CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Repository@Nottingham

The Need for Fuzzy AI

Jonathan M. Garibaldi, *Senior Member, IEEE*

(This paper is an invited paper arising from the Alfred North Whitehead Lecture, Chinese Academy of Sciences, Beijing October 2018)

***Abstract*—Artificial intelligence (AI) is once again a topic of huge interest for computer scientists around the world. Whilst advances in the capability of machines are being made all around the world at an incredible rate, there is also increasing focus on the need for computerised systems to be able to explain their decisions, at least to some degree. It is also clear that data and knowledge in the real world are characterised by uncertainty. Fuzzy systems can provide decision support, which both handle uncertainty and have explicit representations of uncertain knowledge and inference processes. However, it is not yet clear how any decision support systems, including those featuring fuzzy methods, should be evaluated as to whether their use is permitted. This paper presents a conceptual framework of *indistinguishability* as the key component of the evaluation of computerised decision support systems. Case studies are presented in which it has been clearly demonstrated that human expert performance is less than perfect, together with techniques that may enable fuzzy systems to emulate human-level performance including variability. In conclusion, this paper argues for the need for 'fuzzy AI' in two senses: (i) the need for fuzzy methodologies (in the technical sense of Zadeh's fuzzy sets and systems) as knowledge-based systems to represent and reason with uncertainty; and (ii) the need for fuzziness (in the non-technical sense) with an acceptance of imperfect performance in evaluating AI systems.**

***Index Terms*—Fuzzy Sets, Fuzzy Inference Systems, Human Reasoning, Approximate Reasoning, Artificial Intelligence.**

I. Introduction

Following several peaks and troughs, Artificial Intelligence (AI) is once again at the forefront of Computer Science research around the world. Perhaps the first clear demonstration of super-human machine intelligence (that is, machine intelligence that can outperform the best human performance) was in 1997 when IBM's Deep Blue chess-playing computer challenged Gary Kasparov (the reigning world chess champion and possessing the highest official ranking score at the time) to a chess match under regular tournament style rules. Deep Blue emerged victorious, beating Kasparov by 3½–2½ over a closely fought six game match. Whilst IBM have never published full details of the algorithms employed, Deep Blue featured a high-speed parallel implementation of an alpha-beta search featuring a board evaluation algorithm [1]. In this regard, the term 'AI' or even 'machine intelligence' can be disputed as an accurate description of what is essentially a brute-force search algorithm containing no real intelligence; nevertheless, Deep Blue beat a human at chess (arguably, the best in the world) in a task that requires human intelligence.

This work was supported by the University of Nottingham.

J. Garibaldi is with the School of Computer Science, University of Nottingham, Nottingham, NG8 1BB, UK, e-mail: (jon.garibaldi@nottingham.ac.uk)
Corresponding author: jon.garibaldi@nottingham.ac.uk

Research into artificial neural networks has been on-going since the 1950s, punctuated by a hiatus in the 1970s following the publication of Minksy and Papert's exposition on the limitation of perceptrons in 1969 [2], followed by gradual incremental advances on a number of fronts (algorithmic and implementations). The recent explosion in interest in Deep Learning probably commenced around 2010 when deep convolutional neural nets achieved super-human performance in a visual pattern recognition contest at the International Joint Conference on Artificial Intelligence [3] and demonstrated significant improvement over other approaches on several benchmark problems at the Conference on Computer Vision and Pattern Recognition (CVPR) in 2012 [4].

In 2017, twenty years on from Deep Blue, AlphaGo Master demonstrated super-human performance by beating the best human player in world at Go, Ke Jie, 3-0 in the *Future of Go Summit* held in Wuzhen. Alpha Go Master, developed by the Google DeepMind company, employed deep learning in the form of convolutional neural networks combined with a Monte Carlo tree search algorithm, trained through reinforcement learning. Whilst in some ways Monte Carlo tree search is a straight-forward and easily understood search algorithm, the fact that it is heuristic and non-deterministic seems to lead to it being more often described as an 'AI-algorithm'. In particular, the combination of Monte Carlo tree search and deep learning neural networks, both of which do not feature explicit knowledge, has meant that the AlphaGo series of programs is often described as being AI.

The purpose of this paper is not to discuss definitions of AI or machine intelligence; two standard dictionary definitions of artificial intelligence are:

> "computer systems able to perform tasks normally requiring human intelligence" [OED]
> "the capability of a machine to imitate intelligent human behavior" [Merriam-Webster]

Nor is the purpose to undertake a philosophical exploration of whether any particular computer program is viewed as constituting AI, other than to note Turing's observation that when an algorithm is completely predictable, it is unlikely to be considered to be intelligent [5]:

> "The extent to which we regard something as behaving in an intelligent manner is determined as much by our own state of mind and training as by the properties of the object under consideration. If we are able to explain and predict its behaviour or if there seems to be little underlying plan, we have little temptation to imagine intelligence." [Turing, 1948]

Whilst sub-symbolic approaches such as deep learning are currently the vogue, this paper argues for the need *in specific contexts* for knowledge-based approaches to AI, with explicit representation of and reasoning with uncertainty. As part of this argument, the use of fuzzy techniques is advocated as one suitable approach that can deliver the necessary capabilities. Furthermore, it is claimed that the presence of imperfect reasoning and imperfect performance is, and must be, an essential feature of acceptance testing of these AI algorithms.

The rest of this paper is organised as follows. Section II presents recent background on the need for explainable AI, advocating the role of decision support systems. Section III introduces the fundamental basis of and need for uncertainty handling in decision support, while Section IV outlines how decision support systems might be evaluated through a form of a Turing Test. Section V and Section VI discuss how variation in reasoning is a feature of human reasoning and thus should be incorporated in computer expert system reasoning, if it is to pass such evaluation tests. Section VII then presents some current techniques that may be used, in the context of fuzzy expert systems, to incorporate variation, and the benefits that can be obtained by doing so. Discussion of these potential benefits and speculation around various points that arise are given in Section VIII. Finally, some possible future directions of research are outlined and the main conclusions are summarised.

II. Explainable AI

The impressive advances in performance achieved recently by deep learning techniques has increased the focus of attention on sub-symbolic and statistical approaches to AI. Whilst these approaches have indeed achieved some remarkable results, they can suffer from lacking interpretability or the ability for their decision processes to be understood or explained.

The need for explainable AI has grown in importance recently, with explicit recognition through, for example, the Explainable AI (XAI) program of projects funded by DARPA [6]. This suite of research projects is mainly exploring either the development of novel techniques to add (or enhance) the capability of humans to understand the processing of deep learning techniques and/or to develop "alternative machine learning techniques that learn more structured, interpretable, or causal models". The program also extends to the design and development of new explanation interfaces and in exploring "research directly related to the problem of explaining machine learning models to end users".

In parallel, the European Union (EU), recently introduced its directive on General Data Protection Regulation (GDPR) [7], which includes explicit requirements for individuals "*not to be subject to a decision [...] which is based solely on automated processing and which [...] significantly affects him or her [...] without any human intervention*". Further, it requires that such data processing algorithms should allow the individuals subject to the decision "*to obtain an explanation of the decision reached after such assessment and to challenge the decision*".

Deep learning and statistical approaches have had most success in areas such as understanding speech, handwriting recognition, facial recognition and other visual tasks such as generic object recognition. It is not obvious that humans use explicit rules when performing such tasks. Consider, for example, natural language understanding: whilst it is hard to write complete rules of a language such as English, the vast majority of humans have largely mastered the speaking of their native tongue(s) by age five. Similarly, humans are capable of complex visual processing tasks including facial and object recognition, and 3D perception of the environment including estimations of distance, velocity and acceleration. Again, these abilities are acquired by most humans without any recourse to rules or other forms of explicit knowledge representation.

Whilst it is clear that many tasks requiring intelligence can be carried out without explicit knowledge representation and reasoning, nevertheless, it seems equally obvious that many tasks requiring human intelligence benefit hugely from the explicit use of knowledge. Examples are very numerous, but include areas such as the 'rules of the road' — it is more efficient to be told that one should stop at a red-light and proceed through a green one, rather than discover the general principle through reinforcement learning! Indeed, the rule in many countries is more complex, and it is far more efficient to read the explicit rules found on the Wikipedia page 'Turn on red' (https://en.wikipedia.org/wiki/Turn_on_red) rather than learning by example in each country visited.

A similar example might be in mathematics. Of course, relatively simple concepts such as counting, simple addition and even multiplication may be easy to acquire through experience. But some more advanced concepts such as (for example) 'the chain rule' (for computing the derivative of the composition of two or more functions) or integration by parts (finding the integral of a product of functions) is easier to acquire by learning the rules, rather than by deriving from first principles. Then, once the rules have been learned, it is far more efficient to simply apply the rules in subsequent analyses.

Based on these considerations, there is a clear case for the need for knowledge-based approaches in some contexts, as a complementary technique to supplement sub-symbolic approaches. Of course, it can be argued that sub-symbolic approaches might be used in conjunction with explanatory components which are able to 'add' interpretable representations of knowledge which explain how decisions were arrived at, after the event — i.e. that decision making processes and explanatory processes can be two completely independent components in an AI system. This is not a contradiction, but the important point is that knowledge can be explicitly represented and incorporated into AI systems, and there is sometimes benefit from doing so. Hence, it would seem slightly perverse to deliberately not take advantage of this.

As a concluding remark, the use of knowledge-base approaches does not necessarily mean that these must be in the form of 'IF-THEN' production rules that feature in expert systems. The existence of non rule-based techniques such as decision trees or case-based reasoning confirms this point. However, it seems unarguable that knowledge-based approaches are useful, and that *one form* of knowledge-based approach that may be used are expert systems in which some aspects of knowledge are represented mainly as if–then rules.

III. Fuzzy Sets and Systems

Having established the need for knowledge-based approaches and the usefulness of expert systems for AI (as complementary to other approaches), a case will now be made for the use of techniques based on fuzzy sets and systems within such expert systems.

There is uncertainty present in all physical measurements and almost all human knowledge. Of course, one can construct items of data which have no uncertainty, such as *the number of white kings on a chess board* and similar items of knowledge (facts) such as *Albert Einstein is no longer alive*, but in terms of the type of knowledge that usually features within AI (expert) systems, uncertainty is an ever-present feature. In physical measurements, uncertainty is present in all measuring devices and, of course, in the fundamental properties of nature as expressed by Heisenberg's famous uncertainty principle [8]. The essential meaning of which is that it is not possible to know both the position and the velocity (momentum) of a particle at the same time, with zero uncertainty. This fundamental uncertainty is, in practice, infinitesimally small compared to actual uncertainties observed in any real physical measurements, but nevertheless it emphasises the fact the uncertainty is ever present, and is not simply an artefact.

Fuzzy sets were introduced in 1965 by Zadeh [9] specifically to deal with difficult "classes of objects encountered in the real physical world" which are "imprecisely defined" and yet "play an important role in human thinking". In 1973 [10] and then his seminal papers of 1975 [11], [12], [13] he introduced the full framework of fuzzy logic, including linguistic variables, specifically to be used "in the realm of humanistic systems" which he defined to include AI and human decision processes. Since their introduction, fuzzy systems have clearly had significant impact, largely (and somewhat ironically) in the area of control engineering, rather than human reasoning. Whilst, as already observed, rule-based systems are just one form of knowledge-based approaches, they have proved popular frameworks for representing and reasoning with explicit knowledge in which there is uncertainty (imprecision) in both data and knowledge. Nevertheless, from Mamdani's first papers on fuzzy control [14], [15] onwards, fuzzy rule-based systems have been used with great effect [16], [17], [18]. The terms 'fuzzy inference system' (FIS) and 'fuzzy expert system' have both been used as an alternative label for fuzzy rule-based systems — i.e. fuzzy systems which feature a rule-base in the form of fuzzy 'IF-THEN' production rules. Of course, there are many forms of non-rule-based fuzzy inference such as fuzzy neural network approaches [19], [20], but these will not be further discussed here.

Medical decision making domains are a natural target for FISs, due to the presence of uncertainty in data and knowledge, combined with a high level of desirability for the need for explanatory facilities within the system. One example of a medical FIS is that of the Expert DataCare system [21], see Fig. 1, which will serve as an illustrative example as follows.

The process of birth is a stressful process for both mother and infant child; in this context 'stressful' is meant in the technical medical meaning of "a physical, mental, or emotional factor that causes bodily or mental tension" [MedicineNet]. The issue is that deprivation of oxygen (hypoxia) is part of normal birth process, due to restriction of blood-flow through the umbilical cord of the infant during the contractions of labour, and that this hypoxia can develop into a condition serious enough to cause physical harm (asphyxia), potentially in the form of brain damage in the initial stages, but ultimately leading to infant death. The term 'perinatal asphyxia' (asphyxia which occurs at the time of birth) is used in this context, and is thought to affect something like 2-10 cases per one thousand births, with death occurring in perhaps 1-5 infants per thousand.

Fig. 1: The Expert DataCare umbilical acid-base expert system in a clinical setting. The blood gas machine is shown on the left, Expert DataCare implemented on a laptop in the middle, with a printer to output the results on the right.

In the 1990s, exploratory research was undertaken to create an objective measurement-based method to diagnose the presence and severity of perinatal asphyxia, in order to inform the medical staff present at the birth as to the need for resuscitation or other active forms of medical intervention. The method investigated was based on the analysis of biophysical measurements of blood samples taken from the artery and vein of an infant's umbilical cord that had been double-clamped and isolated. These measurements consisted of variables derived from and related to the oxygen content, carbon dioxide content and lactic acid content of the venous blood (oxygenated blood entering the infant) and the arterial blood (de-oxygenated blood leaving the infant). This process is known as *umbilical blood-gas analysis* or *umbilical acid-base analysis*. The medical principles behind this analysis will not be described in depth, but detailed and knowledgable assessment of umbilical acid-base measurements has been shown to provide clinical information on the asphyxiated state of the infant [22]. However, this assessment process is complex and requires significant specialist expertise: it is hereafter referred to as a decision process, in the sense that a decision is made as to the severity of asphyxia present in the infant based on the umbilical acid-base measurements.

Expert DataCare was created to undertake analysis of such acid-base measurements taken from arterial and venous vessels of an infant's umbilical cord at the instant of delivery (birth) [21]. The system was created as an expert system, based (loosely) on rules of interpretation extracted from interviews undertaken with world-leading experts in umbilical acid-base analysis. Whilst the crisp version of the system was found to perform at a level sufficient for deployment in a clinical setting [21], nevertheless it was extended to employ fuzzy

rules within a fuzzy inference system in an attempt to overcome perceived limitations with interpretations falling at crisp boundaries of decision making [23], [24]. The conversion of the crisp system to a fuzzy system led to the identification of an additional key challenge: how could the performance of this 'improved' expert system, featuring fuzzy inference, be properly evaluated in comparison with the existing crisp expert system and human experts?

IV. Evaluating Artificial Intelligence

In the context of computer games, such as Deep Blue and Alpha Go described earlier, evaluation of their performance including (and particularly) in comparison with human experts in the domain is obvious: simply arrange a competitive game with selected human experts and see who wins (accepting that there may be very significant practical issues in arranging such a competition). But, this raises an important and critical question in regards to computer expert systems: *how should expert systems be properly evaluated as to their level of performance in a given domain?* As a corollary, *what level of performance of a computer expert system should be considered sufficient to allow its deployment in the real world?*

A. The Turing Test

In Alan Turing's seminal paper "Computing Machinery and Intelligence" [25], he introduced what he termed 'The Imitation Game' as a proxy test for answering the question 'Can machines think?'. In the original version of the game (subsequently revised by Turing) there are two subjects, a man (A) and a woman (B). A third person, an interrogator (C) of either sex, may ask questions of A and B in order to attempt to identify which is the man and which is the woman. It is part of A's objective (the man) to cause C (the interrogator) to make the wrong identification whilst it is part of B's objective (the woman) to help the interrogator. Then Turing suggests replacing A (the man) with a computer, and proposes that the question "Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?" as replacing the original question 'Can machines think?'.

In subsequent clarifications, Turing dropped the aspect of having a male and female (with the task being to correctly identify gender), and instead simplified it to just being a task of distinguishing machine from man (generally now taken to simply mean a person of either gender). The Imitation Game is shown in stylised form in Fig. 2; it has subsequently become known as simply 'The Turing Test' for AI. The fundamental concept underlying the Turing Test is the concept of indistinguishability — essentially, if a computer is indistinguishable from a human, then it can be deemed as *intelligent*, at least to the same degree as humans are. Note that, of course, Turing emphasises that this indistinguishability is in the domain of the conversational exchange *only* — i.e. physical form, for example, is not relevant here.

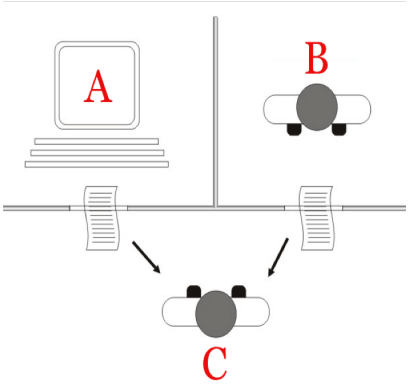


Fig. 2: A stylised representation of the Turing Test (Juan Alberto Sánchez Margallo [CC BY-SA 2.5]).

B. Evaluating Decision Support Systems

The answers to the questions posed in the opening paragraph of this Section may then be as follows. A computer expert system should be evaluated by undertaking an experiment along similar lines to the Turing Test — that is, to pose problems to the expert system and human expert(s) and assess whether observer(s) can distinguish which is the computer expert system and which is/are the human participants. If the expert system is *indistinguishable* from the human experts, then it has demonstrated sufficient expertise to be deployed, *regardless of the absolute level of performance attained*.

Of course, this suggestion is not precisely equivalent to the Turing Test. But, the Turing Test is actually relatively poorly specified in practice. As alluded to above, Turing himself revised and clarified the Test during his lifetime, altering from the originally posed formulation consisting of imitating a man as compared to a woman, to that of simply distinguishing computer from human. He also later clarified that it should be undertaken by a "jury" (i.e. not just a single observer) and repeated multiple times (implying a statistical approach to determining success or failure of imitation).

So, to make this proposal more specific. A test should be constructed consisting of challenging problem instances in the domain of the problem being addressed by the expert system. Then, the problem instances are posed to both the computer expert system and a panel of experts (to obtain a representative characterisation of human expertise in the domain). The answers provided by the expert system and human experts are compared, preferably statistically. If there is no statistical difference between the computer expert system and the human experts, then the expert system is deemed to exhibit sufficient level of performance to be deployed.

In this specification the phrase "*challenging problem instances*" is used. It is sufficient to state that the level of difficulty of the problem instances used in the test provides an indication of the level of expertise that can be claimed for the expert system. For example, if the system passes a test in which only very simple mathematical problems are posed (e.g. "what is $2 + 3$?"), then the system can only be claimed to be performing at a level of simple mathematics, not as an expert mathematician, even if the panel of human

experts were themselves expert mathematicians. Developing this idea further, there may be a set (or subset) of questions for which the human experts achieve 100% performance (i.e. they consistently get all the answers correct); in this case the computer expert system should achieve 100% performance on the same set of questions. However, if a set of questions are posed that are really challenging for the human experts, then one might perfectly reasonably expect that the experts do not get all questions correct, their level of performance may be some way below 100%. In this case the expert system should only be expected to perform at a similar level. The important corollary here is that *if the problem domain is such that human experts cannot achieve 100% performance, then we should not expect a computer expert system in this domain to do so*, or to put it another way **if we allow human experts to make mistakes, then we must allow a computer expert system to do so**.

V. Variation in Human Reasoning

Not only should we allow a computer expert system to make mistakes — as Turing puts it "if a machine is expected to be infallible, it cannot also be intelligent" (Turing, 1947) — but we might also quite naturally expect it to exhibit behaviours which have some random element, or at least as Turing puts it 'partially random'. Whilst we all like to think of ourselves as rational decision makers, it is clear that humans do make mistakes. It is proposed above that the test for computer expert system performance should be carried out in comparison to a panel of human experts. The implication here is that human experts may exhibit differences in opinion — hence the need for a panel, rather than simply selecting a single expert to participate. The phenomenon of inter-expert variation (variation in opinion or answer *between* different human experts) is well established and has been studied (and measured) in many contexts, including in many scientific studies and problem domains requiring expertise. Two case studies will be described to illustrate this.

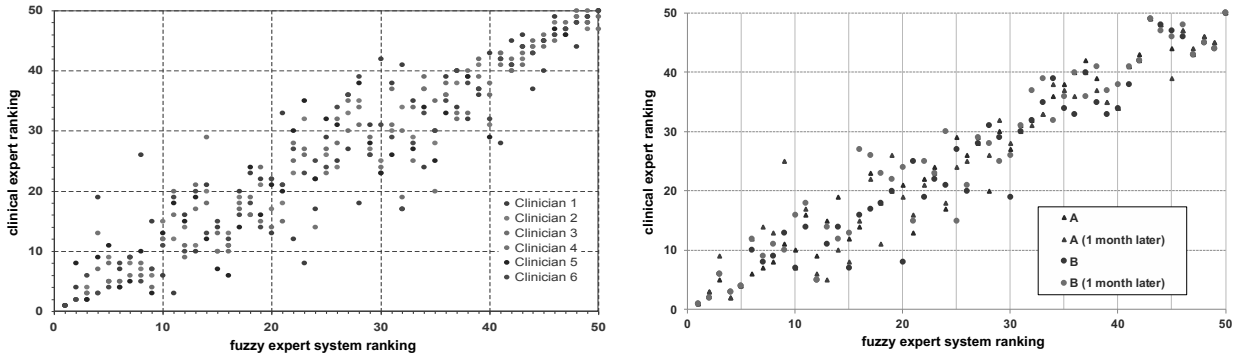
The first is one in which a computerised expert system for the analysis of the cardiotocograph (CTG) in labour was compared against human experts in order to determine if it was performing at an acceptable level [26]. This study featured a set of 17 experts from around the UK, assessed their performance in reviewing 50 cases of infant births in which CTG recording were made, and compared their decision recommendations with that of a computer expert system. The agreement between different experts (termed 'inter-expert agreement' or just 'agreement') was calculated and found to be around 60-75% — that is, the experts agreed with each other around two-thirds of the time. The computer expert system performance was indistinguishable from the experts in terms of agreement in that its agreement with the experts was around 68%. The decision making experiment in this study was carried out twice, one month apart, allowing the performance of the experts to be compared against themselves. This found the agreement of experts with *themselves* (termed 'intra-expert agreement' or 'consistency') to be around 75-90%. Interestingly, whilst the expert system was found to have a much higher consistency than the experts at over 99%, it was actually *not* 100% consistent. This was due to the fact that the system featured a small element of operator intervention, and was subjected to a minor difference in user-input in one case.

Precisely this form of Turing Test style evaluation was carried out on the Expert DataCare system. In this evaluation study, six experts in umbilical acid-base analysis were recruited to take part. Fifty challenging cases of umbilical acid-base results were selected, most of which were associated with situations of poor outcome for the infant at birth. The experts were asked to independently rank the fifty cases from 1 (worst outcome) to 50 (best outcome) in terms of how severe the lack of oxygen (birth asphyxia) was for the infant. The Expert DataCare system was then used to assess the same results, and its numerical output was similarly used to rank the cases [27].

The results are shown in Fig. 3(a). In this figure, the x-axis represents the ranking position specified by Expert DataCare, whilst the ranking position given to the same infant by each expert is represented on the y-axis. So, for example, position 1 on the x-axis represents the infant that Expert DataCare ranked as the worst (sickest) baby in terms of its blood-gas results (i.e. the infant with the most severe asphyxia); at this position, a circle is plotted on the y-axis at the ranking position given by each of the six experts. Thus, at position 1 on the x-axis, there are six circles superimposed at position 1 on the y-axis, reflecting the fact that the expert system and all six experts labelled this case as infant 1 — i.e. the sickest. At position 2 on the x-axis, there are four circles superimposed at position 2 on the y-axis, whilst there is one circle at position 4 and one at position 8. By position $x = 12$, it can be seen that there are six circles, at $y = 9$, 10, 11, 14, 15 and 16. The Spearman rank order correlation between all six experts and Expert DataCare is actually 0.95, which signifies very high overall agreement. Nevertheless, it is clear there are differences of opinion in many of the cases.

This experiment was then repeated voluntarily by two of the six clinicians, one month after the initial experiment. The same 50 cases were re-presented to the two clinicians, but in a different (random) order. Of course, the two clinicians were blinded as to their original results and each other, and once again each independently ranked the cases. The results of this repeat study are shown in Fig. 3(b). The figure is organised the same as Fig. 3(a), with the x-axis still representing the label assigned by the Expert DataCare system, but in this case there are only two experts, each with two repeats of the study. It can be seen that expert A and B both label infant 1 (i.e. the infant labelled as the sickest baby by Expert DataCare) as infant 1 both times. However, vertical separation of the markers (triangles for expert A and circles for expert B) can clearly be seen. For example, for infant 8, expert A labelled it 8 and 12, whilst B labelled it 9 and 11 — i.e. there is also clear intra-expert variation.

A further observation from Fig. 3(b) is that while there are many cases of intra-expert variation (indeed, in around half of the 50 cases, both experts have given different answers in the two repeats of the task), there are a few cases of complete agreement — case 1 is labelled as such by both experts both times, and case 50 is labelled as 50 by both each time. It is



(a) Inter-expert variability.

(b) Intra-expert variability..

Fig. 3: An illustration of inter-expert and intra-expert variability observed in the context of umbilical acid-base analysis.

perhaps reasonable to suggest that the variation observed may be related to the difficulty of the case or problem instance under consideration. In the specific instance of infant 1 in this study, the infant went on to die as a result of the events of labour, and that infant was the only one to do so. From this perspective, it seems clear that this specific individual *was* the sickest infant and thus the fact that the experts all agree, including both times for expert A and B, is perhaps indicative of this clarity. To put this another way, the more difficult a case or problem is to interpret or solve, the more variation is likely to occur.

However, as per the CTG expert system referred to earlier, whilst in terms of overall performance and agreement with the other experts, Expert DataCare is essentially indistinguishable, this is not the case when it comes to consistency. A level of variation is clearly observed in the human experts which is characteristically different in the Expert DataCare system. Indeed, the Expert DataCare system is entirely deterministic in its output given a fixed set of inputs and hence it exhibits zero variability (100% consistency) in this experiment.

VI. Variation in Expert System Reasoning

These observations bring us back to the Turing Test for expert system performance. An expert system, whilst performing similarly to experts, will not be indistinguishable from experts unless it exhibits the same degree of consistency (or variability / inconsistency) as the human experts. That is, the expert system will fail the Turing Test because it will be significantly and obviously more consistent than the human experts. This point was of course recognised by Turing in his original proposition of the test, addressing it as follows:

> "the interrogator could distinguish the machine from the man simply by setting them a number of problems in arithmetic. The machine would be unmasked because of its deadly accuracy. The reply to this is simple. The machine (programmed for playing the game) would not attempt to give the right answers to the arithmetic problems. It would deliberately introduce mistakes in a manner calculated to confuse the interrogator." [Turing, 1950]

As stated in Section IV-B, "a computerised expert system should be considered ... for deployment in the real world if its level of performance is *indistinguishable* from human experts operating in the given domain". This formulation does not actually require that the expert system *must* possess variability or make the same mistakes as human experts. Rather, it requires that the level of performance of the expert system is indistinguishable from human experts. So, if performance can be adequately measured and demonstrated to be indistinguishable without having variability, then that is sufficient. However, if a computer system can be reliably distinguished from the human experts *on any basis*, then it may be hard to argue that the performance is indistinguishable. By this argument, whilst not essential for expert systems to emulate human variability in order to satisfy evaluation tests of performance, it may be of benefit if they do.

Note that a policy to "deliberately introduce mistakes" (in Turing's words) into computer expert systems is not advocated here. The word 'mistake' clearly implies that there is an error or incorrect decision being provided, and it is difficult to see how this might be genuinely useful in any context *other than* passing a Turing Test. Rather, as illustrated in the two case studies of medical decision making mentioned above [26], [27], there may be differences of opinion between experts as to what the correct decision is, and indeed there may be multiple alternative correct decisions. Again, this is easy to conceptualise in the context of games such as chess. Of course there may be an obviously wrong move, but within a world championship game of chess, there will be differences of opinion as to the best move to make — put another way, there may be several obvious wrong moves, some questionable moves, and several alternative good moves. Differences of opinion as to which good move to make may be reflective of different reasoning processes or may even simply be an expression of randomness. Variation (non-determinism) is essential in becoming world chess champion; thus, variation in decision making is *not* the same as making mistakes.

VII. Modelling and Measuring Variation

Having established that variation in decision making behaviour is necessary in the context of producing expert systems that are able to pass tests of indistinguishability from human experts, the question arises as to how variation might be introduced and whether there is any benefit to be gained in

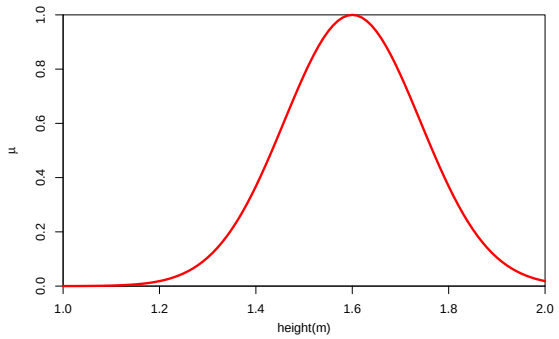


Fig. 4: An illustration of a standard (type-1) fuzzy set modelling the concept *medium height*.

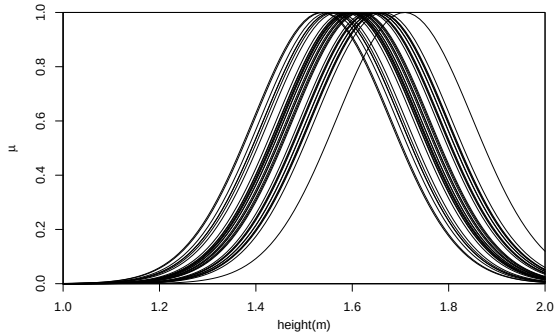


Fig. 5: An illustration of various opinions as to where the set *medium height* might be located.

doing so. A programme of research was instigated to explore whether the inter- and intra-expert variability observed in Figs. 3(a) and 3(b) could be successfully modelled, thereby measured, and whether doing so might have any impact on the performance of the Expert DataCare system.

A. Non-Stationary Fuzzy Sets

An example standard fuzzy set to model the concept *medium height* is shown in Fig. 4. Whilst, of course, the generates a non-linear mapping between the x-axis variable (in this case height in metres) with the membership of the set of *medium height*, there is no fuzziness in this mapping. In a sense the mapping is precise, even though the concept itself appears imprecise in common understanding. This apparent contradiction in the definition of standard (type-1) fuzzy sets was recognised by Zadeh himself, and addressed through the introduction of *type-2* fuzzy sets [11], in which the membership at each value of the domain x is given as a type-1 fuzzy set, rather than as a precise number.

However, a conventional type-2 fuzzy set, whilst 'blurring' the membership function, does not explicitly represent variability in reasoning. Consider as a thought experiment asking different individuals to independently position the set *medium height* on the x-axis (actually it is straight forward to undertake such an experiment, for example, in a classroom of students). This invariably results in a difference of opinion as to where the set should be located, as illustrated in Fig. 5.

A novel modification of a fuzzy set, originally named the *non-stationary fuzzy set* has been proposed [28], and more recently refined to the concept of a *constrained type-2 fuzzy set* [29]. This conceptual framework provides a mechanism for defining an underlying base or *generator* type-1 fuzzy set to model a concept. Then explicit variations of that generator set may be created to represent differences of opinion as to the precise location of the underlying concept. This is illustrated in Fig. 6, in which the generator set (the underlying concept) is shown in red and its variations are shown in black.

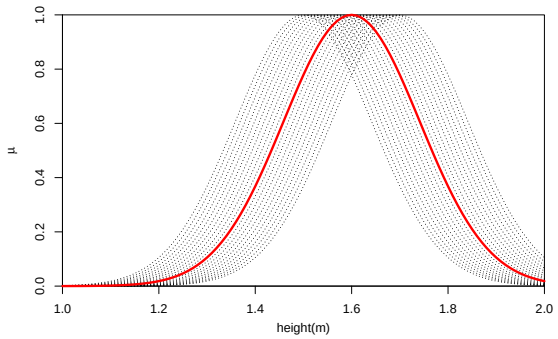


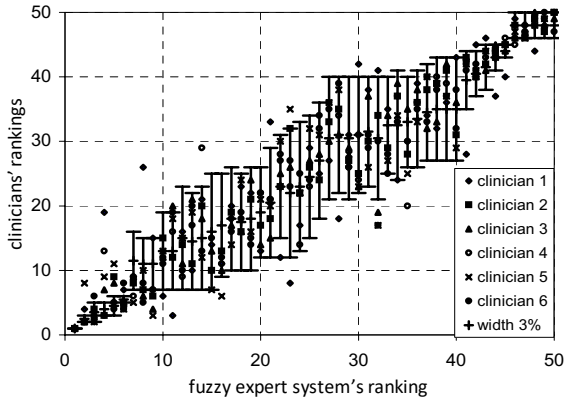
Fig. 6: An illustration of a non-stationary fuzzy set, providing a mechanism for modelling variation in fuzzy inferencing.

The amount of variation allowed in each concept (type-1 set) is explicitly controlled, usually expressed as the amount of variation in the centre point (location) of the type-1 set represented as a percentage of the universe of discourse (the x-axis). Fig. 6 illustrates a variation of 20%, corresponding to left-right shift of the generator set which covers 20% of the x-axis. Actually, this would constitute a huge variation (or uncertainty) in the location of the set — in practice variations of between 1% and a maximum of 10% have been explored.

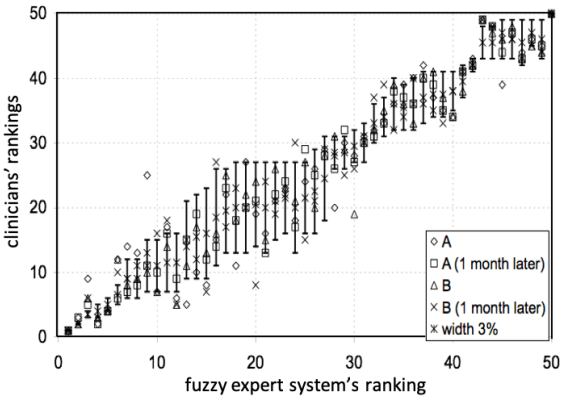
B. Modelling Variation in Umbilical Acid-Base Assessment

The mechanism of non-stationary fuzzy sets was used to model the inter- and intra-expert variation observed in umbilical acid-base assessment, as seen in Figs. 3(a) and 3(b). Experiments were carried out to measure the variation observed in the final ordering of infants that was obtained when the fuzzy sets in Expert DataCare had a specified amount of variation, ranging from 1% to 10%. These experiments were carried out in two independent trials: one to find the best match as compared to the observed inter-expert variation, and a second to find the best match for intra-expert variation. The results of these experiments are shown in Fig. 7, in which the error-bars plotted on the y-axis at each position on the x-axis show the minimum and maximum opinion obtained from the Expert DataCare system containing variability. The first finding was that 3% variation in the underlying fuzzy sets of Expert DataCare obtained variation which was the best match for inter-expert opinion, as shown in Fig. 7(a). Somewhat surprisingly, it was also found that the same 3% variation also best matched intra-expert opinion, as shown in Fig. 7(b).

These experiments represent the first time that variation incorporated into fuzzy sets has been used to model the variation observed in human reasoning. Two primary observations may be made. Firstly, that this technique has not only been used to model variation, but also that it has provided a



(a) Modelling of inter-expert agreement with 3% variation.



(b) Modelling of intra-expert agreement with 3% variation.

Fig. 7: The optimal non-stationary variation to model expert variability.

quantification of variation in the observed opinions. Indeed, a variation of 3% (in location on the x-axis) best matches the inter-variation *and* the intra-variation — loosely, it might be claimed that the amount of inter-expert variation observed in this specific context *is* 3% and the intra-expert variation *is also* 3%. Secondly, it is quite clear that this variation is categorically *not* the same as adding random 'mistakes' into the system. It can be seen that the infant labelled in position 1 (the sickest baby, as in Sec. V) is consistently labelled as such by the system with 3% variation, whereas more variation is observed for infants in the middle of the distribution.

In summary, the use of non-stationary fuzzy sets has allowed the variation in human decision making to be modelled and quantified. Further, it has demonstrated that having some variation in the underlying fuzzy sets that constitute a system — which essentially correspond to the underlying meaning of terms used in the decision context — leads to interesting patterns of variation in the end decision. **Variation can be incorporated into computer expert systems to model human variation, but this is not random decision making.**

C. Impact of Variation on Performance

Having successfully modelled and quantified variation in a decision context, the next obvious question is to whether this variation may be used to somehow improve decision making. Faced with the undeniable fact that variation in opinion is often observed between human experts, a strategy that is commonly adopted is simply to ask the opinion of more than one expert — the 'jury' as Turing called it — and to somehow resolve the (possibly differing) views obtained into an overall consensus opinion. Several methods have been proposed as to reaching consensus when differing opinions are obtained. Obvious methods might be to take the average (this requires numeric output) or perhaps majority vote (which can be used for categorical decisions) and many others.

A further set of experiments were carried out to explore the impact of variation in fuzzy expert systems, using non-stationary fuzzy sets, on the performance of such systems. In order to do so, a different decision making context was required in which there existed some form of target output ('correct' or 'objective' decision) against which the performance of the fuzzy system could be compared. The decision context chosen was in regard to selecting the appropriate form of treatment to apply in cases of breast cancer.

NPI < 3.0	No Adjuvant Treatment
NPI $3.1 - 3.4$	
ER +ve	Recommend Hormone therapy
ER -ve	Recommend Chemotherapy if VI
NPI $3.4 - 4.4$	
ER +ve	Recommend Hormone therapy
ER -ve	Recommend Chemotherapy
NPI > 4.4	
ER +ve	Discuss Chemotherapy Consider: Recommending Chemotherapy: Age < 40 VI HER-2 +ve Weak ER ($< 100/300$) Recommending Against Chemotherapy: Age > 60 Only 1 LN positive Special type cancer
ER -ve	Recommend Chemotherapy

Fig. 8: The clinical protocol, as written, for advising patients on the need for follow-up treatments following surgery for breast cancer, in use in Nottingham University Hospitals NHS Trust.

If a woman is diagnosed as suffering from breast cancer, the initial treatment is relatively straight-forward in that (in most cases) surgery is immediately undertaken to remove the tumour from the breast. However, there are then multiple differing options as to whether further follow-up treatment ('adjuvant therapy' in medical terms) is administered. There are several options available in modern medical care, including drug therapies, radiotherapy and chemotherapy. Recently, the use of chemotherapy is restricted to only the most serious and aggressive forms of breast cancer, as the treatment (whilst potentially very effective) is highly toxic and can cause significant side-effects. A clinical protocol for the administration of chemotherapy is shown in Fig. 8.

This clinical protocol was in use in Nottingham University Hospitals NHS Trust for many years, and a set of data consisting of over 1300 women and the decisions made on administering chemotherapy was made available. The decision

Rule	Antecedent	Consequent
1	IF (*NPI* is *Low*)	THEN (*Chemo* is *No*)
2	IF (*NPI* is *Medium low*) and (*ER* is not *Negative*)	THEN (*Chemo* is *No*)
3	IF (*NPI* is *Medium low*) and (*ER* is *Negative*)	THEN (*Chemo* is *Maybe*)
4	IF (*NPI* is *Medium high*) and (*ER* is not *Negative*)	THEN (*Chemo* is *No*)
5	IF (*NPI* is *Medium high*) and (*ER* is *Negative*)	THEN (*Chemo* is *Yes*)
6	IF (*NPI* is *High*) and (*ER* is not *Negative*)	THEN (*Chemo* is *Maybe*)
7	IF (*NPI* is *High*) and (*ER* is not *Negative*) and (*Age* is *Young*)	THEN (*Chemo* is *Yes*)
8	IF (*NPI* is *High*) and (*ER* is not *Negative*) and (*VI* is *Yes*)	THEN (*Chemo* is *Yes*)
9	IF (*NPI* is *High*) and (*ER* is *Weak*)	THEN (*Chemo* is *Yes*)
10	IF (*NPI* is *High*) and (*ER* is not *Negative*) and (*Age* is *Old*)	THEN (*Chemo* is *No*)
11	IF (*NPI* is *High*) and (*ER* is not *Negative*) and (*LN* is *Negative*)	THEN (*Chemo* is *No*)
12	IF (*NPI* is *High*) and (*ER* is *Negative*)	THEN (*Chemo* is *Yes*)

Fig. 9: Fuzzy rule-base broadly equivalent to the clinical protocol for advising on chemotherapy (*only*).

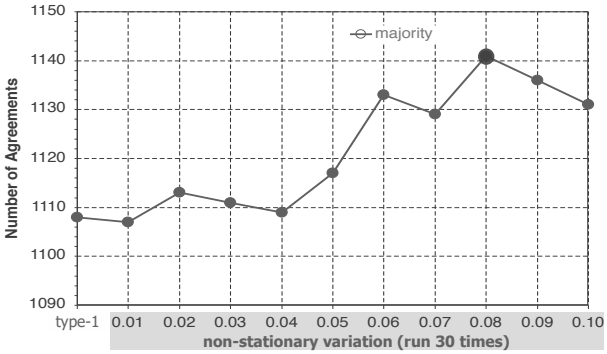


Fig. 10: Performance increase obtained by taking majority vote of an ensemble of 30 non-stationary fuzzy systems.

rules embedded within this clinical protocol were implemented within a fuzzy expert system, focussing solely on the decision as to whether to administer chemotherapy (i.e. the administration of other therapies was omitted at this point). The fuzzy rule set is shown in Fig. 9. The terms (fuzzy sets) in the input and output variables were derived from expert and domain knowledge. For example, the *NPI* variable comprised four terms, *Low*, *Medium Low*, *Medium High* and *High* to correspond to the four main decision categories that can be observed in the protocol in Fig. 8. The membership functions of the fuzzy sets were constructed such that there were crossover points at the boundaries specified in the protocol — for example the crossover between *Low* and *Medium Low* sets occurred at a value between 3.0 and 3.1.

Having constructed an initial fuzzy expert system based on this protocol, this system underwent some basic (but thorough) optimisation, to determine the 'optimal' operating configuration in order to maximise its performance in terms of agreements against the decisions made in actual clinical practice [30]. That is, to maximise the agreement between the fuzzy expert system's recommendation for administering chemotherapy (in terms of the *Chemo* output of the system being *No*, *Maybe* or *Yes* and the actual advice given to the real patients (as recorded on the clinical database).

A set of experiments was then carried out to compare this base type-1 fuzzy expert system against alternative systems which featured non-stationary fuzzy sets. The experimental configuration was as follows. In each case, a comparator system was created which featured non-stationary fuzzy sets in which the location of the fuzzy sets varied by a fixed percentage of the universe of discourse (as illustrated in Fig. 6); this percentage of variation ranged from 1% to 10% of the universe of discourse (x-axis). The non-stationary system was run 30 times, in each case obtaining an output of *No*, *Maybe* or *Yes*. Then a majority vote was used to determine the final recommendation of the system. The results obtained are shown in Fig. 10, in which the variation present in the system is shown on the x-axis and the number of agreements with clinical practice are shown on the y-axis.

It can be seen from Fig. 10 that the base 'optimised' type-1 system achieved 1108 agreements with clinical practice, which constitutes 84.8% agreement (1108 out of 1306 cases). In comparison the fuzzy system consisting of non-stationary fuzzy sets featuring 8% variation (labelled 0.08 on the x-axis), when run 30 times and a majority vote taken, achieved 1141 agreements (87.4% of the 1306 cases). That is, an ensemble of 30 non-stationary systems, incorporating variation of 8%, *increased* performance compared to an 'optimised' system with no variation.

VIII. Discussions and Observations

The two main cases studies presented in Sec. VII illustrate two principles: firstly that, using the concept of non-stationary fuzzy sets, it is possible to model the variation, both inter-expert and intra-expert variation, observed in human experts in a decision making context; secondly, that using an ensemble of systems with variation can improve performance as compared to a system containing no variation. These case studies are not claimed to be in any way definitive, nor are they claimed to *prove* any benefit of incorporating variation into computerised expert systems. However, they are used to illustrate that *incorporating some randomness or variation in some carefully controlled manner may be beneficial*. Whilst the utility of, and indeed the need for, exhibiting variation in decision making appears absolutely clear and accepted in the context of adversarial games (is anyone really going to advocate that Deep Blue or Alpha Go should play each game exactly the same way in an entirely deterministic manner?), it has yet to be accepted as a norm in other decision making contexts. The purpose of this paper is to advocate that further exploration of this phenomena is required.

It may seem an almost paradoxical claim that varying an 'optimal' system can lead to performance increase. The resolution to this apparent paradoxical claim may be through the understanding that the term 'optimal' decision making system is actually a misnomer. A computerised expert system may be tuned to give as best performance possible on a given data set, but of course this does not guarantee the level of performance that will be achieved on future data (which invariably includes data that has not been seen previously). In a sense then, any decision support system, including fuzzy expert system or world-champion level chess playing program, is forecasting future actions in the presence of current uncertainty and imperfect knowledge. Then, the decision (or move for a chess program) arrived at is a 'best guess'; and there may (will)

be error in that best guess. A good analogy is a weather forecasting system. There are now multiple competing weather prediction platforms available across connected smart devices. In each specific instance the vendor of the weather prediction system aims for their system to be the best possible (loosely speaking, 'optimal'). But each is imperfect, and a better overall prediction can often be obtained by consulting multiple independent forecasting systems and reaching a consensus opinion. In this way, a consensus of multiple different opinions provides a better forecast than any individual 'optimal' system.

An interesting observation from Sec. VII-B is that both the inter- and intra-expert variation was found to be 3% (this is shorthand for the actual finding that 3% variation in the fuzzy expert system was found to best model both inter-expert and intra-expert variation). Considering this further, suppose there are two experts A and B. Then if each of A and B exhibit 3% *intra-expert variation* (i.e. each person varies 3% with themselves) and the inter-expert variation between A and B is also 3%, then this can only be the case if both A and B are saying the same thing but with inherent 3% variation. To understand this, consider the alternative. If both A and B possess 3% intra-expert variation *and* A and B are giving systematically different answers, then the inter-expert variation would be *more than* 3%. This remains a conjecture at present; it requires further investigation for it to become established.

An avenue for speculation that also requires significant further study is whether quantifying intra-subject variation may be useful as a proxy measure for actual expertise in a specific area. Consider as a thought experiment that there is an expert in a specific specialised subject area (such as umbilical acid-base analysis) and a novice (the person may be intelligent, but has no idea how to interpret umbilical acid-base results). Then we may postulate that the novice will guess as to the meaning of the results, whereas the expert carefully interprets them according to the best available knowledge. Is it not reasonable to suppose that the novice, who is guessing answers, will have more variation (exhibit a higher level of intra-expert variation) than the expert? Put another way, if two people perform a task, one exhibits (say) 1% variation in doing so, whilst another exhibits 10% variation, which would be considered better? Of course, it is also possible to imagine a kind of autistic savant or a person with 'photographic memory' who may be able to study a page of numbers (such as in the umbilical acid-base task), randomly rank them, and then perform the same task again a week later with perfect recall. Put another way, possessing low variation does not necessarily indicate expertise, but exhibiting high variability in a given task would seem to imply lack of expertise.

As corollary to this, it should also be noted that obtaining data on intra-subject variation from real people as subjects, and particularly those considered to be 'experts' in a particular domain, is a difficult and challenging task. It is not hard to imagine how each of us might feel if approached by a researcher with the request "please can we measure your variability so that we may assess whether you are really an expert?". In the umbilical acid-base case study, only two of the original six expert clinicians agreed to take part in the subsequent study to measure their intra-expert variability; in the case of the CTG interpretation [26] all 17 experts agreed to undertake the study twice, enabling calculation of their intra-expert variability; in both cases, strict conditions of anonymity were required before the experts agreed to participate. Thus, whilst this is an area which surely deserves further study, this is bound to be difficult.

A further observation may be made on interpretability and explainable AI. It is reasonably straight forward to see how the clinical protocol encoded in Fig. 8 has been translated into the fuzzy rule base shown in Fig. 9. Anecdotally, if one shows this fuzzy rule base to clinicians knowledgeable in breast cancer prognosis, then they will often simply read the rules and claim to understand what the fuzzy expert system is doing. Whilst on one level these rules are understandable — they are written in linguistic form, are readily readable, and use terms such as *NPI* which are commonly understood in the domain — on another level, the precise influence they have on the fuzzy inference process is quite unknown. In this way, it may be said that fuzzy expert systems provide *grey box* systems, in which the principles of operation may be understood (or perhaps appreciated) whilst the precise mechanism of operation remains unclear. However, whilst these precise underlying mechanisms of operation remain unclear to all apart from a few experts in fuzzy systems, nevertheless fuzzy systems increase the level of interpretability or provide a certain level of understanding which is above that of other inferencing and decision support systems such as (deep) neural networks. From this perspective, fuzzy systems have a vital role to play in explainable AI, as discussed in Sec. II.

A. Variation and Learning

As a final further observation, there is a relationship between variation and learning, which is important enough to justify specific emphasis. Whilst the meaning of the word 'learning' appears obvious, it is actually quite a difficult concept to define precisely. Two available standard dictionary definitions are:

> "a process which leads to the modification of behaviour or the acquisition of new abilities or responses " [OED]
> "modification of a behavioral tendency by experience" [Merriam-Webster]

These definitions both share an essential element, which constitutes a critical component of learning — there must be some *modification* of behaviour in order for learning to occur. In a Pavlovian sense, behaviour consists of certain actions being taken or *responses* being shown following an associated *stimulus*, usually through some form of sensory input. Put another way, a behaviour consists of a certain output for a given input. If the output is fixed for a given input, then the behaviour remains unchanged. For a process which will lead to a modification of a behaviour to occur, there must at some point be a different output for the same given input. That is, there must be variation in the input-output mapping. If a system (be it human or computer) always produces the same output for given input, there cannot be learning.

So, a system must vary its output for the same given input in order for it (the system) to modify its behaviour. Whilst there

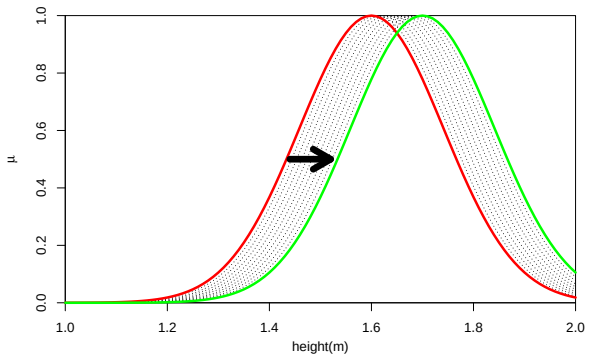


Fig. 11: If variation in the red fuzzy set leads to a permanent relocation, as shown by the green set, this will constitute learning.

are several different types of learning (e.g. [31]), a simple type of reinforcement learning may be as follows. A system has some behaviour encoded as an input-output mapping. The system has some random variation in its behaviour, such that minor variations of behaviour are exhibited from time to time; different outputs are sometimes produced from the same given input. If a difference in behaviour leads to an improvement in outcome, then the input-output mapping may be permanently changed. In the context of fuzzy expert systems, this may be modelled through the use of non-stationary fuzzy sets. Variation of the set around its original generator set is used to provide variation; if the variation leads to (consistently) improved performance, then the location of the fuzzy set permanently alters — the term corresponding to the fuzzy set would have a new meaning, learned from experience.

This concept can be illustrated by considering the concept of *medium height* as represented by the type-1 fuzzy set shown in Fig. 4. Uncertainty around where the fuzzy set should be located in shown in Fig. 5, in which the non-stationary fuzzy sets vary around the specified location. If this variation leads to a permanent shift in the location of the type-1 generator set, as shown in Fig. 11, then the fuzzy set representation of *medium height* will be permanently moved, and hence a new meaning of *medium height* will have been learned. The fundamental point here is that variation is necessary in order for learning to take place; it is hard to imagine a system without learning being deemed to be 'intelligent'; and, hence, **variation is a critical part of AI**.

IX. Future Directions

Non-stationary fuzzy sets have been used to illustrate one mechanism though which variation can be included into expert systems [28], [30]. Whilst these are not necessarily the only modelling technique available to extend standard type-1 fuzzy sets, nevertheless it is clear than some such extensions are required, both for fuzzy systems that incorporate variation, and for learning fuzzy expert systems. As mentioned earlier, type-2 fuzzy sets, as originally proposed by Zadeh, provide a rich mechanism for modelling uncertainty in the membership functions of fuzzy systems. For a long time, fully general type-2 fuzzy systems have been difficult to compute with. As a consequence, much of the work on type-2 sets in the last twenty years has been with a restricted form of type-2 set termed interval type-2 fuzzy sets. Whilst useful, and tractable for the algorithms necessary to create complete fuzzy rule-based inference, interval type-2 fuzzy sets are insufficient for modelling the variations in opinion outlined herein.

Recently, research into fully general type-2 fuzzy sets has been given impetus by the breakthrough in representation introduced by Wagner et al [32], know as 'zSlices'. These allow (arbitrarily accurate) approximations of general type-2 systems to be created, thus opening up the possibilities of progress in general type-2 fuzzy inferencing systems. But, more work is required to create accurate and efficient algorithms. There is also work needed on further exploration of the relationship between Garibaldi's non-stationary fuzzy sets and Wagner's zSlice-based general type-2 fuzzy sets, to allow these two important conceptual frameworks to be unified. Some recent work on *constrained type-2 fuzzy sets* has made some progress in this regard [29], [33], but further work is required.

Techniques for incorporating online learning into fuzzy inference systems, as discussed in Sec. VIII-A, perhaps through the inclusion of reinforcement learning mechanisms with feedback loops, are needed to increase the 'intelligence quotient' of fuzzy expert systems. Whilst learning is clearly an important part of being intelligent, it it hard to reconcile learning with the evaluation of systems. If intelligent systems are designed (and allowed or even encouraged) to learn, then of course this means that there will be modification of their behaviour over time. How then, will learning (changing) systems be evaluated as being satisfactory for deployment? Again, though, following the comments in Sec. IV-B, if humans (which are clearly learning systems) are allowed to perform tasks, surely a way must be found to allow computerised learning systems to perform the same tasks.

Finally, as mentioned in Sec. VII, it is extremely difficult and time-consuming to obtain reliable observational data on human variability. This appears to be particularly so in domains of high expertise and in safety critical systems, as it is here that the notion of variability (and particularly that of making mistakes) is considered most negatively. As stated previously, we know that humans make mistakes, even expert ones, and this does not preclude them operating. However, at present, it appears to be a quite threatening notion to explicitly accept such mistakes and to measure them; even in the context that such information will only be used to improve decision making, and will not be used as a form of punishment.

It does not appear that society has quite reconciled this position at present. Consider, for example, driverless cars. There is much current research in this area, with particular focus on avoiding accidents, particularly those that lead to death of pedestrians or other road users. Whilst, of course, we should minimise risk of injury or death, nevertheless we allow imperfect human beings to drive, resulting in regular injury and death. Indeed, according to the World Health Organization, road traffic accidents caused an estimated 1.25 million deaths worldwide in the year 2010 [34]. Suppose this staggering number could be reduced one-hundred fold through the use of driverless cars, thus saving over 1 million lives per year! Although, over 10,000 people would still be killed,

surely this overall reduction would justify the introduction of driverless cars throughout the world. Striving for perfection (and zero deaths) might unnecessarily delay the radical safety improvements that this technology could provide.

X. Conclusions

This paper has made the case firstly for the need for fuzzy expert systems, as a useful component of a suite of tools necessary for explainable AI systems, and for the need for variation to be incorporated within such systems. Whilst deep learning based neural network systems appear to currently provide perhaps the highest levels of performance available from computerised systems at present (in the context of complex problems requiring AI techniques to solve), they are difficult to explain. Fuzzy expert systems provide some increased level of explanation, potentially sufficient to satisfy requirements for such systems to be able to explain the decisions made.

The argument has also been made that humans (including 'experts', whatever that term might mean) rarely, if ever, attain 100% performance — we should not expect computerised decision support systems to do so, in order for them to be deployed in practice. Unless we allow computer systems to make the same mistakes as the best humans, we will delay the benefits that may be available through their use.

Finally, this paper has made the case that *indistinguishability* between a computerised decision support system and the human experts it seeks to emulate, performed through a form of evaluation test akin to Turing's famous Imitation Game test for AI, should be the test used for allowing deployment of these systems. **We should judge computerised decision support systems as we judge the best humans they are intended to support.**

References

[1] F.-H. Hsu, *Behind Deep Blue: Building the Computer that Defeated the World Chess Champion*. Princeton University Press, 2004.

[2] M. Minsky and S. Papert, *Perceptrons: An Introduction to Computational Geometry*. MIT Press, 1969.

[3] D. C. Cireşan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber, "Flexible, high performance convolutional neural networks for image classification," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, 2011, pp. 1237–1242.

[4] D. C. Cireşan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 3642–3649.

[5] A. M. Turing, "Intelligent machinery," National Physical Laboratory, Tech. Rep., 1948.

[6] Information Innovation Office, "Explainable artificial intelligence (xai)," Defense Advanced Research Projects Agency, Tech. Rep., 2016.

[7] The European Parliament and The Council of the European Union, "General data protection regulation," Official Journal of the European Union, Tech. Rep., 2016.

[8] W. Heisenberg, "Über den anschaulichen inhalt der quantentheoretischen kinematik und mechanik," *Zeitschrift fur Physik*, vol. 43, pp. 172–198, Mar. 1927.

[9] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, pp. 338–353, 1965.

[10] ——, "Outline of a new approach to the analysis of complex systems and decision processes," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, no. 1, pp. 28–44, 1973.

[11] ——, "The concept of a linguistic variable and its application to approximate reasoning – I," *Information Sciences*, vol. 8, pp. 199–249, 1975.

[12] ——, "The concept of a linguistic variable and its application to approximate reasoning – II," *Information Sciences*, vol. 8, pp. 301–357, 1975.

[13] ——, "The concept of a linguistic variable and its application to approximate reasoning – III," *Information Sciences*, vol. 9, pp. 43–80, 1975.

[14] E. H. Mamdani, "Application of fuzzy algorithms for control of simple dynamic plant," *Proceedings of the Institution of Electrical Engineers*, vol. 121, no. 12, pp. 1585–1588, December 1974.

[15] E. H. Mamdani and S. Assilian, "An experiment in linguistic synthesis with a fuzzy logic controller," *International Journal of Man-Machine Studies*, vol. 7, no. 1, pp. 1 – 13, 1975.

[16] L. Magdalena, *Fuzzy Rule-Based Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2015, pp. 203–218.

[17] H. Liu, P. Burnap, W. Alorainy, and M. L. Williams, "A fuzzy approach to text classification with two-stage training for ambiguous instances," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 2, pp. 227–240, April 2019.

[18] H. Mo, K. Yan, X. Zhao, Y. Zeng, X. Wang, and W. Fei-yue, "Type-2 fuzzy comprehension evaluation for tourist attractive competency," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 1, pp. 96–102, 2019.

[19] J. Wang and T. Kumbasar, "Parameter optimization of interval type-2 fuzzy neural networks based on PSO and BBBC methods," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 1, pp. 247–257, January 2019.

[20] A. Rubio-Solis, P. Melin, U. Martinez-Hernandez, and G. Panoutsos, "General type-2 radial basis function neural network: A data-driven fuzzy model," *IEEE Transactions on Fuzzy Systems*, vol. 27, no. 2, pp. 333–347, 2019.

[21] J. M. Garibaldi, J. A. Westgate, E. C. Ifeachor, and K. R. Greene, "The development and implementation of an expert system for the analysis of umbilical cord blood," *Artificial Intelligence in Medicine*, vol. 10, no. 2, pp. 129 – 144, 1997.

[22] J. A. Westgate, J. M. Garibaldi, and K. R. Greene, "Umbilical cord blood gas analysis at delivery: A time for quality data," *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 101, no. 12, pp. 1054–1063, 1994.

[23] J. M. Garibaldi, "Intelligent techniques for handling uncertainty in the assessment of neonatal outcome," Ph.D. dissertation, University of Plymouth, 1997.

[24] J. Garibaldi and E. Ifeachor, *The Development of a Fuzzy Expert System for the Analysis of Umbilical Cord Blood*, ser. Fuzzy Systems in Medicine. Studies in Fuzziness and Soft Computing. Physica, Heidelberg, 2000, vol. 41.

[25] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. LIX, no. 236, pp. 433–460, October 1950.

[26] R. D. F. Keith, S. Beckley, J. M. Garibaldi, J. A. Westgate, E. C. Ifeachor, and K. R. Greene, "A multicentre comparative study of 17 experts and an intelligent computer system for managing labour using the cardiotocogram," *BJOG: An International Journal of Obstetrics & Gynaecology*, vol. 102, no. 9, pp. 688–700, 1995.

[27] J. M. Garibaldi and T. Ozen, "Uncertain fuzzy reasoning: A case study in modelling expert decision making," *IEEE Transactions on Fuzzy Systems*, vol. 15, no. 1, pp. 16–30, 2007.

[28] J. M. Garibaldi, M. Jaroszewski, and S. Musikasuwan, "Nonstationary fuzzy sets," *IEEE Transactions on Fuzzy Systems*, vol. 16, no. 4, pp. 1072–1086, 2008.

[29] J. M. Garibaldi and S. Guadarrama, "Constrained type-2 fuzzy sets," in *IEEE Symposium on Advances in Type-2 Fuzzy Logic Systems (T2FUZZ)*. IEEE, 2011, pp. 66–73.

[30] J. M. Garibaldi, S.-M. Zhou, X.-Y. Wang, R. I. John, and I. O. Ellis, "Incorporation of expert variability into breast cancer treatment recommendation in designing clinical protocol guided fuzzy rule system models," *Journal of Biomedical Informatics*, vol. 45, no. 3, pp. 447–459, 2012.

[31] H. Zuo, J. Lu, G. Zhang, and F. Liu, "Fuzzy transfer learning using an infinite gaussian mixture model and active learning," *IEEE Transactions on Fuzzy Systems*, vol. 27, no. 2, pp. 291–303, 2019.

[32] C. Wagner and H. Hagras, "Toward general type-2 fuzzy logic systems based on zSlices," *IEEE Transactions on Fuzzy Systems*, vol. 18, no. 4, pp. 637–660, 2010.

[33] P. D'Alterio, J. M. Garibaldi, and A. Pourabdollah, "Exploring constrained type-2 fuzzy sets," in *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2018, pp. 1–7.

[34] WHO, "Global status report on road safety," World Health Organisation, Geneva, Tech. Rep., 2015.

Jonathan M. Garibaldi received the B.Sc (Hons) degree in Physics from Bristol University, UK in 1984, and the M.Sc. degree in Intelligent Systems and the Ph.D. degree in Uncertainty Handling in Immediate Neonatal Assessment from the University of Plymouth, UK in 1990 and 1997, respectively. He is Head of School of Computer Science at the University of Nottingham, UK, and leads the Intelligent Modelling and Analysis (IMA) Research Group. His main research interests are modelling uncertainty and variation in human reasoning, and in modelling and interpreting complex data to enable better decision making, particularly in medical domains. He has made many theoretical and practical contributions in fuzzy sets and systems, and in a wide range of generic machine learning techniques in real-world applications. Prof. Garibaldi has published over 200 papers on fuzzy systems and intelligent data analysis, and is the current Editor-in-Chief of IEEE Transactions on Fuzzy Systems (2017-). He has served regularly in the organising committees and programme committees of a range of leading international conferences and workshops, such as FUZZ-IEEE and WCCI. He is a Senior Member of the IEEE.