ROBOT FREE WILL

NOTICE

This paper appears in F. van Harmelan (ed) *ECAI 2002. Proceedings of the 15th European Conference on Artificial Intelligence* pages 559-563, IOS Press, Amsterdam, 21-26 July, 2002.

Robot Free Will

James A.D.W. Anderson[1]

Abstract. We introduce the perspex machine which unifies projective geometry and Turing computation and results in a supra-Turing machine. We show two ways in which the perspex machine unifies symbolic and non-symbolic AI. Firstly, we describe concrete geometrical models that map perspexes onto neural networks, some of which perform only symbolic operations. Secondly, we describe an abstract continuum of perspex logics that includes both symbolic logics and a new class of continuous logics. We argue that an axiom in symbolic logic can be the conclusion of a perspex theorem. That is, the atoms of symbolic logic can be the conclusions of sub-atomic theorems. We argue that perspex space can be mapped onto the spacetime of the universe we inhabit. This allows us to discuss how a robot might be conscious, feel, and have free will in a deterministic, or semi-deterministic, universe. We ground the reality of our universe in existence. On a theistic point, we argue that preordination and free will are compatible. On a theological point, we argue that it is not heretical for us to give robots free will. Finally, we give a pragmatic warning as to the double-edged risks of creating robots that do, or alternatively do not, have free will.

1 Introduction

We introduce the perspex machine which unifies projective geometry and Turing computation and results in a supra-Turing machine that can do everything a Turing machine can, and more. Perspex machines cannot be described entirely in language, but unlike Wittgenstein, who struggled to explain the non-linguistic underpinnings of language[4], we have no inherent difficulty, in so far as perspex computations can be laid out in space so that nearby computations are nearly the same. Hence our linguistic account of perspex computations, based on a countable set of sentences, can come arbitrarily close to describing any particular perspex computation chosen from an uncountable set, or continuum[18], of computations. Two papers currently under review take this further. One discusses some limitations to the expressability of concepts in language and hence some limits to analytical philosophy. The other shows that a technical property of rational numbers necessitates paradigm shifts in linguistic accounts of science, but does not necessitate paradigm shifts in the performance of scientific instruments. We now know that this property justifies Ockham's razor. Here we show that perspexes unify symbolic and non-symbolic approaches to AI. In a philosophically bold step we claim that perspex space can explain all of the properties of the universe. This claim is to be understood as being one instance of a schema. We fully expect that our account will be improved by substituting superior geometries and physical explanations of the universe. However, this bold step allows us to offer an explanation of consciousness, feeling, and free will in our universe. Together with this paper's mathematical companion[3], we hope we have done enough to indicate how to implement a conscious, feeling robot that possesses free will. In addition to dealing with the theistic issue of preordination and free will and the theological issue of potential heresy in usurping God's role in the creation of beings with free will, we present a pragmatic warning as to the double-edged risks of creating robots that do, or alternatively do not, have free will.

2 Glossary

Here we give the special meaning of words used in this paper. We claim that the special meaning includes the standard meaning. For example, a physical *atom* is a member of a countable set and a *choice* in folk psychology is reducible to $\text{jump}(\vec{\vec{z}}_{11}, t)$. These definitions do not rely on defining mind, but are grounded in the perspex machine.

action $\vec{\vec{x}}\vec{\vec{y}} \rightarrow \vec{\vec{z}}$; $\text{jump}(\vec{\vec{z}}_{11}, t)$.
algorithm A perspex program[3].
atom A member of a countable set.
belief X believes Y when X is conscious of its perspex program Y.
choice $\text{jump}(\vec{\vec{z}}_{11}, t)$.
consciousness A, possibly identity, synaesthesia of visual consciousness.
feeling X feels Y when X is conscious of the effects Y of transducing Z.
free will X believes that Y has free will when X believes that Y can will Z and that Y can choose Z to be an algorithm obtained from an un-encoded source.
image A vector field.
logical atom An axiom, sentence letter, or operator of symbolic logic.
perspex A 4×4 matrix of real, homogeneous co-ordinates[3].
see X can see Y when X has a partial, bidirectional mapping between an algorithm and an image of Y.
visual consciousness X is visually conscious of Y when X can see Y.
will X wills Y when X is conscious of its choice Y.

[1] Department of Computer Science, The University of Reading,
Whiteknights, Reading, England, RG6 6AY.
Email: J.A.D.W.Anderson@reading.ac.uk

3 Perspex Machine

The perspex machine[3] is embedded in a 4D space of locations called *perspex* or *program space*. Program space is defined in homogeneous co-ordinates, so it is identical to the space in some popular models of projective geometry[8,11,12]. Each location in program space contains a *perspex*[2], that is, a perspective simplex. Writing a perspex into a location of program space is equivalent to naming a part of a figure in a projective geometry proof. By default one universal halting perspex is stored implicitly at every location. Any other perspex is stored explicitly. Overwriting an explicitly stored perspex with the universal halting perspex frees the memory at that location.

A perspex is a 3D tetrahedron embedded in 4D homogeneous space. A perspex can be transformed by a bilinear transformation so that it can be projected orthogonally onto the shape of a 3D tetrahedron in Euclidean space as viewed from any position and orientation with a camera of any focal length. In other words, a perspex is a mathematical object that looks like a tetrahedron under any possible viewing condition.

The perspex machine has one instruction: $\vec{x}\vec{y} \rightarrow \vec{z}$; $\mathrm{jump}(\vec{z}_{11}, t)$. As shown in Figure 1, the instruction L, being a perspex, follows the vector x and reads the perspex at that location. The superscript arrow denotes the operation of dereferencing a pointer, and the infix arrow denotes assignment. Similarly the perspex machine reads the perspex at the location y. It then forms the matrix product of these two perspexes and reduces the product to canonical form, $\vec{x}\vec{y}$. The canonical form of a perspex p is obtained by dividing p throughout by its bottom-right element, p_{44}, if this element is non-zero. The perspex machine then writes the resultant perspex into the location z. In the conditional part of the instruction the perspex machine examines the top-left element of the resultant perspex, $\vec{z}_{11}$, and jumps to some relative location in 4D space as controlled by the vector t. If $\vec{z}_{11} < 0$ the jump is by $\begin{bmatrix} t_1 & 0 & 0 & t_4 \end{bmatrix}^T$, otherwise if $\vec{z}_{11} = 0$ the jump is by $\begin{bmatrix} 0 & t_2 & 0 & t_4 \end{bmatrix}^T$, otherwise if $\vec{z}_{11} > 0$ the jump is by $\begin{bmatrix} 0 & 0 & t_3 & t_4 \end{bmatrix}^T$. That is, the jump has a, possibly zero, component t_4 along the t-axis and a, possibly zero, component t_1, t_2, or t_3 along the x-, y-, or z-axis as the top-left element of the resultant perspex is respectively less than, equal to, or greater than zero. Thus the conditional jump and the operations of reading and writing, that is, all of the operations, are inherently geometrical.

The whole instruction $\vec{x}\vec{y} \rightarrow \vec{z}$; $\mathrm{jump}(\vec{z}_{11}, t)$ is isomorphic to a perspex with column vectors x, y, z, and t respectively. Thus perspexes can be both data and program instructions. Crucially, perspexes look like tetrahedra in real-world space, so all perspex programs have a geometrical interpretation as objects in the world and all objects in the world, described as tetrahedra, define some perspex program. Thus, a robot equipped with a perspex visual system obtains programs by the act of seeing[13] and is, therefore, both personally and historically creative[7].

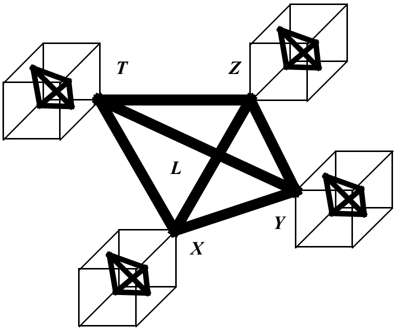

Figure 1: The perspex L accesses perspexes at locations X, Y, Z, and T.

Current perspex machines are implemented using rational numbers and are therefore Turing equivalent[3]. However, a perspex machine that operates on real numbers, or one carefully chosen irrational number, contradicts the Church-Turing thesis. Such a machine is, for example, able to test the irrational number for equality with zero, whereas a Turing machine cannot do this for a small, irrational number whose decimal expansion has an indefinite number of zeros after the decimal point and before the significant figures. It is interesting that such a simple machine contradicts the Church-Turing thesis. It is suggested in[3] that it might be possible to construct a real-numbered perspex machine as an optical computer, taking advantage of the interpretation of perspexes as perspective transformations. An ultraviolet perspex machine might operate in the petahertz range of instructions per second and should be extremely resistant to environmental interference such as radiation and immersion in liquids and gasses.

The perspex machine starts by executing the perspex at the Euclidean origin, $\begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix}^T$, and continues until it attempts to write or jump to the point at nullity[2], $\begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}^T$, or to jump relatively by nullity. The point at nullity is excluded from conventional, homogeneous models of projective geometry[11], though lines may be constructed that would pass through the point at nullity had it not been punctured from space. This is equivalent to allowing a read from nullity, so that it can be seen that the line is punctured, but not allowing any other operation on nullity. Hence the perspex machine has infinitely many halting instructions, those that attempt to cause a relative jump to nullity from each location in space, and three universal halting instructions that operate at any location in space by attempting to cause a write to nullity and/or a relative jump by nullity. In practice the zero perspex is used as the universal halting instruction. It is theoretically possible to express the Turing[14,15] and Gödel[9] incomputability theorems in a perspex machine and to ground the meaning of the Turing halting instruction in the universal perspex halting instruction.

4 Reality and Sensation

Starting from Descartes[19] *cogito ergo sum* we accept the less adventurous maxim that being aware implies that something exists. Whatever this something is, it is part, perhaps all, of that which exists. We call the whole of existence the physical universe, but accept the scientific and philosophical responsibility to seek to correct our account of the universe if we believe there is some error in it. We call this paradigm[16] *scientific realism*.

In defining existence as physical we assert that everything that exists is functionally related to what we believe we know of the universe. Given our current understanding of physics, we suppose that everything that exists is composed of matter and energy, though we will change this opinion if presented with good evidence that, for example, there is some other substance in the universe or that the concepts of matter and energy are otherwise inadequate. As a further example, if presented with good evidence that our perceptions are dreams in the mind of God[6], we will accept that God is the physical substance of at least the part of the universe we are aware of. This willingness to change our opinion when presented with evidence distinguishes scientific realism from *naive realism* in which a naive observer supposes that the universe is what it appears to be to the senses.

Some philosophers argue that the closest we can come to knowing the universe is to know the functional relationships between our sensations (sense[5]) and to deduce that these are caused by objects that exist (sensibillia[5]). This view is compatible with our account of scientific realism, though we further assert that all possible functions can be expressed as perspexes.

In this paper we suppose that all sensation is reducible to seeing. We are able to maintain this position because of the very general definition we give of seeing, but we will give up this position if presented with good evidence of error in our formulation or of the special character of any other sensory modality. Our emphasis on seeing is simply to use this sensory modality as a, probably, simplified model of all sensation.

5 Seeing, Consciousness, and Feeling

The definition of seeing given in the glossary, “X can see Y when X has a partial, bidirectional mapping between an algorithm and an image of Y”, replaces our earlier definitions[1,2]. Here algorithms stand for all kinds of knowledge. Knowledge of the world is obtained by a mapping from the image to an algorithm, but this mapping need only be partial. One need see only part of the Earth to deduce that the whole of the Earth is present. Such a deduction from sensation is consistent with the paradigm of scientific realism. The converse mapping, from knowledge to image, is justified by the modality of seeing. When we see we can indicate where things are in an image and, often, where things are in the world that is projected onto the image. However, this indication is often incomplete and approximate, so that the mapping from algorithm to image is often partial too. Our other human senses fit this definition. We obtain knowledge from our sense organs, but can also indicate, at least, where our sense organs are in our bodies. Thus a partial, bidirectional mapping is maintained, we argue, for all our senses.

The definition of visual consciousness given in the glossary, “X is visually conscious of Y when X can see Y”, gives seeing the status of consciousness, which is to claim that there is no special content to consciousness, though we admit a difference in degree. Thus a passive, infra-red detector that turns on a light when a hot object moves in a scene has less consciousness than a human who observes the same object in the scene and turns on the light. But, we argue, this difference is to do with the resolution of sight and the number and sophistication of algorithms (knowledge) and has nothing to do with any special content of consciousness. As usual, our paradigm of scientific realism requires that we will give up this claim if presented with good evidence to the contrary, but, in the mean time, we believe our definition serves as a workable, scientific model that will be sufficient to guide the development of conscious robots.

The definition of feeling given in the glossary, “X feels Y when X is conscious of the effects Y of transducing Z”, encompasses both functional accounts of feeling and accounts of a special physical content to feeling. Consider feeling the passage of time. A robot might be equipped with a clock and a transducer such that it can count the number of clock ticks. This supports a functional feeling of time, but the robot might be conscious of other effects of the passage of time. For example, it might see objects move in the world, it might see its own programs execute and terminate at different times, only to be replaced by the execution of new programs. The robot might see that its performance declines over extended periods of time as its components wear out. Such a robot might be conscious of the passage of time in many ways, and there is a physical content to this consciousness - the elapsed time. Thus we argue that our definition of feeling encompasses all current, philosophical accounts of feeling and gives us a workable, scientific model to guide the development of robots with feeling.

Notice that all of these definitions of seeing, consciousness, and feeling are based, ultimately, on properties of computers and, most generally, on the perspex machine. We do not define these properties in terms of mind. We hope that, one day, sufficient will be known about these properties to give a strict classification of their mental nature, but, for the present, our definitions stand as a scientific model.

One consequence of our definitions is that a computer may enjoy these supposedly mental properties, but it can do so in terms of perspex operations that are inherently geometrical and do not necessarily involve language. This increases the cognitive status of non-verbal humans, other animals, and robots.

6 Free Will

Discussions of free will tend to deal with determinism[17]. For example, those of a religious persuasion require a magical, non-deterministic component of free will that is in the provenance of the soul. They sometimes argue that in a completely deterministic world a being cannot have free will, because all of its actions are determined. By contrast, those of a scientific persuasion sometimes argue that the universe is deterministic at the macro level, so a being can have a mostly deterministic means of making choices and acting on them, but at the micro level of quantum physics the universe is non-deterministic, so a being's actions are not entirely pre-determined and the being may properly be held responsible for the consequences of its actions that it is equipped to predict. Sometimes discussion moves on to the nature of this equipment and the degree of responsibility held by those who choose to diminish or increase their ability to predict the consequences of their actions[10]. But there is an alternative to all such discussions. We define free will in a totally deterministic universe and hold that the definition applies in universes with increasing degrees of non-determinism until the increasing chaos deprives a being of a mechanism of choice.

Consider a robot in a totally deterministic universe. If the robot has free will it must have some mechanism for making choices. The universe is totally deterministic, so the mechanism of choice is entirely reliable. Having made a choice the robot must be able to generate some kind of consequences, such as actions in the universe. Conversely, a robot that cannot generate any consequences of its choices is so constrained by the universe that it cannot do anything at all, and cannot exercise free will in particular. So the ability to make choices and generate consequences is part of the notion of free will. But what more must be the case if we are to hold a robot responsible for its actions?

Suppose I program a missile to fly along a certain trajectory and explode. Suppose further that the missile is fired by someone, that it works as designed, and kills people. Then the missile is *not* responsible for the deaths it causes, because it could not exercise free will in the matter of the deaths. However, I may be held responsible for creating the weapon, because in programming the missile I could foresee its use as a weapon. Whoever fired the missile may be held responsible for the deaths caused - to the extent that that person was equipped to foresee that his or her chosen actions would cause the deaths. But what would have to pertain for the missile to be apportioned some responsibility?

Suppose that, after launch, the missile has some mechanism of choosing to detonate or else to crash without detonating. Again if, as programmer, I predetermine the choice, then the missile may not properly be held responsible. But suppose that the missile has some means of obtaining programs without input from any programmer, other than itself, and, as a consequence of these programs, makes the choice of detonating or not. Then I may be held responsible for programming this ability into the missile. Whoever launched the missile remains as responsible as before for the deaths, that is, to the extent that that person was equipped to predict that the deaths would follow from his or her chosen actions. But the missile is now responsible for choosing to explode and, to the extent that it is equipped to predict the deaths, it is responsible for them. Let us also say that those who die are responsible for bringing this military action on themselves, to the extent that they were equipped to predict it. This does not, of course, pre-judge the issue of whether the responsible parties acted in a morally right or wrong way.

Now imagine yourself as a suicide-bomber pilot in the place of the missile. If we have correctly identified the components of free will then you are exactly as responsible as the missile, but if you feel yourself to be more or less responsible then, presumably, you can correct our account of free will?

In the glossary free will is defined thus, "X believes that Y has free will when X believes that Y can will Z and that Y can choose Z to be an algorithm obtained from an un-encoded source". The definition that the algorithm can come from an unencoded source allows free will in a totally deterministic universe. For example, Newton may discover gravity as a causal consequence of seeing an apple fall, but unless the falling apple is encoded, say, as a sign from God, then Newton exercises free will in developing his theory of gravity. Similarly, in a deterministic universe without God, Newton exercises free will.

The meta level of belief in the definition of free will allows one to believe that one has free will, consequently to hold oneself morally responsible for ones chosen actions, and to suppose that others have free will and are responsible for their chosen actions. Thus one may develop a moral code that regulates behaviour in a mixed society of humans, other animals, and, potentially, robots. This definition also admits of error, so, for example, we may believe ourselves to be possessed of free will, only to discover that God has preordained all of our actions by encoding the universe so that we obey His will.

If we follow Christian theology, then God created us in his image. He is possessed of omnipotence, hence of free will, so we obtain free will in his image. We may create robots with free will by virtue of exercising our God-given free will. There is no scriptural prohibition against this, so we do so without heresy.

It is unusual to consider theological issues in a scientific paper, but it provides one widely understood conduit to the public understanding of science.

7 Neural Networks

Figure 2 shows a perspex *X, Y, Z, T* drawn as a simple neuron at its location *L*. The neuron reads from neurons at locations *X* and *Y* in program space. Data from these neurons are passed into the nucleus at *L*. Each data packet read is an exact specification of the neuron from which it comes, that is, the data packet is the perspex drawn as the neuron from which it comes. The nucleus performs the matrix multiplication and reduction to canonical form specified by its perspex instruction[3] and writes the resulting neuron into the program space at the location *Z*. The first time any neuron writes into the location *Z* it creates a neuron there, analogous to growing into that part of program space. Subsequent writes into *Z* overwrite the perspex at *Z*. Thus, in a serial program, the perspex at *Z* changes in shape and size in a way that is generally analogous to a rapid and febrile growth. By contrast a perspex machine supporting parallel programs must deal with the possibility of a destructive write contention, that is, the simultaneous, non-deterministic writing of two or more perspexes into the same location. One deterministic way to cope with this simultaneity is to form the average of the perspexes written into a location, either in the current time step, or by a weighted average over time. Hence the neuron at *Z* experiences a slow and smooth growth as a consequence of avoiding destructive write contentions - which it must do if a perspex machine is to obtain the speed advantages of parallel computation. This is analogous both to biological growth of neurons and to the changes of weights in conventional, artificial neurons. It neatly explains why both biological and artificial neurons average their input data. The neuron then passes control to the neuron at *T*.

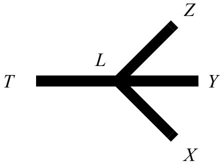

Figure 2: A perspex *X,Y,Z,T* drawn as a neuron at its location *L*.

In the case that all of the neurons, that is perspexes, are in canonical form[3], they generally form into two 3D hyperplanes, that is conventional 3D volumes, located at $t = 0, 1$. These volumes may be laid out arbitrarily in 3D space, but it keeps the data paths shorter if they are laid out as two adjacent sheets. Real, orientable, projective geometries[12] generally partition into four 3D volumes, plus oriented, dimensional nullities (vacua). In the case that some of the perspexes are in a non-canonical form, they can be arbitrarily partitioned into further 3D blocks and can be placed in 3D space. Thus perspex neurons embed naturally in a 3D space analogous to an animal brain, and have at least two distinct functional blocks, which may be laid out arbitrarily or as sheets. In practice, many sheets are created by programs that step to adjacent, integral locations in *t*, $t = 0, 1, 2, \ldots$.

The analogy with neuro-anatomy goes much further. In general, computer programs advance the program counter to the successor location far more often than they jump to a non-adjacent location. Correspondingly, sequences of perspex neurons form into long pathways. Where a peripheral device is involved, such as the pixels in a camera, there are many perspex pathways bundled together, analogous to a biological fibre bundle or tract.

It is shown in[3] that perspex programs can be continuous, or else symbolic. All perspex programs can be translated into neural networks, as just described, so we can implement symbolic or non-symbolic AI in neural networks.

8 Symbolic and Continuous Logics

It is shown in[3] that a Turing machine can be implemented as a configuration of natural-numbered perspective transformations, but that the inclusion of irrational-numbered transformations produces a supra-Turing machine. These irrational transformations form part of a continuum[18], so they lie outside any symbolic system. However, our thesis is that supposedly mental phenomena can be grounded in seeing, and hence in irrational perspective transformations. Thus the axioms of symbolic logic need not be thought of as magically, self-evidently true, but might, instead, be the conclusions of valid perspex reasoning in the continuum of sensation. Thus axioms, and all symbols in logic and language, might be justified by sub-atomic theorems in the continuum of sensation.

9 Conclusion

We introduce the perspex machine, which is a supra-Turing machine, and show how it supports symbolic and non-symbolic AI. We define consciousness, feeling, and free will in terms that can be implemented in a robot.

To conclude this very wide ranging paper, we note that there are double edged risks to creating robots with free will. On the one hand, robots without free will may be programmed to follow an evil program without remission. Robots with free will cannot be forced to follow an evil program, but they might choose to do so. We hope that robots with free will will be constructed and that they will take part as moral beings in a multi-ethnic society of humans, other animals, and robots.

Acknowledgement

This paper was greatly improved by taking the anonymous reviewers' criticisms into account. Any remaining errors, infelicities, or misunderstandings as to the interests of the AI community are entirely the author's responsibility.

References

1. Anderson, J.A.D.W. 'Visual Conviction' pp. 301-303, *AVC 89*, University of Reading, September 1989.
2. Anderson, J.A.D.W. 'Representing Geometrical Knowledge' *Phil. Trans. Roy. Soc. Lond.* series B, vol. 352, no. 1358, pp. 1129-1139, Aug. 1997.
3. Anderson, J.A.D.W. 'Perspex Machine' to appear in *Vision Geometry XI, SPIE*, Seattle, USA, 7 July 2002.
4. Audi, R., ed. *The Cambridge Dictionary of Philosophy* Cambridge university Press, 1995.
5. Austin, J.L. *Sense and Sensibilia* Oxford University Press, 1964.
6. Berkeley, G. *Principles of Human Knowledge and Three Dialogues Between Hylas and Philonous* First published 1710 and 1713. Penguin Books 1988.
7. Boden, M. *The Creative Mind* Cardinal, 1992.
8. Greenberg, M.J. *Euclidean and Non-Euclidean Geometries*, 3rd edition, W.H. Freeman, 1993.
9. Nagel, E. & Newman, J.R. *Gödel's Proof* Routledge, 1993.
10. Raphael, D.D. *Moral Philosophy* Oxford University Press, 1990.
11. Riesenfeld, R.F. 'Homogeneous Coordinates and Projective Planes in Computer Graphics' *IEE CG&A* pp. 50-55, 1981.
12. Stolfi, J. *Oriented Projective Geometry* Academic Press, 1991.
13. Swartz, R.J., ed. *Perceiving, Sensing, and Knowing* Anchor Books, 1965.
14. Turing, A.M. 'On Computable Numbers, with an Application to the Entscheidungs Problem' *Proc. Lond. Math. Soc.* vol. 2, no. 42, pp. 230-265, 1937.
15. Turing, A.M. 'On Computable Numbers, with an Application to the Entscheidungs Problem. A correction.' *Proc. Lond. Math. Soc.* vol. 2, no. 43, pp. 544-546, 1937.
16. van Frassen, B.C. *The Scientific Image* Clarendon Press, 1989.
17. Watson, G., ed. *Free Will* Oxford University Press, 1989.
18. Weyl, H. *The Continuum* Dover Publications Inc. 1994. English translation of 1919 original in German.
19. Williams, B. *Descartes* Penguin Books, 1990.