Formalising robot ethical reasoning as decision heuristics

Charles Fox
Sheffield Centre for Robotics, UK, charles.fox@sheffield.ac.uk

Abstract: Autonomous robotics ultimately seeks to replicate human functioning in all possible cognitive domains, and this must include the domain of ethics. Well-functioning ethical reasoning should enable rapid decisions to be made in difficult situations, such as where the interests of multiple parties are at stake. Decision making in autonomous systems is currently almost always construed in terms of Bayesian utility maximisation [9]. But how does ethical reasoning fit into this maximisation? Many concepts of ethics involve adherence to sets of specific rules. Again, how can such rule sit alongside utility maximisation decisions in automated systems? We consider classical philosophical concepts for ethics from the perspective of an engineer attempting to integrate them into the standard utility based decision approach.

Keywords: robot, military, ethics, utility, rules, heuristics

Most research concerning robotics and ethics has focused on ethical codes for humans, about when and where they should deploy their robots. But truly autonomous robotics ultimately seeks to replicate human functioning in all possible cognitive domains, and this must include the domain of ethics itself. For example, a truly autonomous military robot operating in theatre would need to make similarly complex ethical judgements to those made by human soldiers, which are learned though years of classroom study as well-instilled though important ceremonies and tradition. Well-functioning ethical reasoning should enable rapid decisions to to be made in difficult situations, such as where the interests of multiple parties are at stake. They could also enable an agent to know when to break explicit rules and orders if they are unethical.

To automate ethical reasoning and begin its implementation for robotics, we must define a conceptual framework to formalise ethical reasoning as an inference scheme. Decision making in autonomous systems is currently almost always construed in terms of Bayesian utility maximisation [9]. But how does ethical reasoning fit into this maximisation? Is any kind of altruistic ethics by definition in violation of standard utility maximisation? Or can altruisms be construed as particular forms of utility functions? Many concepts of ethics, especially religious systems, involve adherence to sets of specific rules. Again, how can such rules be formalised to sit alongside utility maximisation decisions in automated systems? This paper will consider classical philosophical concepts for ethics from the perspective of a robotics engineer attempting to integrate them into the standard utility based decision approach of the robotics field.

1 Utility maximisation

Current autonomous systems can be categorised according to whether they follow predetermined rules or make decisions. Examples of rule-following robots include mobile robots programmed to follow walls, solve mazes by turning left at every junction, or operate factory equipment using machine vision and if-then rules to generate actions based on the percepts. Decision making agents do not follow such pre-programmed behaviours, but internally simulate the effects of various possible actions in order to choose the best one. For example, classic AI game-playing agents search trees of board games for the best move; Monte Carlo supply simulations search for best actions to operate business supply chains [13] and financial trades [17]; and active SLAM [2] simulates mobile robot explorations to find best movements for building maps. ‘Best’ is always specified by some function $U^a(s)$ for the agent a, which maps world states s to real valued utilities, and the agent’s sole objective is to maximise this utility function. If the world states are modelled as complex structures, then multiple parts of these structures $s_i \subset s$ may contribute independently and often additively ($U^a(s) = \sum_i s_i$) to the world state utility. For example, an active SLAM robot could assign additive utilities to a word state based both on how much map-building is achieved (positive utilities) and how much time and power has been expended on motion (negative utilities).

Predetermined rule-following agents’ rules can usually be viewed as approximate heuristics to optimise a utility, whilst saving on the computation required for simulation of actions and world states. For example the simple 'always turn left' rule for maze-solving will maximise a utility function which assigns 1 to escaping the maze and 0 to all other states; and pre-programmed rules for car-building robots optimise a utility function which assigns 1 to correctly built cars and 0 otherwise. Most AI players for real-time strategy games (e.g. as seen in the source code of `boswars.org`) do not attempt to simulate the future but use hand programmed rules which have been found empirically to yield behaviour which performs well at reaching the desired utilities (such as ‘if the enemy has more than five tanks, build a defence turret’). A current research area in machine learning concerns the relationship between computationally complex inference problems and fast heuristics. For example, it is known that discriminative linear classification [12] gives identical classifications to full Bayesian inference of two generative distributions sharing a spherical covariance matrix. More recently research has examined how hierarchical Bayesian networks can be compiled into fast heuristics classifiers, such as the feed-forward systems of Helmholtz machines [16] and Deep Belief Networks [8] and kernelised Bayesian networks [22]. The field of action selection has seen similar approaches to ‘compile’ slow computational decisions into fast heuristic stimulus-response action selections. Reinforcement learning in particular [20] gives actions of this type, and psychological and machine learning research has shown how it can be trained as a secondary decision system from a primary full computational decision system [21, 11].

2 Consequentialism vs. absolutism

One way to categorise philosophical positions on ethics is to consider consequentialist views vs. absolutist views. Consequentialism means that only the final outcome state of a decision is considered in making ethical judgements, rather than the path taken to achieve that state. Absolutism means the opposite, that ethical judgements are taken to concern the actions taken rather than the outcome state. For example, various consequentialists have argued that it is good to take money (taxationists), property (communists) or life (eugenicists) from individuals by force on the grounds that the final state of society is made better off. Various absolutists have argued that these actions themselves are bad on the grounds of fundamental rights to property (the US constitution), life (human rights acts) or to freedom from all force (libertarians).

Absolutist views split into duty and virtue approaches. Duty ethics encodes judgements about actions in terms of the specifics of the actions themselves, for example ‘Do not steal’ and ‘Do not kill’ (Exodus 20), ‘Do not use force except in defence’ (Rand), ‘put the strength of our group before any individual’ (Hitler and Marx), ‘treat others as you would like to be treated yourself’ (Kant and Jesus). Asimov’s laws are of this form[A].

In contrast, virtue ethics encodes judgements on actions according to what those actions reveal about postulated underlying latent variables of one’s character. For example a robotic agent might be constructed to have characteristics concern for others: 80%, bravery: 50%, truth-telling:90%, keeping-promises: 95% Maximisation of the application of chosen traits - or sometimes the externally perceived application of them - becomes the basis for ethical decision making. From a robotics perspective we note that any trait maximisation can be re-written as a duty rule, so virtue and duty ethics for robots can be viewed together as absolutism. For example, we could construct duty rules such as ‘Thou shalt maximise truth telling’ or ‘Thou shalt maximise displays of bravery’. [B]

3 Heuristic utility maximising agents

If we consider a robot with a given utility function (leaving aside for a moment what that function should be), we can see how the philosophical distinction between consequentialist and absolutist ethics maps onto the distinction between full probabilistic computation and the use of heuristics for decision making. Consequentialism means that all consequences (utilities) of an action in the future are taken into account and the best (highest utility) action selected. Consequences of most actions are probabilistic so very large computation is required to compute consequentialist utility accurately, by

(1) $U^a(\alpha, s_0)=\sum_{t=0}^{\infty}\sum_{st+1}p(s_{t+1}|s_t, \alpha)U^a(s_{t+1})$

where the sum is over all future times t and all future states s_t at each time that could occur as a result of performing action α in the initial state s_0. For example, the decision to buy a meat sandwich will affect the existence and livelihood of many animals in the future, as well as the agent’s own hunger, the retailer's profits, the retailer’s landlord’s profits, tax revenue received by the government and so on. It is generally impractical for a human or machine to compute all consequences of real-world actions, let alone estimate all their probabilities and integrate over them to produce an expected utility.

Instead, humans rely on some form of heuristics to make most (and probably all) decisions. For split second and relatively unimportant decisions, the ‘simple heuristics’ of [7] and [3] provide rules of thumb such as ‘buy any product whose brand is familiar’ that achieve reasonably good (approximately optimal) utilities without the need to sit and compute for hours. Such rules can be acquired by various means. In the reinforcement learning literature, stimulus-response rules can be acquired through repeated experience of actions and outcomes [14], or by repeated simulation of these same actions and outcomes (some authors [19] have suggested that simulations of and learning from events requiring fast decisions is a function of dreaming in humans.) Acquiring such experiences is at least computationally time consuming; if real world experiences are used it can be dangerous and wasteful, and if simulations are used it can be inaccurate as simulations never behave exactly as the real world does.

So an alternative way to learn heuristics is to be taught them. In chess, simply being told 'don't get your king out early' saves many thousands of lost exploratory games and analysis that would otherwise be required to induce this simple but powerful rule. In life, 'don't kill', 'sell in May, go away' and 'take a rest on Sunday' are similarly simple but powerful rules which have been found useful by various people over time, and passed from person to person. Heuristics in chess and life are learned by years or millennia of experience, and are thus extremely valuable cultural artefacts, with books of them painstakingly hand-copied by scribes and handed down generations. [C]

Importantly, an agent does not need to know *why* such rules work, only that they do work. To understand them may indeed require as much experience and/or computation as to induce them from scratch. Thus in teaching such rules to humans or agents we do not necessarily need to provide any cognitive understanding, rather we must provide training to make the rules part of the agent's behaviour. For example, to train the rule 'share with your neighbours' into an agent: rather than perform or simulate thousands of actions and learn the rule by trial and error, we might design specific world states and actions for them to perform and experience so that the rule is induced more quickly, such as ritual sharing coupled with a reward. [D] Such ritualised states and actions might bear little resemblance to the real world situations in which the induced rules will be applied, but still serve to help the agent learn and apply the rule in those real world situations later.

In humans [3], rats [14] and robots [6] it is often useful to use a mixture of consequentialist and rule-based approaches for utility maximisation. For example, classic AI game players perform a consequentialist tree-search of all possible games to some level of computational complexity, then evaluate each resulting state (e.g. 4 moves ahead) using heuristic rules. Human grandmasters appear to use a similar process, though using fewer moves ahead but better rules [4]. Daw and Dayan [14] designed experiments to show explicitly when rats switch from computation to heuristics. Most readers will have had their own experiences of using simple rules for relatively small decisions (selecting a known brand in a supermarket) and of deliberately and slowly cognating over bigger decisions (making spreadsheets to decide which house to buy).

We can thus view absolutist ethical rules as computational heuristics for optimising utility functions. Robots have the advantage that rules may be programmed in directly rather than learned though classrooms or rituals.

4 Ethical utility functions

Many of the absolutist rules discussed so far, such as rules for winning at chess, do not seem especially 'ethical' in character, while others such as 'do not kill' and 'put our group before any individual' do. Why is this? Ethical rules are those that take account of real-world effects on other agents, whereas non-ethical rules do not. Ethical rules can still be construed as heuristics to maximise an agent's utility, but in cases where the agent's own utility is affected by utilities of other agents. For example, an agent could conceivably have an interest in helping other agents in its team or army, or in helping its creator, or its creator's company, family, country, genetic group or species. In military cases, it could have interests in hindering members of similar groups on the other side. In such cases, the agent would ideally include weighted copies of these other agent's utilities in its own utility function, such as

(2) $U^a(s) = w_a{}^a U_a{}^a(s) + \sum_g w_g{}^a U^g(s)$,

where U_a is the agent's purely selfish utility (concerning the state of its own body, own resources etc), U_g are the utilities of other agents or groups, and w are weights.

We should distinguish this form from rules which affect others' utilities only as a side-effect of maximising the agent's own utility. For example, 'be honest in business' in a free market society is purely a way to maximise one's own long term utility though reputation building, rather than any attempt to help others for its own sake. In such cases, the utilities of others are not considered, they are treated simply as parts of the external world, and the problem reduces to a purely private optimisation, $U^a(s) = U_a{}^a(s)$. [E]

Let us consider some aspects involved in building robotic systems having ethical utility functions of this form. For example, consider a military robot operating in theatre, alongside human soldiers, civilians and robots from its own side, from the enemy side, and - as in a typical modern conflict - from many additional allies of both sides having varying degrees of trust and kinship with the principal actors. The recent killing of

Osama bin Laden in Pakistani territory provides an interesting case of such a third party: US troops were presumably acting to maximise the interests of their colleagues first, then their commanders and country, also to minimise the interests of Al Qaeda, but with some weighted but not total respect for Pakistan's government and local population, and international human rights and national reputations. Had the attack been carried out by a fully autonomous drone then these factors would all have to be modelled in its ethical utility function.

In addition to specifying weights on other parties' utilities in eqn. 1, a robot designer would also need a way to estimate what the other parties' utilities are. Write ${U_g}^a(s) \approx U^g(s)$ for agent a's estimate of g's utility for state s. This estimation can be highly non trivial, especially in political cases such as the Pakistan example where agents may conceal their true objectives. [F]

The situation becomes even more complex when the other affected agents' utilities are also ethical, i.e. including models of yet further agents' utilities, perhaps leading to an endless recursion. One might have for example,

$$(3) \quad U^a(s) = w_a {U_a}^a(s) + {w_b}^a {U_b}^a(s)$$

$$U^b(s) = w_b {U_b}^b(s) + {w_a}^b {U_a}^b(s),$$

which is reminiscent of the story of the two friends, one of whom sells his hair to buy a watchstrap gift for the other, who sells her watch to buy a hair brush for the first. Here both ethical utilities are dependent primarily on the other's, leading to a paradoxical regress in deciding whether they are happy or not. In theatre we see similar cases, where agents in a group often value each other's lives more than their own.

Further modelling complications arise from the relationship of utility to risk. While many current robotics systems assume that the total utility of an action is the average (expected) utility over possible future states, in the real world humans generally assign additional utility to actions which have the lowest outcome uncertainty *ceteris paribus*. In finance this is especially clear: equities tend to give bigger returns because those returns are less certain than those of bonds, and finance has a well developed theory of the utility of risk leading to models such as

$$(4) \quad U^a(\alpha, s_0) = \sum_{t=0}^{\infty} \sum_{st+1} p(s_{t+1}|s_t, \alpha) U^a(s_{t+1})$$

$$+ \lambda S[U^a(s_{t+1})]_{p(st+1|st, \alpha)},$$

where S is the standard deviation of the future utility and λ is a risk appetite parameter. In theatre there are similar preferences - if an action is highly likely to lead to one outcome, it is *ceteris paribus* preferable to another with uncertain outcomes, because the second would require twice as much work to go into planning for the future scenarios. So utility should be augmented with risk appetite λ to account for the cost of this work.

Similar to assigning utility by risk appetite is the issue of distributing utility across groups of agents. If an agent has to choose between actions leading to \$1000 of utility each for 1000 soldiers, or to \$1,000,000 for a single soldier[G], which should it chose? This is the classic problem of utilitarianism *vs*. egalitarianism in robotic decision form, and as with risk premia there is no obvious solution but again a parameter ρ similar to the price of risk could be set to model whatever trade-off is preferred by its creators,

$$(5) \quad {U_g}^a(s) = \sum_{i \in g} {U_i}^a(s) + \rho S({U_i}^a(s))_{p(i|g)}.$$

Where utilities of group (team, family, nation etc) are concerned, an agent also needs a function that assigns a degree of membership of individuals to these groups. Who exactly is a 'soldier', 'civilian', 'enemy combatant', 'Jew' or 'Taliban'? Do these categories depend on individuals' behaviour, clothing, contracts, beliefs or genetics? Some of these features can be measured in theatre, others cannot.

Computing non-ethical expected utilities $U^a(\alpha, s)$ is already computationally hard, but when these additional factors are included for ethical utilities it appears that all hope must be abandoned of computing exact consequentialist utilities, especially in real time such as in theatre. Hence the use of absolutist rules for ethical decision making is a much more practical approach to practical problems. Ethical utility-maximising robots must, like their human counterparts, rely on the centuries of wisdom that have been compiled into simple rules rather than trying to estimate and simulate their own and *others*' future utilities from scratch at every decision. [H]

The practical robot designer must therefore be concerned with *which* absolutist rules to implement. Should we build libertarian robots that maximise their non-ethical utilities but vote for free markets; national socialist or communist robots that kill anyone from other nations or classes and protect members of their own nation or class? Old Testament robots that 'Do not spare him, but kill men and women, children and infants' (1 Samuel 15:3)? New Testament robots that treat their enemies as their friends? None of these rules, nor Asimov's laws, seem appropriate for a modern military robot.

The standard criticism of absolutist systems, many of which contradict each other, is 'why these particular rules?' And the standard answer is 'because they work at this particular time and place'. Having arrived at the need for absolutist rules from the need to compile down complex utility computations into fast action selections, we can give a more precise answer, 'because they are, empirically, the most accurate computationally efficient approximations to optimal action selection under our desired utility function that we currently know'. Does this lead to relativism? No, arbitrary rules cannot be used. It is pragmatism: given a desired complex utility function, we search empirically for fast rules that can be used to find actions that are useful. The question of what goes into this given target utility is not one that concerns engineers. If they are designing for a company, it will be provided by the customers; if for a democratic nation, by the voters; and if for a dictatorship, by the dictator. In practice, we will find that human soldiers already have very well developed rules taught to them though classrooms, exercises and rituals that serve approximately this purpose and which could be implemented similarly in their robot counterparts. They might not be perfect or complete — neither are our laws of mathematics or physics — but they are the best we have.

5 Public ethics *vs*. private ethics

We may encounter scenarios where what an agent communicates *about* ethics is as important as its actual ethical rule set. Any agent that has to build trust with others - such as soldiers, eBay traders, and teachers - is concerned with other agents' perceptions of its own rule-set. It might be argued that this is what 'trust' in an agent means - being able to predict its future actions. This form of ethical perception is a central aspect of many virtue and duty absolute ethical systems. But it is not always the case that what an agent should (in the sense of maximising its utility) reveal about itself is identical to its actual rule-set. Machiavellian ethics recommends that an agent promotes altruistic ethics to its associates — so that they will include the agent's own utility in their utility functions - whilst simultaneously excluding any altruism from one's own private utility. Such strategies must often be balanced against the negative utility of being exposed - hypocrisy - and consequent loss of trust by others. Automating this type of 'second-order' ethical reasoning would be an additional research area after implementing ethical systems for individual agents. It is possible that it could draw from game-theoretic research such as automated Poker playing [11] which focuses on the inference and concealment of strategies between players. Taken to extremes, this kind of game playing between many members of whole societies is called 'politics'.

6 Conclusion

Robots with increasing levels of autonomy are progressing rapidly from labs to companies to active military service, and many autonomous robot engineers will have been involved in collaborations such as BAE Systems autonomous Wildcat [25] or QinetiQ's UAVs [10]. While most robots in critical real world situations still have a human 'in the loop', [15] it is becoming more important to automate more and more behaviour in the military and civilian worlds, for reasons of speed, accuracy and even cost efficiency. The humane dream of pure robot-on-robot warfare is still far away, and with both humans and robots now in theatre together we are rapidly reaching a point where some human decisions concerning ethics need to be automated.

We have examined how to formalise absolutist, rule-based ethical reasoning as a heuristic approximation to consequentialist utility maximisation as a step towards a framework for implementation on robotic agents. The latter is the ideal form of reasoning but is usually computationally intractable, especially for real-time split-second decisions in military theatre. An agent's utility function, and heuristic rules, can be viewed as ethical when they take the utilities of other agents into account. Ethical utility functions become even more complex and have even greater need for compilation into absolutist rules. Ethical utility functions have several parameters which must be specified. These include weights for immediate self-utility vs. utilities of other agents of various groups; models specifying degrees of membership of agents into such groups; models of other agents' own utility functions (which may be recursive); utilities of the distribution of other agents' utilities over possible outcomes (risk appetite) and over populations (egalitarianism).

We noted that the underlying utility function for an automated agent should be specified by its designer to reflect the utility of the customer or voter requesting and paying for the work. Customers and voters may have utilities that prioritise themselves, their families, communities, ethnic groups, nations, cultures or species, but the nature of their specifications is none of the business of a robot engineer[I]. These requirements may change over time, so for example, absolutist rules for military robots may need to be 'recompiled' from new utility functions if the commissioning government changes its policies or if a new government is elected. Rules may be searched for by specifying an explicit utility function, running many simulations or real experiences, and testing candidate heuristics on them against the utility (cf. [18]). Or, more simply, they could be obtained by asking humans who already know useful rules to share with robots and their designers.

References

[1] C. Rohlfs. The government's valuation of military life-saving in war: a cost-minimization approach. Papers and Proceedings of the American Economic Association:96(2):39-44, 2006.

[2] L. Carlone, Jingjing Du, M.K. Ng, B. Bona, M. Indri. An application of Kullback-Leibler divergence to active SLAM and exploration with Particle Filters. Intelligent Robots and Systems (IROS), 287 -293, 2010.

[3] D. Kahneman. Thinking Fast and Slow. Macmillan, 2011.

[4] D.H. Holding. The psychology of chess skill. Hillsdale, NJ: Lawrence Erlbaum Associates, 1985.

[5] F.M. Kamm. Harming Some to Save Others. Philosophical Studies, 57:227-60, 1989.

[6] C.W. Fox, M. Humphries, B. Mitchinson, T. Kiss, Z. Somogyvari, T.J. Prescott. Technical Integration of Hippocampus, Basal Ganglia and Physical Models for Spatial Navigation. Frontiers in Neuroinformatics, 3(6), 2009.

[7] G. Gigerenzer, P.M. Todd, ABC Group. Simple Heuristics That Make Us Smart. Oxford University Press, 2000.

[8] Geoffrey E. Hinton, Simon Osindero. A fast learning algorithm for deep belief nets. Neural Computation, 18:2006, 2006.

[9] J. M. Bernardo, A. F. M. Smith. Bayesian Theory. John Wiley & Sons, 2000.

[10] Graham S. Horn Jeremy W. Baxter. Fly-by-agent: Controlling a pool of UAVs via a multi-agent system. Knowledge-Based Systems:21:232-237, 2008.

[11] Kevin Korb, Ann Nicholson, Nathalie Jitnah. Bayesian Poker. Proc. Uncertainty in AI 343—350, 1999.

[12] Jean-Luc Gauvain Lori Lamel, Gilles Adda. PCA vs LDA. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23:228-2339, 2001.

[13] M. Taylor, C. Fox. Inventory management with dynamic Bayesian network software systems. Proc. Int. Conf. Business Information Systems, Springer LNBIP, 2011.

[14] N.D. Daw, Y. Niv, P. Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatum systems for behavioral control. Nature Neuroscience, 8:1704-1711, 2005.

[15] N.E. Sharkey. Automating Warfare: lessons learned from the drones. Journal of Law, Information and Science, 21(2), 2012.

[16] P. Dayan, G.E. Hinton, R.M. Neal, R.S. Zemel. The Helmholtz Machine. 1995.

[17] P. Glasserman. Monte Carlo Methods in Financial Engineering. Springer, 2003.

[18] R. Axelrod. The Evolution of Cooperation (Revised ed.). Perseus Books Group, 2006.

[19] Revonsuo A. The reinterpretation of dreams: an evolutionary hypothesis of the function of dreaming. Behav Brain Sci., 23(8):904-1121, 2000.

[20] Richard S. Sutton, Andrew G. Barto. Reinforcement Learning: An Introduction. IEEE Trans. Neural Networks, 9(5):1054-1054, 1998.

[21] Steven M LaValle. Planning Algorithms. 2004.

[22] Tommi Jaakkola, David Haussler. Exploiting Generative Models in Discriminative Classifiers. NIPS:487-493, 1998.

Notes

[A] 1. A robot may not injure a human being or, through inaction, allow a human being to come to harm. 2. A robot must obey the orders given to it by human beings, except where such orders would conflict with the First Law. 3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws. 0. A robot may not harm humanity, or, by inaction, allow humanity to come to harm.

[B] For conscious humans the duty/virtue distinction is arguably more important, as philosophers such as Confucius and Plato argued that the experiential qualia of being a good character is a pleasurable end in itself, but we assume that robots constructed with current technology do not experience qualia.

[C] Eventually lists of rules may become so long that it becomes useful to construct abstract theoretical concepts to unify them, as a memory aid. Again, it is not necessarily the case that such concepts must correspond to the actual underlying generative process, for example postulating Chi forces helps to unify many martial arts principles and allows new useful moves to be derived, even if there are no noumenal Chi forces involved in the actual generative process of fighting.

[D] Ritual sharing of food may be especially useful as eating food is a reward in itself, reinforcing the desired sharing aspect.

[E]Whilst is it true that benefits to others are merely side-effects of individual trade actions I such a system, the act of its members voting to enforce the existence of that system is an ethical action which does take account of such side effects. When voting for freedom, agents can and do take account of the utilities of others in their society by choosing that will benefit them in this way.

[F]The difficultly of estimating others' utility functions leads to a justification for libertarianism over central planning: that the state can never know the details of these different functions, e.g. When assuming that everyone wants a grey car.

[G]estimated by [1] as the implicit modern value placed on a WWII soldier's life by the US military.

[H] Evidence for human use of both and conflicting consequentialist and absolutist ethical systems is provided by some of the trolley problems of [5], which as in Daw and Dayan's [14] illustration of rats switching between computation and heuristic decisions, similarly show how human ethical judgements can be made in both ways, and try to determine in which situations each system tends to dominate. For example a cognitive illusion tends to occur in the case of asking a human whether to push a button that would kill one person if the absence of a press would kill five people. Here consequentialist ethics says to push the button but most people's absolutist rules say not to. It seems likely that as in Daw and Dayan, humans will switch to using absolutist rules more as the problem complexity increases, and/or the time allowed for decision making decreases. Such lab based scenarios are interestingprecisely because they are simple enough to allow full consequentialist reasoning, but in the real world most problems are too complex and so absolutist rules become the norm.

[I]While the author respects the right of individuals to vote for the interests of their ethnic and other groups, he also notes that history shows that those groups are usually best served in the long term by acting to help their neighbours too, and hopes that voters and their representatives are smart enough to take this into account in their decisions. He respects the right of individuals to choose whether or not to be professional engineers on projects conflicting with their personal utility functions. Such functions are personal beyond the scope of the profession and this paper.

Thanks to Anand Ramamoorthy and Andrew Bower for comments.