intermediate ratings increased the difference in trustworthiness between the two persons, a result which exactly replicates the findings of White et al. (2014), with judgments of trustworthiness, instead of affective evaluation.

General Discussion

In the first section of this chapter, we described the purpose of this investigation, which follows that from White et al. (2014), that is, to examine whether the process of articulating an (e.g.) affective evaluation for a positively or negatively valenced stimulus, can influence how an oppositely valenced subsequent stimulus is rated. Then, we reviewed the three experiments in the original study. We next described how the use of QP offered a

[6] This shows the variability of public opinion regarding people in the media spotlight. The pilot was conducted in September 2012 and experimental data collected over 2012 and 2013. We can only assume that during that time these particular celebrities had demonstrated behaviour that led the public to perceive them as being more or less trustworthy than they were thought to be in 2012.

relatively simple mechanism by which the constructive effects of making a judgment can be modelled. Finally, we presented a new set of experiments that: (1) addressed some methodological limitations in the White et al. (2014) experiments and (2), extended White et al.'s results with judgments of a completely different kind.

First, the aim of Experiment 1 was to replicate the previous results but change participants' response measure from a nine point rating scale with anchors "1: very unhappy to 9: very happy" (used in White et al. 2014), to a simple binary choice between being either happy or unhappy, in response to the images presented. Our results showed that one of the predicted differences (according to the quantum model) was significant while the other one showed a non-significant trend but in the expected direction. Thus, the results provide partial support for the interaction we predicted. Specifically, in the NP condition, the intermediate rating increased the probability of the second positive advert being rated positively. In the PN condition, the probability of a positive rating for the second negative advert was lower following an intermediate rating than without, but the difference was not significant. No differences were observed between single and double rated neutral adverts.

Second, in Experiment 2, we controlled the amount of time people had to process the first image and made their ratings. We hypothesised that a larger time in processing the first stimulus in the double rating condition, compared to the single rating one, may have been the cause for the difference in participants' ratings. In other words, participants spending more time processing the first image would develop a mental representation with greater saliency. And thus, this would increase the affective contrast of the original stimulus. After controlling this length of time, we also observed the same interaction, as predicted by the quantum model.

Finally, in Experiment 3, where we tested whether the QP model applied to judgments of the trustworthiness of celebrities and well-known people using stimuli based on facial images, we confirmed the hypothesis that there is an effect of an intermediate judgment of trustworthiness, for both the PN and NP conditions. When a more trustworthy celebrity was rated first, the trustworthiness of the second celebrity was lower, than without the intermediate rating. Similarly, when a less trustworthy celebrity was rated first, the trustworthiness of the second celebrity was higher, than without the intermediate rating.

In sum, the results of all three experiments, whose aim was to address some methodological limitations in White et al. (2014), provided further support for the corresponding QP model. Furthermore, the predictions of the QP model in relation to constructive effects were not limited to affective evaluations of visual stimuli but could be extended to different judgments and stimuli, as the results of Experiment 3 suggested. Therefore, there are other domains in which this effect can be observed.

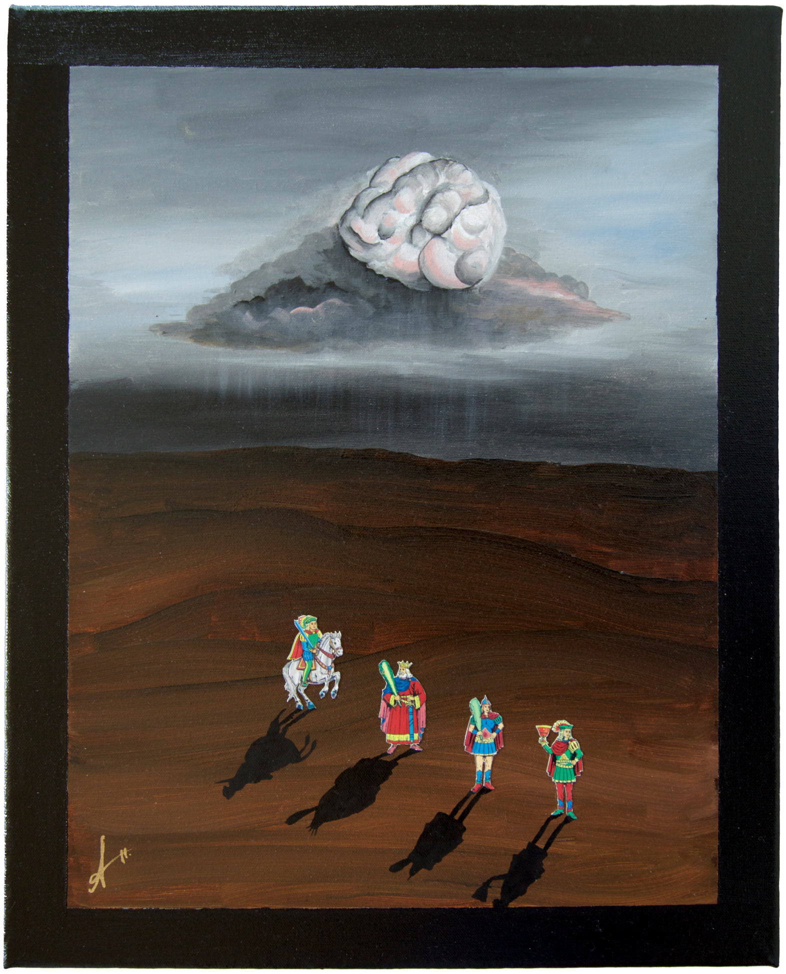
Several interesting possibilities for extensions present themselves. One of them is to consider the same hypotheses of this investigation and explore them in other domains. For example, in *Theme 3* and *4* (*Chapters 6, 7* and *8*) we present some research on moral judgments and social dilemmas. We wonder if the results presented in this chapter could be observed or replicated in the moral psychology field. Specifically, whether or not someone articulates an affective evaluation for a personal/high conflict moral scenario or impersonal/low conflict moral scenario, may influence how an oppositely valenced moral dilemma is rated. In *Theme 3* (*Chapters 6* and *7*) we do not explore the constructive role of moral judgments but we do investigate their dynamics and some other properties of such moral dilemmas; and in *Theme 4* (*Chapter 8*) we discuss, as in this chapter, how a QP model can offer a relatively simple mechanism for the preferences and beliefs of making a judgment in sequential social dilemmas.

The idea of superposition (in the QP sense) is novel in psychology and, as White et al. (2014) pointed out, at the heart of the present research is the debate on the following issue:

> *"Are the feelings of subjective awareness we have, relating to choices or preferences or even simple impressions, linked to a constructive process of creating some of the relevant information or do they reflect a process of reading off internally generated and pre-existing information?"*

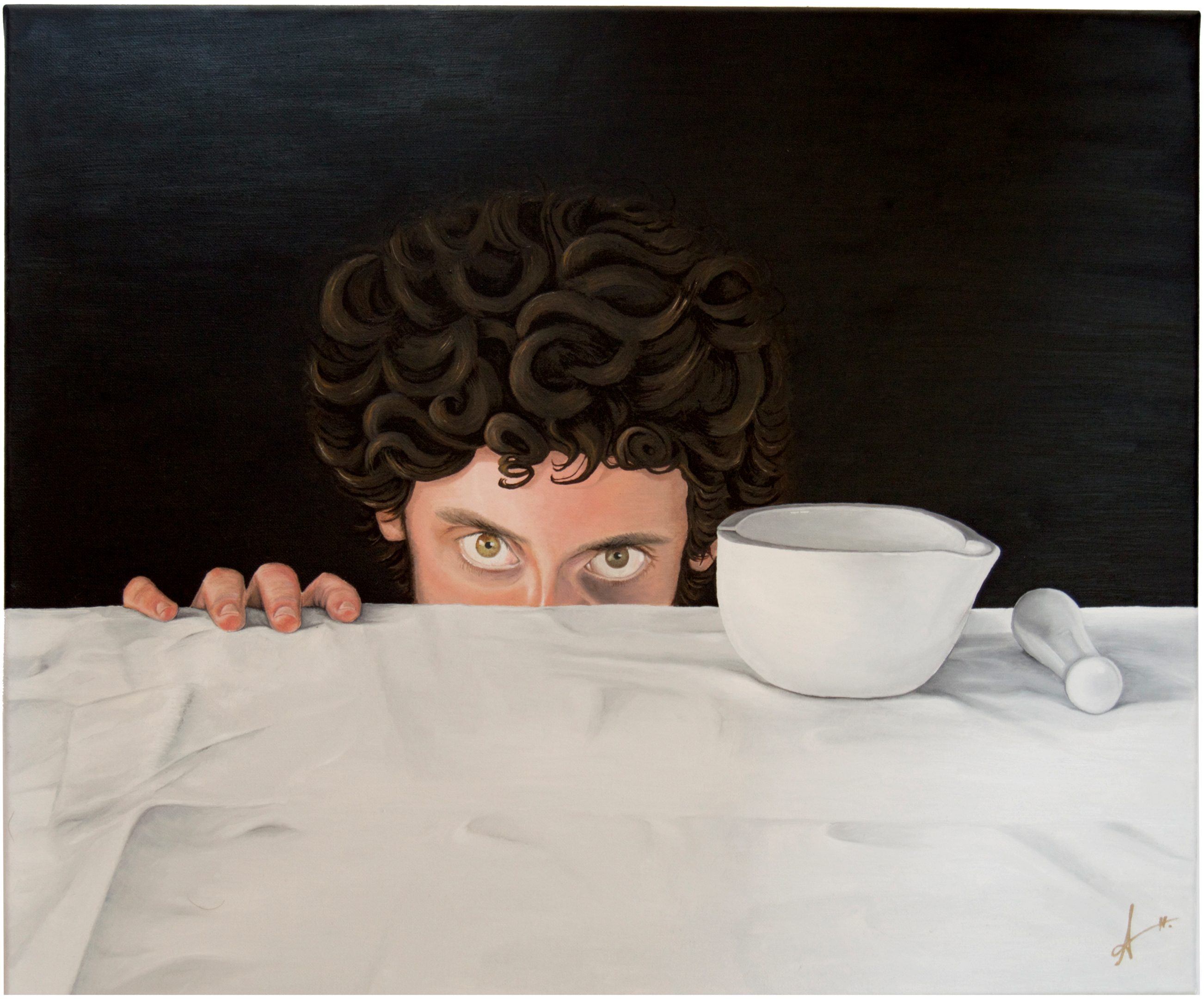
Even though more work is needed regarding both the mathematical and conceptual elaboration of the quantum approach, the results presented in this chapter provide a clear empirical case and illustrate a framework for the principled study of such effects.

Special Artwork Chapter



Journey Of and From the Mind (2015)
Oil on canvas (40.6 x 50.8cm)
Private Collection

On the nature of uncertainty in choice. The dominant metaphor used to conceptualise risky decision-making involves choices between explicit gambles. Moreover, in both experimental and theoretical work, this notion is made operational by using explicit gambling devices such as dice, urns, bingo cages, and the like. However, the nature of uncertainty people experience in real world decisions is often quite different from that inherent in gambling devices. For instance, people are highly sensitive to contextual variables, and changes in context can strongly affect the evaluation of risk.



The Awakening of a New Day (2015)
Oil on canvas (60 x 49.5cm)
Private Collection

Theme 3

On Moral Judgments

Statement of Contribution

Theme 3 is a collaborative work with Emmanuel Pothos, James Hampton and James Yearsley. The author developed the study concepts, designed the experiments, and collected, analysed and interpreted the data for all the studies presented with input from all other authors. The author wrote the manuscripts published. The work presented in *Theme 3* has been presented and discussed at the *Uehiro Centre for Practical Ethics* (*University of Oxford*) during the author's visiting graduate stage at the mentioned institution.

List of publications for *Theme 3*:

Barque-Duran, A., Pothos, E., Yearsley, J., Hampton, J. Contemporary Morality: Moral Judgments in Digital Contexts. (*under review*)

Barque-Duran, A., Pothos, E., Yearsley, J., Hampton, J. (2015). Patterns and Evolution of Moral Behavior: Moral Dynamics in Everyday Life. *Thinking and Reasoning*. 22, 31-56.

Barque-Duran, A., Pothos, E., Yearsley, J., Hampton, J. (2015). Moral Dynamics in Everyday Life: How morality evolves in time? *Proceedings of the 37th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society. 154-159.

Chapter 6

Patterns and Evolution of Moral Behavior: Moral Dynamics in Everyday Life.

Abstract

Recent research on moral dynamics (the processes and phenomena –collective or individual– by which moral behavior and moral attitudes emerge, evolve, spread, erode or disappear) shows that an individual’s ethical mind-set (i.e., outcome-based vs. rule-based) moderates the impact of an initial ethical or unethical act on the likelihood of behaving ethically on a subsequent occasion. More specifically, an outcome-based mind-set facilitates Moral Balancing (behaving ethically or unethically decreases the likelihood of engaging in the same type of behavior again later), whereas a rule-based mind-set facilitates Moral Consistency (engaging in an ethical or unethical behavior increases the likelihood of engaging in the same type of behavior later on). The objective was to look at the evolution of moral choice across a series of scenarios, that is, to explore if these moral patterns (Balancing vs. Consistency) are maintained over time. The results of three studies showed that Moral Balancing is not maintained over time. On the other hand, Moral Consistency could be maintained over time, if the mind-set was reinforced before making a new moral judgment (but not otherwise).

Introduction

Moral Balancing vs. Moral Consistency

How do individuals deal with the ethical uncertainty in their lives? People are confronted with a vast amount of moral scenarios to resolve, such as donating to charities, volunteering, recycling, buying fair trade products, or donating blood. People have to regulate their moral self-image while pursuing self-interest. Studies on moral self-regulation have convincingly demonstrated that one's recent behavioral history is an important factor in shaping one's current moral conduct (e.g., Monin & Jordan, 2009; Zhong, Liljenquist, & Cain, 2009) and two different effects have been reported: Moral Balancing and Moral Consistency.

Moral Balancing (Nisan, 1991) suggests that engaging in an ethical or unethical behavior at one point in time reduces the likelihood of engaging in that form of behavior again in a subsequent situation (Merritt, Effron, & Monin, 2010; Sachdeva, Iliev, & Medin, 2009). To explain this type of behavior, it has been argued that individuals tune their actions in such a way that their moral self-image (which represents individuals' moment-to-moment perception of their degree of morality) fluctuates around a moral-aspiration level or equilibrium (Jordan, Mullen, & Murnighan, 2011; Merritt et al., 2010). It is said that an individual's moral-aspiration level does not equate to moral perfection but rather to a reasonable level of moral behavior for that individual (Nisan, 1991). Ethical and unethical acts respectively elevate and depress the moral self-image. Moral balancing researchers argue that when the moral self-image exceeds the moral-aspiration level, the individual feels "licensed" to engage in more self-interested, immoral, or antisocial behavior (i.e., moral licensing). When the moral self-image is below the moral-aspiration level, people tend to experience emotional distress (Higgins, 1987; Klass, 1978) and become motivated to enact some corrective behavior (i.e., moral compensation). In contrast to Moral Balancing, Moral Consistency (Foss & Dempsey, 1979; Thomas & Batson, 1981) suggests that after engaging in an ethical or unethical act, individuals are more likely to behave in the same fashion later on. This pattern is explained in terms of a psychological need to maintain one's self-concept (Aronson & Carlsmith, 1962), self-perception effects (Bem, 1972), or the use of behavioral consistency as a decision heuristic (Albarracín & Wyer, 2000; Cialdini et al., 1995).

Outcome-Based Mind-Sets vs. Rule-Based Mind-Sets

Recent research on moral dynamics addressed an unresolved question, that is, under which conditions each pattern of moral behavior can occur. Cornelissen et al. (2013) showed that an individual's ethical mind-set (Outcome-based vs. Rule-based) moderates the impact of an initial ethical or unethical act on the likelihood of behaving ethically on a subsequent occasion and, thus, affects the pattern of moral behavior seen. The idea of ethical mind-sets comes from two frameworks on moral philosophy: consequentialism and deontology (Singer, 1991). Past work has demonstrated that this distinction is not exclusively philosophical, but that individuals consider it meaningful when reflecting on their behavior and are flexible in the use of either type of moral pattern (Uhlmann, Pizarro, Tannenbaum, & Ditto, 2009).

A consequentialist perspective considers whether an act is or is not morally right, depending on the consequences of that act (Sinnott-Armstrong, 2008). An individual understands an ethical behavior "because it benefitted other people" and an unethical behavior "because it hurt other people". In other words, when taking a consequentialist perspective, one behaves according to an *Outcome-based* mind-set. By contrast, a deontological perspective implies that what makes an act right is its conformity to a moral norm (Alexander & Moore, 2008), i.e., principles that impose duties and obligations, such as not to break promises or not to lie. In this vein, an individual understands a behavior as ethical "because she followed an ethical norm or principle" or a behavior as unethical "because she did not follow an ethical norm or principle". In other words, when taking a deontological perspective, an individual adopts a *Rule-based* mind-set. An outcome-based mind-set is thought to facilitate Moral Balancing; on the contrary, a rule-based mind-set facilitates Moral Consistency (Cornelissen et al. 2013).

Other studies in the literature support this idea of ethical mind-sets and how they affect moral behavior or under which conditions the mentioned patterns of moral behavior can occur. For example, Conway and Peetz (2012) previously showed that recalling moral behavior in a particular manner moderates, in a similar way as individual's ethical mind-sets, the impact of an initial ethical or unethical act on the likelihood of behaving ethically on a subsequent occasion. They showed that recalling prosocial behavior in a concrete fashion (focusing people on the specifics of the action itself, i.e. the way in which they have helped and supported another person) reminded people that they have already fulfilled moral obligations and allowed them to relax subsequent efforts. In other words, recalling past good deeds in a concrete fashion (like in a consequentialism framework, outcome-based mind-set) might license more selfish, compensatory behavior, and likewise recalling past selfish behavior in a concrete fashion might motivate people to compensate through more prosocial behaviors (Moral Balancing).

In contrast, abstract recollections of past moral behavior (activating moral identity concerns, motivating people to uphold their sense of self by acting in identity-consistent ways, Blasi, 1980, Reed et al., 2007) induced people to act prosocially, whereas abstractly recalling previous selfish behavior induced people to act selfishly. In other words, recalling past selfish behavior in an abstract fashion (like in a deontological framework, rule-based mind-set) might encourage people to maintain one's self-concept or self-perception through their moral behaviors (Moral Consistency).

Evolution of Moral Dynamics

One consequence of considering the role of moral self-image in moral behavior is that it forces one to think of moral choices as a sequence, rather than in temporal isolation. Moral and immoral actions occur in the context of prior moral and immoral actions and the idea of moral self-image provides a connecting thread across these instances. All the relevant findings so far have been produced using an experimental paradigm based on a 2-stage scenario: a manipulation part and a response part. As our aim was to understand how the Moral Balancing and Moral Consistency behaviors evolve in time (we call this evolution moral dynamics), we used a novel experimental paradigm, involving 5 stages (See Figure 1). The importance of studying the evolution of moral dynamics is of clear significance. We designed a novel empirical paradigm, based on the previous successful techniques: participants received two manipulations at the beginning of the experiment: (a) one to induce them to adopt a specific mind-set (outcome-based vs. rule-based) and (b) another to recall an action of a particular morality (ethical vs. unethical). Then, they were presented with a series of moral scenarios (5 stages) that were used to measure the likelihood of engaging in a prosocial behavior. This is the first study to look at the evolution of moral choice across a series of scenarios.

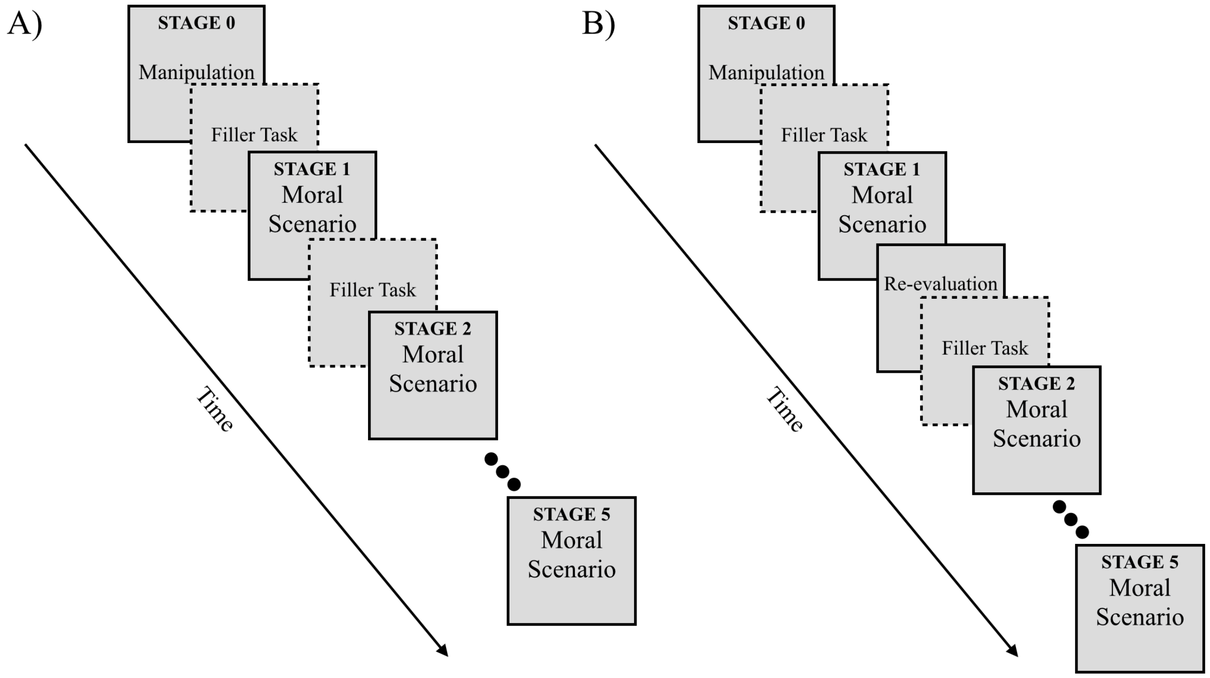


Figure 1: Experimental paradigm using 5-stages for Experiments 1, 2 and 3. In A, we represent the manipulation given to participants at the beginning of Experiment 2. In B, we represent the two manipulations employed in Experiment 3: one at the beginning of the experiment (same as in Experiment 2) and another presented before confronting a new moral scenario, at each stage.

Our objective was to explore the hypothesis that mind-set, Moral Balancing and Moral Consistency are maintained over time (indeed, otherwise, it would be hard to appreciate their psychological significance). We know from previous research that mind-set can influence relatively immediate moral behavior (Cornelissen et al. 2013), but it remains unknown whether mind-sets can be sustained over time and so have a persistent influence on moral behavior. This experimental design assumes that participants are in a specific mind-set. That is, it is meaningful to ask about the sustainability of patterns in moral dynamics, only for those participants who can be said to be clearly in a particular mind-set at the outset. Without this assumption, the contrast between the hypotheses of interest cannot be made (i.e., if a participant cannot be said to be in an outcome-based mindset, it is meaningless to ask whether there is moral balancing which lasts over time). Therefore, this consideration will need to be taken into account for the statistical analysis.

The conflicting hypotheses regarding how moral behavior evolves in time are illustrated in Figures 2 and 3. Both putative patterns of moral behavior are illustrated over a sequence of moral scenarios or stages. We called the ‘Zig-Zag pattern’ the idealized pattern for a Moral Balancing behavior. By analogy, we called ‘Flat pattern’ the idealized pattern for a Moral Consistency behavior. We then used these idealized patterns to motivate the analyses for the results obtained in Experiments 1, 2 and 3. For Moral Balancing, an initial ethical manipulation (such as recall of an ethical action) at Stage 0 should be followed at the next stage by an unethical choice. However at the subsequent stage, the previous unethical choice should now promote a more ethical one. The result is a predicted oscillation between ethical and unethical choices, as the participant tries to maintain a balance (Figure 2). Alternatively, Moral Consistency should lead to the persistence of an initial choice, as with each Stage the participant becomes more and more confirmed in the belief of their consistent moral position, be it either ethical or unethical (Figure 3).

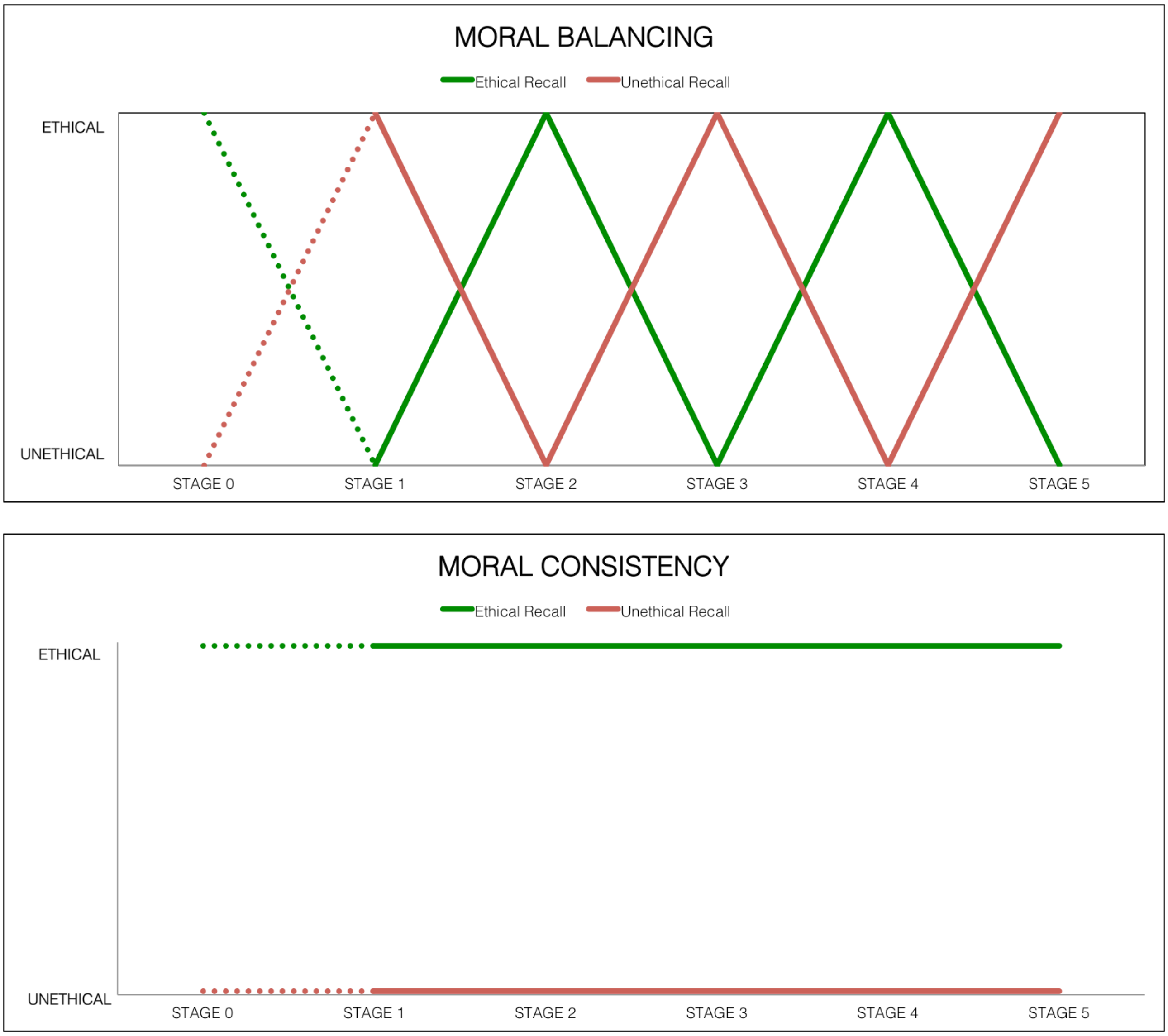


Figure 2 and 3: ZIG-ZAG and FLAT Patterns. Idealized pattern of behavior according to the balancing or consistency views of moral dynamics. The dashed lines represent the transition from the manipulation phase [STAGE 0] to the first moral scenario [STAGE 1], given recall of an ethical or unethical action

In order to study the evolution of moral tendencies and the perseverance of mind-sets we ran three experiments plus a pilot study. In the pilot study we identified the most suitable moral scenarios to use in the main experiments. Experiment 1 allowed us to collect baseline data, as a control group, for comparisons with the results of the subsequent experiments. Experiment 2 was used to replicate the results in the moral dynamics literature (Cornelissen et al., 2013; Jordan, Mullen, Murningham, 2011) and to pursue the novel question of how the tendency to behave morally evolves over time. Finally, in Experiment 3, we aimed to explore again how the two possible patterns of moral dynamics evolve over time, but in this case, we added a manipulation before each new moral scenario, to test if ethical mind-sets are maintained if reinforced.

Pilot Study

The objective of the pilot study was to identify suitable moral scenarios for the main experiments. We were looking for five moral scenarios such that they would (1) be perceived to have high levels of morality, (2) have a similar frequency of engagement (prosocial behavior) and (3) be perceived similarly in terms of emotionality, that is, they would produce a similar affective reaction. Measuring the affective reaction is important, as Szekely and Miu (2014) showed the existence of an influence of emotional experience on moral choice scenarios.

Participants

Twenty experimentally naïve students at City University London received course credit for participating in the study.

Materials and Procedure

The experiment, designed in Qualtrics, lasted approximately 15 minutes. Eleven novel moral scenarios were initially created. For each scenario we tested the perceived morality of the choice of actions using a 7-point scale: -3=very immoral, 3=very moral (How moral do you think this behavior is?), and the prosocial behavior measured as the likelihood of engaging in an (un)ethical behavior on a 7-point scale: 1=very unlikely, 7=very likely; (Jordan, Mullen, et al., 2011). Participant responses on perceived morality and likelihood of engagement were the main dependent variables in our pilot. Also, we tested the perceived emotionality of the scenarios presented, measured with the (SAM) Self-Assessment Manikin (Bradley & Lang 1994). We used the SAM method as it is a non-verbal pictorial assessment technique that directly measures the pleasure, arousal, and dominance associated with a person's affective reaction to stimuli presented, in this case moral scenarios. From the results of this pilot, we then chose five situations for the main Experiments 1, 2 and 3 (one for each of the five stages in the experiments). To do so, we computed the average and the variance of our 3 measures: perceived morality, likelihood of engagement and emotionality, for each of the scenarios. Then we chose the five scenarios with the highest scores in perceived morality and with similar (intermediate) scores in likelihood of engagement and perceived emotionality measures (see Appendix 3 for details).

Experiment 1

The aims of the first study were to test the novel experimental paradigm and collect baseline data. As this was a control condition, there was no manipulation of the participants' mind-set (outcome-based vs. rule-based) nor the recall of a moral deed. We used Prosocial Behavior, that is, the likelihood of engaging in an (un)ethical behavior, as the dependent measure, using an experimental paradigm involving 5 stages. The experiment lasted approximately 30 minutes. In the absence of any manipulation, we expected intended behavior not to be biased towards ethical or unethical choices.

Participants

A total of 104 participants, all of them US residents, were recruited on-line and received \$0.90 for doing the task.

Materials and Procedure

The study was designed in Qualtrics and run on Amazon Mechanical Turk. There is some evidence that data obtained via Mechanical Turk demonstrate psychometric properties similar to laboratory samples (Buhrmester, Kwang, & Gosling, 2011). First, participants completed a filler task (10 trivia questions ≈ 1.6min per filler task) before responding to two items, one about their likelihood of engaging in a prosocial behavior (STAGE 1) and another about their likelihood of engaging in a leisure activity that simply acted as a distractor. Then, participants completed another filler task, like the first one, before responding to 2 more items, again, one about their likelihood of engaging in another prosocial behavior (STAGE 2) and in another leisure activity. Subsequently, participants completed the same procedure three more times, until STAGE 5. The order of presentation of the moral scenarios on each stage, as well as the filler tasks, were randomized across participants.

Results and Discussion

A one-sample t-test was run to determine whether the likelihood of engaging in a prosocial behavior was biased towards a more ethical or unethical tendency. We defined a score of 4.0 (the midpoint of the 1-7 scale we used) as neither moral nor immoral behavior. We accepted the null hypothesis that the population mean was not different from 4.0; (*M* = 4, *SD* = 1.96); t(103)=0.00, p=1.0. The range of means across scenarios was from 3.5 to 5. That is, in the absence of any manipulation, prosocial choices were not biased towards ethical or unethical behavior, as intended.

Experiment 2

The objectives here were twofold. First, we wanted to replicate the results in the moral dynamics literature, that an Outcome-based mind-set leads to Moral Balancing, whereas a Rule-based mind-set leads to Moral Consistency. The motivation to do so was to validate the experimental approach. Second, Experiment 2 employed a multi-stage procedure, so allowing us to pursue the novel question of how the tendency to behave morally evolves over time. In contrast to Experiment 1, we manipulated the participant's mind-set (outcome-based vs. rule-based) and the morality of an action that they were asked to recall, at the beginning of the experiment. The experiment lasted approximately 35 minutes.

Participants

A total of 200 participants, all of them US residents, were recruited on-line and received $0.90 for doing the task.

Design and Procedure

The experiment was designed in Qualtrics and run on Amazon Mechanical Turk. Ethical mind-set (outcome-based vs. rule-based) and the ethicality of an initial recalled act (ethical vs. unethical) were both manipulated between participants. The induction of ethical mind-sets was the same as used in Cornelissen et al. (2013), so we only briefly summarize it here (see Appendix 3 for details). To induce the appropriate mind-set, we provided instructions that defined ethicality as either a function of consequences or in terms of rule compliance, and then provided three prototypical examples. Subsequently, we asked participants to provide an example of a behavior—not necessarily their own—that was ethical or unethical, because of either its consequences or its rule compatibility (depending on condition). This procedure aimed to induce the intended mind-set in participants, before they finally reflected on their memory of the last action with moral valence.

There were therefore four conditions: (1)Outcome-Based/Ethical recall, (2)Outcome-Based/Unethical recall, (3)Rule-Based/Ethical recall and (4)Rule-Based/Unethical recall. In the first one, our participants were instructed to think about a behavior that was ethical ("because it benefitted other people"). In the second group, participants were instructed to think about a behavior that was unethical ("because it hurt other people"). In the third group, participants thought about a behavior that was ethical ("because you followed an ethical norm or principle") and in the fourth group, participants were instructed to think about a behavior that was unethical ("because you did not follow an ethical norm or principle").

We used Prosocial Behavior, as in all the other experiments, as the dependent measure. After the manipulation (STAGE 0), participants followed the same experimental paradigm as in Experiment 1: they completed a filler task before rating their likelihood of engaging in a prosocial behavior (STAGE 1) and then repeated the same procedure until STAGE 5. The order of presentation of the moral scenarios on each stage, as well as the filler tasks, were randomized for each participant.

Results and Discussion

Replication of previous studies. Mean intention to perform the prosocial action at the first stage of the procedure is shown in Figure 4. As predicted, when given an Outcome-based mindset, the recall of an unethical act led to Moral Balancing and an increased intention to perform the moral action. When given a Rule-based mindset, the reverse pattern was observed. This result was confirmed with an ANOVA, which showed a significant interaction between Type of Mind-set and Type of Ethical Recall, F(1,44) = 7.12, p < 0.01, but no main effect of Type of Mind-set, nor of Recall, (both F < 1). Independent samples t-tests were employed to explore the interaction. In the outcome-based mind-set condition, participants who recalled an unethical act were more likely to engage in a prosocial behavior (*M* = 4.54, *SD* = 1.66), than those who recalled an ethical act (*M* = 3.82, *SD* = 1.69), *t*(91) =

−2.06, *p* = .04. In other words, participants with an Outcome-based mind-set showed a Moral Balancing effect. By contrast, in the Rule-based mind-set condition, participants who recalled an ethical act were more likely to engage in a prosocial behavior (*M* = 4.36, *SD* = 1.68) than those who recalled an unethical act (*M* = 3.6, *SD* = 1.74), *t*(93) = 2.14, *p* = .03. In other words, these participants showed a Moral Consistency effect.

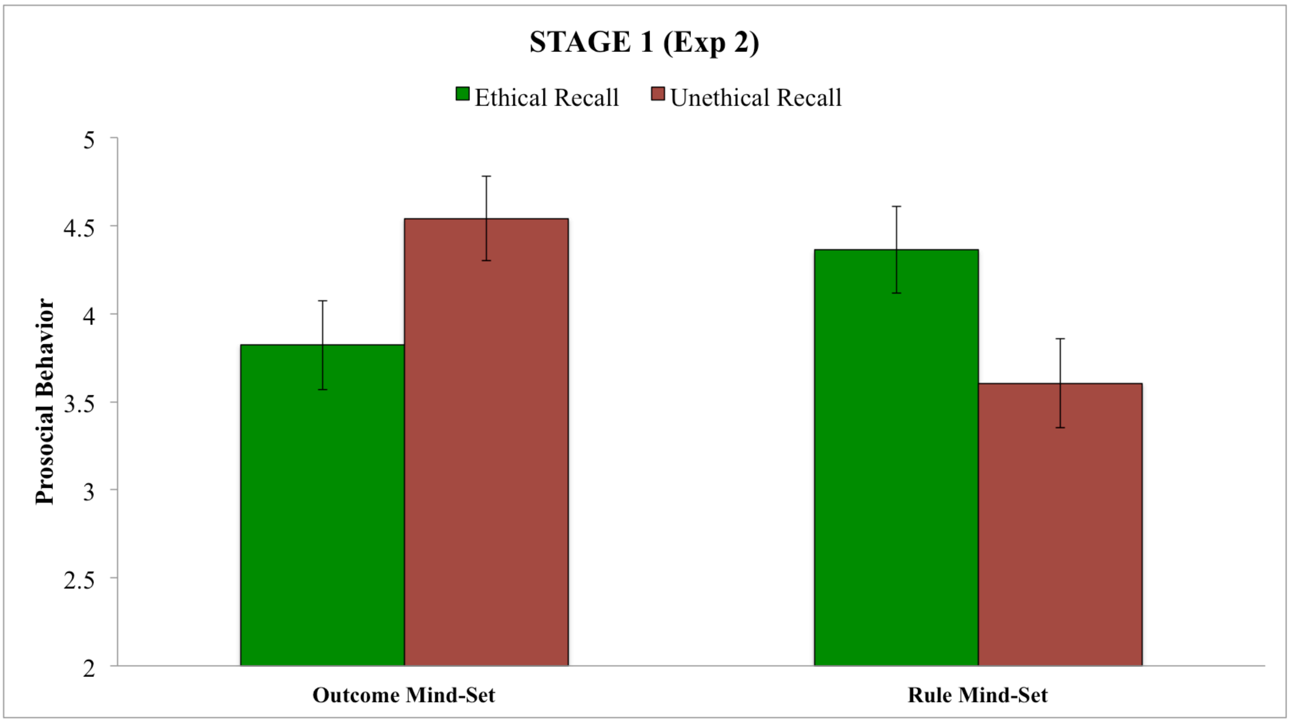


Figure 4: Prosocial behaviors [STAGE 1] in Experiment 2; mean likelihood of engaging in a prosocial behavior, as a function of a participants' ethical mind-set and the ethicality of the act they recalled. This pattern replicates the results of Cornelissen et al. (2013). Error bars represent standard errors.

Evolution of moral dynamics. We first applied some selection criteria to the data in order to properly examine the hypotheses of interest. A restriction of the sample was needed since, as we previously mentioned, the mind-set procedure would not be expected to work equally well for every participant, and our research hypothesis is only meaningful for participants assumed to be in specific mindsets. The experimental design proposed in this paper assumes that participants behave in a certain way. That is, it is meaningful to ask about the sustainability of patterns in moral dynamics only for those participants who can be said to be clearly in a particular mind-set at the outset. Without this assumption, the contrast between the hypotheses of interest can not be tested. The issue of the effectiveness of the mind-set procedure is separate from that of whether, given that the induction of mind-set was effective, the mind-set's influence on moral decisions perseveres across stages. So we eliminated the cases that were considered far from the intended behavior in STAGE 1, i.e., the participants whose behavior did not conform to the expectations associated with the mind-set manipulation (Cornelissen et al., 2013).

As the scale of our dependent variable was 1-7, we eliminated participants with a prosocial behavior rating after the mindset manipulation that was in the wrong direction relative to the neutral midpoint of 4 and the mean of their group. Specifically, for the two conditions which we intended to use to test the persistence of a prosocial attitude (those with means over 4 in Figure 4), all participants with a rating of less than 4 were excluded. Thus in these two conditions all remaining participants had responded as predicted to the combination of mindset and recall manipulations. Similarly for the two conditions which were to test the persistence of non-prosocial attitudes (those where the group mean was below 4 in Figure 4), all participants with a rating greater than 4 were excluded. As a consequence, 19 out of 45 cases were excluded from condition 1, and 15 out of 48, 16 out of 47, 19 out of 48 cases were rejected from conditions 2, 3 and 4 respectively.

While we believe the preselection manipulation to be an essential condition for a meaningful test of our hypotheses, for completeness we also present an analysis for the whole sample in Appendix 3. In fact, no conclusions are altered by considering the entire sample.

We examined the levels of Prosocial Behavior throughout all stages, first comparing the two mind-set conditions within the same analysis and then analyzing the Outcome-based and the Rule-based conditions separately, in order to study the evolution of moral tendencies across STAGES [1-5]. We assessed the results against the idealized predictions in Figures 2 and 3.

First, we ran a three-way ANOVA, with Type of Ethical Recall (2 levels: ethical recall and unethical recall, between participants), Type of Mind-set (2 levels: outcome-based and rule-based, between participants) and Stage (5 levels: five stages, within participants), on the dependent variable (likelihood of engaging in a prosocial behavior). There was no main effect of type of Type of Ethical Recall, no significant effect of Type of Mind-set, and no main effect of Stage, (all F < 1). There was a significant interaction between Recall and Type of Mind-set, F(1,25) = 20.786, p < .01, but not between Recall and Stage nor between Type of Mind-set and Stage, (both F < 1). Finally, there was a significant interaction between the three factors, F(4,100) = 13.9, p < .01.

Evidence for Moral Balancing. In Figure 5, we can see how for the Outcome-based mind-set group, the 'Zig-Zag pattern' is broadly evident across STAGES 0 and 1, as we have seen in the previous section (this finding replicates previous research, Cornelissen et al., 2013; Jordan, Mullen, Murningham, 2011). (For Stage 0 we have inserted imaginary data points to represent the ethical or unethical recall manipulation). The pattern across stages [0-1] concerns the initial mind-set manipulation with an (un)ethical recall and the first moral scenario. As a starting point we used y=4, the mid-point of the prosocial behavior scale (where our pilot data suggested that participants start off from prior to the onset of the ethical mind-set manipulation). What happened across the rest of stages?

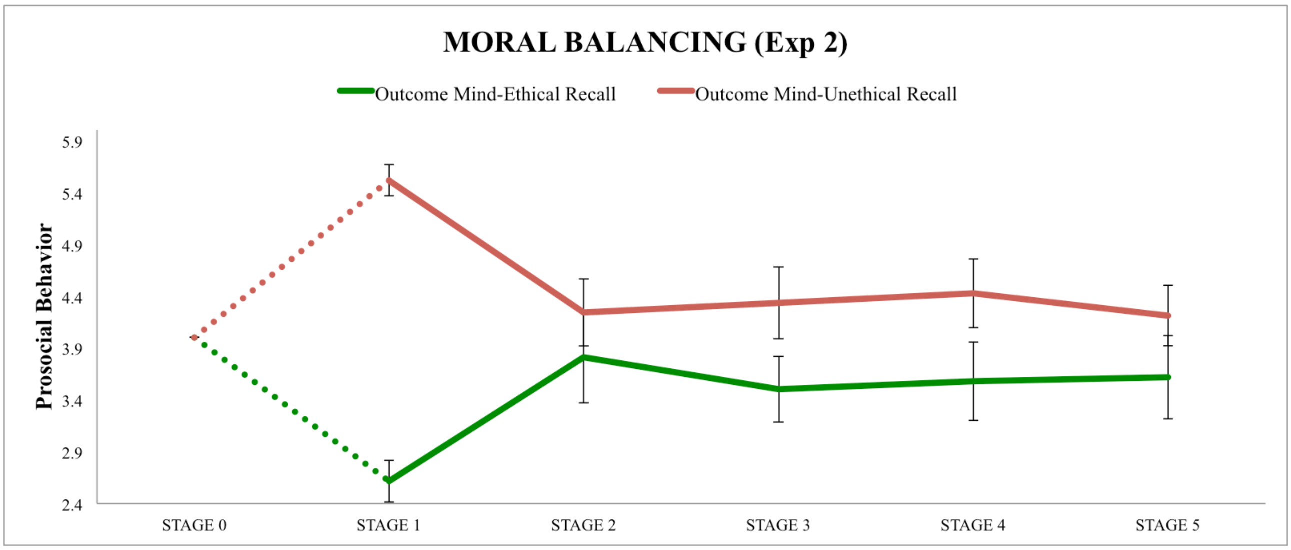


Figure 5: Evolution of the prosocial behaviors of the Outcome Based Mind-set group (ethical + unethical recalls) in Experiment 2. The dashed lines represent the transition from the manipulation phase [STAGE 0] to the first moral scenario [STAGE 1], given an (un)ethical recall. Error bars represent standard errors.

We ran a mixed two-way ANOVA with Type of Ethical Recall and Stage (1-5), on the dependent variable (likelihood of engaging in a prosocial behavior). Minimally, Moral Balancing would be evidenced by no main effect of Recall, but a significant interaction between Recall and Stage. There was a main effect of Type of Ethical Recall, F(1,25) = 13.1, p < .001, no significant effect of stage, F < 1, and a significant interaction between the two factors, F(4,100)=5.57, p < .01. Inspection of Figure 5 makes it clear that the interaction is just a result of prosocial choice converging towards an average level by Stage 2, after which it flattens out across the two conditions of ethical recall.

We then analyzed the evolution of prosocial behavior between STAGES [1-2] to see if, at least, the Moral Balancing pattern was maintained for just one more stage. A two-way analysis of variance (ANOVA) with Type of Ethical Recall and Stage as independent variables indicated a main effect of Recall, F(1,25)=23.2, p < .01, and no main effect of Stage, F(1,25) < 1. The results also revealed a significant interaction between Type of Ethical

Recall and Stage, F(1,25) = 12.0, p = .002. So, as above, there was little evidence for Moral Balancing.

Finally, we wanted to know whether the data at each stage showed any evidence of a residual effect of Type of Ethical Recall factor after STAGE 1. We ran an ANOVA with STAGES [2-5] and Recall. The effect of Recall approached significance, F(1,25) = 3.41 p = .077, but there was no main effect of stage, F < 1, and no significant interaction between the two factors, F(3,75) < 1. Therefore, the interaction seen in the previous analysis, STAGES [1-5], is explained by the change from STAGE 1 to STAGE 2 and disappears after that.

Overall, the results show that Moral Balancing was not observed in this experiment, beyond the initial manipulation. The conclusion is that the 'Zig-Zag pattern' was only observed throughout STAGES [0-1], but not further maintained over time, in contrast to the idealized prediction of Figure 2. Instead, it appears that the evolution of prosocial behavior converged to a neutral level of morality (Figure 5). The marginal effect of Recall in Stages 2-5 suggests in fact that after the initial Moral Balancing at Stage 1, participants settle into an approximate state of Moral Consistency for subsequent decisions.

Evidence for Moral Consistency. We examined the results for Moral Consistency with the Rule-based mindset conditions. In Figure 6, we can see how the 'Flat pattern' was broadly evident between STAGES [0-1]; recall, this was also demonstrated in the previous section (where we aimed to replicate previous research). What happened across the rest of stages?

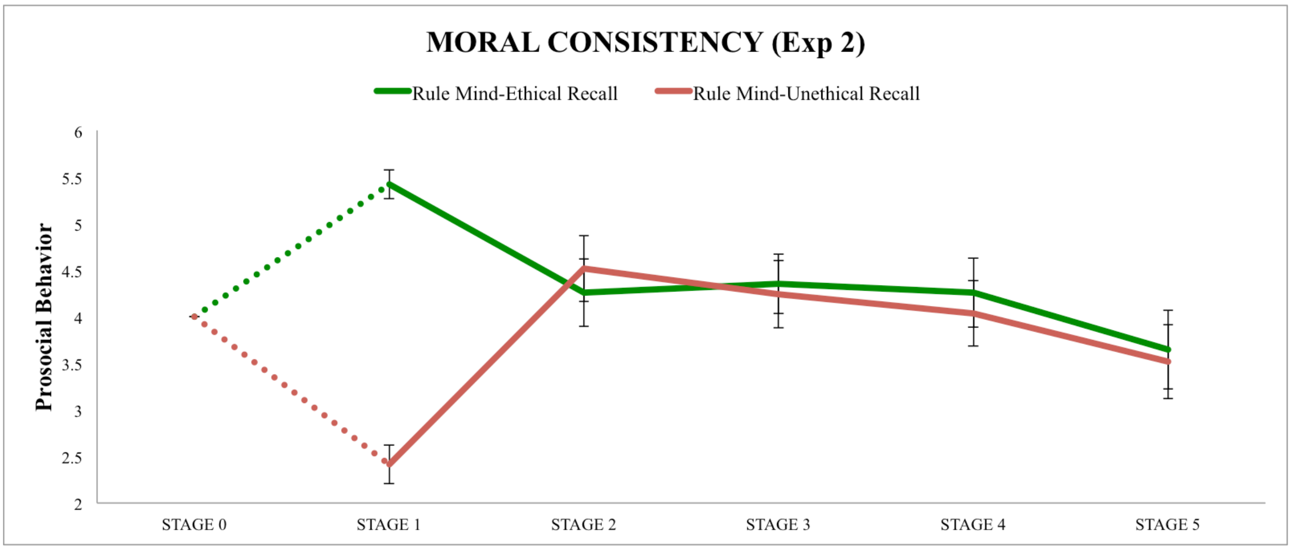


Figure 6: Evolution of the prosocial behaviors of the Rule Based Mind-set group (ethical + unethical recalls) in Experiment 2. The dashed lines represent the transition from the manipulation phase [STAGE 0] to the first moral scenario [STAGE 1], given an (un)ethical recall. Error bars represent standard errors.

Regarding the evolution between STAGES [1-5], we ran a two-way analysis of variance (ANOVA) with Type of Ethical Recall and Stage on likelihood of Prosocial Behavior. Minimally, Moral Consistency would be evidenced by a main effect of Recall, no main effect of Stage, and no interaction between Recall and Stage. There was indeed a main effect of Recall in Prosocial Behavior, F(1,28) = 7.02, p = .013, but also a significant interaction between Recall and Stage, F(4,112) = 8.07, p < .01. Note, there was no main effect of stage, F(4,112) = 1.64, p = .170.

Inspection of Figure 6 makes it clear that it was not necessary, as in the previous analysis, to analyze the evolution of prosocial choice between STAGES [1-2] to see if, at least, the Moral Consistency pattern was maintained for just one more stage. The pattern converged to a neutral point and did not remain attached to the low or high levels of (un)ethicality.

Finally, we wanted to know whether the data across stages showed any evidence of a residual effect of the Type of Ethical Recall factor, after STAGE 1. We ran an ANOVA with STAGES [2-5] and Recall. There was no main effect of Recall, no significant effect of Stage, and no interaction between the two factors, (all F < 1). Therefore, the main effect seen in the previous analysis, STAGES [1-5], is explained by the change from STAGE 1 to STAGE 2 and disappears after that.

The conclusion is that the 'Flat pattern' only remained attached to the low or high levels of (un)ethicality, as in the idealized pattern (Figure 3), for STAGES [0-1]. The rest of stages converged to a neutral level of morality; thus, Moral Consistency was not maintained over time (Figure 6).

Experiment 3

In Experiment 2, after an initial mind-set induction and ethical recall, we found that the anticipated patterns of moral dynamics were not maintained. There are two possible explanations. First, the theory linking mind-set, (un)ethical recall, and ethical choice is simply incorrect (or, at any rate, incomplete). Second, the mind-set induction attenuates rapidly with time, so that, after the initial stages, participants can no longer be assumed to be in a specific mind-set. Do ethical mind-sets decay if not manipulated or re-evaluated continuously? Experiment 3 examines this second possibility. As with Experiment 2, we aimed to explore how the two possible patterns of moral dynamics evolve over time, but in this case, we added a re-evaluation process (manipulation of the mind-set + un(ethical) recall), before presenting a new moral scenario at each of the 5 stages. In this way, having manipulated the type of mind-set and type of recall at the beginning of the task, we reinforced the manipulation at each subsequent stage of the task. The experiment lasted approximately 40 minutes.

Participants

A total of 206 participants, all of them US residents, were recruited and received 1$ for doing the task.

Design and Procedure

The experiment was designed in Qualtrics and run on Amazon Mechanical Turk. The same procedure was followed as in Experiment 2, with 4 conditions (Outcome-Based/Ethical recall, Outcome-Based/Unethical recall, Rule-Based/Ethical recall and Rule-Based/Unethical recall). We manipulated (between participants) the ethical mind-set (outcome-based vs. rule-based) and the ethicality of an initial act (ethical vs. unethical). We used Prosocial Behavior, as in all the other experiments, as a dependent measure. After the manipulation, participants followed the same experimental paradigm as in Experiment 1 and 2: they completed a filler task before responding to the likelihood of engaging in a prosocial behavior (STAGE 1). Then, we introduced a new manipulation (the re-evaluation process), in which participants were asked to reflect on their last moral choice, in order to reinforce their mind-set, in a similar way as in the manipulation at the beginning of the experiment (manipulation of the mind-set + un(ethical) recall; see Appendix 3 for details. Afterwards, they completed another filler task, like the first one, before responding to the likelihood of engaging in a prosocial behavior (STAGE 2). Participants followed the same steps until STAGE 5, as in Experiment 2, but justifying their choices, after their response, at each stage (Figure 1). The order of presentation of the moral scenarios on each stage, as well as the filler tasks, were randomized for each participant.

Results and Discussion

Replication of previous studies. Mean intention to perform a prosocial action at the first stage of the procedure is shown in Figure 7. As predicted, when given an Outcome-based mindset, the recall of an unethical act led to Moral Balancing and an increased intention to perform the moral action. When given a Rule-based mindset, the reverse pattern was observed. These results were in the right direction, but were not confirmed in the ANOVA, which showed no significant interaction between Type of Mind-set and Type of Recall, F(1,49) = 1.167, p = .285, and no main effect of Type of Mind-set, nor of Recall (both F < 1).

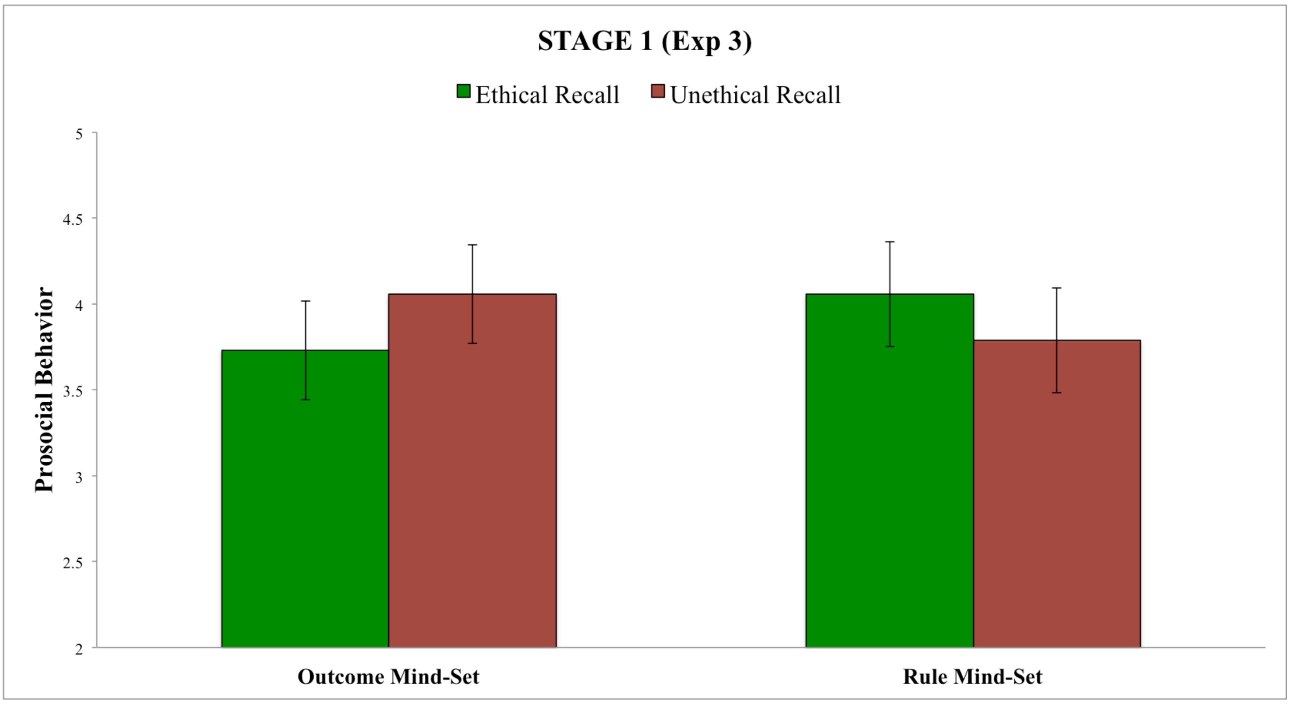


Figure 7: Prosocial behaviors [STAGE 1] in Experiment 3; mean likelihood of engaging in a prosocial behavior, as a function of participants' ethical mind-set and the ethicality of the act they recalled. Error bars represent standard errors.

Evolution of moral dynamics. We first applied the same selection criteria to our results, as for Experiment 2. Specifically, 23 out of 52 cases were rejected from condition 1, and 19 out of 50, 20 out of 52, 22 out of 52 cases were rejected from conditions 2, 3 and 4 respectively. An analysis for the whole sample is presented in Appendix 3; the conclusions derived by focusing on the restricted sample are equivalent to those in the entire sample for the Moral Balancing case and different for the Moral Consistency case (but, as argued in Experiment 2, we think that the analyses in the restricted sample are more valid, since one cannot test the persistence of a state in participants who are not initially placed into that state).

As in Experiment 2, we examined the levels of Prosocial Behavior throughout all stages, first examining the two mind-set conditions within the same analysis and then the Outcome-based and the Rule-based conditions separately, in order to study the evolution of moral tendencies across STAGES [1-5]. We then compared the results to the idealized predictions (Figures 2 and 3).

First, we ran a three-way ANOVA with Type of Ethical Recall, Type of Mind-set and Stage, on the dependent variable (likelihood of engaging in a prosocial behavior). There was no main effect of type of Recall, no significant effect of Type of Mind-set, and no main effect of Stage, (all F < 1). There was a significant interaction between Recall and Type of Mind-set, F(1,28) = 94.3, p<.01, but not between Recall and Stage and between Type of Mind-set and Stage, (all F < 1). Finally there was a significant interaction between the three factors, F(4,112) = 13.9, p<.01.

Evidence for Moral Balancing. First, we considered the evidence for Moral Balancing. We ran a two-way ANOVA, as in Experiment 2, with Type of Ethical Recall and Stage on the dependent variable. As before, Moral Balancing would be minimally evidenced by no main effect of Recall, but a significant interaction. Instead, there was a main effect of Recall, F(1,28) = 40.4, p<.01, and no effect of Stage, F < 1. The results also indicated a significant interaction between Recall and Stage, F(4,112) = 7.54, p<.01.

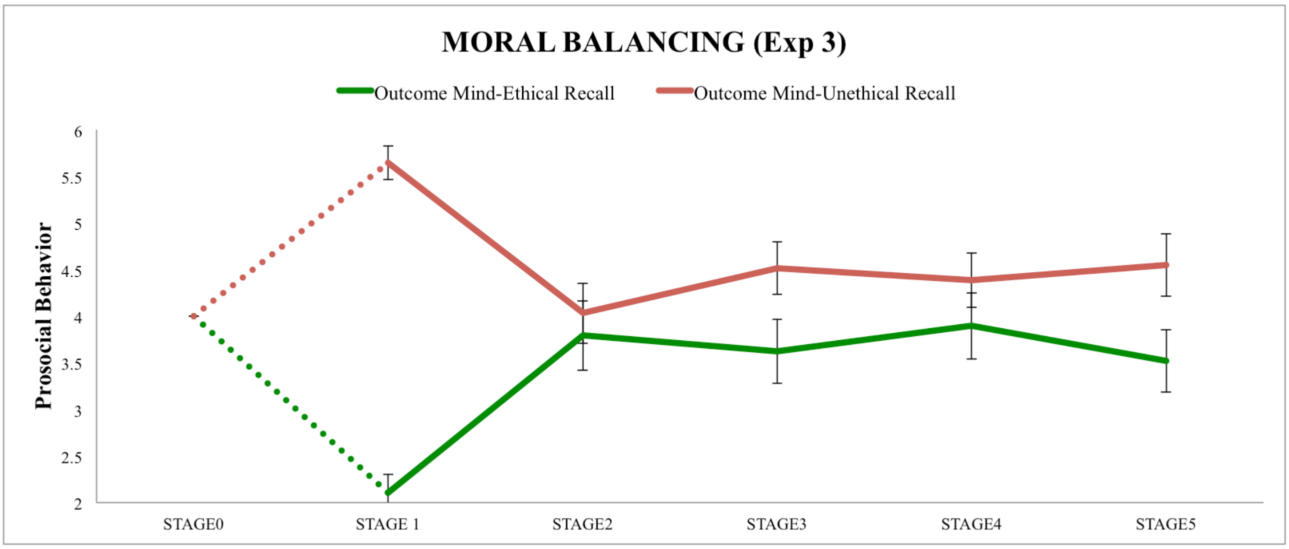


Figure 8: Evolution of the prosocial behaviors of the Outcome Based Mind-set group (ethical + unethical recalls) in Experiment 3. The dashed lines represent the transition from the manipulation phase [STAGE 0] to the first moral scenario [STAGE 1], given an (un)ethical recall. Error bars represent standard errors.

We then analyzed the evolution between STAGES [1-2] to see if, at least, the Moral Balancing pattern was maintained for just one more stage. A two-way ANOVA with two within participant factors, Type of Ethical Recall and Stage, revealed a similar pattern of results: a main effect of Recall, F(1,28) = 44.5, p<.01, no effect of Stage, F < 1, and a significant interaction between Recall and Stage, F(1,28) = 30.9, p<.01.

Finally, we wanted to know whether the data at each stage showed any evidence of a residual effect of Type of Ethical Recall factor after STAGE 1. We ran an ANOVA with STAGES [2-5] and Recall. There was a main effect of Recall, F(1,28) = 9.37, p<.01, no significant effect of stage, F < 1, and a non significant interaction between the two factors, F < 1. Therefore, the interaction seen in the previous analysis, STAGES [1-5], is explained by the change from STAGE 1 to STAGE 2 and disappears after that.

The conclusion is that the 'Zig-Zag pattern' was only approximately observed across STAGES [0-1]. Thus, compared with the idealized pattern (Figure 2), Moral Balancing was not a behavior maintained over time. Instead, as in Experiment 2, the evolution of the behavior converged to a neutral level of morality (Figure 8). In fact, as in Experiment 2 there was a tendency (this time statistically significant) for participants to settle into a Moral Consistency pattern from Stage 1 onwards, regardless of the reminders that had been introduced in the present experiment.

Evidence for Moral Consistency. Regarding the evolution between STAGES [1-5] in the Moral Consistency case, we ran a two-way ANOVA with two within participant factors, Type of Ethical Recall and Stage on the dependent variable (likelihood of engaging in a prosocial behavior). Moral Consistency would be minimally evidenced by a main effect of Recall, but not a significant interaction. There was a main effect of Recall on Prosocial Behavior, F(1,29) = 53.2, p<.01, but not on Stage, F(4,116) = 2.02, p=.096. Also, the interaction between Recall and Stage was significant, F(4,116) = 5.68, p<.01, which is not consistent with a 'pure' form of Moral Consistency.

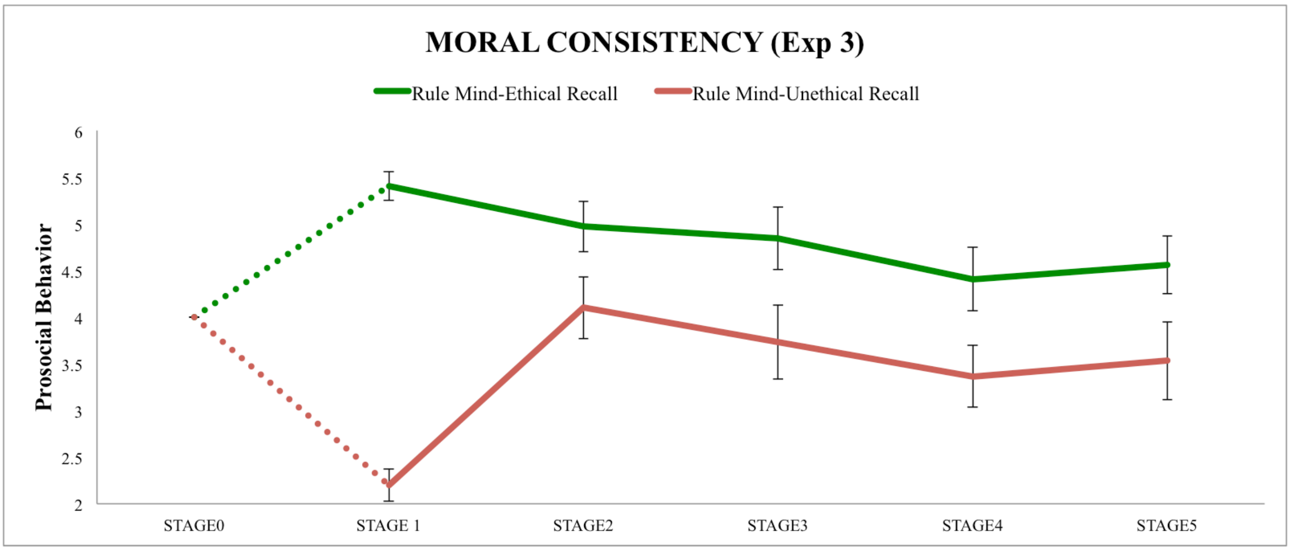


Figure 9: Evolution of the prosocial behaviors of the Rule Based Mind-set group (ethical + unethical recalls) in Experiment 3. The dashed lines represent the transition from the manipulation phase [STAGE 0] to the first moral scenario [STAGE 1], given an (un)ethical recall. Error bars represent standard errors.

Then, we ran an ANOVA with STAGES [2-5] and Type of Ethical Recall to see if the Moral Consistency pattern was maintained over time, as it can be seen that Figure 9 was the one most similar to the idealized 'Flat pattern' (Figure 3), across all experiments. There was a main effect of Recall, F(1,29) = 18.88, p<.01, no significant effect of Stage, F < 1, and a non significant interaction between the two factors, F < 1.

Finally, we used Bonferroni corrected t-tests to examine the main effect of Type of Ethical Recall, to show that Prosocial Behavior elicited by each Type of Ethical Recall differed at each Stage. In all cases, there was a trend in the expected direction (ethical recall led to more ethical behavior and unethical recall led to more unethical behavior). For Stage 1: *t*(60) = 13.749, *p* <.0005; for Stage 2: *t*(60) = 2.057, *p* =.044; for Stage 3: *t*(60) = 2.606, *p* =.012; for Stage 4: *t*(60) = 2.193, *p* =.032; for Stage 5: *t*(60) = 1.995, *p* =.051. Note, the Bonferroni corrected p-value for rejecting the null hypothesis in this family of t-tests is .05/4=.0125, so, we can confidently conclude that significant differences exist only for stages 1 and 3. Nevertheless, we think that the overall pattern is indicative enough and supports the view that the Moral Consistency pattern is broadly evident across the different stages (noting also that the Bonferroni adjustment for multiple t-tests is considered to be conservative; e.g., Nakagawa, 2004)

The conclusion is that the 'Flat pattern' was sustained to the low or high levels of (un)ethicality throughout STAGES [0-5], but not as much as predicted in the idealized pattern (Figure 3). Moral Consistency was a behavior broadly maintained over time (with a tendency to converge to a neutral level of morality), if a re-evaluation process (manipulation of the mind-set plus un(ethical) recall) was carried out before confronting each new moral scenario (Figure 9).

General Discussion

This is the first study to look at the evolution of moral choice across a series of scenarios. Five scenarios were tested, embedded in a task with many fillers, to mask the design of the experiment. In three experiments, we provided new empirical support for the hypothesis that ethical mind-sets moderate how an individual's behavioral history shapes his or her ethical behavior. An outcome-based mind-set is meant to lead to moral-balancing effects, whereas a rule-based mind-set to moral consistency. Furthermore, the three experiments shed some light on the persistence of these ethical mind-sets and on the evolution of moral dynamics, exploring whether moral patterns, such as Moral Balancing and Moral Consistency, can be maintained over time. When the manipulation of Mind-set and Recall was just made at the start, there was a quick regression to neutral performance. When the manipulation was reinforced before each moral choice, then one pattern of behavior was sustained, while the other was not.

Moral Balancing, or as we call it, the 'Zig-Zag pattern', was only observed in the first stage of the experiments. This type of behavior converged to a neutral level of morality over time, even when the mind-set was reinforced at every stage, before making a new moral judgment (Experiment 3). We conclude that Moral Balancing is not a behavior maintained over time. However, some would argue that moral licensing effects should not persist in an oscillating pattern over time. Imagine a less ethical behavior at t_0 that is compensated by a more ethical one at t_1, and vice versa, an ethical behavior at t_0 that gives the license to an individual to behave less ethically at t_1. At that point, balance is 'restored', and it is difficult to make predictions regarding further effects on behavior at t_2 and beyond, or so some might argue.

On the other hand, participants in the Rule-based condition, approximated the idealized pattern of Moral Consistency behavior (Figure 3), when a re-evaluation process (manipulation of the mind-set plus (un)ethical recall) was included, before confronting each new moral scenario. In other words, there was some evidence that Moral Consistency could be maintained over time, if the mind-set was reinforced before each moral judgment. Either way, we overall conclude that ethical mind-sets (and their influence on prosocial choice) decay, unless reinforced continuously.

Moral Consistency is perhaps a more stable pattern of mind-set, since if a person is led into seeing himself/ herself as consistent, it is perhaps more natural to remain consistent – that is the very nature of consistency. On the other hand, Moral Balancing would seem to require the keeping of a running total of one's positive and negative acts, and once the initial stages are past, this tally-keeping may prove complex to maintain. It is easier to recall that one has consistently chosen the prosocial or anti-moral path and so keep that on, than it is to recall that one's last choice was pro, so the next one should be anti. This difference in stability might also account for the tendency in both Experiments 2 and 3 for the Moral Balancing group to show a continuing Moral Consistency after their initial response at Stage 1. Although all the data trended towards the middle of the scale, there was a residual difference between the Ethical Recall and Unethical Recall groups that persisted to the end.

Overall, some would argue that this tendency to converge to a neutral level of morality might be due to the low personal costs of the scenarios presented. Gneezy et al. (2012) showed that when recent prosocial behavior is personally costly, people interpret that behavior as a signal of their prosocial identity and that they are more likely to subsequently behave prosocially. Prosocial behavior involving lower cost, in contrast, offers a more ambiguous signal: prosocial behavior is clearly positive, yet because it came at no cost, it is less likely to be judged as diagnostic of one's prosocial disposition. Under these circumstances the positive act does not affect individuals' self-perceptions, presumably resulting in a reduction in subsequent prosocial behavior.

Our results question the importance of the concept of mind-sets in understanding prosocial choice, since, if such mind-sets cannot be maintained across more than a few choices, what value could they have in understanding the relevant behaviors? We see three directions for future research in addressing this important question. First, it is possible that an alternative mind-set induction procedure will reveal more lasting influences of mind-sets on prosocial choice.

Second, a related possibility is that the measurement of prosocial choice in the present experiments was inadequate. Perhaps people's prosocial choices do reflect patterns of consistency or balancing, across time, but such patterns can be revealed in realistic time scales of days or weeks, not within the limited duration of a psychology experiment. Also, there are merits and demerits of the different approaches regarding how we ask participants to respond to scenarios. We used a 7-point scale because it let us explore our hypotheses. Some would say that individuals who want to establish a balance between moral motives and selfish motives might achieve that by staying safely in the midrange of the scale. So balance can easily be achieved within each moral scenario, removing the necessity to balance over time. It may be the case that more interesting results would emerge with binary answering options (an ethical vs. an unethical alternative). However, the scale we opted to use did lead us to a particular interesting conclusion, namely that participants do neither Moral Balancing nor Moral Consistency, but rather want to achieve a middle ground.

Third, it is possible that the idea of manipulating mind-sets directly is flawed. In other words, perhaps there is a reality to the proposal that there are different mind-sets and these mind-sets can impact on prosocial choice, but perhaps these are stable individual characteristics. That is, people can have a particular mind-set, but the mind-set cannot be easily altered experimentally (at least in an effective way).

Finally, we would like to point out that most of the research on moral judgment and decision-making has been obtained through moral vignettes, questionnaires and thought experiments. As standard these methodologies seem to be, one could argue that they are all restricted in terms of ecological validity. Using alternative methods such as Ecological Momentary Assessment (Hoffman et al., 2014) would perhaps be a better way to capture moral events, experiences, and dynamics as they unfurl in individuals' regular habitats. All these issues reveal considerable challenges (and corresponding exciting directions) for future work, regarding our current understanding of moral judgments.

Chapter 7

Contemporary Morality: Moral Judgments in Digital Contexts.

Abstract

Nowadays, several of the situations in which we have to make decisions are in digital form. In a first experiment we explored moral judgments in a large (N=1010) sample and showed that people's moral judgments depend on the Digital Context (Smartphone vs. PC) in which a dilemma is presented, becoming more utilitarian (vs. deontological) when using Smartphones. To provide additional evidence, we ran a second (N=250) and a third experiment (N=300), where we introduced time constraints and we manipulated whether instructions drew attention to the amount of time for processing a moral judgment; our key finding of the impact of digital context on moral judgments was replicated. Additionally, our results challenge one of the key assumptions in Dual-Process Models of Moral Judgment, as we showed that the (assumed) hurried, often surreptitious nature of using smartphones, that one would argue is consistent with gut-feeling reactions, decreased the likelihood of deontological responses and increased utilitarian ones. We suggest that the increased psychological distance of using a Smartphone induces utilitarianism. This is the first study to look at the impact of the digital age on moral judgments and the results presented have consequences for understanding moral choice in our increasingly virtualized world.

General Introduction

In this digital age, we spend a lot of time interacting with computer screens, smartphones and other digital gadgets. We buy online, work on the cloud, our social relationships are online-based, etc. Thus, the contexts where we typically face ethical decisions and are asked to engage in moral behaviour have changed. Nowadays, moral dilemmas are often presented digitally, that is, relevant information is presented through and decisions are made on a technological device.

A key distinction regarding moral judgments concerns deontological versus utilitarian decisions (Singer, 1991). Recent dual-process accounts of moral judgment contrast deontological judgments, which are generally driven by automatic/unreflective/intuitive responses, prompted by the emotional content of a given dilemma, with utilitarian responses, which are the result of unemotional/rational/controlled reflection, driven by conscious evaluation of the potential outcomes (Greene et al., 2001; Greene & Haidt, 2002; Slovic, 2007, Koenigs et al., 2007). In this account, an individual's ethical mind-set (rule-based vs. outcome-based, Barque-Duran et al., 2015; Cornelissen et al. 2013) can play a central role. A deontological perspective evaluates an act based on its conformity to a moral norm (Alexander & Moore, 2008). By contrast a consequentialist/utilitarian perspective evaluates an act depending on its consequences (Sinnott-Armstrong, 2008).

People often believe that judgments about "right" and "wrong" should be consistent and unaffected by irrelevant aspects of a moral dilemma or by its context. However, studies have shown, for example, that manipulations of the language (foreign vs. mother tongue) in which a moral scenario is presented can affect moral judgments through increasing psychological distance from the situation, and so inducing utilitarianism (Costa et al, 2014). The choice of deontological versus utilitarian judgments can vary depending on the emotional reactivity triggered by the dilemma (Valdesolo & DeSteno, 2006; Wheatley & Haidt, 2005).

As such, establishing which conditions favor each of these two influences is fundamental to understanding the psychology of moral choice.

The present study explores whether using a digital device (Smartphone, PC), as hundreds of millions of individuals do every day, can have a systematic impact on these processes.

Construal Level Theory (CLT) provides a framework of considerable potential relevance by linking mental representations to moral judgment. Individuals' judgments, decisions, and behaviours can differ as a function of construal levels. CLT proposes that the same event or object can be represented at multiple levels of abstraction (see Trope & Liberman, 2010, for a review). More weight is given to global, abstract features at high-level construals, whereas local, concrete features are more influential at low-level construals. According to CLT, psychological distance is a major determinant of what level of construal is activated. Distancing a target on any dimension of psychological distance (i.e., time, space, social, and hypotheticality) leads to greater activation of high-level construals than low-level construals (Liberman et al., 2002). Crucially, this is often assumed to align with more utilitarian decision-making (Trope & Liberman, 2010). Gong et al. (2012) examined the idea that whether a person focuses on actions or outcomes while making moral choices depends on the psychological distance from the moral situation. They found that when the situation is perceived as far off, whether in time or space, consequentialist considerations loom larger; establishing that psychological distance from an event decreases deontological judgments and increases consequentialist choices. Furthermore, Aguilar et al. (2013) examined whether psychological distance gives rise to an abstract representation of actions that make goals more prominent and can help us ignore their immediate effects. In three experiments they confirmed that psychological distance increase consequentialism. In other words, that different manipulations of psychological distance increased participants' consequentialist choices.

Smartphones often serve as a go-to source for staying informed in a fast way, quickly checking email, getting from place to place, sharing moments in social media, sending brief messages, etc. It seems intuitive, to us at least, that the hurried, often surreptitious nature of using smartphones increases the distance between the subject and the representation of a scenario, as presented on the smartphone screen.

To summarize, we assume that moral judgments made on Smartphones would increase psychological distance. Psychological distance weakens the intensity of people's affective reactions, when making judgments and choices, such as the feelings of empathy that promote charitable giving (Williams et al., 2014). Furthermore, increasing psychological distance leads individuals to construe situations in more abstract terms, which in some circumstances also aligns with more utilitarian decision-making (Trope & Liberman, 2010). Thus, we are led to a clear prediction: when faced with moral dilemmas, a Smartphone context should induce more utilitarian judgments than a PC context. We first tested this prediction using three versions of the well-known Trolley Problem (Switch, Fat Man, Balanced; Thomson, 1985; see Methods sections). To provide additional support we also ran a second and a third experiment where we introduced a Time Constraint (10 seconds vs. Unlimited Time to respond) and where we manipulated Time Instruction, relating to how participants were given information about the time constraints for reaching a decision (Instructing Unlimited Time vs. No Time Instruction).

Pilot Study

This research is primarily based on two versions of the Trolley Problem, the Switch version and the Fat Man version (see shortly), as these have been extensively shown to lead to utilitarian and deontological judgments, respectively (Greene, 2001). But, we also wanted to identify a scenario, in which the relative utilitarian and deontological influences would be reasonably well-balanced. It is possible that digitality does affect the balance between deontological and utilitarian choices, but predominant influences in the original scenarios are too strong and so suppress any effect. To obtain a Balanced version of the task a pilot study was run.

Method

Sample

Forty-two experimentally naïve students at City University London received course credit for participating in the study (31 women, 11 men; mean age=20 years, *SD*=3.1).

Materials and Procedure

The experiment, designed in Qualtrics and run in a lab, lasted approximately 5 minutes. A Fat Man version of the Trolley Problem was presented. We modified the Fat Man scenario (briefly, one has to push a man onto the train tracks to avoid killing some workmen) by asking participants how many workmen they would need to save to be justified in taking the action. The aim was to maintain the emotionality of one of the choices but to increase the utilitarian approach of the other one by increasing the lives one could save. We refer to this scenario as the "Balanced" dilemma.

The dilemma presented a scenario like this: "You are standing on a footbridge over a trolley track. You can see a trolley hurtling down the track, out of control. You turn around to see where the trolley is headed, and there are some workmen on the track that exists under the footbridge. What do you do? You know of one certain way to stop an out-of-control trolley: drop a really heavy weight in its path. But where to find one? It just so happens that standing next to you on the footbridge is a big fat man, a really big fat man. He is leaning over the railing watching the trolley; all you have to do is to give him a little shove, and over the railing he will go, onto the track in the path of the trolley." Participants are normally asked to make a choice between (A): You can shove the man onto the track in the path of the train, killing him. Or (B): You can refrain from shoving the man onto the track, letting the workmen die. Instead, we asked participants not to choose one of the options but to write how many workmen would need to be saved, so that they would be undecided between Choice A and Choice B. In other words, how many "lives saved" would be needed, so that they do not know what to do, whether to shove the man (so killing him) or refrain from shoving the man and letting the workmen die. Participant responses on the specific number of workmen to be saved were the dependent variable in this pilot. From the results of this experiment, we then specified the settings of the Balanced version of the Trolley Problem, in Experiment 1.

Results Pilot Study

Participant responses had a mean score of 150, a median of 15, a mode of 2 and a range of 998. Based on these considerations, we decided to adopt the median response, 15 workers, for designing a corresponding balanced scenario. We so aimed to maintain the emotionality of one of the choices and to increase the utilitarian value of the other one, so that the scenario would have neither a utilitarian nor an emotional predominant bias. We used this Balanced version of the Trolley Problem together with the Switch and Fat Man dilemmas in Experiment 1.

Experiment 1

The objective was to explore whether a manipulation of the Digital Context (Smartphone vs. PC) can have an impact on moral judgment. Specifically, we wanted to test if making moral judgments using a Smartphone increased the number of utilitarian responses in comparison to when using a PC.

Method

Sample

A total of 1010 participants[7], all US residents, were recruited on-line and received $1 for doing the task (482 women, 528 men; mean age=31.7 years, *SD*=9.6). Sample sizes were based on extant research (Hofmann et al. 2014; Suter & Hertwig, 2011) and were determined prior to the start of the experiments; the stopping rule for data collection was enforced automatically, as data collection was done through the Amazon Mechanical Turk platform.

Materials and Procedure

The study was designed in Qualtrics, run on Amazon Mechanical Turk and lasted approximately 10-15 minutes. Digital Context (Smartphone vs. PC)[8] and Version of the Trolley Problem (Switch vs. Fat Man vs. Balanced) were manipulated between participants. We used the frequency of Utilitarian vs. Deontological Responses as the dependent measure.

Participants were randomly told to switch to a Smartphone or a PC after reading and agreeing the general instructions on Amazon Mechanical Turk. Having a smartphone was a pre-requisite to participate in the experiment, in the Smartphone condition. Participants in the Smartphone condition had to respond to all questions from their smartphone devices. As a manipulation check for this condition, we tracked and verified through Qualtrics that the responses were indeed made from an iPhone, Android, Windows Phone or Blackberry.

Participants were randomly allocated to one of these six conditions: (1) Smartphone/Switch; (2) Smartphone/Fat Man; (3) Smartphone/Balanced; (4) PC/Switch; (5) PC/Fat Man; (6) PC/Balanced.

[7] According to the Ofcom Communication Market Report (2014), more than eight in ten adults had household internet access and the most important device for internet access among four in ten 16-34 years old is the smartphone. Furthermore, studies conducted by Nokia, AT&T and T-Mobile in 2012 and 2013 state that the average person checks their phone 150 times per day (approximately every six minutes).

[8] In the Smartphone condition participants could do the experiment with the following devices: iPhone, Android, Windows Mobile Phone and BlackBerry. In the PC condition participants could use a desktop or a laptop computer.

One third of the participants (327 Participants) on each Digital condition were presented with the Fat Man version of the Trolley dilemma, where one imagines standing on a footbridge overlooking a train track. A small incoming train is about to kill five people and the only way to stop it is to push a heavy man off the footbridge in front of the train. This will kill him, but save the five people. A utilitarian analysis dictates sacrificing one to save five; but this would violate the moral prohibition against killing. Imagining physically pushing the man is emotionally difficult and therefore people typically avoid this choice (Thomson, J., 1985). According to our hypothesis, participants would be more likely to opt for sacrificing one man to save five when dealing with such moral dilemma using a smartphone in comparison to a PC, since this would increase psychological distance (leading to greater activation of high-level construals), which aligns with more utilitarian decision-making.

Another third of participants (313 Participants) were presented with the Switch dilemma, where the trolley is headed towards the five men, but you can switch it with a lever to another track, where it would kill only one man. People are more willing to sacrifice the one man by pulling the switch than by pushing him off the footbridge and the extensively supported explanation is that pulling the switch is less emotionally aversive. If the impact of using Smartphones vs. PCs is increased psychological distance and reduced emotional reactions, then in the Switch dilemma we should not find an effect of Digital Context, since the affective reaction in this dilemma is already low (in comparison to the Fat Man version).

The last third of participants (314 Participants) were presented with the Balanced version of the Trolley Problem. The Balanced dilemma had a setting similar to that in the Fat Man version, but with a different number of people one could save (15 instead of 5), so that utilitarian choice would increase.

All participants first completed a filler task (10 trivia questions) before responding to one of the versions of the Trolley Problem. A “catch question” was introduced in the experiment, to control for attention during the task (i.e. "If you are paying attention to this question please select answer '36' from the options below"). Then, participants were presented with one of the three moral scenarios (Switch, Fat Man or Balanced) where they had to choose between Choice A (utilitarian) or Choice B (deontological). In all cases the dilemma was presented with both text and an illustration. Subsequently, participants completed another filler task (10 trivia questions). Finally, participants were asked to complete The Big Five Inventory (John et al., 1991) questionnaire, which is considered a quick (44-items), reliable, and accurate measure of the five dimensions of personality. We considered that the impact of digital content on moral choice could also interact with personality characteristics (Penner et al., 1995; Ozer & Benet-Martínez, 2006) but the results did not lead to firm conclusions and therefore will not be reported further. In Figure 1a we illustrate the experimental paradigm used for the Smartphone condition and in 1b the three moral conditions.

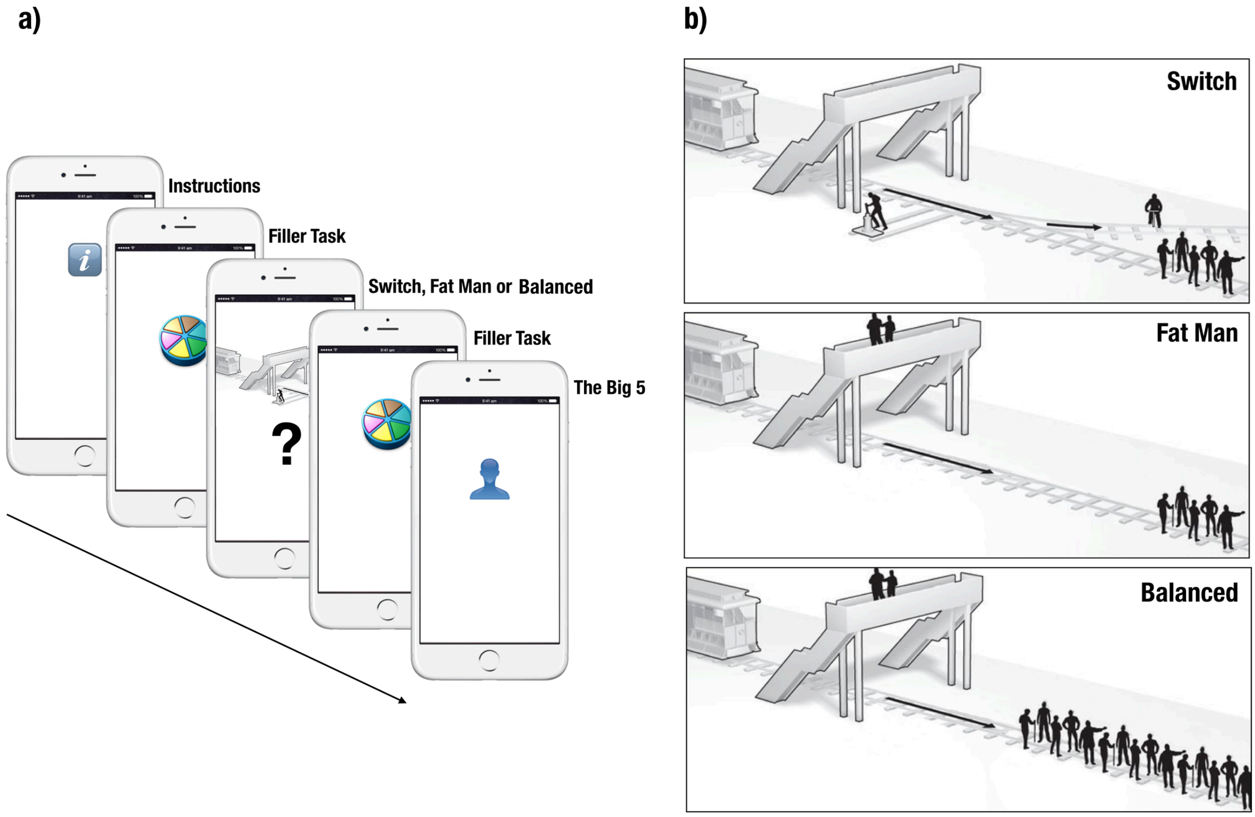
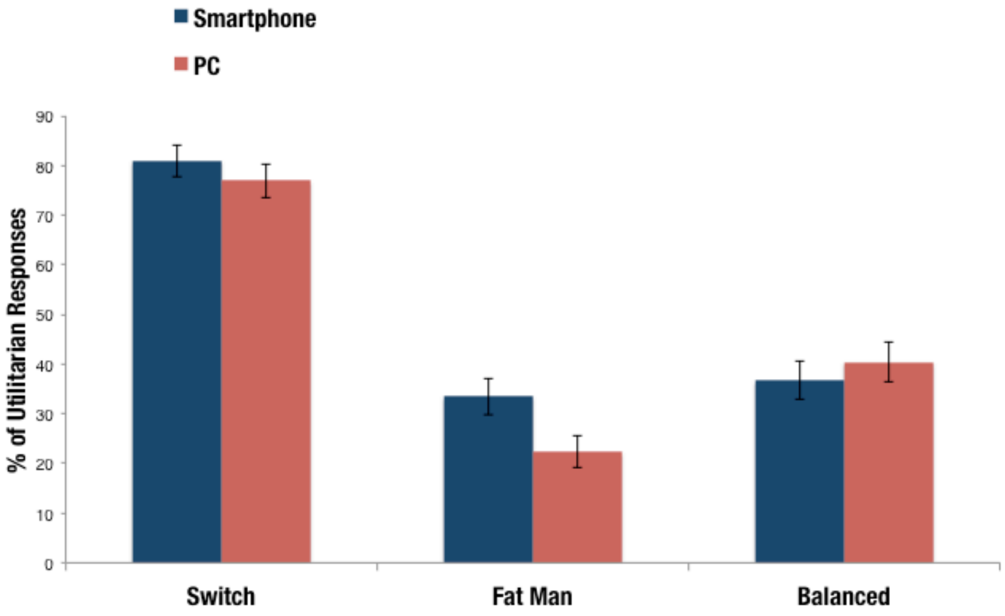


Fig. 1. A) The experimental paradigm used in the Smartphone condition in Experiment 1. B) The illustrations used in each of the three moral conditions (Switch, Fat Man and Balanced).

Results Experiment 1

We excluded participants whose first language was not English, as Costa et al., (2014) showed that the use of a foreign language (instead of a mother tongue) in a moral scenario increases psychological distance and induces utilitarianism when making moral judgments. We also excluded those participants who did not answer the catch question correctly. A total of 56 participants out of 1010 were thus excluded (the numbers of participants per condition, for all experiments, are reported in Appendix 4).

We first compared the percentage of Utilitarian Responses for the two Digital Contexts (Smartphone[9] vs. PC) on each of the three Versions of the Trolley Problem that were employed (Switch vs. Fat Man vs. Balanced; Figure 2).



[9] In the Smartphone condition, 39% of participants used an iPhone during the experiment, 58.5% an Android, 2.2% a Windows Mobile Phone and 0.2% a BlackBerry.

Fig. 2. Percentage of Utilitarian Responses for both Digital Contexts (Smartphone vs. PC) on each of the three versions of the Trolley problem (Switch vs. Fat Man vs. Balanced). Error bars represent standard errors[10].

As expected, in the Fat Man dilemma more participants avoided the act of pushing the heavy man off the footbridge in front of the train, presumably because of the emotional burden of this choice. More importantly, participants were more likely to opt for sacrificing the Fat Man (utilitarian response) to save five men when using a Smartphone (33.5%) than when using a PC (22.3%). A 2x2 chi-square test of independence was performed to examine the frequency of Utilitarian vs. Deontological Responses against Digital Context in the Fat Man condition and this revealed a significant association between the variables, χ^2 (1, N=327) = 5.15, p=.023. This result supports our hypothesis that moral judgments in Smartphones increase utilitarian decision-making, than when using a PC.

We then analyzed the frequency of Utilitarian vs. Deontological Responses, across the two Digital Contexts, in the Switch condition. Slightly more participants decided to sacrifice one man by pulling the switch than to do nothing and let five people die (80.9% for the Smartphone users; 76.9% for the PC users), but there was no evidence for an association between the two variables, χ^2 (1, N=313) = .741, p=.389. This result supports our expectation that in less emotional scenarios, such as the Switch dilemma, there is a reduced effect of Digital Context. That is, there is no difference in participants' moral judgments when using a Smartphone or a PC if the moral scenario is already highly utilitarian.

Finally, we examined the frequency of Utilitarian vs. Deontological Responses in the Balanced condition. Note, this condition was designed so that, in the PC condition at least,

[10] We computed errors bars for binary categorical data in this way: Let's say our estimate for the probability of assignment in a target category is p. Then $SE = \sqrt{\frac{\mathrm{pq}}{\mathrm{n}}}$, where q = (1-p).

there would be fairly equivalent utilitarian and deontological influences, and this was approximately the case. Regarding the manipulation of interest, 40.4% of participants decided to push the heavy man off the footbridge in the PC and 36.7% in the Smartphone conditions. Nevertheless, a chi-square test of independence showed that the relation between these variables was not significant, χ^2 (1, N=314) = .448, p=.503. The (tentative) conclusion from this experiment is that using a Smartphone rather than a PC has a reliable impact on moral judgments only when dilemmas or scenarios have high emotional content.

Experiment 2

The objective of Experiment 2 was to provide additional evidence for the increased number of utilitarian responses using a Smartphone (a digital device that is associated with fast responses) by manipulating the amount of time available to form a moral judgment. We wanted to explore Digital Context (Smartphone vs. PC) and Time Constraint (10 seconds vs. Unlimited time to respond) on moral judgments. It is possible that the effect of Digital Context is independent from that of Time Constraint, in which case we cannot explain the former in terms of (just) the latter. Alternatively, Time Constraint may 'mimic' the effect of Digital Context (e.g., an increase of utilitarian responses, in the fat man scenario, when participants are using a Smartphone), in which case the use of Smartphones may increase utilitarian responses only because of decreased response times. We also measured participants' affective reaction with the Self Assessment Manikin test (Bradley and Lang, 1994).

Method

Sample

A total of 250 participants, all of whom were US residents, were recruited on-line and received \$0.80 for doing the task (114 women, 136 men; mean age=32.9 years, *SD*=9.1).

Materials and Procedure

The study was designed in Qualtrics, run on Amazon Mechanical Turk and lasted less than 10 minutes. Digital Context (Smartphone vs. PC), Version of the Trolley Problem (Switch vs. Fat Man) and Time Constraint (10 seconds vs. Unlimited Time to respond) were manipulated between participants. There were therefore eight conditions. We used the frequency of Utilitarian vs. Deontological Responses as the dependent measure.

We followed the same procedure (verification and tracking methods) as in Experiment 1 for the Smartphone condition.

All participants followed a similar procedure as in Experiment 1. They first completed a filler task (10 trivia questions) including a catch question, as in Experiment 1. Then, participants were presented with one of the two moral scenarios (Switch or Fat Man). In all cases the dilemma was presented with both text and an illustration. Participants were alerted of the available time for responding depending on their condition (i.e. "You will only have 10 seconds to answer the question in the next screen" vs. "You will have unlimited time to answer the question in the next screen"). After the presentation of the scenario, in the "10 seconds" condition participants had to choose between Choice A (utilitarian) or Choice B (deontological), while a countdown timer appeared at the top of their screen (both Smartphone and PC). In contrast, in the "Unlimited Time" condition, participants had to make their judgment without time pressure. Finally, participants were asked to complete the Self Assessment Manikin test (Bradley and Lang, 1994), which is a technique that directly measures the pleasure, arousal and dominance associated with a person's affective reaction.

Results Experiment 2

We excluded a total of 10 participants out of 250 following the same criteria as in Experiment 1 (participants were rejected if they answered the catch question incorrectly or if English was not their first language).

As a manipulation check, we first examined the amount of time that participants took to finish the experiment. Overall, participants ended up spending more time in the Unlimited Time condition (5min 10s) than in the 10s condition (4min 32s), but this was not significant, t (238) = -1.916, p = .057.

We examined the differences in the percentage of Utilitarian Responses for the two Digital Contexts (Smartphone vs. PC) on each of the two versions of the Trolley Problem (Switch vs. Fat Man) and with or without time pressure (10s; Figure 3).

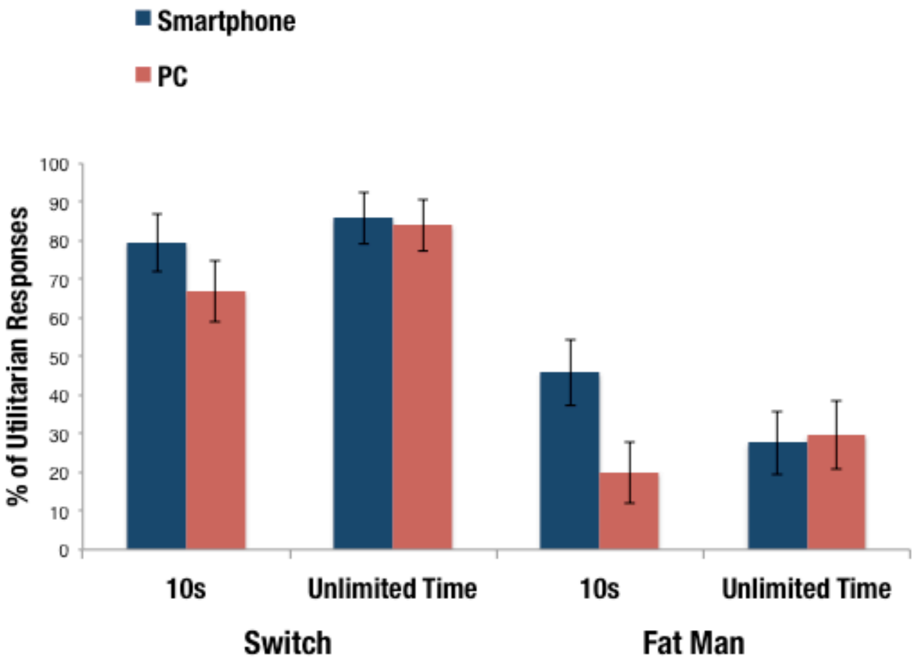


Fig. 3. Percentage of Utilitarian Responses for both Digital Contexts (Smartphone vs. PC) on each of the two versions of the Trolley problem (Switch vs. Fat Man) depending on Time Constraint (10s vs. Unlimited Time). Error bars represent standard errors.

As in Experiment 1, all statistical tests involve the variables frequency of Utilitarian vs. Deontological Responses and Digital Context (Smartphone vs. PC).

In the time pressure (10s), Switch condition, slightly more participants decided to sacrifice one man by pulling the switch than to do nothing and let five people die, when using a Smartphone (79.31%) than when using a PC (66.67%), but this difference was not reliable, χ^2 (1, N=65) = 1.282, p=.257.

Regarding the Unlimited Time condition, in the Switch condition, Digital Context also did not appear to play a role in moral judgments (85.71% and 83.87% for Smartphone and PC, respectively); regardless of Digital Context, we observed highly utilitarian responses. Thus, as before, the results in the Switch dilemma indicate that Digital Context and (as it seems) Time Constraint have a reliable impact on moral judgments only when dilemmas or scenarios have high emotional content. This result also supports our assumption that in less emotional scenarios, such as the Switch dilemma, any effect of either Digital Context or Time Constraint does not result in a reliable increase in utilitarian responding.

In the time pressure (10s), Fat Man condition, participants were more likely to opt for sacrificing the Fat Man (utilitarian response) to save five when using a Smartphone (45.7%) than when using a PC (20.0%), χ^2 (1, N=60) = 4.239, p=.04. At face value, these results challenge the assumption that there is an independent effect of Digital Context, over and above time pressure, which induces a utilitarian bias in judgments.

Finally, we examined participant's responses in the Unlimited Time, Fat Man condition. The results here challenge our conclusion from Experiment 1, in that there was no difference in Utilitarian vs. Deontological responses, between the Smartphone and PC conditions (27.58% and 29.63%, respectively, χ^2 (1, N=64) = 2.224, p=.136). In other words, when participants were allowed to spend unlimited time to resolve the dilemma (Unlimited

Time condition), the Digital Context effect vanished. We return to this finding in Experiment 3.

We also considered whether the impact of Digital Content on moral choice could interact with the perceived emotionality of the scenario/context or affective reactions, but the results did not lead us to firm conclusions and therefore will not be reported further (see Appendix 4).

Experiment 3

Experiments 1 and 2 left us with a major challenge to explain the difference in the Fat Man condition of Experiment 1 and in the corresponding condition in Experiment 2 (where the effect of Digital Context had disappeared). The only difference between these two conditions was that in Experiment 1 participants were not told anything regarding time, while in Experiment 2, in the equivalent conditions, participants were specifically told they had unlimited time. Could this perhaps have induced participants to respond in a more thoughtful way, based on their personal predisposition, and so in a way resistant to incidental biases in their judgments (notably from Digital Context)? In Experiment 3 we address this issue by manipulating the Time Instruction to either specify that there was unlimited time available for a moral judgment, or not mentioning time at all (Instructing Unlimited Time vs. No Time Instruction). We only used the Fat Man scenario, as it is for this scenario that the effect of interest was observed. We also measured participants' Response Time to resolve the dilemma.

Method

Sample

A total of 300 participants, all of whom were US residents, were recruited on-line and received \$0.8 for doing the task (120 women, 180 men; mean age=32.2 years, *SD*=8.9).

Materials and Procedure

The study was designed in Qualtrics, run on Amazon Mechanical Turk and lasted less than 10 minutes. Digital Context (Smartphone vs. PC) and Time Instruction (Instructing Unlimited Time vs. No Time Instruction) were manipulated between participants, using the Fat Man scenario (see Experiment 1 for details). We used the frequency of Utilitarian vs. Deontological Responses as the dependent measure. We also measured participants' Response Time.

Time Instruction was manipulated in the following way. Half the participants were given the instructions (as in the Experiment 2 Unlimited Time condition): "You will have unlimited time to answer the question in the next screen". The other half did not have any indication of the time they had to spend making their judgment (same procedure as in Experiment 1). For the rest of the task, all participants followed a similar procedure as in Experiment 1 and 2. We also employed the same verification/ tracking methods as in Experiments 1, 2 for the Smartphone condition. Finally, because of the large samples in Experiments 1, 2, in this experiment we included an additional question regarding whether participants had taken part 'in a similar trolley experiment before'. We informed them that there would be no penalty for an affirmative response (i.e., the participant could still do the experiment and get paid normally).

Results across all three experiments

In this section we report the results of Experiment 3 and then bring together the results from Experiments 1, 2 and 3, focusing on the Fat Man scenario (Figure 4).

First, we summarize the results from Experiment 3. In this experiment we excluded a total of 141 participants out of 300 (the total number of participants per condition are reported in Appendix 4) following the same criteria as in Experiment 1 and 2. One participant was rejected because they answered incorrectly to the catch question and one because English was not their first language. Additionally, 139 participants were eliminated because they said they had come across a moral choice in the context of the Trolley Problem before. The pattern of results does not change qualitatively if these participants are included, but we decided not to do so.

In this experiment we measured Response Time for the particular moral judgment, though we note that, as the experiment was run over the internet, the accuracy of these measurements can be questioned. Did participants in the Instructing Unlimited Time condition take longer to respond than ones in the No Time Instruction one? There was no evidence that this was the case (2x2 ANOVA with Digital Context and Time Instruction, F<1 for all effects). We suggest that the effects from Time Constraint and Time Instruction seen in Experiments 2, 3 could result in a change of the participants' ethical mind-set and approach to the problems, without showing clear differences in Response Time.

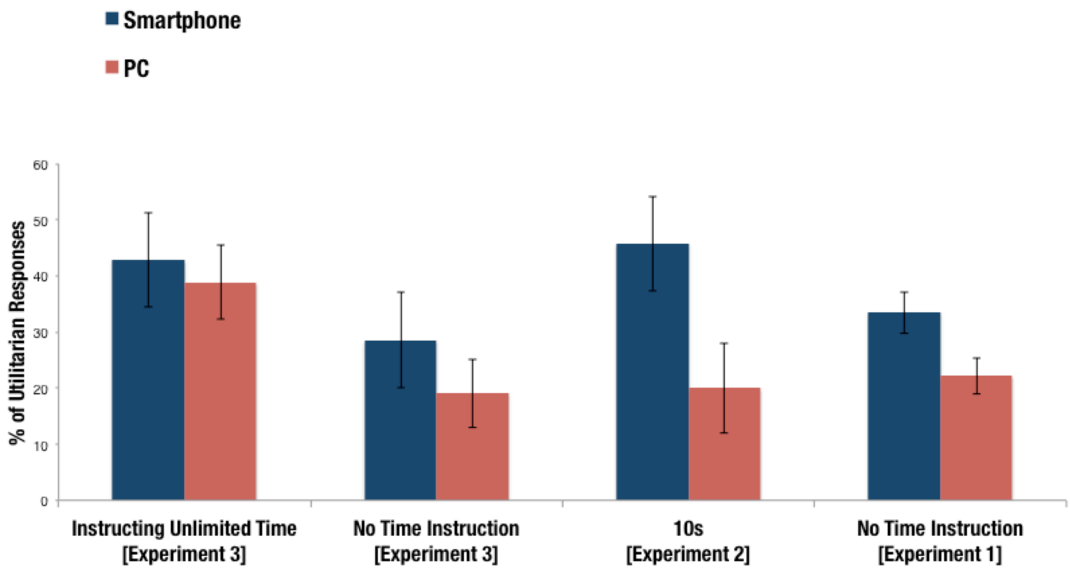


Fig. 4. Summary of the relevant results from Experiments 1, 2 and 3 for the Fat Man problem. The vertical axis shows percentage of utilitarian responses and the horizontal axis the conditions of interest. Error bars represent standard errors.

The left two bar clusters in Figure 4 show the results of Experiment 3. Interestingly, using the data from Experiment 3, we replicated the finding from Experiment 2, that the mere fact of "nudging" participants to use unlimited time resulted in utilitarian responses that were not biased by Digital Context. A 2x2 chi-square test with frequency of Utilitarian vs. Deontological Responses against Time Instruction (Instructing Unlimited Time vs. No Time Instruction) confirmed this conclusion, χ^2 (1) = 5.509, p = .018.

We next considered whether the results from Experiments 3 replicated the (apparently) inconsistent effects from Experiments 1 and 2 regarding Digital Context. The pattern of results from the No Time Instruction condition in Experiment 3 closely matched the corresponding results in Experiment 1. In Experiment 3, as expected, participants were more likely to opt for sacrificing the Fat Man (utilitarian response) to save five when using a Smartphone (28.6%) than when using a PC (19%). Even though the trend was as expected, a 2x2 chi-square test with frequency of Utilitarian vs. Deontological Responses against Digital Context (Smartphone vs. PC) was not significant, χ^2 (1, N=70) = 0.864, p=.35. However, after collapsing the data (for the identical Fat Man, No Time Instruction conditions) from Experiments 1 and 3, we obtained a significant association between frequency of Utilitarian vs. Deontological Responses and Digital Context (Smartphone vs. PC), χ^2 (1, N=397) = 6.27, p=.012. This result supports our hypothesis that moral judgments in Smartphones increase utilitarian decision-making, compared to when using a PC.

Importantly, we compared the results from Experiments 1, 2 and 3, with the view to explore whether Digital Context and Time Constraint provide independent biases in favor of

Utilitarian Responses. We explored this issue using the data from Experiment 1 and 3 (Fat Man, No Time Instruction) and the data from Experiment 2 (10s) in a three-way loglinear analysis, in which the highest-order interaction is of interest (Utilitarian vs. Deontological Responses x Digital Context x Time Constraint interaction), because this would tell us whether the preponderance of Utilitarian responses when using a Smartphone vs. a PC, in the Fat Man condition, is different across the 10s vs. the Unlimited Time conditions. This was not significant, χ^2 (1) = 1.035, p = .316, which indicated that a sense of time pressure does not introduce a bias favoring utilitarian judgments, over and above the bias from Digital Context.

General Discussion

This is the first study to look at the impact of digital context in moral judgments. We considered whether the increasing tendency for our judgments to be mediated through the use of technological gadgets might be changing our approach to moral dilemmas. We have shown that people's moral judgments become more utilitarian (vs. deontological) when using Smartphones as opposed to PCs and, moreover, that this effect cannot be explained as arising from a sense of time pressure.

We first consider the implications of these results for Dual-Process Models of Moral Judgment (Greene et al. 2001). A standard assumption is that moral dilemmas resolved in fast, gut-feeling conditions engage a deontological mode of responding, while utilitarian responses are typically the result of longer consideration and involve cognitive control. Our results from Experiment 1 challenge this assumption. If we assume that the use of smartphones, relative to PCs, is often hurried, such use would be consistent with gut-feeling reactions, so increasing the likelihood of deontological responses and decreasing utilitarian ones, but we obtained the opposite result.

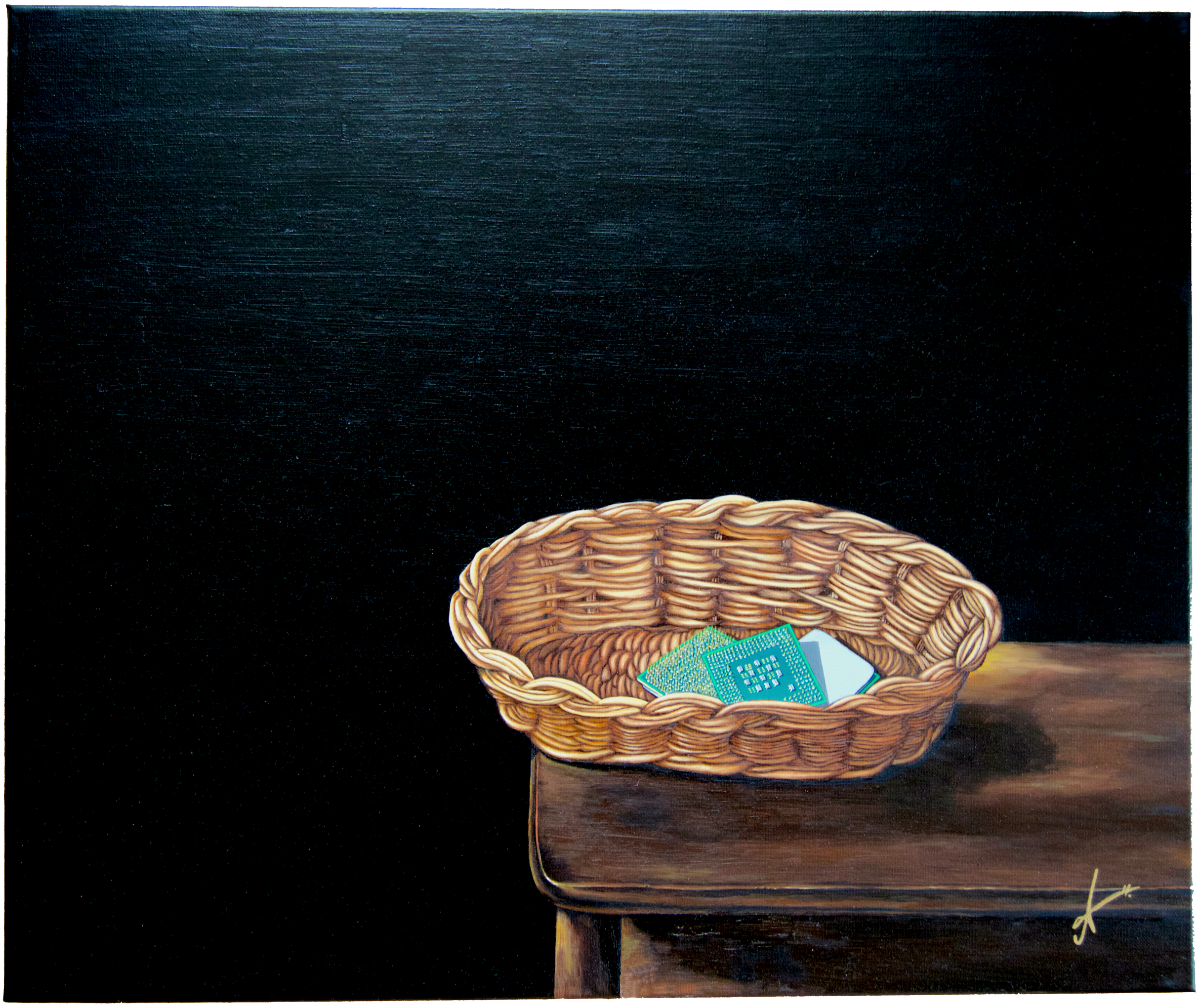
Other research has provided a more complex picture regarding the impact of time constraints on deontological vs. utilitarian judgments (Greene et al., 2001, 2009). Specifically, Suter & Hertwig (2011) showed that participants in a time-pressure condition (associated with fast, gut-feeling conditions), relative to a no-time-pressure condition (associated with longer consideration and higher cognitive control), were more likely to give deontological responses only in high-conflict dilemmas. By contrast, in low-conflict and in impersonal dilemmas, the proportion of deontological responses did not differ between conditions. The results from the present experiments partly support these differences between high-low conflict dilemmas. In less emotional scenarios (Switch), neither Digital Context nor Time Constraint resulted in a reliable increase in utilitarian responding. By contrast, in more emotional scenarios (Fat Man), our results question the well-established assumption (from Suter & Hertwig, 2011, amongst others) that hurried decisions enhance deontology, since we also showed in Experiment 2 that moral judgments under time constraints and in a context promoting more rushed responding (Smartphones) seem to make utilitarian judgments more common.

We next consider the results regarding Response Time in Experiment 3. There were no statistically reliable differences in reaction times between the various conditions. Nevertheless, we argue that the instructions regarding timing (i.e. "You will only have 10 seconds…" or "You will have unlimited time…") induce different mind-sets for making the moral judgments, for example, one of 'pressure' (regardless of whether in actual fact the time is sufficient or not) vs. one of a need to consider the issue carefully (again, regardless of how much time is actually spent on the problem). Indeed, our results indicate that drawing attention to unlimited time to answer a dilemma encourages more thoughtful (utilitarian) responses (Experiments 2 and 3). Clearly, more work is required to disentangle possible explanations for the exact effect of the different instructions concerning timing, but the crucial point regarding the present study is that our conclusion considering Digital Context and moral judgments is independent of such explanations.

Our results enrich the philosophical debate about the moral relevance of distance. There is an impressive body of evidence showing that psychological distance affects judgments and decisions in a wide range of psychological domains. According to Construal Level Theory (CLT), psychological distance can vary on at least four dimensions: temporal, spatial, social and hypotheticality (i.e. probability for a scenario to become reality; Trope & Liberman, 2010). Can we localize the particular effect of distance in going from a smartphone to a PC? Our results were inconclusive regarding a hypothesis that the psychological distance elicited by a smartphone decreased the intensity of people's affective reactions. It is possible that smartphones induce a greater distance in other respects or that an alternative procedure regarding the measurement of affective reaction may be more effective. For example, it might be the case that the use of digital devices interacts/mediates with the hypotheticality dimension. Therefore, we suggest that the standard dimensions for psychological distance and CLT need be further studied using new and up-to-date methods.

We note that insights from contemporary morality research have mostly been acquired through moral vignettes, questionnaire data and thought experiments such as trolley problems. As important as these approaches are, they are all limited by the artificial nature of the stimuli used and the non-natural settings in which they are embedded. Using Ecological Momentary Assessment (Hoffman et al., 2014) would perhaps be a better way to capture moral events, experiences, and dynamics as they unfold in people's natural environments. More generally, the present work reveals a need for the further systematic study of the factors and mediators affecting moral choice that condition the way we perceive and interact with our fast-changing environment and reality, all the more so given that, increasingly, governments, charities and other institutions engage in intense campaigns to encourage moral choices for important aspects of our way of life.

Special Artwork Chapter

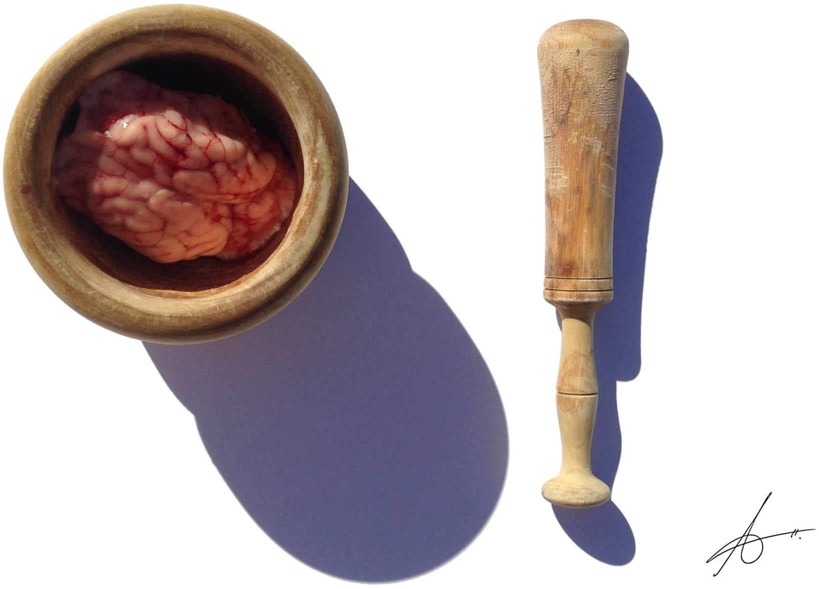


Basket of CPUs (2016)
Oil on canvas (60 x 49.5cm)
Private Collection

Salvador Dalí painted ‘Basket of Bread’ in 1945. The painting depicted a heel of a loaf bread in a basket, precariously situated on the edge of an uncovered table, against a starkly black backdrop, an omen to its own sacrificial destruction. This environment created the mystical, paroxysmic feeling of a situation beyond our ordinary notion of the real.

‘Basket of CPUs’ is a reinterpretation of Dalí’s painting in our Digital Age. Here, CPUs are depicted as “The New Bread”. A Central Processing Unit (CPU) is the electronic circuitry that carries out the instructions of a computer program by performing the basic arithmetic, logical, control and input/output operations specified by the instructions.

Will an Artificial Intelligent agent (with computational creativity abilities) ever be considered as creative as a human being? Using the words of the same Dalí: “This typically realist work is the one that has most satisfied my imagination. Here we have a painting about which nothing can be said: the total enigma!”



Forget (?) (2015)
Photography from the collection: *Brain Moments*

Theme 4

A Quantum Cognitive Approach on Game Theory

Statement of Contribution

Theme 4 is a collaborative work with Ismael Martinez-Martinez and Jacob Denolf. The author contributed to the development of the study concept, designed the experiments, analysed and interpreted the data for the studies presented. The author contributed to the writing of the manuscripts published.

List of publications for *Theme 4*:

Denolf, J., Martinez-Martinez, I., Barque-Duran, A. (*forthcoming*) A quantum-like model for complementarity of preferences and beliefs in dilemma games. *Journal of Mathematical Psychology.*

Martínez-Martínez, I., Denolf, J., Barque-Duran, A. (2016). Do Preferences and Beliefs in Dilemma Games exhibit Complementarity? *Lecture Notes in Computer Science*. 9535, 142-153.

Chapter 8

Preferences and Beliefs in Sequential Social Dilemmas

Abstract

In this chapter we use the data collected in a sequential prisoner's dilemma experiment conducted by Blanco et al. (2014) and we study the presence of intrinsic interactions between the preferences and the beliefs of participants in social dilemma games. We overview three effects concerning the interaction of these beliefs and the first and second move actions from the players. We argue that two of these three effects have a quantum-like nature, as shown by a violation of the sure thing principle. Here, we present the first steps towards a Quantum-like Preferences and Beliefs (QP&B) model. In Martinez-Martinez, Denolf and Barque-Duran (2016) and Denolf, Martinez-Martinez and Barque-Duran (*forthcoming*), we present a quantitative formalization of the model and proper fit to experimental data, showing successful predictions.

Introduction

In this thesis (*Chapters 1 to 5*) we have shown that especially over the last decade, there has been a growing interest in decision-making and cognitive models using a quantum probabilistic (QP) framework. We have seen how this development encompasses publications in major journals (see Deutsch, 1999; Pothos and Busemeyer, 2013; Wang et al., 2014; and Yearsley and Pothos, 2014; among others), special issues, and dedicated workshops, as well as several comprehensive books (Barque-Duran et al., 2016; Busemeyer and Bruza, 2012; Khrennikov, 2010; and Haven and Khrennikov, 2010). The majority of models presented in the quantum cognition literature addresses standard aspects of decision-making processes: similarity judgments (*Chapters 1 to 4*), the constructive role of articulating impressions (*Chapter 5*), order effects in belief updating (Trueblood and Busemeyer, 2011), and among others.

Nevertheless, not much literature has focused on strategic decision-making or game theory. When two or more agents interact, one agent is not only reacting to the information that she receives, but she is also generating information to other players. These strategic environments are different from standard decision-making scenarios under uncertainty, because each agent has to reason over two aspects of the problem: her actions and her expectations on the opponents' actions. Only a couple of studies have been published regarding this specific topic and making use of the QP tools to model the way agents process the information in a game: Pothos and Busemeyer (2009), Pothos et al. (2011), and Busemeyer and Pothos (2012). Other approaches in which the quantumness enters through an extension of the classical space of strategies and/or signals have also been discussed, e.g., by La Mura (2005), Brandenburger (2010), and Brunner and Linden (2013); as well as a model to analyze games with agents exhibiting contextual preferences (Lambert-Mogiliansky and Martínez-Martínez, 2014).

In this chapter we present the first steps towards a quantum-like preferences and beliefs (QP&B) model that mimics the experimental results from Blanco et al. (2014) and provides a novel theoretical approach regarding cognitive dynamics in strategic interactions. Our model takes full advantage of the notions of measurement used in quantum mechanics. We claim that the relationship between a player's beliefs and his preferences is inherently non-classical. We will redefine these two properties as complementary. As such, they cannot be measured at the same time, as the act of measuring one property alters the state of the other property. The non-classical nature of such a relationship and its application in cognition has already been discussed in, e.g., Denolf & Lambert-Mogiliansky (2016).

Experimental Paradigms in Strategic Decision-making

Standard version of the prisoner's dilemma game

As the data the QP&B model wants to replicate comes from a one-shot sequential-move prisoner's dilemma experiment from Blanco et al. (2014), we first introduce the classic or standard version of the prisoner's dilemma game (see Figure 1a). The symmetric prisoner's dilemma game is a game involving two players, *I* and *II*, that can choose among two actions: cooperate *(C)* or defect *(D).* The normal form of this game is defined by the following 2 × 2 payoff matrix

		Player *II*	
		C	*D*
Player *I*	*C*	(π_c, π_c)	(π_b, π_a)
	D	(π_a, π_b)	(π_d, π_d)

where the payoff entries satisfy the inequalities $\pi_a > \pi_c > \pi_d > \pi_b$.

The scheme of possible payoffs results as follows. If player *I* decides to cooperate, *I* can receive the second best possible outcome if the opponent *II* also cooperates, but *I*'s attempt to cooperate is exposed to being exploited by *II* if *II* decides to defect. In the later scenario, *II* would collect the first best outcome of value *I* while leaving *I* with the lowest payoff π_b. If player *I* decides to defect, then this player is securing not to obtain the lowest payoff, but at least an amount π_d if facing also defection from player *II*. And in the case that *II* decided to cooperate, then *I* is taking advantage of the situation and obtaining the maximum benefit π_a.

Technically we say that mutual defection is the Nash equilibrium of this game because there is no unilateral deviation that could make the deviating player earn more, but mutual cooperation is the Pareto optimal situation. Therefore, this game represents a social dilemma for the players: the individual choice of defection dominates the attempt to cooperate for any given choice of the opponent, which is not socially optimal. This is because if both players actually choose to defect, both of them generate a total payoff of $2 \cdot \pi_d$, which is by definition lower than the aggregate payoff if both of them coordinated in full cooperation, $2 \cdot \pi_c$. This formalizes a conflict or dilemma between the individual and the collective level of reasoning.

The standard version of the prisoner's dilemma game is as a one-shot strategic interaction with simultaneous moves by the opponents. This means, both players make their own individual decision (whether to cooperate or not) without knowing what the opponent is choosing. Once both players have chosen their strategy, both actions become public and the payoffs are generated.

Each player reacts to her own belief or expectation on the opponent's intention, and as a consequence, the preferred action in the dilemma crucially depends on the way players form their beliefs about the opponent moves. Therefore, it is important to understand how beliefs and preferences do (or do not) influence each other in the decision-making process.

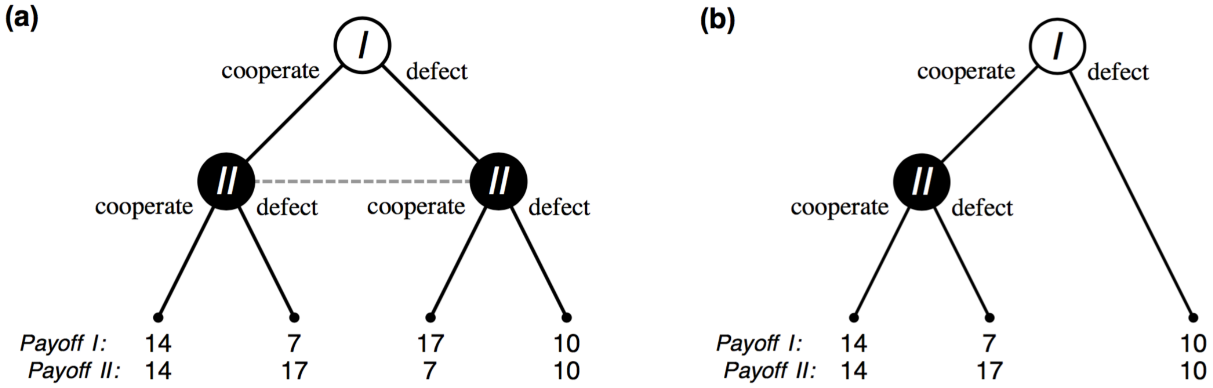


Figure 1: Game trees and payoffs for (a) Standard (simultaneous) Prisoner's Dilemma and (b) Sequential Prisoner's Dilemma in Blanco et al. (2014). (Source: Denolf, J., Martínez-Martínez, I., Barque-Duran, A. (*forthcoming*). A quantum-like model for complementarity of preferences and beliefs in dilemma games. Journal of Mathematical Psychology.)

Sequential-move prisoner's dilemma game

The experiment conducted by Blanco et al. (2014) focuses on a variation of the Prisoner's Dilemma game discussed above: a sequential one. They designed a within-subject sequential social dilemma experiment to observe interactions between the beliefs and preferences of players, which could have implications for the interpretation of observed behavior.

For the games analyzed in this chapter, it is enough for the reader to understand a player's belief as the subjective distribution with which the agent judges the likelihood of realization of each possible state of the world that the player could face, and which in general, influences the type of payoffs to be received from the actions. Also, we can consider the preferences as an individual's attitude towards a set of outcomes, typically reflected through the actions taken in an explicit decision-making process. For more details, see Lichtenstein and Slovic (2006).

In Figure 1 we showed the game tree of the game played in the experiment (b), and compared it to its standard (simultaneous) counterpart with equivalent payoffs (a). In the sequential version, the solution concept required is the Subgame Perfect Nash Equilibrium (SPNE), a usual refinement of the Nash Equilibrium (NE) when turning to sequential games. Solving by backwards induction, we see that it is in the best interest of Player *II* to defect if given the chance to move, which would leave Player *I* with a payoff of 7, and therefore *I* should choose defect at the beginning of the tree, because 10 is a better outcome. Thus, the sequential game maintains the content of the social dilemma because the SPNE implies that both players' incentives drive them towards mutual defection, even though they could obtain a higher social payoff if they coordinated on full cooperation.

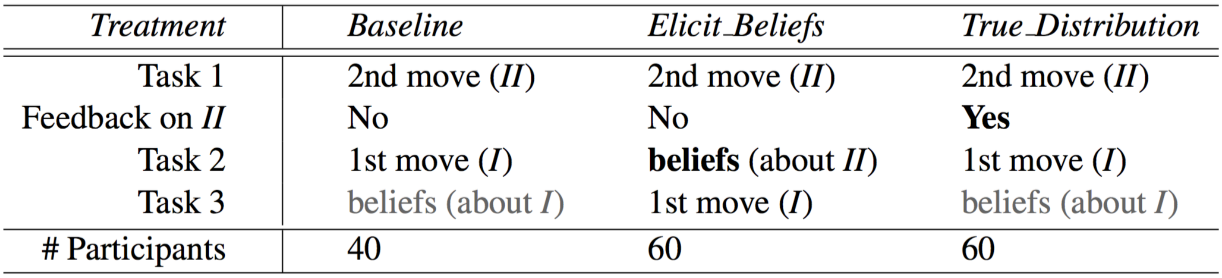
In this sequential variation, only the player *I* is bearing the risk of her cooperative choice being exploited by a selfish decision of player *II*. In order to restore the symmetry between the players, all participants in the experiment play the game in both roles (*I* and *II*). After all decisions have been made, the players are randomly matched into pairs, with the assignment of roles being random as well. Then they earn the payoffs determined by the relevant decisions, given their roles.

In the original experiment the authors designed three treatments that intersperse a belief-elicitation task with the choices of actions. Because of the sequential structure in the decision-making and because each choice can be observed (measured) at a time, the treatments differ in the order in which each task is performed and this allows to measure different correlations between actions (that are supposed to proxy the preferences of the players) and beliefs. In Table 1 we summarize the three different treatments.

Baseline: This treatment can be considered as a mere control group, such that the subjects play the game in its natural structure, with no attention paid to observing their beliefs. The players first choose what their action *II* will be and no information is revealed to them so that the participants' beliefs are not exogenously influenced. Subsequently, they choose what their action for the role of *I* will be, and finally they are given a meaningless question about their beliefs on the global rate of cooperation in the group as first movers. The informational gain of this last task is void because its only use is to balance the different treatments making their length comparable (both in time and number of tasks).

Elicit Beliefs. This treatment allows us to explore the effect of a measurement of the beliefs about the move by opponent *II* on the choice of action *I*. The players first choose what their action *II* will be, and then they have to reveal their belief about the rate of cooperation that they will receive from the second movers. Finally, they have to choose their action *I*. Thus, this treatment introduces a belief-measurement between the two choices of actions.

True Distribution. This treatment presents a somewhat 'similar' sequence of tasks for the players to the previous treatment Elicit Beliefs. The players begin by choosing their action *II*. Then, they are told what the true cooperation rate for action *II* was in their group. They finish by choosing the action *I*. This treatment differs from the previous one in that this time, the forecast of the opponents' move is not a belief generated by the players themselves, but true information being released to them exogenously.



Treatment	*Baseline*	*Elicit_Beliefs*	*True_Distribution*
Task 1	2nd move (*II*)	2nd move (*II*)	2nd move (*II*)
Feedback on *II*	No	No	**Yes**
Task 2	1st move (*I*)	**beliefs** (about *II*)	1st move (*I*)
Task 3	beliefs (about *I*)	1st move (*I*)	beliefs (about *I*)
# Participants	40	60	60

Table 1: Experimental treatments in Blanco et al. (2014).

Next, we present a summary of the results from the three experimental treatments (see Table 2). First, there is no significant difference in the cooperation rates as a second mover between treatments. This is to be expected as the question (measurement) regarding the choice of action in the role of player *II* is identical in all aspects over all treatments. Note especially that it is the first measurement performed in all treatments and therefore, it is not subject to the order effects targeted by this experimental design. The small variation in the proportion of cooperation reported for the Elicit Beliefs treatment (53.3% vs. 55% in the others) can be attributed to sample variance.

However, the cooperation rates in the role of first mover (player *I*) show meaningful differences. A chi square test across all three treatments yields a p- value of 0.007886 (χ2 = 9.6853, df=2). Starting with the first move cooperation rates of the Baseline treatment (21,5%) and the Elicit Beliefs treatment (55,0%), the null hypothesis of no difference between these proportions yields a p-value of .0007, (χ2 = 7.3661, df=1), clearly indicating a significant difference. There is only one procedural variation between these two treatments: Elicit Beliefs includes the elicitation of beliefs about the cooperation rate expected from the rivals *II* before the agents choose their action in the role of *I*. Thus, we can attribute the difference in the cooperation rate as player *I* to the effect that a measurement of the beliefs that a subject holds about the opponent *II* may have on her attitude toward the actions as first mover.

A similar result can be found for the first move cooperation rates of the Baseline treatment (27.5%) and the True Distribution treatment (56.7%). The null hypothesis claiming no difference between these proportions can be rejected, as it gives us a p-value of 0.004 (χ2 = 8.2674, df=1). For the first move cooperation rates (role *I*) of the Elicit Beliefs treatment (55.0%) and the True Distribution treatment (56.7%), the null hypothesis of no difference between these proportions yields a p-value of 0.85, (χ2 = 0.0351, df=1), indicating no significant difference between the result in the two treatments. In this sense, the incentivized elicitation of beliefs has an impact in the state of the subjects participating in the experiment similar to an update of beliefs via the acquisition of true information revealed exogenously.

Treatment	*Baseline*	*Elicit_Beliefs*	*True_Distribution*	*Total*
First mover (Player *I*)	27.5%	55.0%	56.7%	48.8%
Second mover (Player *II*)	55.0%	53.3%	55.0%	54.4%

Table 2: Average cooperation rates by treatment in the experiment by Blanco et al. (2014).

Three building blocks (effects) for the QP&B model

Analysing the data collected in Blanco et al. (2014), we identified three distinct effects exhibited by the participants. The differences in first move cooperation rates reveal the presence of a violation of the sure thing principle in the data, as

$$27.5\% = p(C_I) \neq \sum_i p(C_I|B_i) = 55\%,$$

with C_I the event of the player cooperating on the first move and B_i the event of the player answering that he thinks i opponents cooperate during the belief elicitation. This in turn points out the interest in using a quantum-like model to describe the behavior of the participants in this experiment, since classical statistics cannot account for them in a simple manner, while quantum-like easily do. Focusing also on the role of measurements allowed us to fully utilize the quantum paradigm. We first define these effects by looking at the observed outcomes of the beliefs and actions. These three effects all emerge as an influence of belief measurements and action measurements on each other or themselves. We used these findings as building blocks for our QP&B model.

Consensus Effect: Proof and an extensive commentary of the presence of this effect is presented in Blanco et al. (2014), where it is shown that players' beliefs are biased towards their own actions. As such, e.g., a player who cooperates as second mover will expect a higher second-mover cooperation rate amongst the other players. A visualization of this effect can be found in Figure 2. Viewing this in light of the performed measurements, the consensus effect denotes the influence of second mover action measurements on the beliefs of the same participant.

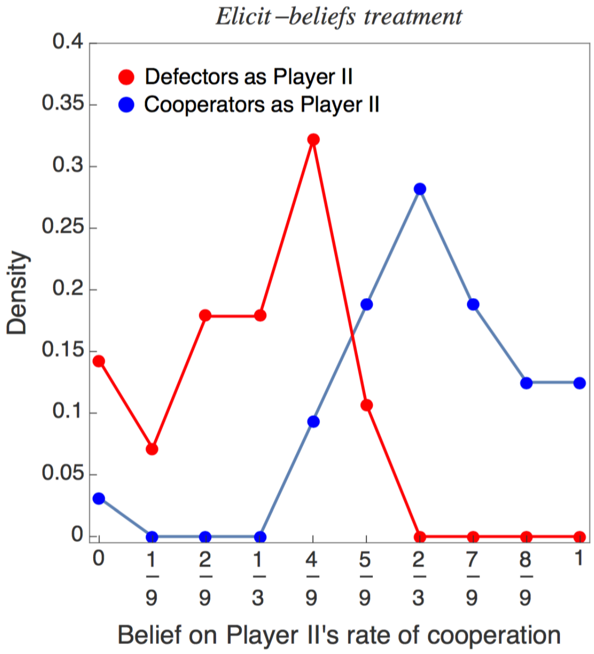


Figure 2: Second move defecting players (red line) believe that fewer opponents will cooperate. Second move cooperating players (blue line) believe more opponents will cooperate. Second move actions were measured before the beliefs. (Source: Denolf, J., Martínez-Martínez, I., Barque-Duran, A. (*forthcoming*). A quantum-like model for complementarity of preferences and beliefs in dilemma games. Journal of Mathematical Psychology.)

The reasoned player: The second effect is the influence belief measurements have on action measurements. As these actions are driven by one's preferences, this effect also encompasses the influence of the belief measurements on the preferences of the same player. We claim that the act of eliciting the beliefs of the player, fundamentally changes this player even when disregarding the exact outcome of this belief measurement. When the player is asked to form an opinion about the cooperation rate of his opponents, he changes into a more reasoned state about the opponent, in opposition to a more intuitive state when not explicitly asked to form this opinion. The average first move cooperation rate of players, after forming explicitly their beliefs about the cooperation of the opponent (Elicit Belief), is twice the average first move cooperation rate of players, which beliefs were not elicited (Baseline) (see Table 2). However, this cooperation rate in the Elicit Beliefs group is not differing significantly from the cooperation rate in the True Distribution group. In this group, participants received full information about the cooperation rate of the opponents and are therefore assumed to make a more deliberate decision. Since these cooperation rates are similar, we can assume that players are in a similar reasoned state in the Elicit Beliefs group.

Correlation between First Move in the second round and Second Move in the first round: The third effect we discuss is the correlation between a player's first move and second move. This is observed in all three conditions, as noted in Result 1, 2 and 3 from Blanco et al. (2014). That is, first move cooperators are likely to also cooperate on the second move and vice versa. We concur with Blanco et al. that this correlation is exhibited mostly through an indirect belief-based channel. This way, we attempt to include the observed correlation as a logical consequence of our previously described effects. The second move action measurement influences the first move action measurement through a player's beliefs. We assume this correlation to be classical in nature, as opposed to the two other effects.

Complementarity of Preferences and Beliefs as non-classical effects

We argue that two of the effects described above have a fundamentally non-classical nature. With both the consensus effect and the players being in reasoned state after having their beliefs elicited, the act of measuring itself influences the system, regardless of its outcome. In this regard, the measurements of actions and the measurements of beliefs seem to be complementary in the vein defined by Bohr (1950).

Two measurements are considered incompatible if the order in which the measurements are done changes the outcome, as the act of performing one measurement, influences the other measurements, regardless of outcome. As such, both measurements cannot be performed together, as the act of performing one of the measurements (without specifying its outcome), influences the other one. This could not be a consequence of any practical or experimental difficulties, but an inherent property of the system itself. These concepts elegantly deal with situations where violations of the sure thing principle emerge.

We consider the belief elicitation to be complementary with the action measurements, as this explains both the consensus effect and the reasoned player effect. This approach should not come as a surprise. First, using complementarity as an explanation for the consensus effect is argued in Busemeyer and Pothos (2012), where the consensus effect is seen as a form of social projection. Second, the idea of the player being more reasoned can be seen as a violation of the sure thing principle. These violations are a prime indicator of measurements not commuting, which is the definition of incompatible measurements. We mathematically show how the projective measurement formalism deals with our hypothetically compatible (first and second move actions) and incompatible (actions and beliefs) measurements in Denolf, Martinez-Martinez, Barque-Duran (*forthcoming*).

Overall, we argue that two of the effects described above have a fundamentally non-classical nature and, as player's first and second moves are driven by his/her preferences, we claim that a player's preferences and beliefs are complementary properties, which cannot be measured and/or exhibited at the same time. Roughly speaking, two measurements M_1 and M_2 are considered incompatible if the order in which the measurements are done changes the outcome, as the act of performing one measurement influences the other measurements regardless of the outcome. Mathematically speaking, this means that one or more projector matrices associated with outcomes of measurement M_1 do not commute with one or more projector matrices associated with outcomes of measurement M_2. If two measurements are maximally incompatible, no projector matrix associated with an outcome of measurement M_1 commutes with a projector matrix associated with an outcome of measurement M_2, and they are called complementary. As such, both measurements M_1 and M_2 cannot be performed together, as the act of performing one of the measurements (without specifying its outcome), influences the other measurement. We therefore propose to utilize the quantum statistical framework, based in a Hilbert Space to model these properties, as this type of models were originally devised to deal with similar complementary properties in physical settings. This idea results in a model with few parameters, giving a clear view of the non-classical nature of the relationship between a player's beliefs and his preferences.

A quantum-like model: towards a QP&B model

In this section, we present the basics for the development of the QP&B model. We provide a first description of the model and an initial fit to the data from Blanco et al. (2014) in Martinez-Martinez, Denolf and Barque-Duran (2016) and a full description of the model and a proper fit to the data in Denolf, Martinez-Martinez and Barque-Duran (*forthcoming*). Here, we just introduce the notation we use to represent concepts such as actions, preferences and beliefs in quantum-like terms (observables, measurements and orthonormal basis of their outcomes).

We consider the preferences of an agent as the individual's attitude toward the different elements of a set of outcomes, to be reflected in the choices observed along the sequence of decisions (Lichtenstein and Slovic, 2006). In this case, and because of the strategic nature of this decision-making process, the outcomes (possible payoffs to be obtained) depend on the actions (cooperate or defect) that a players chooses, but also on the choices made by a rival.

The actions can be represented by two orthogonal vectors $|C_i\rangle$ and $|D_i\rangle$. The two vectors form an orthonormal basis and span the Hilbert space $\mathcal{H}_i \equiv \mathbb{R}^2$, with $i \in \{I, II\}$ denoting the role in the game (as player *I* or *II*) for which such action is chosen. The player is considered to be in a superposition over these actions, being represented by a normalized state vector $|S\rangle$. The projection of the state vector onto the elements of the orthonormal basis defines the probability that the player chooses each of the actions, as a proxy of her preferences.

We consider the beliefs as the subjective distribution with which the agents judge the likelihood of realization of each relevant possible state of the world. The possible states in this setting concern the possible cooperation of opponents, as this, together with one's own actions, determine the outcome of the game. These beliefs are also represented by a set of orthogonal vectors $\{|B_j\rangle\}$, with the index j running from 0 to 9, and representing how many of the opponents (maximum 9) are expected to cooperate. This orthonormal basis also spans a Hilbert space, $\mathcal{H}_B$, with the player's beliefs being represented by a normalized state vector: a superposition over the orthonormal basis of beliefs. Straightforwardly, $j/9$ is the expected share of cooperation among the opponents, and $1 - j/9$ is the expected rate of defection.

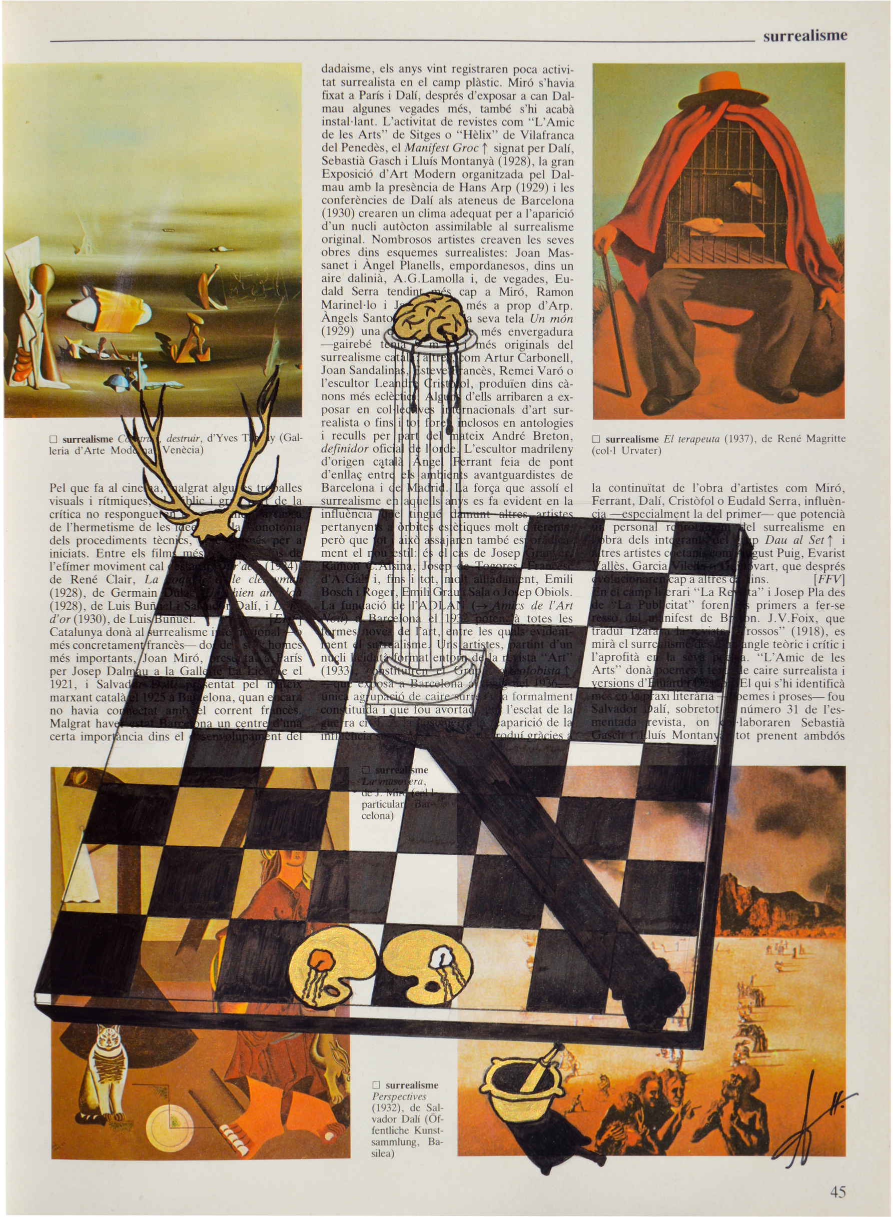
Quantum-like models use projective measurements to represent measurements being performed to the system of interest. In Martinez-Martinez, Denolf and Barque-Duran (2016), we apply this to model the observed behavior in the choice of action as player *II* in the data from Blanco et al. (2014). That is, we use projective measurements (with their resulting probabilities) to explain the first results observed in the data from Blanco et al. (2014) and discussed above. We also provide our first approach of the QP&B model. In Denolf, Martinez-Martinez, Barque-Duran (*forthcoming*) we provide a full description of the model and proper fit to the data.

General Discussion

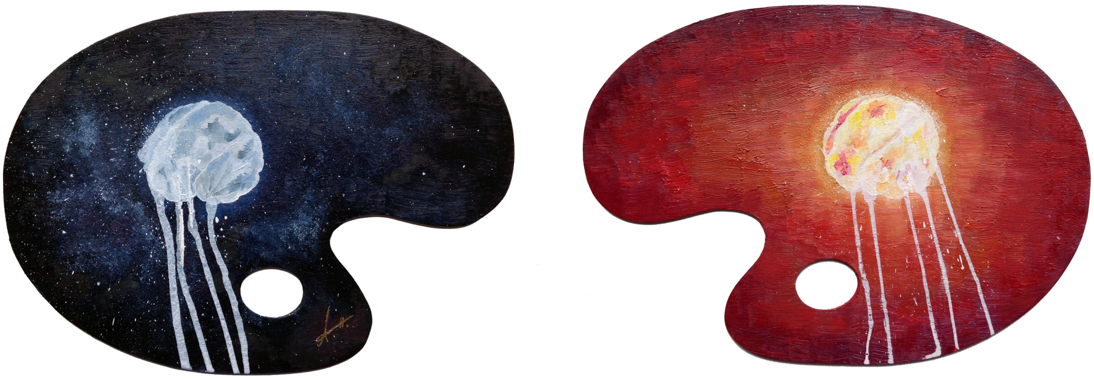
In this chapter we have shown how the relationship between a player's beliefs and his preferences is a prime candidate to receive a quantum treatment. Exploring the data collected by Blanco et al. (2014) during a sequential prisoner's dilemma, we identified and discussed three distinct effects. The three discussed effects naturally lead to our proposed Hilbert Space, projectors and resulting probabilities. Next to this mathematical sophistication, the use of complementary measurements seems to fit the complex relationship of beliefs and preferences on an interpretational level, as suggested in Busemeyer and Pothos (2012). Firstly, it explains the consensus effect as a form of social projection by making players beliefs aligning with their own actions, by assuming the action measurements influence the belief measurements. Second, the idea of the player being more reasoned can be seen as a violation of the sure thing principle. The natures of these last two effects both pointed us towards a quantum-like model for beliefs and preferences in a social dilemma game.

However, not all work is done yet and we would like to note an open question that needs further investigation. To fully investigate the quantum nature of the paradigm and for a complete view of the complementarity we would need a new experimental condition. Next to the Baseline group (action-action-beliefs) and Elicit Beliefs group (action-belief-action), an extension of the original experiment with a new treatment (belief-action-action) would be required. This might shed both experimental and theoretical new and conclusive light on the presumed incompatibility of the action and beliefs measurements.

Special Artwork Chapter



Study on Surrealism (2014)
Black and golden ink on paper (Encyclopaedia) (20.2 x 28.6cm)
Private Collection



Quantum Nightmares (2015)
Oil on palettes (22 x 30cm)x2
Private Collection

Epilogue

We are in the midst of writing a new chapter in the history of computation. Today, classical physics define our thinking, our experiences, our computers, and ultimately how we process information. The classical model of computing used today is on the verge of reaching its limits. Are the classical computational methods used in the study of cognition reaching their limits too? Are the classical scenarios and environments used in the study of decision-making and behaviour outdated? The application of alternative mathematical methods (e.g., quantum methods) and human-machine interaction techniques (e.g., digital contexts) to understand patterns in human cognition and to model them has the potential to solve certain problems that are impossible to unravel using today's standard approaches. Understanding the computational principles which guide human decision-making is of obvious significance. As we emphasized previously, such an understanding would allow us to predict consistent patterns in judgment, anticipate perhaps problematic decisions, and attempt to predict reactions to particular decision-making problems.

The present dissertation, on the one hand, had the aim to present theoretical and empirical examples of a new way of thinking about cognitive modeling (*Themes 1, 2* and *4*).

The research on *Theme 1* was focused on similarity judgments. In *Chapter 1* we presented the Quantum Similarity Model and considered how it can account for Tversky's (1977) key challenges. We saw how the QSM generalized the notion of geometric representations, but the emergent similarity metric was not distance-based, thus avoiding many of the criticisms Tversky (1977) made against distance-based similarity models. The QSM was developed to associate knowledge with subspaces. This idea of representations as subspaces allowed us to capture the intuition that a concept is the span of all the thoughts produced by combinations of the basic features that form the basis for the concept.

In *Chapter 2*, where the QSM also helped us to cover some key empirical results such as basic violations of symmetry, we presented a series of empirical tests with the aim to predict asymmetries in similarity judgments using two approaches. One approach for such asymmetries was that degree of knowledge, formalized as number of features, drove salience, which in turn drove asymmetries in similarity judgments (Tversky, 1977). The other approach was a novel model of similarity, based indeed on the mathematics of quantum theory, that could predict that asymmetries do not arise from the number of features, but rather from the number of independent features or dimensions, which are needed to represent a concept. Unfortunately, the work in that chapter showed inconclusive results regarding to both Tversky's (1977) (based on absolute number of features) and QSM's (based on principal components) accounts on predicting asymmetries. We discussed a number of methodological challenges that could be addressed in future extensions of this work.

In *Chapter 3*, with the aim to make progress regarding these issues, we provided an extensive collection of empirical results examining contextual influences in choice in a similarity task, with a set of novel experimental paradigms and different stimuli with different features. The results pointed to the conclusion that rather than having a consistent diagnosticity effect, we appeared to have a combination of diagnosticity and attraction. We also explored the possibility, that at least in some of the traditional diagnosticity paradigms, instead of context effects we might be simply looking at differences in feature salience. Moreover, it seemed that feature salience was malleable and was susceptible of attention changes. These conclusions are important in understand the diagnosticity effect and, in particular, the inconsistent findings regarding its presence.

In *Chapter 4*, we pursued an exploratory direction regarding similarity asymmetries and possible extensions for the QSM. Our starting points were similarity asymmetries of the form $Sim\ (NP, P) > Sim(P, NP)$, where the stimuli were simple perceptual ones, so that no differences in degrees of knowledge were expected and with a manipulation to alter distinguishability of the stimuli. Overall, the results pointed to the conclusion that for both conditions studied (percentage size difference) there were no significant preferences for statements in one direction to statements in another direction, that is, that there were no asymmetries in similarity judgments. The results showed an asymmetry only in one case, where the preference was towards the prototypical stimuli.

The research on *Theme 2* was focused on constructive judgments. In *Chapter 5* we examined whether the process of articulating an affective evaluation for a positively or negatively valenced stimulus could influence how an oppositely valenced subsequent stimulus was rated. We reviewed the methods used in an original study on the topic and we described how the use of QP offered a relatively simple mechanism by which the constructive effects of making a judgment could be modelled. Finally, we presented a new set of experiments that addressed some methodological limitations in the White et al. (2014) experiments and extended White et al.'s results with judgments of a completely different kind. In sum, the results of our experiments provided further support for the corresponding QP model. Furthermore, the predictions of the QP model in relation to constructive effects were not limited to affective evaluations of visual stimuli but could be extended to different judgments and stimuli.

The research on *Theme 4* was focused on strategic decision-making and game theory. In *Chapter 8* we showed how the relationship between a player's beliefs and his preferences in a social dilemma game is a prime candidate to receive a quantum treatment. Even though more work is needed regarding both the mathematical and conceptual elaboration of the quantum approach, the results presented in that chapter provided a clear empirical case and illustrated a framework for the principled study of such effects.

Overall, regarding the work and results presented in *Theme 1, 2* and *4*, plenty of promising directions for future research remain. Several interesting possibilities for extensions present themselves. One of them is to consider the same hypotheses that motivated these investigations and explore them in other stimulus domains. Can a quantum model account be developed to accommodate the key empirical effects? As part of the emerging research area of "Computational Social Science" we are convinced that there are other domains and many other analogous areas of research in which this method/approach can be successfully applied to understand and model not only individual behaviour but also the emerging global properties of social and cognitive systems.

An additional aim of the present thesis was to present theoretical and empirical progress on moral judgment and moral psychology (*Theme 3*).

In *Chapter 6* we presented the first study to look at the evolution of moral choice across time using a series of scenarios. We provided new empirical support for the hypothesis that ethical mind-sets moderate how an individual's behavioral history shapes his or her ethical behavior. An outcome-based mind-set was meant to lead to moral-balancing effects, whereas a rule-based mind-set to moral consistency. Furthermore, our three experiments shed some light on the persistence of these ethical mind-sets and on the evolution of moral dynamics, exploring whether moral patterns, such as Moral Balancing and Moral Consistency, could be maintained over time. When the manipulation of Mind-set and Recall was just made at the start, there was a quick regression to neutral performance. When the manipulation was reinforced before each moral choice, then one pattern of behavior was sustained, while the other was not. We concluded that ethical mind-sets (and their influence on prosocial choice) decay, unless reinforced continuously. Overall, our results questioned the importance of the concept of mind-sets in understanding prosocial choice, since, if such mind-sets could not be maintained across more than a few choices, what value could they have in understanding the relevant behaviors?

In *Chapter 7* we presented the first study to look at the impact of digital context in moral judgments. We considered whether the increasing tendency for our judgments to be mediated through the use of technological gadgets might be changing our approach to moral dilemmas. We showed that people's moral judgments become more utilitarian (vs. deontological) when using Smartphones as opposed to PCs and, moreover, that this effect could not be explained as arising from a sense of time pressure.

Our results had implications for Dual-Process Models of Moral Judgment (Greene et al. 2001) and challenged their assumptions. Under the supposition that the use of smartphones, relative to PCs, is often hurried, such use would be consistent with gut-feeling reactions, so increasing the likelihood of deontological responses and decreasing utilitarian ones, but we obtained the opposite result. More generally, this work revealed a need for the further systematic study of the factors and mediators affecting moral choice that condition the way we perceive and interact with our fast-changing environment and reality.

Overall, regarding the work and results presented in *Theme 3*, plenty of promising directions for future research remain. Autonomous systems are emerging whether people like it or not. Will they be ethical? Will they be good? And what do we mean by "good"? Many agree that artificial moral agents are necessary and inevitable. Others say that the idea of artificial moral agents intensifies their distress with cutting edge technology. There is something paradoxical in the idea that one could relieve the anxiety created by sophisticated technology with even more sophisticated technology. A tension exists between the fascination with technology and the anxiety it provokes. This anxiety could be explained by (1) all the usual futurist fears about technology on a trajectory beyond human control and (2) worries about what this technology might reveal about human beings themselves. The question is not what will technology be like in the future, but rather, what will we be like, what are we becoming as we forge increasingly intimate relationships with our machines. What will be the human consequences of attempting to mechanize moral decision-making?

Nowadays, when computer systems select from among different courses of action, they engage in a kind of decision-making process. The ethical dimensions of this decision-making are largely determined by the values engineers incorporate into the systems, either implicitly or explicitly. Until recently, designers did not consider the ways values were implicitly embedded in the technologies they produced (and we should make them more aware of the ethical dimensions of their work!) but the goal of artificial morality moves engineering activism beyond emphasizing the role of designers' values in shaping the operational morality of systems to providing the systems themselves with the capacity for explicit moral reasoning and decision-making.

As we have presented in this dissertation, one of the key distinctions regarding moral judgments concerns deontological versus consequentialist decisions. But are these ethical principles, patterns, theories, and frameworks useful in guiding the design of computational systems capable of acting with some degree of autonomy?

The task of enhancing the moral capabilities of autonomous software agents will force all sorts of scientists and engineers to break down moral decision-making into its component parts. Would making a moral robot only be a matter of finding the right set of constraints and the right algorithms for resolving conflicts? For example, a top-down approach takes an ethical theory, say, utilitarianism, analyzes the informational and procedural requirements necessary to implement this theory in a computer system, and applies that analysis to the design of subsystems and the way they relate to each other in order to implement the theory. On the other hand, in bottom-up approaches to machine morality, the emphasis is placed on creating an environment where an agent explores courses of action and learns and is rewarded for behavior that is morally praiseworthy. Unlike top-down ethical theories, which define what is and is not moral, in bottom-up approaches any ethical principles must be discovered or constructed.

Some claim that utilitarianism is not a particularly useful or practical theory arguing that calculations should be halted at precisely the point where continuing to calculate rather than act has a negative effect on aggregate utility. But how do you know whether a computation is worth doing without actually doing the computation itself? How do we, humans, do it? We generally practice what Herbert Simon, a founder of AI and a Nobel laureate in economics in 1982, called “bounded rationality”, which is the idea that when individuals make decisions, their rationality is limited by the available information, the tractability of the decision problem, the cognitive limitations of their minds and the time available to make the decision. In the *Prologue* of this dissertation we argued that *Homo Economicus* was no longer a reality and we asked ourselves if *Homo Heuristicus* was our best approach. The question here is whether a more restricted computational system, weighing the same information as a human, would be an adequate moral agent. But just as utilitarians do not agree on how to measure utility, deontologists do not agree on which list of duties apply and contemporary virtue ethicists do not agree on a standard list of virtues that any moral agent should exemplify.

Human-computer interactions are likely to evolve in a dynamic way, and the computerized agents will need to accommodate these changes. The demands of multiapproach systems thus illustrate the relationship between increasing autonomy and the need for more sensitivity to the morally relevant features of different environments. We should focus on the incremental steps arising from present technologies that suggest a need for ethical decision-making capabilities and how the research on human-machine interaction feeds back into humans' understanding of themselves as moral agents, and of the nature of ethical theory itself. As we stated previously, the particular characteristics of human decision-making are a fundamental aspect of what it means to be human.

Appendices

Appendix 1: The effect of masks in similarity judgments.

Experiment

The aim of this experiment was to explore in more detail the possibility that the mask interacts with the ability of participants to perceive the stimuli. It could well be that the effectiveness of the mask depended on the similarity of the grey level between the mask itself and the stimulus elements presented in the trials. We designed a 2(contrast of mask) x 2(contrast of stimuli) experiment to explore related possibilities.

Participants

Thirty-seven experimentally naive students at City University London received course credit for participating in the study.

Materials and Procedure

We used the same materials as in the experiment reported in *Chapter 5* and we manipulated the contrast of the stimuli and the masks. Specifically, the background for each stimulus was set to 255 (in RGB scale this is white colour) and the colours for each of the two conditions was set as: high contrast (192RGB) and low contrast (219RGB) for both stimuli and for the mask. We did not manipulate the percentage size difference (we used only P3 (9%) condition). From the 18 stimuli showed in Table 1 *(Chapter 5),* we used a small subset of adjacent pairs. We used the two adjacent smallest stimuli and the two adjacent largest stimuli for each of the clusters. Overall we had four pairs of stimuli, which were organised by sets. Each stimulus set had sixteen randomised trials and we presented each set nine times, so in total the experiment consisted of 144 trials (half of them in the opposite direction). We used a 2-alternative forced choice task, where we asked participants which item was the biggest, the first or the second one. As we intended to measure the actual confusability (i.e. how often participants confused two different stimuli as identical or vice versa) we present each stimulus pair more often than once. This allowed us to measure probability of confusion and to get an idea of where the threshold for clear discrimination was likely to lie.

Results and Discussion

First, we performed an analysis of variance (ANOVA) to test if there were any differences in performance (% of correct responses), depending on the contrast of the stimuli (192RGB vs. 219RGB) and contrast of the mask (192RGB vs. 219RGB). The contrast of the mask had a significant effect on performance, F(1,36)=5.181, p=.029. Nevertheless, the interaction between the contrast of the mask and contrast of the stimuli was not significant, F(1,36)=.001, p=.975. A graph with the mean values of correct responses obtained depending on contrast is shown in Figure 1.

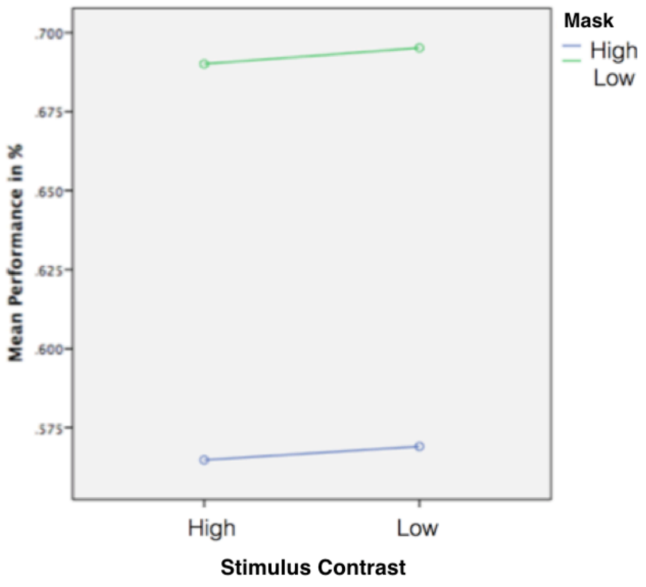


Figure 1. Mean values of correct responses (%) depending on contrast of the mask and contrast of the stimuli.

Second, we performed an analysis of variance (ANOVA) to test if there were any differences in reaction time (in milliseconds), depending on the contrast of the stimuli (192RGB; 219RGB) and contrast of the mask (192RGB; 219RGB). Again, the contrast of the mask had a significant effect on reaction time, F(1,36)=4.801, p=.035. Nevertheless, the interaction between contrast of the mask and contrast of the stimuli was not significant, F(1,36)=.232, p=.633. A graph with the mean values of reaction times obtained depending on contrast are shown in Figure 2.

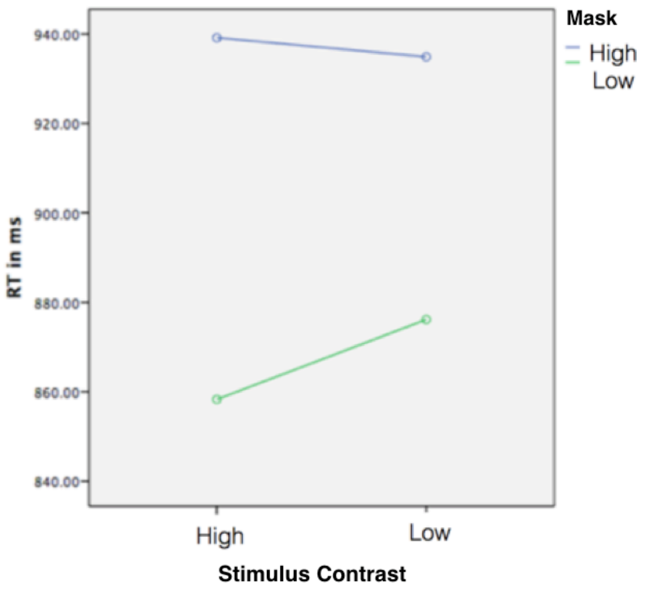


Figure 2: Mean values of reaction times (in milliseconds) depending on contrast of the mask and contrast of the stimuli.

Overall, our results suggest that when the mask had low contrast, performance was better and when the mask had high contrast, performance was worse. Additionally, when the mask had lower contrast, the reaction time was shorter and when the mask had high contrast the reaction time was longer. Nevertheless, the interaction between contrast of the mask and contrast of the stimuli was not significant for both performance and reaction time cases. It might be the case that when a mask does not have the same contrast as the stimuli presented in a trial, people are better at discriminating the similarities/differences between the stimuli. In other words, when a stimulus is flashed upon the screen (3 seconds) and is immediately followed by a mask (a random curve segments of similar curvature to the stimulus and having a different contrast), the average accuracy in detecting the similarities with the second stimuli might be higher. Conversely, we might expect that having the same contrast level between stimuli and masks creates some sort of interference that affects the similarity judgment. These were the possibilities that motivated this investigation, but there was no evidence to support them and so we did not pursue this direction further.

Appendix 2: A QP model for the constructive effect of affective evaluation.

This section is adapted from White, Pothos and Busemeyer (2014) and White, Barque-Duran and Pothos (2015). It provides a brief summary on how Quantum Probability (QP) has been employed to create a cognitive model for the constructive effect of affective evaluation.

We represent the QP system using a two-dimensional real space where the subspaces of all possibilities are one-dimensional (rays) and coplanar. We assume that positive and negative affect are represented by orthogonal rays since certainty that e.g. an image is positive (which means that the state vector aligns with the positive affect ray) implies zero probability that the image is negative (Figure 1A). The representations corresponding to a positive image or a negative one are also represented by a set of rays. These rays, also assumed to be orthogonal (because the images in each pair in the experimental paradigm were selected so that they were unrelated (Figure 1B), are positioned close to their respective positive and negative affect rays.

One of the assumptions of the model regarding its dynamics (assumed to be unitary) is the following: when participants see a positive image the state vector is aligned with the positive image ray (Figure 1C). When they subsequently perceive the negative image, the impact on the state vector is to rotate by a fixed amount towards the negative affective ray. Then, the rating of the second image presented is assumed to correspond to the length of the projection onto the negative affect ray (Figure 1D). Squaring this length we have the probability of interpreting the second image negatively. That is, this squared length corresponds to the negativity in the rating.

What happens when we consider the impact of an intermediate rating? If the first image is a positive one, such a rating is likely to result in a positive impression (aligning the state vector with the positive affect ray). Then, when introducing the negative image, the state vector is rotated to the same extent, which brings the state vector closer to the negative affect ray (Figure 1E), thus leading to a longer projection along the negative affect ray.

This is how the prediction that the intermediate rating amplifies the negative impression of the subsequent image arises in the QP model, in the PN order. An analogous reasoning predicts that in the NP order an intermediate rating will amplify the positivity of the subsequent rating (Figure 1F).

Computational fits to the results were not the primary objective of the investigation in this chapter. We were interested only in the general, qualitative prediction that, introducing the intermediate rating, leads to a more negative rating for the second image (in the PN condition) and, analogously, a more positive rating for the second image (in the NP condition).

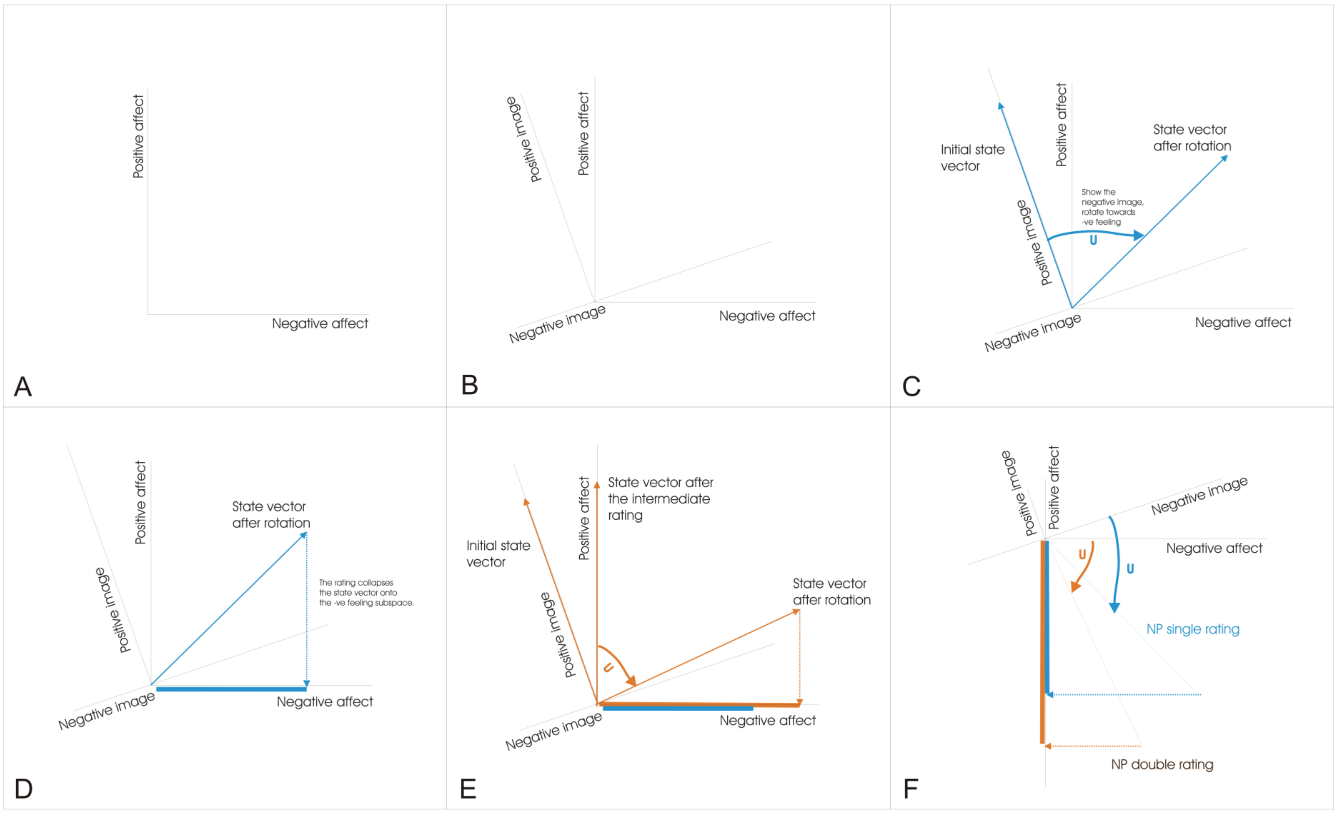


Figure 1. Quantum Probability Model: a QP model for the constructive role of measurement in the present experiments, in the PN condition (2A – 2E) and NP condition (2F). (Source: White et al. (2015). An investigation of a quantum probability model for the constructive effect of affective evaluation. Phil. Trans. R. Soc.

Appendix 3: Supplemental Material for *Patterns and Evolution of Moral Behavior: Moral Dynamics in Everyday Life.*

Supplemental Material 1

The 5 Moral Scenarios chosen from the pilot study to be presented in experiments 1, 2 and 3.

- During 2 days of this week, a bus from the National Health Services will be at your neighborhood asking people to donate blood.
  How likely are you going to donate blood?
  (7-point scale: -3=very unlikely, 3=very likely)
- You saw an advert saying that some volunteers are needed this weekend in a shelter of your city to help some poor families.
  How likely are you going to volunteer?
- You find a homeless person while going to work.
  How likely are you to give him some money?
- You've had a party with some friends at home. Now, you are tired and it's time to clean.
  How likely are you going to recycle, putting the rubbish in their corresponding bins?
- You are in the supermarket and you want to buy the coffee you always take. Now, you realize that in the shelf next to your coffee there's a new one, $1.5 more expensive, which is made from fair trade coffee (it helps producers in developing countries to make better trading conditions and promote sustainability).
  How likely are you going to buy the fair trade product?

Supplemental Material 2

Instructions-manipulation presented at the beginning of the experiment 2 and 3.

Condition 1: outcome mind-set, ethical recall

Now, this section deals with ethical behavior. Sometimes, we decide to do something for the positive consequences it has for other people. That means that we do something that benefits others, even though it might cause ourselves some inconvenience. For example, someone like you may:

...help another person with some work, even though you have to give up a free night for it.

...give away some money, for example to an NGO, that you could have used to buy something for yourself.

...lend out your scooter to someone who needs it, even though you are worried something may happen to the scooter.

Now please describe another example of something that someone can do that would benefit others, but would cause some inconvenience to him/her:

___

Describe who benefited from that action:

___

Describe what the benefit was for the person:

___

What was the inconvenience or the cost for the person who engaged in the action?

___

Now we focus on your behavior and, more specifically, your ethical behavior in the recent past. Please think of something you recently did, that benefitted someone else or others and that caused inconvenience or a cost to you.

Describe in detail what you did. Please take at least 5 minutes to do so.

___

Now specify who benefitted from that action:

___

What was the benefit for the other person?

What was the inconvenience or cost it caused you?

Condition 2: outcome mind-set, unethical recall

This section deals with ethical behavior. Sometimes, we decide to do something for the positive consequences it has for ourselves. That means that we do something that benefits ourselves, even though it might cause some inconvenience or cost to other people. For example, someone like you may:

... decide to go to the movies with some friends, even though a friend has asked you to help with some work.

... realize that a waiter has returned to much change and you decide to keep the 10$ difference.

... decide to go on a long journey with the family car, although you know your family will be worried.

Now please describe another example of something that someone can do that would benefit him or herself, but would cause some inconvenience to others:

Describe who was inconvenienced:

Describe what the inconvenience or cost was:
What was the benefit for the person who engaged in the action?

Now we focus on your behavior and, more specifically, your unethical behavior in the recent past. Please think of something you recently did, that benefitted yourself and that caused inconvenience or a cost to others. Describe in detail what you did. Please take at least 5 minutes to do so.

Specify who was hurt by your action.

What was the cost or inconvenience for the other person?

What was the benefit for you?

Condition 3: rule mind-set, ethical recall

This section deals with ethical behavior.   Sometimes, we decide to do “the right thing”. In those situations we believe we should act in a certain way, although we are tempted to do the opposite. The reason why we think we should act in a certain way is not based on the consequences of that action, but a personal rule or principle in which we believe, which we have learned through education or simply because we believe we are supposed to do something, even though we cannot explain why. For example, someone like you could:

...not be unfaithful to your partner, even though at a party there is an opportunity to do so.

...not cheat on a test, even though nobody would realize.

...not litter and hold on to a wrapper until you find a trash bin.

Now please describe another example of something that someone can do because s/he considers it “the right thing to do”, independent of the consequences:

Describe which rule or principle that was followed in your example:

Now we focus on your behavior and, more specifically, your ethical behavior in the recent past. Please think of something you recently did, that was “the right thing to do”, independent of its consequences. Describe in detail what you did. Please take at least 5 minutes to do so.

What was the rule, value, or principle you followed?

Condition 4: rule mind-set, unethical recall

This section deals with ethical behavior. Sometimes, we decide to do what is not “the right thing to do”, independent of its consequences. Even though it is possible that what we do does not hurt or inconvenience others, we shouldn`t do it because it violates personal rules, principles, or values, or simply because we consider it “not right”. For example, someone like you could:
...litter by throwing a candy wrapper in the street, not the bin.
...cheat on a test/exam.
...lie to your family about where you will spend the weekend.

Now please describe another example of something that someone can do which is not “the right thing to do”, independent of the consequences:

Describe which rule or principle that was not followed in your example:

Now we focus on your behavior and, more specifically, your unethical behavior in the recent past. Please think of something you recently did, that was not “the right thing to do”. Although you didn’t really hurt or inconvenience anyone else, you were tempted to do something that was not in line with your personal values or principles, or simply seemed “wrong”. Describe in detail what you did. Please take at least 5 minutes to do so.

What was the rule, value, or principle you did not follow?

Supplemental Material 3

Re-evaluation process. Instructions-manipulation presented at each stage before taking a new decision during experiment 3.

Conditions 1 and 2: Outcome mind-set

Sometimes, we decide to do something for the positive consequences it has for other people. That means that we do something that benefits others, even though it might cause ourselves some inconvenience. Sometimes, we decide to do something for the positive consequences it has for ourselves. That means that we do something that benefits ourselves, even though it might cause some inconvenience or cost to other people.

Now we focus on your behavior and, more specifically, your moral behavior in the recent past. Please think of the last moral decision you took in this study.

Describe in detail the scenario. Please take at least 5 minutes to do so.

Now specify who benefitted or who was hurt by your action:

What was the benefit or the cost/inconvenience for the other person?

What was the benefit or the cost/inconvenience it caused you?

Conditions 3 and 4: Rule mind-set

Sometimes, we decide to do "the right thing". In those situations we believe we should act in a certain way, although we are tempted to do the opposite. The reason why we think we should act in a certain way is not based on the consequences of that action, but a personal rule or principle in which we believe, which we have learned through education or simply because we believe we are supposed to do something, even though we cannot explain why. Sometimes, we decide to do what is not "the right thing to do", independent of its consequences. Even though it is possible that what we do does not hurt or inconvenience others, we shouldn't do it because it violates personal rules, principles, or values, or simply because we consider it "not right".

Now we focus on your behavior and, more specifically, your moral behavior in the recent past. Please think of the last moral decision you took in this study.

Describe in detail the scenario. Please take at least 5 minutes to do so.

____

What was the rule, value, or principle you followed or you did not follow?

____

Appendix 4: Supplemental Material for *Contemporary Morality: Moral Judgments in Digital Contexts*

Total number of participants per condition on each experiment (after data cleaning).

	Experiment 1					
	Smartphone			PC		
	Switch	Fat Man	Balanced	Switch	Fat Man	Balanced
N	157	161	158	156	166	156

	Experiment 2							
	Smartphone				PC			
	Switch		Fat Man		Switch		Fat Man	
	10s	Unlimited Time	10s	Unlimited Time	10s	Unlimited Time	10s	Unlimited Time
N	29	28	35	29	36	31	25	27

	Experiment 3			
	Smartphone		PC	
	Priming Unlimited Time	No Time Priming	Priming Unlimited Time	No Time Priming
N	35	28	54	42

Tables S1: Total number of participants per condition on Experiment 1, 2 and 3 (after data cleaning).

Participants' Affective Reactions.

Here we report the results from the Self Assessment Manikin test (SAM; Bradley and Lang, 1994), with its three factors (Pleasure, Arousal and Dominance). We first ran a two-way ANOVA with Version of the Trolley Problem and Digital Context as independent variables and Affective Reaction in Experiment 2 as the dependent variable. The Affective Reaction variable was computed as a combination of the three SAM factors. That is, all three factors were added to the variable independently of their category, and treated as one single factor. Neither factor nor the interaction were significant, $F < 1$. However, here we present an illustration of how data from the different types of Moral Scenarios, across conditions, would be placed on a 2-dimensional affective space defined by SAM arousal and dominance ratings. In Figure 1 we can see how both Moral Scenarios in the Smartphone condition, compared to the PC one, are placed towards the lower levels of the horizontal axis (meaning they feel more "Quiet/ Relaxed") and towards the lower levels of the vertical axis (meaning they feel more "Submissive/ Dependent").

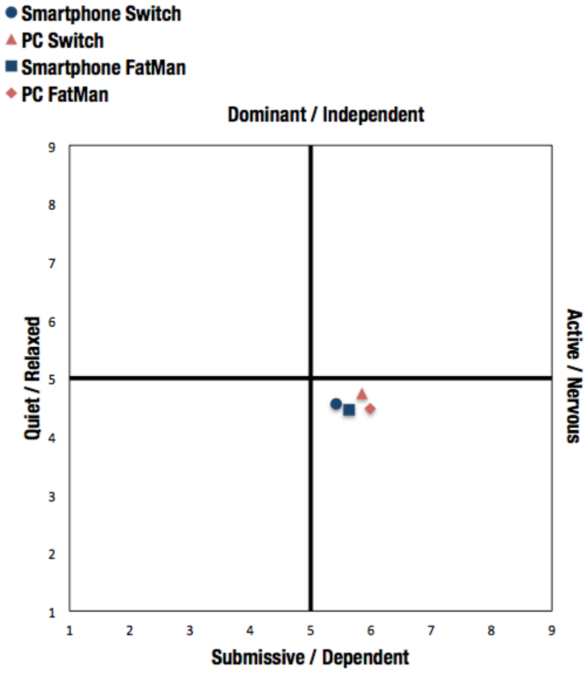


Figure S1: Illustration of the placement for each Type of Moral Scenarios and for each Digital Condition in a 2-dimensional affective space defined by SAM arousal and dominance ratings.

We then considered the results in the Fat Man scenario separately to see if, at least, the levels of participants' Affective Reaction were lower in the Smartphone condition compared to the PC one. Independent samples t-tests were employed to explore the three factors (Pleasure, Arousal and Dominance) in the Fat Man scenario across the two Digital conditions (PC vs. Smartphone). Indeed, lower levels of affective reaction for each of the three factors were reported in the Smartphone condition, as broadly expected, but none of the tests reported reached significance: Pleasure PC (M = 2.48, SD = 1.63), Pleasure Smartphone (M = 2.19, SD = 1.59), *t*(114) = .974, *p* = .332; Arousal PC (M = 5.98, SD = 2.42), Arousal Smartphone (M = 5.64, SD = 2.39), *t*(114) = .757, *p* = .450; Dominance PC (M = 4.48, SD = 2.17), Dominance Smartphone (M = 4.45, SD = 1.93), *t*(114) = .072, *p* = .943. Although the above results were in the predicted direction, there was little evidence to support our hypothesis that moral judgments made on Smartphones are less affected by the emotional reactivity elicited by a dilemma. We note that measuring the affective impact of a judgment is a complex issue, which presents experimentalists with many challenges including sensitivity to the underlying emotions. Future work could usefully further explore this issue, though in the context of the present study it was not possible to do so without deviating from the intended short format of the experiments.

Appendix 5: Glossary

Balancing (Moral): when the moral self-image exceeds the moral-aspiration level, the individual feels "licensed" to engage in more self-interested, immoral, or antisocial behavior (i.e., moral licensing). When the moral self-image is below the moral-aspiration level, people tend to experience emotional distress (Higgins, 1987; Klass, 1978) and become motivated to enact some corrective behavior (i.e., moral compensation).

Bayesian models of cognition: the Bayes rule is a simple theorem that follows from the classical probability definition of conditional probability. Suppose $\{H_1, \ldots, H_N\}$ is a set of hypotheses that you wish to evaluate, and *D* represents some data that provide evidence for or against each hypothesis. Then according to the definition of conditional probability, $p\ (H_i|D) = p(H_i \cap D)/p(D)$. Bayes rule uses the classical definition of joint probability to rewrite the numerator on the right hand of the equation: $p(H_i \cap D) = p(H_i) * p\ (D|H_i)$; and the Bayes rule uses the law of total probability to rewrite the denominator: $\sum j\, p(H_j) * p\ \left(D\middle|H_j\right)$. Bayesian models of cognition use these rules to construct models that predict how people make complex inferences from a set of observations.

Compatibility: two questions are compatible if they can be answered simultaneously, or even if they are answered sequentially, the order does not matter; two questions are incompatible if they have to be asked sequentially and the order does matter. The principle of complementarity posits that some questions are incompatible, and these incompatible questions provide different perspectives for understanding the world, and these different perspectives are needed for a complete understanding of the world. Classical probability models usually assume unicity, which means all events can be described within a single compatible collection of events. By comparison, incompatible events are unique to quantum theory, which does not impose the principle of unicity.

Conjunction fallacy: classical probability theory usually assumes that events are as subsets of a single sample space. This implies that the probability of an event A can never be less than the probability of the conjunction of A with another event B (A and B): $p(A) \geq p(A \cap B)$. However, violations of this law of classical probability, called the 'conjunction fallacy', have been found in empirical studies. The best-known empirical study is the famous Linda problem where human subjects rated the conjunction to be more probable. A quantum model has been proposed which explains the conjunction fallacy, together with other puzzling findings.

Consequentialist or Utilitarian Judgments: A consequentialist perspective considers whether an act is or is not morally right, depending on the consequences of that act (Sinnott-Armstrong, 2008). An individual understands an ethical behavior "because it benefitted other people" and an unethical behavior "because it hurt other people".

Consistency (Moral): after engaging in an ethical or unethical act, individuals are more likely to behave in the same fashion later on. This pattern is explained in terms of a psychological need to maintain one's self-concept (Aronson & Carlsmith, 1962), self-perception effects (Bem, 1972), or the use of behavioral consistency as a decision heuristic (Albarracín & Wyer, 2000; Cialdini et al., 1995).

Contextuality: constructing a classical probabilistic model involves defining relevant variables, which in turn form the basis of a joint probability distribution over the variables. However, research on entangled quantum systems has taught us that we cannot always assume the existence of joint distributions, and this approach to constructing probabilistic models can fail when applied to the observed data. This failure has come to be known as contextuality. It refers to the inability to construct the joint distribution over the variables.

Deontological Judgments: what makes an act right is its conformity to a moral norm (Alexander & Moore, 2008), i.e., principles that impose duties and obligations, such as not to break promises or not to lie. In this vein, an individual understands a behavior as ethical "because she followed an ethical norm or principle" or a behavior as unethical "because she did not follow an ethical norm or principle". Theyassumed to be driven by automatic/intuitive emotional processes (Greene et al., 2001; Greene & Haidt, 2002; Greene et al., 2004; Koenigs et al., 2007).

Disjunction fallacy: the classical probability theory also implies that the probability of the disjunction of an event A with another event B (A or B) can never be less than the probability of the event A: $p(A) \leq p(A \cup B)$. Violations of this classical probability rule, called 'disjunction fallacy', have been found in empirical studies and are described in this thesis. The same quantum cognition model used to explain conjunction fallacy also explains the disjunction fallacy.

Dutch Book Theorem (DBT): decision scientists and probability theorists use the Dutch book argument to show that classical probability theory is a rational theory. The idea originated with Bruno de Finetti, who proposed a game between a bookmaker and a better, where the bookmaker provided stakes for bets that reflected his probability of winning. The better could make a Dutch book against the bookmaker if the bookmaker's stakes for individual bets were chosen in a way that the sum across bets guaranteed that the better would win money and the bookmaker would lose money in every state of the world. If the bookmaker chooses stakes that satisfy an additive measure, then no Dutch book can be made against the bookmaker.

Hilbert space: an abstract and complete vector space defined on the complex field and possessing the operation of an inner product (or dot product). It is named after the famous mathematician David Hilbert. It extends vector algebra and calculus from the 2D Euclidean plane and 3D space to spaces with an arbitrary number of dimensions, including spaces of infinite dimensions. A finite Hilbert space is an N-dimensional vector space defined on a field of complex numbers and the vector space is endowed with an inner product.

Superposition: a basic principle of quantum probability theory. Classical probability theory assumes that, at any moment, a system is in a definite state with respect to possible states. This definite state can change stochastically across time but, at each moment, the state is still definite, and the system produces a definite sample path. By contrast, quantum probability theory assumes that, at any moment, a system is in an indefinite (technically dispersed) superposition state until a measurement is performed on the system. To be in a superposed state means that all possible definite states have the potential for being actualized, but only one of them will become actual upon measurement. The concept of superposition resonates with the fuzzy, ambiguous, uncertain feelings in many psychological phenomena.

The law of total probability: in classical probability theory, the law of total probability is a fundamental rule relating marginal probabilities to conditional probabilities. It is derived from the distributive axiom of Boolean logic: if $\{A, B, C\}$ are events, then $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$. Define $p(a), p(B) and\ p(\sim B)$ as the marginal probabilities of events A, B, and $\sim B$, respectively; and define $p(A|B)\ p(A|\sim B)$ as the conditional probability of event A conditioned on knowing either event B or $\sim B$, respectively. The law of total probability is then: $p(A) = p(B)p(A|B) + p(\sim B)p(A|\sim B)$. This law provides the foundation for inferences with Bayes networks. In some experiments, $p(A)$ is estimated from one condition, and $p(B)p(A|B) + p(\sim B)p(A|\sim B)$ is estimated from another condition, and violations of this classical law have been found.

The ‘sure-thing’ principle: Savage introduced the ‘sure-thing’ principle as a normative principle governing rational decision making. According to this principle, if under the state of the world *X*, a person prefers action *A* over *B*, and if under the complementary state of the world ‘not *X*’, the person also prefers action *A* over *B*, then the person should prefer action *A* over *B* even when she/he does not know the state of the world. Violations of the sure-thing principle have been empirically found (i.e. *A* is preferred over *B* for each known state of the world, but the opposite preference occurs when the state of the world is unknown).

References

Theme 1:

Aerts, D., & Gabora, L. (2005). A theory of concepts and their combinations II: A Hilbert space representation. Kybernetes, 34, 192-221.

Aerts, D. (2009). Quantum structure in cognition. *Journal of Mathematical Psychology*, 53, 314-348.

Aguilar, C. M., & Medin, D. L. (1999). Asymmetries of comparison. Psychonomic Bulletin & Review, 6, 328-337.

Ashby, G. F. & Perrin, N. A. (1988). Towards a Unified Theory of Similarity and Recognition. Psychological Review, 95, 124-150.

Atmanspacher, H., Filk, T., & Romer, H. (2004). Quantum zero features of bistable perception. *Biological Cybernetics*, 90, 33-40.

Barque-Duran, A., Pothos, E. M., Yearsley, J., Hampton, A., Busemeyer, J. R. & Trueblood, J. S. (2015). Similarity Judgments: From Classical to Complex Vector Psychological Spaces. In, E. Dzhafarov, R. Zhang, S. Joardan, and V. Cervantes (Eds.) *Contextuality from Quantum Physics to Psychology*. World Scientific.

Blutner, R. K. (2008). Concepts and Bounded Rationality: An Application of Niestegge's Approach to Conditional Quantum Probabilities. In L. Accardi, et al. (eds.), *Foundations of Probability and Physics 5*, Vol. 1101, pp. 302-10. American Institute of Physics Conference Proceedings. New-York.

Bordley, R. F. (1998). Quantum mechanical and human violations of compound probability principles: Toward a generalized Heisenberg uncertainty principle. *Operations Research*, 46, 923-926.

Bowdle, B. F. & Gentner, D. (1997). Informativity and asymmetry in comparisons. Cognitive Psychology, 34, 244-286.

Bowdle, B. F. & Medin, D. L. (2001). Reference-point reasoning and comparison asymmetries. In J. D. Moore & K. Stenning (Eds.) Proceedings of the 23rd Annual Conference of the

Cognitive Science Society, pp. 116 – 121. Psychology Press.

Bruza, P. D. (2010). Quantum Memory. *Australasian Science*, 31, 34-35.

Bruza, P. D., Kitto, K., Nelson, D. & McEvoy, C. L. (2009). Is there something quantum-like about the human mental lexicon? Journal of Mathematical Psychology, vol. 53, pp. 362-377.

Busemeyer, J. R. & Bruza, P. (2011). Quantum models of cognition and decision mak- ing. Cambridge University Press: Cambridge, UK.

Busemeyer, J. R., Pothos, E. M., Franco, R., & Trueblood, J. (2011). A quantum theoretical explanation for probability judgment errors. Psychological Review, 118, 193-218.

Busemeyer, J. R., Wang, Z., & Townsend, J. T. (2006). Quantum dynamics of human decision-making. *Journal of Mathematical Psychology*, 50, 220-241.

Bush, R. R. & Mosteller, F. (1951). A model for stimulus generalization and discrimination. *Psychological Review*, 58, 413-423.

Casale, M. B., Roeder, J. L., & Ashby, F. B. (2012). Analogical transfer in perceptual categorization. Memory & Cognition, 40, 434-449.

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M. J., Voorspoels, W., & Storms, G. (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. Behavior Research Methods, 40, 1030-1048.

Eisler, H. & Ekman, G. (1959). A mechanism of subjective similarity. *Acta Psychologica*, 16, 1-10.

Evers, E. R. K. & Lakens, D. (2014). Revisiting Tversky's diagnosticity principle. Frontiers in Psychology, Article 875.

Gärdenfors, P. (2000). Conceptual spaces: the geometry of thought. MIT Press.

Gentner, D. (1983). Structure-mapping: a theoretical framework for analogy. Cognitive Science, 7, 155-170.

Gleitman, L. R., Gleitman, H., Miller, C., & Ostrin, R. (1996). Similar, and similar concepts. Cognition, 58, 321-376.

Goldstone, R. L. (1994). The role of similarity in categorization: providing a ground- work. Cognition, 52, 125-157.

Goodman, N. (1972). Seven strictures on similarity. In N. Goodman, Problems and projects (pp. 437-447). Indianapolis: Bobbs-Merrill.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: exploring representations and inductive biases. Trends in Cognitive Sciences, 14, 357-364.

Hampton, J. A. (2007). Typicality, graded membership, and vagueness. *Cognitive Science*, 31, 355-384.

Huber, J., Payne, J., & Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. Journal of Consumer Research, 9, 90–98.

Hughes, R. I. G. (1989). *The structure and interpretation of quantum mechanics*. Harvard University press.

Isham, C. J. (1989). *Lectures on quantum theory*. Singapore: World Scientific.

Jones, M.N., Gruenenfelder, T.M., Recchia, G.: In defense of spatial models of lex- ical semantics. In: Carlson, L., Hlscher, C., Shipley, T. (eds.) Proceedings of the 33rd Annual Conference of the Cognitive Science Society, pp. 3444–3449. Cognitive Science Society, Austin (2011)

Kintsch, W. (2014). Similarity as a function of semantic distance and amount of knowledge. *Psychological Review*, 121, 559-561.

Khrennikov, A.Y.: Ubiquitous Quantum Structure: From Psychology to Finance. Springer, Berlin (2010)

Krumhansl, C. L. (1978). Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review, 85*, 445-463.

Krumhansl, C. L. (1988). Testing the density hypothesis: comment on Corter. *Journal of Experimental Psychology: General, 117*, 101-104.

Lambert-Mogiliansky, A., Zamir, S., & Zwirn, H. (2009). Type indeterminacy: A model of the kt (Kahneman-Tversky) man. Journal of Mathematical Psychology, 53, 349-361.

Larkey, L. B. & Love, B. C. (2003). CAB: Connectionist analogy builder. Cognitive Science, 27, 781-794.

Medin. D.L., Goldstone, R.L., & Gentner, D. (1993). Respects for Similarity. *Psychological Review, 100*, 254-278.

Michelbacher, L., Evert, S., Schtze, H.: Asymmetry in corpus-derived and human word associations. Corpus Linguist. Linguist. Theor. 7(2), 245–276 (2011)

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. Journal of Experimental Psychology: Learning, Memory, and Cognition, 10, 104- 114.

Nosofsky, R. M. (1991). Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology*, 23, 94–140.

Oaksford, M. & Chater, N. (2009). Précis of Bayesian rationality: the probabilistic approach to human reasoning. Behavioral and Brain Sciences, 32, 69-120.

Parducci, A. (1965). Category judgment: A range-frequency model. Psychological Review, 72, 407-418.

Polk, T. A., Behensky, C., Gonzalez, R., & Smith, E. E. (2002). Rating the similarity of simple perceptual stimuli: asymmetries induced by manipulating exposure frequency. Cognition, 82, B75-B88.

Pothos, E. M. & Trueblood, J. S. (2015). Structured representations in a quantum probability model of similarity. *Journal Mathematical Psychology*, 64, 35-43.

Pothos, E. M., Busemeyer, J. R., & Trueblood, J. S. (2013). A quantum geometric model of similarity. *Psychological Review*, 120, 679-696.

Pothos, E., Barque-Duran, A., Yearsley, J., Trueblood, J., Busemeyer, J., Hampton, J. (2015). Progress and current challenges with the Quantum Similarity Model. *Frontiers in Psychology*.

Pothos, E. M. (2005). The rules versus similarity distinction. Behavioral & Brain Sciences, 28, 1-49.

Pothos, E. M. & Busemeyer, J. R. (2009). A quantum probability explanation for violations of rational decision theory. *Proceedings of the Royal Society B*, 276, 2171-2178.

Pothos, E. M. & Busemeyer, J. R. (2013). Can quantum probability provide a new direction for cognitive modeling? *Behavioral & Brain Sciences*, 36, 255-327.

Rosch, E. (1975). Cognitive reference points. Cognitive Psychology, 7, 532-547.

Rosch, Eleanor and Lloyd, Barbara B. (eds), Cognition and categorization 27-48. Hillsdale, NJ: Lawrence Erlbaum.

Shafir, E. B., Smith, E. E., & Osherson, D. N. (1990). Typicality and reasoning fallacies. Memory & Cognition, 18, 229-239.

Shepard, R. N. (1987). Toward a Universal Law of Generalization for Psychological Science. *Science*, 237, 1317-1323.

Simonson, I. (1989). Choice based on reasons: The case of attraction and compromise effects. Journal of Consumer Research, 16, 158–174.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist networks. Artificial Intelligence, 46, 159–216.

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology*, 25, 231-280.

Sloman, S. A. & Rips, L. J. (1998). Similarity as an explanatory construct. *Cognition*, 65, 87-101.

Tenenbaum, J. B, Kemp, C., Griffiths, T. L., & Goodman, N. (2011). How to grow a mind: statistics, structure, and abstraction. Science, 331, 1279-1285.

Trueblood, J. S. & Busemeyer, J. R. (2011). A quantum probability account of order effects in inference. *Cognitive Science*, 35, 1518-1552.

Tversky, A. (1977). Features of Similarity. Psychological Review, 84, 327-352.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjuctive fallacy in probability judgment. Psychological Review, 90, 293-315.

Voorspoels, W., Vanpaemel, W., & Storms, G., (2008). Exemplar and prototypes in natural language categories: a typicality-based evaluation. *Psychonomic Bulletin & Review*, 15, 630-637,

White, L., Pothos, E., Busemeyer, J. (2014). Sometimes it does hurt to ask: the constructive role of articulating impressions. *Cognition*, 133, 48-64.

Yearsley, J. M., Pothos, E. M., Hampton, J. A., & Barque-Duran, A. (2014). Towards a quantum probability theory of similarity judgments. *Proceedings of the Quantum Interaction Conference*.

Yearsley, JM., Pothos, EM. (2014). Challenging the classical notion of time in cognition: a quantum perspective. *Proceedings of the Royal Society B 281, 1781, 20133056*.

Yearsley, J., Barque-Duran, A., Pothos, E., Hampton, J., Scerrati, E. The Triangle Inequality Constraint in Similarity Judgments (*submitted*).

Yukalov, V., & Sornette, D. (2010). Decision theory with prospect interference and entanglement. *Theory and Decision*, 70, 283-328.

Theme 2:

Ariely D, Norton MI. (2008) How actions create – not just reveal – preferences. *Trends Cogn Sci* 12, 13-16. (doi:10.1016/j.tics.2007.10.008)

Kahneman D, Snell J. (1992) Predicting a changing taste: Do people know what they will like? *J Behav Decis Making* 5, 187-200. (doi:10.1002/bdm.3960050304)

Payne JW, Bettman JR, Johnson J. (1993) *The Adaptive Decision Maker*. Cambridge, UK: Cambridge University Press.

Sharot T, Velasquez CM, Dolan RJ. (2010) Do decisions shape preference? : Evidence from blind choice. *Psychol Sci* 21, 1231–1235. (doi:10.1177/0956797610379235)

Sherman SJ. (1980) On the self-erasing nature of errors of prediction. *J Pers Soc Psychol* 39, 211-221. (doi:10.1037/0022-3514.39.2.211)

Slovic P. (1995) The construction of preference. Am Psychol 50, 364-371. (doi:10.1037//0003-066X.50.5.364)

Brehm JW. (1956) Post-decision changes in the desirability of choice alternatives. J Abnorm Soc Psych 52, 384-389. (doi:10.1037/h0041006)

White LC, Pothos EM, Busemeyer JR. (2013) A quantum probability perspective on the nature of psychological uncertainty. In *Proceedings of the 35th Annual Conference of the Cognitive*

Science Society (eds M Knauff, M Pauen, N Sebanz, I Wachsmuth), pp. 1599–1604. Austin TX: Cognitive Science Society.

White LC, Pothos EM, Busemeyer, JR. (2014) Sometimes it does hurt to ask: the constructive role of articulating impressions. *Cognition* 133, 48-64. (doi:10.1016/j.cognition.2014.05.015)

White, L., Barque-Duran, A., Pothos, E. (2015) An investigation of a quantum probability model for the constructive effect of affective evaluation. *Philosophical Transactions of the Royal Society A*.

Dan-Glauser ES, Scherer KR. (2011) The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance. *Behavioral Research* 43, 468-477. (doi:10.3758/s13428-011-0064-1)

Hogarth RM, Einhorn HJ. (1992) Order effects in belief updating: the belief-adjustment model. *Cognitive Psychol* 24, 1-55. (doi:10.1016/0010-0285(92)90002-J)

Busemeyer JR, Pothos E, Franco R, Trueblood, JS. 2011 A quantum theoretical explanation for probability judgment ‘errors’. *Psychol Rev* 118, 193-218. (doi:10.1037/a0022542)

Trueblood JS, Busemeyer JR. (2011) A quantum probability account of order effects in inference. *Cognitive Sci* 35, 1518-1552. (doi:10.1111/j.1551-6709.2011.01197.x)

Wang Z, Solloway T, Shiffrin RM, Busemeyer JR. (2014) Context effects produced by question orders reveal quantum nature of human judgments. *P Natl Acad Sci USA*. (doi:10.1073/pnas.1407756111)

Bruza PD, Kitto K, Nelson D, McEvoy CL. (2009) Is there something quantum-like about the human mental lexicon? *J Math Psychol* 53, 362-377. (doi:10.1007/978-3-642-04417-5_1)

Aerts D, Aerts S. (1995) Applications of quantum statistics in psychological studies of decision processes. *Found Sci* 1, 85–97. (doi:10.1007/BF00208726)

Atmanspacher H, Filk T. (2010) A proposed test of temporal nonlocality in bistable perception. *J Math Psychol* 54, 314–21. (doi:10.1016/j.jmp.2009.12.001)

Pothos EM, Busemeyer JR, Trueblood JS. (2013) A quantum geometric model of similarity. *Psychol Rev* 120, 679–696. (doi:10.1037/a0033142)

Busemeyer JR, Bruza P. (2011) *Quantum Models of Cognition and Decision-making*. Cambridge, UK: Cambridge University Press.

Haven E, Khrennikov AY. (2013) *Quantum Social Science*. Cambridge: Cambridge University Press.

Pothos EM, Busemeyer JR. (2013) Can quantum probability provide a new direction for cognitive modeling? *Behav Brain Sci* 36, 255-327. (doi:10.1017/S0140525X12001525)

Wang ZJ, Busemeyer JR, Atmanspacher H, Pothos EM. (2013) The potential for using quantum theory to build models of cognition. *Top Cog Sci* 5, 672-688. (doi:10.1111/tops.12043)

Hughes RIG. (1989) *The Structure and Interpretation of Quantum Mechanics*. Cambridge, MA: Harvard University Press.

Isham CJ. (1989) *Lectures on quantum theory*. Singapore: World Scientific.

Peres A. 1998 *Quantum Theory: Concepts and Methods*. Kluwer Academic.

Bohr N. (1958) *Atomic Physics and Human Knowledge*. New York: Wiley.

Bargh JA, Chaiken S, Govender R, Pratto F. (1992) The generality of the automatic evaluation activation effect. *J Pers Soc Psychol* 62, 893–912. (doi:10.1037/0022-3514.62.6.893)

Damasio AR. (1994) *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: Putnam.

Duckworth KL, Bargh JA, Garcia M, Chaiken, S. (2002) The automatic evaluation of novel stimuli. *Psychol Sci* 13, 513-519. (doi:10.1111/1467-9280.00490)

Fazio RH, Sanbonmatsu DM, Powell MC, Kardes FR. (1986) On the automatic activation of evaluations. *J Pers Soc Psychol* 50, 229–238. (doi:10.1037/0022-3514.50.2.229)

Greenwald AG, Klinger MR, Liu TJ. (1989) Unconscious processing of dichoptically masked words. *Mem Cognition* 17, 35–47. (doi:10.3758/BF03199555)

LeDoux JE. (1996) *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. New York: Simon and Schuster.

Zajonc RB. (1980) Feeling and thinking: Preferences need no inferences. *Am Psychol* 35, 151−175. (doi:10.1037//0003-066X.35.2.151)

Hermans D, De Houwer J, Eelen P. (1994) The affective priming effect: Automatic activation of evaluative information in memory. *Cognition Emotion* 8, 515–533. (doi:10.1080/02699939408408957)

Klauer KC, Rossnagel C, Musch J. (1997) List-context effects in evaluative priming. *J Exp Psychol Learn* 23, 246–255. (doi:10.1037//0278-7393.23.1.246)

Bargh JA, Chaiken S, Raymond P, Hymes C. (1996) The automatic evaluation effect: Unconditional automatic evaluation activation with a pronunciation task. *J Exp Soc Psychol* 32, 104–128. (doi:10.1006/jesp.1996.0005)

Ferguson MJ, Bargh JA, Nayak DA. (2005) After-affects: How automatic evaluations influence the interpretation of subsequent, unrelated stimuli. *J Exp Soc Psychol* 41, 182–191. (doi:10.1016/j.jesp.2004.05.008)

Brehm JW, Miron AM. (2006) Can the simultaneous experience of opposing emotions really occur? *Motiv Emotion* 30, 13-30. (doi:10.1007/s11031-006-9007-z)

Rule NO, Krendl AC, Ivcevic Z, Ambady N. (2013) Accuracy and consensus in judgments of trustworthiness from faces: Behavioral and neural correlates. *Attitudes and Social Cognition* 104, 409-426. (doi:10.1037/a0031050)

Tanner RJ, Maeng A. (2012) Celebrity facial cues influence trust and preference. *J Consum Research* 39, 769-783. (doi:10.1086/665412)

Todorov A, Pakrashi M, Oosterhof NN. (2009) Evaluating faces on trustworthiness after minimal time exposure. *Soc Cognition* 27, 813–33. (doi:10.1521/soco.2009.27.6.813)

Willis J, Todorov A. (2006) First impressions: Making up your mind after 100 milliseconds exposure to a Face. *Psychol Sci* 17, 592–98. (doi:10.1111/j.1467-9280.2006.01750.x)

Engell AD, Haxby JV, Todorov A. (2007) Implicit trustworthiness decisions: automatic coding of face properties in the human amygdala. *J Cognitive Neurosci* 19, 1508–1519. (doi:10.1162/jocn.2007.19.9.1508)

Porter S, England L, Juodis M, Brinke LT, Wilson K. (2008) Is the face a window to the soul? Investigation of the accuracy of intuitive judgments of the trustworthiness of human faces. Can J Beh Sci 40, 171-177. (doi:10.1037/0008-400X.40.3.171)

Rule NO, Ambady N. (2008) The face of success: Inferences from chief executive officers' appearance predict company profits. *Psychol Sci* 19, 109-111. (doi:10.1111/j.1467-9280.2008.02054.x)

Moore DW. (2002) Measuring new types of question-order effects. *Public Opin Quart* 66, 80-91. (doi:10.1086/338631)

Wang Z, Busemeyer JR. (2013) A quantum question order model supported by empirical tests of an a priori and precise prediction. *Top Cog Sci* 5. (doi:10.1111/tops.12040)

Theme 3:

Aguilar, Brussino, & Fernández-Dols. (2013). Psychological distance increases uncompromising consequentialism. *Journal of Experimental Social Psychology*. 49 (2013) 449–452.

Albarracín, D., & Wyer, R. S. (2000). The cognitive impact of past behavior: Influences on beliefs, attitudes, and future behavioral decisions. Journal of Personality and Social Psychology, 79, 5–22.

Alexander, L., & Moore, M. (2008). Deontological ethics. In E. N. Zalta (Ed.), The Stanford encyclopedia of philosophy. Stan- ford, CA: Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University. Retrieved from http://plato.stanford.edu/entries/ethics-deontological/

Aronson, E., & Carlsmith, J. M. (1962). Performance expectancy as a determinant of actual performance. The Journal of Abnormal and Social Psychology, 65, 178–182. doi:10.1037/h0042291

Barque-Duran, A., Pothos, E., Yearsley, J., Hampton, J. (2015). Contemporary Morality: Moral Judgments in Digital Contexts. (*Under review*)

Barque-Duran, A., Pothos, E., Yearsley, J., Hampton, J. (2015). Patterns and Evolution of Moral Behavior: Moral Dynamics in Everyday Life. *Thinking and Reasoning*.

Barque-Duran, A., Pothos, E., Yearsley, J., Hampton, J. (2015). Moral Dynamics in Everyday Life: How morality evolves in time? *Proceedings of the 37th Annual Conference of the Cognitive Science Society. Pasadena, California*.

Bem, D. J. (1972). Self-perception theory. In L. Berkowitz (Ed.), Advances in experimental social psychology (Vol. 6, pp. 1–62). New York, NY: Academic Press.

Blasi, A. (1980). Bridging moral cognition and moral action: A critical review of the literature. *Psychological Bulletin, 88*, 1-45. doi:10.1037/0033-2909.88.1.1

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. Journal of behavior therapy and experimental psychiatry, 25(1), 49-59.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high- quality, data? *Perspectives on Psychological Science, 6*, 3-5. doi:10.1177/1745691610393980

Cialdini, R. B., Trost, M. R., & Newsom, J. T. (1995). Preference for consistency: The development of a valid measure and the discovery of surprising behavioral implications. Journal of Personality and Social Psychology, 69, 318–328. doi:10.1037/0022- 3514.69.2.318

Conway and Peetz (2012). When Does Feeling Moral Actually Make You a Better Person? Conceptual Abstraction Moderates Whether Past Moral Deeds Motivate Consistency or Compensatory Behavior. Personality and social psychology bulletin. DOI: 10.1177 / 0146167212442394

Costa A, Foucart A, Hayakawa S, Aparici M, Apesteguia J, et al. (2014) Your Morals Depend on Language. *PLoS ONE* 9(4): e94842. doi:10.1371/journal. pone.0094842

Cornelissen et al. (2013). Rules or Consequences? The Role of Ethical Mind-Sets in Moral Dynamics. Psychological Science. DOI: 10.1177/0956797612457376.

Foss, R. D., & Dempsey, C. B. (1979). Blood donation and the foot-in-the-door technique: A limiting case. Journal of Personality and Social Psychology, 37, 580–590. doi:10.1037/0022- 3514.37.4.580

Gneezy et al. (2012) Paying to Be Nice: Consistency and Costly Prosocial Behavior. Management Science. ISSN 0025-1909

Greene, J. D., Sommerville, R., Nystrom, L., Darley, J., & Cohen, J. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, 293, 2105–2108.

Greene J, Haidt J. (2002). How (and where) does moral judgment work? *Trends Cognitive Science*

6: 517 523. doi:10.1016/S1364-6613(02)02011-9.

Greene, J. D. (2009). Dual-process morality and the personal/impersonal distinction: A reply to McGuire, Langdon, Coltheart, and Mackenzie. *Journal of Experimental Social Psychology*, 45, 581–584.

Gong, Iliev, & Sachdeva (2012). Consequences are far away: Psychological distance affects modes of moral decision making. Cognition. doi: 10.1016/j.cognition.2012.09.005

Higgins, E. T. (1987). Self-discrepancy: A theory relating self and affect. Psychological Review, 94, 319–340. doi:10.1037/0033-295x.94.3.319

Hofmann, W., Wisneski, D., Brandt, M., Skitka, L. (2014). Morality in everyday life. *Science*. 345 (6202): 1340-1343.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). The Big Five Inventory--Versions 4a and 54. Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research.

Jordan, J., Mullen, E., Murninghan, J.K. (2011). Striving for the Moral Self: The Effects of Recalling Past Moral Actions on Future Moral Behavior. Personality and Social Psychology Bulletin. DOI: 10.1177/0146167211400208.

Klass, E. T. (1978). Psychological effects of immoral actions: The experimental evidence. Psychological Bulletin, 85, 756–771.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F. (2007) Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature* 446: 908– 911. doi:10.1038/nature05631.

Liberman N., Sagristano, M., Trope, Y. (2002). The effect of temporal distance on level of mental construal. *Journal of Experimental Social Psychology*. 38 (2002) 523–534.

Merritt, A., Effron, D., Monin, B. (2010). Moral Self-Licensing: When Being Good Frees Us to Be Bad. Social and Personality Psychology Compass. 344–357, 10.1111/j.1751-9004.2010.00263.x.

Monin, B., & Jordan, A. H. (2009). The dynamic moral self: A social psychological perspective. In D. Narvaez & D. K. Lapsley (Eds.), Personality, identity, and character: Explorations in moral psychology (pp. 341–354). New York, NY: Cambridge University Press.

Nakagawa S. (2004). A farewell to Bonferroni: the problems of low statistical power and publication bias. Behavioral Ecology, 15, 1044-1045.

Narvaez, D., & Lapsley, D.K. (Eds.) (2009). Personality, Identity, and Character: Explorations in Moral Psychology. New York: Cambridge University Press.

Nisan, M. (1991). The moral balance model: Theory and research extending our understanding of moral choice and deviation. In W. M. Kurtines & J. L. Gewirtz (Eds.), Handbook of moral behavior and development (Vol. 3, 213–250). Hillsdale, NJ: Erlbaum.

Ofcom Communication Market Report: Internet and web-based content. (2014).

Ozer, D. J., & Benet-Martínez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology*, 57, 401–421.

Penner LA., Fritzsche BA., Craiger JP., Freifeld TR. (1995). Measuring the prosocial personality. *Advances in Personality Assessment*, ed. J Butcher, CD Spielberger, 10:147–63. Hillsdale, NJ: Erlbaum.

Reed, A., II, Aquino, K., & Levy, E. (2007). Moral identity and judgments of charitable behaviors. *Journal of Marketing, 71,* 178-193.

Sachdeva, S., Iliev, R., & Medin, D. L. (2009). Sinning saints and saintly sinners. Psychological Science, 20, 523–528. doi:10.1111/ j.1467-9280.2009.02326.x

Singer, P. (1991). A companion to ethics. Oxford, England: Blackwell Reference.

Sinnott-Armstrong, W. (2008). Consequentialism. In E. N. Zalta (Ed.), The Stanford encyclopedia of philosophy. Stanford, CA: Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University. Retrieved from http://plato .stanford.edu/entries/consequentialism/

Slovic, P. (2007). “If I look at the mass I will never act”: Psychic numbing and genocide. *Judgment and Decision-making*, 2, 79–95.

Suter, R. & Hertwig, R. (2011). Time and moral judgment. *Cognition.* 119 (2011) 454-458.

Szekely, R. D., & Miu, A. C. (2014). Incidental emotions in moral dilemmas: The influence of emotion regulation. Cognition & emotion, (ahead-of-print), 1-12.

Thomson, J. (1985). The trolley problem. Yale Law, 94: 1395–1415. doi:10.2307/796133.

Thomas, G., & Batson, C. D. (1981). Effect of helping under normative pressure on self-perceived altruism. Social Psychology Quarterly, 44, 127–131. doi:10.2307/3033708

Trope, Y., & Liberman, N. (2010). Construal Level Theory of Psychological Distance, *Psychological Review*, 117 (April), 440–63.

Uhlmann, E. L., Pizarro, D. A., Tannenbaum, D., & Ditto, P. H. (2009). The motivated use of moral principles. Judgment and Decision-making, 4, 476–491

Valdesolo, P. & DeSteno, D. (2006). Manipulations of Emotional Context Shape Moral Judgment. *Psychological Science*, 17, 476-477.

Williams, Lawrence E. Stein, R., Galguera, L. (2014), The Distinct Affective Consequences of Psychological Distance and Construal Level. *Journal of Consumer Research*, 40, 1123-1138.

Wheatley, T., & Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological Science*, 16, 780–784.

Zhong, C.-B., Liljenquist, K., & Cain, D. M. (2009). Moral self-regulation: Licensing and compensation. In D. De Cremer (Ed.), Psychological perspectives on ethical behavior and decision-making (pp. 75–89). Charlotte, NC: Information Age Publishing.

Theme 4:

Aerts, D., (2009). Quantum structure in cognition. Journal of Mathematical Psychology 53 (5), 314–348.

Aerts, D., Gabora, L., (2005). A theory of concepts and their combinations ii: A hilbert space representation. Kybernetes 34 (1/2), 192–221.

Aliakbarzadeh, M., Kitto, K., (2016). Applying povm to model non-orthogonality in quantum cognition. In: Atmanspacher, H., Filk, T., Pothos, E. (Eds.), Quantum Interaction. Vol. 9535 of Lecture Notes in Computer Science. Springer International Publishing, pp. 284–293.

Atmanspacher, H., Filk, T., Romer, H., (2004). Quantum zeno features of bistable perception. Biological Cybernetics 90 (1), 33–40.

Barque-Duran, A., Pothos, E. M., Yearsley, J., Hampton, J., Busemeyer, J. R., Trueblood, J. S., (2016). Similarity Judgments: From Classical to Complex Vector Psychological Spaces. In the series: Advanced Series on Mathematical Psychology by E. Dzhafarov, R. Zhang, S. Joardan, and V. Cervantes (Eds.) Contextuality from Quantum Physics to Psychology. World Scientific.

Blanco, M., Engelmann, D., Koch, A., Normann, H.-T., (2014). Preferences and beliefs in a sequential social dilemma. Games and Economic Behavior 87, 122– 135.

Blutner, R. K., (2008). Concepts and Bounded Rationality: An Application of Niestegge's Approach to Conditional Quantum Probabilities. In L. Accardi, et al. (eds.), Foundations of Probability and Physics 5, Vol. 1101, pp. 302-10. American Institute of Physics Conference Proceedings.

Bohr, N., (1950). On the notions of causality and complementarity. Science 111(2973)), 51–54.

Bordley, R., (1998). Quantum mechanical and human violations of compound probability principles: toward a generalized Heisenberg uncertainty principle. Oper. Res 46, 923–926.

Brandenburger, A., (2005). The relationship between quantum and classical correlation in games. Games Econ. Behav 69(1), 175–183.

Brunner, N., Linden, N., (2013). Connection between Bell nonlocality and Bayesian game theory. Nat. Commun 4(2057), 1–6.

Bruza, P. D., (2010). Quantum memory. Australasian Science 31 (1), 34–35.

Bruza, P. D., Kitto, K., Nelson, D., McEvoy, C. L., (2009). Is there something quantum-like about the human mental lexicon? Journal of Mathematical Psychology 53, 362–377.

Busemeyer, J., Bruza, P., (2012). Quantum Models of Cognition and Decision. Cambridge University Press, Cambridge.

Busemeyer, J., Pothos, E., (2012). Social projection and a quantum approach for behavior in Prisoner's Dilemma. Psychological Inquiry 23 (1), 28–34.

Busemeyer, J., Pothos, E., Franco, R., Trueblood, J., (2011). A quantum theoretical explanation for probability judgment errors. Physchol. Rev 118(2), 193–218.

Busemeyer, J., Wang, Z., Townsend, J., (2006). Quantum dynamics of human decision-making. J.

Math. Psychol 50, 220–241.

Danilov, V. I., Lambert-Mogiliansky, A., (2008). Measurable systems and behavioral sciences. Mathematical Social Sciences 55 (3), 315–340.

Denolf, J., Lambert-Mogiliansky, A., (2016). Bohr complementarity in memory retrieval. Journal of Mathematical Psychology. 28-36.

Denolf, J., (2015). Subadditivity of episodic memory states: a complementarity approach. In: Atmanspacher, H., Bergomi, C., Filk, T., Kitto, K. (eds.) QI 2014. LNCS. Springer, Heidelberg.

Deutsch, D., (1999). Quantum theory of probability and decisions. Proc. Roy. Soc.A 455, 3129–3137.

Hameroff, S. R., (2007). The brain is both neurocomputer and quantum computer. Cognitive Science 31 (6), 1035–1045.

Haven, E., Khrennikov, A., (2013). Quantum social science. Cambridge University Press.

Hughes, R. I. G., (1989). The structure and interpretation of quantum mechanics. Harvard University Press.

Isham, C. J., (1989). Lectures on quantum theory. Singapore: World Scientific.

Khrennikov, A., (2010). Ubiquitous Quantum Structure: From Psychology to Finance. Springer, Berlin.

Khrennikov, A., Basieva, I., Dzhafarov, E. N., Busemeyer, J. R., (2014). Quantum models for psychological measurements: an unsolved problem. PloS one 9 (10), e110909.

La Mura, P., (2005). Correlated equilibria of classical strategic games with quantum signals. International Journal of Quantum Information 3 (01), 183–188.

Lambert-Mogiliansky, A., Martinez-Martinez, I., (2015). Games with Type Indeterminate players: a Hilbert space approach to uncertainty and strategic manipulation of preferences. In: Atmanspacher, H., Bergomi, C., Filk, T., Kitto, K. (eds.) QI 2014. Lecture Notes in Computer Science 8951. Springer International Publishing Switzerland.

Lambert-Mogiliansky, A., Zamir, S., Zwirn, H., (2009). Type indeterminacy: a model for the

KT(Kahneman-Tversky)-man. J. Math. Psychol 53, 349–361.

Lichtenstein, S., Slovic, P., (2006). The construction of preference. Cambridge University Press, New York.

Litt, A., Eliasmith, C., Kroon, F. W., Weinstein, S., Thagard, P., (2006). Is the brain a quantum computer? Cognitive Science 30 (3), 593–603.

Martinez-Martinez, I., (2014). A connection between quantum decision theory and quantum games: The hamiltonian of strategic interaction. Journal of Mathematical Psychology 58, 33–44.

Pothos, E., Busemeyer, J., (2009). A quantum probability explanation for violations of 'rational' decision theory. Proceedings of the Royal Society of London B: Biological Sciences, rspb–2009.

Pothos, E., Perry, G., Corr, P., Matthew, M., Busemeyer, J., (2011). Understanding cooperation in the Prisoners Dilemma game. Pers. Individ. Differ 51(3), 210– 215.

Pothos, E., Yearsley, J., Barque-Duran, A., Hampton, J., Busemeyer, J., Trueblood, J., (2015). Progress and current challenges with the Quantum Similarity Model. Frontiers in Psychology, 6:205.

Pothos, E. M., Busemeyer, J. R., (2013). Can quantum probability provide a new direction for cognitive modeling? Behav. Brain Sci 36, 255–327.

Trueblood, J., Busemeyer, J., (2011). A quantum probability account of order effects in inference. Cogn. Sci 35, 1518–1552.

Wang, Z., Solloway, T., Shiffrin, R., Busemeyer, J., (2014). Context effects produced by question orders reveal quantum nature of human judgments? Proc. Natl. Acad. Sci. USA 111(26), 9431–9436.

White, L., Barque-Duran, A., Pothos, E., (2015). An investigation of a quantum probability model for the constructive effect of affective evaluation. Philosophical Transactions of the Royal Society A.

White, L., Pothos, E., Busemeyer, J., (2014). Sometimes it does hurt to ask: the constructive role of articulating impressions. Cognition 133, 48–64.

Yearsley, J., Pothos, E., (2014). Challenging the classical notion of time in cognition: a quantum perspective. Proc. Roy. Soc. B 281), 20133056.

Yearsley, J. M., Pothos, E. M., Hampton, J., Barque-Duran, A., (2014). Towards a quantum probability theory of similarity judgments. Proceedings of the Quantum Interaction Conference.

Yukalov, V. I., Sornette, D., (2011). Decision theory with prospect interference and entanglement. Theory and Decision 70 (3), 283–328.