Chapter 12
Reinforcing Trust in Autonomous Systems: A Quantum Cognitive Approach

Peter D. Bruza and Eduard C. Hoenkamp

12.1 Introduction

Bad decisions can have dire consequences. From high exposure events such as an oil spill or a plane crash, to the smaller scale drama of a patient who dies on the operating table, the unavoidable question soon follows: Was this mechanical failure or human error? Yet, in a society where people increasingly base their decisions on autonomous systems such as search engines, recommender systems, or social media, the distinction becomes blurred. Although these systems are based on algorithms (less material, but nonetheless mechanical) people will have to process and consider the provided information, thus becoming the weakest link in the decision chain. In general, mechanical failure, once discovered, seems more easily addressed than human error. So if autonomous systems could be made aware of how humans judge information, they could become more judicious in advising humans, and more proactive in the way they present their information. Currently this is not the case. To change this, we have looked into decades of research about human judgement (For the case of how judgement of a particular system is shaped by some of its properties see Sect. 7.6 of this book.). We found a whole range of human judgement that deviates substantially from what would be normatively correct according to logic and probability theory. As example take the famous experiment in which Tversky and Kahneman [35] presented participants with the following text:

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also

P. D. Bruza (✉) · E. C. Hoenkamp
Information Systems School, Queensland University of Technology (QUT), Brisbane, Australia
e-mail: p.bruza@qut.edu.au

E. C. Hoenkamp
Institute for Computing and Information Sciences, Radboud University, Nijmegen, The Netherlands
e-mail: hoenkamp@acm.org

© The Author(s) 2018
H. A. Abbass et al. (eds.), Foundations of Trusted Autonomy, Studies in Systems, Decision and Control 117, https://doi.org/10.1007/978-3-319-64816-3_12
participated in anti-nuclear demonstrations Which is more probable:
(a) Linda is a bank teller, or
(b) Linda is a bank teller and is active in the feminist movement?

In this experiment, the participants consistently rated option (b) as more probable than (a). However, according to the axioms of probability theory, the conjunction of events is less probable than a single event, so (b) must be less probable than (a).

This type of judgement errors has been found so invariably that it became known as the conjunction fallacy (see [31, 34] for other such fallacies). These findings are not widely known, let alone implemented, by programmers designing the communication part of autonomous systems. The latter systems so far adhere to the laws of probability and logic, which is their strength. Their weakness, however, is to not account for how humans make decisions. For example, an intelligent system that would correctly answer (a) in the experiment above, might bewilder the human who was expecting answer (b). In turn this could erode that human’s trust in such a system.

Let us take another, more concrete, example from technology soon to become reality. Suppose you arrive late in the evening at a meeting. As the street is clearly indicated as a tow zone, you let your self-driving car find a parking space elsewhere to park in your stead. Later you come back and find that it parked in the first free spot in that same street. Would this not be the time to reconsider your trust in the autonomy of the car? And so you instruct it not to do this again. But then it politely explains [24] that the tow zone only applies during rush hour, thus restoring your trust. An even better scenario would be that when you leave the car to park itself, it would tell you that yes, this is a tow zone, but only during rush hour. Machines have become smart enough to do the first part, in this case finding a parking space and park. But then, is it not time to work on the second part, where the machine proactively explains its actions from the human’s point of view? Or the part where it can foresee a human error because it knows how a human would reason in a particular case? This stands to hugely help humans put trust in autonomous systems, and in the current presentation we show a direction one could go.

To elaborate the problem we want to address, before trying to solve it, consider Wittgenstein’s often cited remark “If a lion could talk, we could not understand him” ([37], p. 223). The remark has been food for much thought and speculation (that it is probably true) for half a century, notably in the philosophy of language. We wonder, however, that if we cannot even understand a lion, whence the confidence that we will understand machines when it comes to communicating with them verbally? Here we think that Wittgenstein’s own, less quoted, comment can lead the way [37]:

If language is to be a means of communication there must be agreement not only in definitions but also (queer as this may sound) in judgements. This seems to abolish logic, but does not do so — It is one thing to describe methods of measurement, and another to state results of measurement. (p. 391, our italics)

This is an important vantage point for the current presentation: first it emphasizes the role of judgement, second it distinguishes the method of measurement from its result, and third it challenges the role of logic.
We will express these notions in the language of quantum cognition, which derives terminology and computations from quantum mechanics. But whereas quantum mechanics deals with physical states, quantum cognition describes cognitive (or mental) states. The Linda experiment can be described in this language, as we showed already in the book we wrote about quantum cognition [10]. For further details we will refer the reader to that book, so that we can use the space here to describe a new experiment. The experiment is also modeled using quantum cognition, laying a foundation for its implementation in future autonomous systems. Incorporating such models can proactively help the user avoid mistakes that are inherent in human judgment and thus prevent an erosion of trust. We contend that in this way the interactions between humans and future autonomous system will become more effective.

12.2 Compatible and Incompatible States

The conjunction fallacy does not mean that people always judge the probability of a conjunction as higher than each of its conjuncts. That would indeed be counter to probability theory. Just imagine that the choice between (a) and (b) was presented without the story about Linda. Then one would expect everyone with some notion of probability to choose (a) over (b), as confirmed in [35]. But when asked the question after first hearing the story, even people schooled in statistics fall victim to the conjunction fallacy. Why is this? The question generated a host of publications with possible explanations over the last several decades (see [19] for an overview). Among the many kinds of explanations offered, two stand out in particular. One assumes that words such as ‘and’ and ‘probability’ are misunderstood by the participants, or at least not understood in their formal interpretation. The other assumes a reasoning bias. A recent overview [30] concludes that the latter has the best support of the two. But this answer begs the question: if there is a reasoning bias, where does that reasoning bias originate?

Indeed, we are not satisfied with just knowing there is such a bias, rather we would like to describe how that bias unfolds as a cognitive process. To do so, let us formulate the participants’ judgements as the outcome of a decision process. The explanations in the literature almost invariably involve two competing states, one in which Linda is a bankteller, and another where she is a feminist. To most participants in the experiment these states are not compatible, and whether bias or reasoning, each can in principle tip the scale in favor of one state or the other. In order to make headway, we have to take a closer look at the notion of compatible and incompatible states.

In this presentation we will formulate states in the language of quantum cognition. Especially for the reader who is not already familiar with this approach, we will first recall some concepts from quantum mechanics. One such concept is the (formal) notion of compatible and incompatible states. Incompatibility lies at the heart of Heisenberg’s famous uncertainty principle. It holds that when we are certain about a quantum particle’s momentum, we are necessarily uncertain about its position, and vice versa. Position and momentum are therefore called incompatible states. On the other hand, again given the momentum of that particle, we can still measure its kinetic energy with certainty, momentum and kinetic energy thus being compatible states.
12.3 A Quantum Cognition Model for the Emergence of Trust

There is a vision of nature that readers may have entertained themselves at one time or another, namely that “all phenomena could be explained mechanically if only we knew enough.” Those readers are in good company, as this is a direct quote from Leibniz’s writing of 1695 [29]. Yet the problems with this vision are several. First, we are not omniscient and second it is uncertain if we will ever have enough computing power to do the explaining. But even if is this may become an issue for intergalactic travel, why would it apply to the Linda experiment? After all, everything the participants in the experiment need to know is given in the instructions, and the computing power needed is an unassuming application of probability theory. So why is an explanation for the experimental findings still wanting?

There is a third, perhaps more fundamental, problem with Leibniz’s position: His mechanical view of the laws of nature has been drastically undermined with the advent of quantum mechanics. It turns out that some phenomena require a probability calculus, but different than given by classical probability theory. For example, probabilities may not always add up to 1.

The problem does not just apply to the description of physical systems, but also to cognitive systems, and more in particular those that play a role in interactions between humans and autonomous systems. We can safely assume that such interactions require decisions under uncertainty. For decades cognitive scientists have studied how humans make judgments under such conditions. The theories they produced can be roughly divided into a heuristic and a rational approach.

The heuristic approach is founded on so-called bounded rationality [33]. It assumes that in order to make decisions, humans use simple heuristics such as representativeness, anchoring-and-adjustment, and base-rate neglect. They support powerful (often inductive) processes that may depend on the environment [22]. In contrast, the rational approach conforms to rules drawn from a theory, most notably Bayes’ rule [13] or expected utility theory [32]. In this approach the same basic axioms can be used to derive inferences and utilities across all environmental conditions. Recently, a third approach, called quantum cognition has emerged [10, 11, 36]. In common with the heuristic approach, it assumes that the human decision maker is subject to bounded rationality. And in common with the rational approach, inferences used for decisions are derived from basic axioms that define a probability theory. However, the axioms are different from those employed by the Bayesian approach, and consequently, so are the decisions that follow from it.

Two core concepts underpinning quantum cognition are incompatibility and contextuality. We will briefly come back to ‘contextuality’ later in this chapter, but right now we will continue with ‘incompatibility’ from Sect. 12.2 in more detail. Both concepts play an important role in formalizing people’s perception of trust.

At the time of writing this chapter, a term that gained notoriety was the term fake news. Many readers may have realized their quandary over what news can be trusted any longer. We will describe an experiment in which we induced such a quandary
and measured how it influenced people’s degree of trust. We will then offer a formal explanation in terms of quantum cognition, which at the same time shows that we need to distinguish between two forms of trust.

We could perhaps have presented participants with pieces of news of varying degrees of trustworthiness. However, note the many variables that should then be kept under control, such as the participant’s background knowledge or the ephemerality of news. Instead we used pictures of which they had to assess the trustworthiness on a five point scale from untrustworthy to very trustworthy [18]. (Trustworthy was defined as an image that gave an accurate representation of a situation, person or object.) The participants were also asked to supply the reasons for their rating. An example of one such a picture was a smiling face of Putin, which can be seen by following the URL in [7].

The image of Putin was deliberately chosen. Many participants will know that he is a former KGB agent and probably not predisposed to smiling. Would this then lead to uncertainty whether the image had been photoshopped, and a consequent lack of trust that the image is a true an accurate depiction of Putin? When analyzing the qualitative data a curious phenomenon emerged. A considerable number of participants appeared to confuse the decision regarding trust of the image with a decision on whether they trust the content of the image. For example, “This is Vladimir Putin, a world leader I associate with dishonesty and distrust, who works to his own agenda and doesn’t worry about other people”, “I wouldn’t trust the person, but the photo is fine”, and “I really could not separate what I know about this man from his image”.

Assuming the confounding of the trustworthiness decision is a robust cognitive effect, how can it be explained? Quantum cognition offers an explanation based on incompatible decision perspectives. Consider diagram (a) on the left in Fig. 12.1.

This figure depicts two decision perspectives. One is a two-dimensional vector space featuring two orthogonal basis vectors (in black) corresponding to the decision

![Diagram](image)

**Fig. 12.1** Incompatible decision perspectives explain why trust in the image is sometimes confused with trust in its content (elaborated in the text)
that participant trusts, or doesn’t trust the image. We will call this the “image”
decision space. The other decision perspective is a two-dimensional vector space
(dotted basis vectors) which models the decision whether the participant trusts, or
doesn’t trust the content of the image. We will call this the “content” decision space.
This vector space is rotated with respect to the vector space modeling the decision of
the trustworthiness of the image. The red vector represents the cognitive state of the
participant when the image is first presented. Note that the cognitive state-vector is
superposed with respect to both decision perspectives. If this participant’s cognitive
state is suspended between the basis vectors of both decision spaces. Superposition
in quantum cognition is a major departure to mixed state in a standard probabilistic
model. A mixed state implies that the participant is always in one basis state or the the
other. For example, with respect to the decision regarding the image, the participant
will be either in the state corresponding to the decision that they trust the image or be
in the alternative state in which they don’t trust the image. We may not know what
state they are in, but they must be in one of these two states. It may be the case that
in the course of considering the image the cognitive state of the participant moves
between these states. Superposition allows the participant to be in both
states at once.

As we shall see, this has a marked effect on the probabilities of the associated states.
Quantum cognition is defined by quantum probabilities which are related geometric-
ally, not by an underlying Boolean algebra over the event space. More formally,
let \( T \) denote the decision that the participant trusts the image and let \( \overline{T} \) denote the
decision that the participant does not trust the image. These are the basis vectors of
the image subspace. Similarly, let \( C \) and \( \overline{C} \) denote the basis vectors of the content
subspace. The cognitive state \( \psi \) is superposed between the two decisions in both
the image and content subspaces. The probability that the participant decides they
trust the content portrayed in the image is equal to the square of the projection of
the cognitive state \( \psi \) onto the basis vector \( C \), denoted \( \| P_C \psi \|^2 \). We can see from
the diagram that the length of this projection is small, hence the probability is small,
which reflects the weak predisposition of the participant to trust Vladimir Putin. Con-
versely, the predisposition not to trust Putin is high because of the long projection of
the cognitive state vector \( \psi \) onto the basis vector \( \overline{C} \). Hence the associated probability
\( \| P_{\overline{C}} \psi \|^2 = 1 - \| P_C \psi \|^2 \) is high.

Consider diagram (b) in Fig. 12.1. Note that the cognitive state vector now lies
on the basis vector. This models the situation in which the participants have decided
that they do not trust the content of the image (Vladimir Putin). Therefore, the
cognitive state \( \psi \) is no longer superposed with respect to the decision perspective
regarding the trustworthiness of the content. Note, however, that the cognitive state
vector is necessarily superposed with respect the decision perspective regarding the
trustworthiness of the image. The reason for this is the degree of rotation between the
two decision subspaces. Because these subspaces are not orthogonal to each other
the decision perspectives are incompatible. Incompatibility between the decision
perspectives explains why participants confuse the decision on the trustworthiness
of the image versus content of the image. The reason is that in case of incompatibility the
law of total probability does not hold. For example, the probability that the participant
will trust the content of the image in terms of the two decision perspectives is as follows:

\[
p(C) = \| P_C \psi \|^2 = \| P_C \cdot (P_T + P_T^T) \psi \|^2 \]

(12.1)

(12.2)

(12.3)

(12.4)

\[
\text{Int}
\]

The preceding rendered in standard probability theory looks like the law of total probability, which is being modified by extra term \text{Int}:

\[
p(C) = p(C, T) + p(C, \overline{T}) + \text{Int}
\]

(12.5)

\(\text{Int}\) is referred to as the “interference term”. The interference term can positively or negatively contribute to the probability \(p(C)\). For this reason, incompatible subspaces have been put forward as a natural explanation why human beings do not adhere to the law of total probability like in the conjunction and other so called fallacies in human decision making [12]. When the interference term is zero, the law of total probability holds. This happens when the decision perspectives are compatible.

Incompatible decision perspectives are a recent development in cognitive modeling and their striking characteristic is the use of “quantum” probabilities. By quantum probabilities, we mean that the decision event space is modelled as a vector space rather than a Boolean algebra of sets. A key differentiator is the use of the interference term. When this term is non-zero, violations of the law of total probability occur. The interference term has been used in models of the perception of gestalt images \([15, 28]\), models of the conjunction and other decision fallacies \([12, 14]\), modeling violations of rational decision theory \([6, 28, 31]\), modeling belief dynamics \([34]\) and conceptual processing \([3–5, 20, 21]\). In this presentation there is no need to go any deeper into the fine points where quantum cognition and the Bayesian approach part company. Therefore we only present Table 12.1 as a summary, and refer the interested reader to \([9]\) for further details.

Placed in a psychological context, the uncertainty principle becomes relevant because a person’s understanding of two events, such as two different perspectives

| Table 12.1 Comparing Bayesian and Quantum cognition |
|-----------------------------------------------|
| Bayesian cognition | Quantum cognition |
| Human is in a definite cognitive state | Human can be in a superposed state |
| Events are compatible | Events may be incompatible |
| Law of total probability | May violate law of total probability |
| Models are non-contextual | Models may be contextual |
on a matter, requires changing from one point of view to another, and the two points of view can imply incompatibility. In other words, the uncertainty principle entails that it is not possible to be simultaneously decided on both the image and content with respect to assessing trustworthiness. Just like it is not possible to form a joint probability of both the position and momentum of a quantum particle, it is not possible for the human participant to form a joint probability across decisions, e.g., whether they both trust the image and the content of the image. Not being able to form joint probabilities signals the presence of contextuality. A well studied example of contextuality is the curious phenomenon of entanglement: Empirical observations are collected in four measurement settings of a system of two quantum particles such as photons. Each of the four settings yields a pairwise joint probability distribution which models the observations made in that measurement setting. An entangled system is deemed “contextual” because it is not possible to combine these four pairwise joint probability distributions into a single probabilistic model such that the four pairwise empirical distributions are marginal distributions of this global model. Even though contextuality manifests within the sub-atomic realm, there is a growing body of research which is exploring whether contextuality manifests in cognition and related areas (e.g., [1, 2, 8, 9, 16, 17]). In the context of our example, contextuality arises because the image and content decision perspectives cannot be meaningfully combined into a single joint distribution.

12.4 Conclusion

In the foreseeable future, humans and autonomous systems will engage in shared decision making. Given the discrepancies between the way they arrive at decisions, whose form of rationality should be given precedence: human or machine?

One form of rationality can be termed “Bayesian rationality” in which Bayesian probability theory provides such powerful models of both human and machine cognition, that it is sometimes called “Bayesian fundamentalism” [26]. We made the point in this chapter that humans often do not adhere to Bayesian rationality, but rather to a “quantum rationality,” as it is based on the the same Dirac-von Neumann axioms as quantum theory. Neither of these two rationalities should be given precedence. We argued that quantum rationality is often more suited to model human decision making, and Bayesian rationality more to model decision making by machines. So the more pertinent question is how to best align these rationalities so that shared decision making between human and machine becomes more effectual than that of each in isolation.

There are good examples where traditional (often brute force) mechanisms were unable to solve difficult problems, and which became tractable when augmented with cognitive elements. One famous example is the breaking of the Enigma code by augmenting brute force methods with simple aspects of human communication [23]. A second example is the incorporation of human reasoning in chess programs. And finally we have already seen this clearly confirmed in our own research in areas outside that of this chapter [24, 25].
Therefore, we propose to augment the rationality already available in the current systems with quantum rationality. We expect that future autonomous systems can compute decisions based on both rationalities and hence detect situations when these decisions do not align. In such cases the machine could make the human aware of the discrepancy, thus preventing a potential erosion of trust between human and machine.

References

1. S. Abramsky, Contextual semantics: from quantum mechanics to logic, databases, constraints, and complexity. Bull. EATCS 2 (113) (2015)
2. S. Abramsky, M. Sadrzadeh, Semantic unification: a sheaf theoretic approach to natural language, in Categories and Types in Logic, Language and Physics: Essays Dedicated to Jim Lambek, ed. by C. Casadio, B. Coecke, M. Moorgat, P. Scott (Springer, Berlin, 2014)
3. D. Aerts, Quantum structure in cognition. J. Math. Psychol. 53(5), 314–348 (2009)
4. D. Aerts, L. Gabora, S. Sozzo, Concepts and their dynamics: a quantum-theoretic modeling of human thought. Top. Cogn. Sci. 5(4), 737–772 (2013)
5. R. Blutner, E.M. Pothos, P.D. Bruza, A quantum probability perspective on borderline vagueness. Top. Cogn. Sci. 5(4), 711–736 (2013)
6. R.F. Bordley, Violations of compound probability principles: toward a generalized Heisenberg uncertainty principle. Oper. Res. 46, 923–926 (1998)
7. A. Borowitz, Putin announces historic G1 summit http://goo.gl/Hd8Jhp. Apr 2017
8. P.D. Bruza, K. Kitto, B. Ramm, L. Sitbon, A probabilistic framework for analysing the compositionality of conceptual combinations. J. Math. Psychol. 67, 26–38 (2015)
9. P.D. Bruza, Z. Wang, J.R. Busemeyer, Quantum cognition: a new theoretical approach to psychology. Trends Cogn. Sci. 19(7), 383–393 (2015)
10. J. Busemeyer, P.D. Bruza, Quantum Cognition and Decision (Cambridge University Press, Cambridge, 2012)
11. J. Busemeyer, E.M. Pothos, Can quantum probability provide a new direction for cognitive modeling? Behav. Brain Sci. 36, 255–327 (2013)
12. J. Busemeyer, E.M. Pothos, R. Franco, J.S. Trueblood, A quantum theoretical explanation for probability judgment errors. Psychol. Rev. 118(2), 193–218 (2011)
13. N. Chater, J.B. Tenenbaum, A. Yuille, Probabilistic models of cognition: conceptual foundations. Trends Cogn. Sci. 10(7), 287–291 (2006)
14. E. Conte, A. Khrennikov, O. Todarelo, R. De Robertis, A. Federici, J. Zbilut, On the possibility that we think in a quantum mechanical manner: an experimental verification of existing quantum interference effects in cognitive anomaly of conjunction fallacy. Chaos Complex. Lett. 4(3), 123–136 (2011)
15. E. Conte, O. Todarelo, A. Federici, F. Vitiello, M. Lopane, A. Khrennikov, J. Zbilut, Some remarks on an experiment suggesting quantum-like behaviour of cognitive entities and formulation of an abstract quantum mechanical formalism to describe cognitive entity and dynamics. Chaos, Solitons & Fractals 33, 1076–1088 (2007)
16. E. Dzhafarov, J. Kujala, Contextuality is about identity of random variables. Physica Scripta T163, 014009 (2014)
17. E. Dzhafarov, R. Zhang, J. Kujala, Is there contextuality in behavioral and social systems? Philos. Trans. Royal Soc. A 374, 20150099 (2015)
18. L. Fell, P.D. Bruza, S.K. Devitt, G. Oliver, M. Gradwell, H. Partridge, The cognitive decision space of trust: an exploratory study of image trustworthiness and the propensity to deceive, http://eprints.qut.edu.au/102009/ (2016)
19. J.E. Fisk, Conjunction fallacy, in Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory, ed. by R.F. Pohl (Psychology Press, Hove, UK, 2004), pp. 23–42
20. L. Gabora, D. Aerts, Contextualizing concepts using a mathematical generalization of the
quantum formalism. J. Exp. Theor. Artif. Intell. 14, 327–358 (2002)
21. L. Gabora, E. Rosch, D. Aerts, Toward an ecological theory of concepts. Ecol. Psychol. 20,
84–116 (2008)
22. G. Gigerenzer, P.M. Todd, Simple Heuristics That Make Us Smart (Oxford University Press,
Oxford, 1999)
23. A. Hodges, Alan Turing: The Enigma (Walker, New York, 2000)
24. E.C.M. Hoenkamp, P.D. Bruza, How everyday language can and will boost effective informa-
tion retrieval. J. Am. Soc. Inf. Technol. 66(8), 1546–1558 (2015)
25. E. Hoenkamp, Why information retrieval needs cognitive science: a call to arms, in Proceedings
of the 27th Annual Conference of the Cognitive Science Society, pp. 965–970 (2005)
26. M. Jones, B.C. Love, Bayesian fundamentalism or enlightenment? on the explanatory status
and theoretical contributions of Bayesian models of cognition. Behav. Brain Sci. 34, 169–231
(2011)
27. D. Kahneman, Thinking, Fast and Slow (Farrar, Straus and Giroux, New York, 2011)
28. A. Khrennikov, Ubiquitous Quantum Structure: From Psychology to Finance (Springer, Berlin,
2010)
29. G.W. Leibniz, Specimen dynamicum, in Philosophical Papers and Letters, ed. by L.E. Loemker
(Springer, The Netherlands, Dordrecht, 1989), pp. 435–452
30. R. Moro, On the nature of the conjunction fallacy. Synthese 171(1), 1–24 (2009)
31. E.M. Pothos, J. Busemeyer, A quantum probability explanation for violations of ‘rational’
decision theory. Proc. Royal Soc. B 276(1165), 2171–2178 (2009)
32. L.J. Savage, The Foundations of Statistics (Wiley, Hoboken, 1954)
33. H.A. Simon, Models of Man: Social and Rational (Wiley, Hoboken, 1957)
34. J.S. Trueblood, J. Busemeyer, A quantum probability explanation for order effects on inference.
Cogn. Sci. 35, 1518–1552 (2011)
35. A. Tversky, D. Kahneman, Extensional versus intuitive reasoning: the conjunctive fallacy in
probability judgment. Psychol. Rev. 90, 293–315 (1983)
36. Z. Wang, J. Busemeyer, The potential of using quantum theory to build models of cognition.
Topics Cogn. Sci. 5(4), 672–688 (2013)
37. L. Wittgenstein, Philosophical Investigations (Basil Blackwell, Oxford, 1953)