Kantian Ethics in the Age of Artificial Intelligence and Robotics

Ozlem Ulgen[*]

1. *Introduction*

Artificial intelligence and robotics is pervasive in daily life and set to expand to new levels potentially replacing human decision-making and action. Self-driving cars, home and healthcare robots, and autonomous weapons are some examples. A distinction appears to be emerging between potentially benevolent civilian uses of the technology (eg unmanned aerial vehicles delivering medicines), and potentially malevolent military uses (eg lethal autonomous weapons killing human combatants). Machine-mediated human interaction challenges the philosophical basis of human existence and ethical conduct. Aside from technical challenges of ensuring ethical conduct in artificial intelligence and robotics, there are moral questions about the desirability of replacing human functions and the human mind with such technology. How will artificial intelligence and robotics engage in moral reasoning in order to act ethically? Is there a need for a new set of moral rules? What happens to human interaction when it is mediated by technology? Should such technology be used to end human life? Who bears responsibility for wrongdoing or harmful conduct by artificial intelligence and robotics?

Whilst Kant may be familiar to international lawyers for setting restraints on the use of force and rules for perpetual peace, his foundational work on ethics provides an inclusive moral philosophy for assessing ethical conduct of individuals and states and, thus, is relevant to discussions on the use and development of artificial intelligence and robotics. His philosophy is inclusive because it incorporates justifications for morals and legitimate responses to immoral conduct, and applies to all human agents irrespective of whether they are wrongdoers, unlawful

[*] Visiting Fellow at the Lauterpacht Centre for International Law, University of Cambridge and Visiting Fellow at Wolfson College, Cambridge. Senior Lecturer in Law, School of Law, Birmingham City University, UK.

combatants, or unjust enemies. Humans are at the centre of rational thinking, action, and norm-creation so that the rationale for restraints on methods and means of warfare, for example, is based on preserving human dignity as well as ensuring conditions for perpetual peace among states. Unlike utilitarian arguments which favour use of autonomous weapons on the basis of cost-benefit reasoning or the potential to save lives, Kantian ethics establish non-consequentialist and deontological rules which are good in themselves to follow and not dependent on expediency or achieving a greater public good.

Kantian ethics make two distinct contributions to the debate. First, they provide a human-centric ethical framework whereby human existence and capacity are at the centre of a norm-creating moral philosophy guiding our understanding of moral conduct. Second, the ultimate aim of Kantian ethics is practical philosophy that is relevant and applicable to achieving moral conduct.

I will seek to address the moral questions outlined above by exploring how core elements of Kantian ethics relate to use of artificial intelligence and robotics in the civilian and military spheres. Section 2 sets out and examines core elements of Kantian ethics: the categorical imperative; autonomy of the will; rational beings and rational thinking capacity; and human dignity and humanity as an end in itself. Sections 3-7 consider how these core elements apply to artificial intelligence and robotics with discussion of fully autonomous and human-machine rule-generating approaches; types of moral reasoning; the difference between 'human will' and 'machine will'; and respecting human dignity.

2. *Core elements of Kantian ethics*

Kantian ethics provide a human-centric ethical framework placing human existence and capacity at the centre of a norm-creating philosophy that guides our understanding of moral conduct.[1] Kant's works may

[1] See generally I Kant, *The Moral Law: Kant's Groundwork of the Metaphysic of Morals* (HJ Paton tr, Hutchinson & Co 1969); I Kant, *The Metaphysics of Morals* (Mary Gregor tr and ed, CUP 1996); I Kant, 'Toward Perpetual Peace' in M Gregor (ed and tr), *Practical Philosophy* (CUP 1996); I Kant, *Critique of Pure Reason* (Paul Guyer and

be criticised for being dense and opaque, but his ultimate aim was practical philosophy that may lead to the creation of norms or rules capable of practical implementation. The following core elements of Kantian ethics, which establish the human-centric ethical framework, are examined: the categorical imperative; autonomy of the will; rational beings and rational thinking capacity; and human dignity and humanity as an end in itself.

2.1. *Moral rules capable of universalisation – the categorical imperative*

A fundamental starting point of Kantian ethics is Kant's 'categorical imperative' concept underpinning every moral judgment and determining moral duties. This helps us understand how morality of action in international society can be judged on the basis of underlying rules. The categorical imperative is a rule of intrinsic value[2] that is based on reason, prescribing objectives and constraints on conduct whether we desire these or not.[3] Kant's main formulation of the categorical imperative – 'act only on that maxim through which you can at the same time will that it should become a universal law'[4] – sets a test for identifying rules of intrinsic value: if a person acts according to a maxim believing it to be morally correct and desiring it to become universal law, the maxim may constitute a categorical imperative.[5] Another formulation is – 'act in such a way that you always treat humanity, whether in your own per-

Allen Wood trs, CUP 1998); I Kant, *Critique of the Power of Judgment* (Paul Guyer ed, Paul Guyer and Eric Matthews trs, CUP 2000).

[2] Kant expresses the intrinsic value in categorical imperatives as 'if the action is represented as *good in itself* and therefore as necessary' (emphasis added), Kant, *The Moral Law* (n 1) 78 para 414.

[3] Kant refers to an 'imperative' as a 'command of reason' which produces objective principles that any rational person would follow if reason had full control over their choice. A 'categorical' imperative has intrinsic value because it is not concerned with the matter of the action and its presumed results 'but with its form and with the principle from which it follows; and what is essentially good in the action consists in the mental disposition, let the consequences be what they may'. Kant, *The Moral Law* (n 1) 77 para 413, 78 para 414, 80 para 416.

[4] ibid 84 para 421.

[5] D McNaughton, P Rawling, 'Deontology' in David Copp (ed), *The Oxford Handbook of Ethical Theory* (OUP 2007) 436-437.

son or in the person of any other, never simply as a means, but always at the same time as an end'.[6] The categorical imperative sets objectives to create certainty of action, imposes constraints on subjective desires, and is capable of universalisation.[7] It is distinguishable from 'hypothetical imperatives' which are based on personal free will and subjective desires, without any constraints, and therefore not capable of universalisation.[8]

A rule capable of universalisation derives from human rational thinking capacity[9] and constrained free will[10] (more on these concepts later). It is fundamentally beneficial to humankind (good *qua* humankind),[11] and excludes rules of personal choice without wider appeal. This makes it more difficult for any individual or group to exert personal or particular beliefs disguised as moral rules. The rule needs to be inherently desirable, doable, and valuable for it to be capable of universalisation. O'Neill's principle of followability in relation to practical reasoning explains how a rule becomes 'doable'. First, the rule must be followable by others in thought; it must be intelligible to them. Second, the rule must be action guiding; it must also aim to recommend or pre-

[6] Kant, *The Moral Law* (n 1) 91 para 429.

[7] 'Universalisation' in this context means a rule that becomes morally permissible for everyone to act on. See Kant's references to 'universal law' and 'universal law of nature', Kant, *The Moral Law* (n 1) 83-86 paras 421-423. A modern version of universalisation is Rawl's 'original position' behind a 'veil of ignorance', J Rawls, *A Theory of Justice* (OUP 1999) 15-19.

[8] Kant, *The Moral Law* (n 1) 78-80 paras 414-417.

[9] For Kant only rational agents have the capacity to act according to their ideas of laws because they are able to set objectives and act on them, Kant, *The Moral Law* (n 1) 76 para 412, 99 para 437.

[10] The idea of constrained free will relates to rational agents being free to make decisions and take actions based on morals, which act as constraints on purely subjective or personal motives, Kant, *The Moral Law* (n 1) 93-94 paras 431-432, 107-109 paras 446-449.

[11] Note different conceptions of 'good' in utilitarianism (eg a right action is one that maximises the greater good) and Aristotelian virtue ethics (eg human good from a lifetime of virtuous achievements leading to human flourishing): J Bentham, *An Introduction to the Principles of Morals and Legislation* (first published [1781], Batoche Books 2000); J Stuart Mill, 'Utilitarianism' in M Warnock (ed), *Utilitarianism* (Fontana 1973); Aristotle, *Nicomachean Ethics* Book I (H Rackham tr, Harvard UP 2014) 33 para 16; TD Roche, 'Happiness and the External Goods' in Ronald Polansky (ed), *Cambridge Companion to Aristotle's Nicomachean Ethics* (CUP 2014) ch 3.

scribe action, to warn against or proscribe action.[12] Some examples of rules capable of universalisation may assist here.

There is a general duty not to harm others.[13] It is inherently desirable for humans not to be harmed in the normal course of interaction so that they can freely exist and function properly. It is inherently doable because, apart from exceptional circumstances of warfare, emergency, medical intervention, and self-defence, harm is not a necessary condition for human existence or fulfilment. The duty not to harm others is fundamentally beneficial to humankind, protecting our physical and mental well-being, and valuing our existence as rational beings with free will.

The duty not to steal is another rule. I believe that stealing is wrong and therefore I will not steal. Is my personal belief capable of universalisation as a rule? The answer must be yes. I may have personal reasons for not stealing, such as maintaining an honest reputation, but my belief is also fundamentally beneficial to humankind. If stealing were morally correct there would be lack of trust in human interaction, unpredictability of taking anything from anyone at any time, insecurity of ownership, and instability of not knowing what belongs to us.

Lying promises is an example of something that is not capable of universalisation.[14] I make a promise but I am lying when I do so because I have no intention of keeping the promise. Maybe that is part of my belief system. Maybe I believe it is morally correct. But is my belief capable of universalisation as a rule? The answer must be no. If everyone conducted their affairs by making promises then breaking them it would defeat the purpose of making promises in the first place. People would soon realize that promises are worthless and would not trust one

[12] O O'Neill, *Towards Justice and Virtue* (CUP 1996) 57.

[13] See generally, Kant, *The Moral Law* (n 1); Kant, *The Metaphysics* (n 1) 209-210 paras 6:462-6:464 on the general duty of respect towards others and prohibition of 'disgraceful punishments that dishonour humanity itself'; Rawls (n 7) 98-101 on 'natural duties' including the duty not to harm or injure others; A Linklater, *The Problem of Harm in World Politics: Theoretical Investigations* (CUP 2011) on the principle of humaneness and prevention of unnecessary harm in global ethics; A Linklater, 'Cosmopolitan Harm Conventions' in S Vertovec and R Cohen (eds), *Conceiving Cosmopolitanism: Theory, Context, and Practice* (OUP 2002) on duty to prevent harm in international society under 'cosmopolitan harm conventions'.

[14] See Kant, *The Moral Law* (n 1) 85 para 422 using the example of false promises to illustrate a maxim that cannot be universalised because it would necessarily destroy itself.

another, making it difficult to interact and enter into contractual agreements.

Some criticism may be levelled at the categorical imperative.[15] First, it seems an indeterminate and potentially chaotic means of creating rules if any self-determined personal conduct can be elevated to the status of a moral rule. It is unclear how wider acceptance or 'universalisation' of the rule is determined. Is it determined by the person acting according to a maxim they believe to be moral and capable of universalisation? Or does it require public affirmation and confirmation through conduct or some other means? Kant's main formulation of the categorical imperative is intended to create normativity in moral conduct so that it would not make sense to restrict the determination of universalisation to personal or private assessment. Thus, categorical imperative rules must be capable of being 'public and shareable'.[16] One's own reasoning and the reasoning of others are of equal importance in assessing whether a particular moral rule is capable of universalisation. In this sense Kant's categorical imperative is more restrictive and stringent than initially appears.

Second, Kant conceptualised morality on the basis of rational human beings; that the world consists of humans *capable* of acting rationally. Yet we know humans do not always act rationally, and one person's belief in the morality of their conduct does not necessarily extend to others. Closer scrutiny of Kant's works reveals an understanding and accommodation for the possibility of irrational conduct and wrongdoing, not detrimental to the human-centric ethical framework.[17] It is the *capacity* for rational conduct rather than *actual* rational conduct that enables rules capable of universalisation to emerge.

Third, Kant's human-centric approach appears to provide limited scope for establishing rules governing conduct towards non-human animals and inanimate objects (eg cultural heritage; property; personal possessions; the environment). But again, Kant makes reference to rules prohibiting wanton cruelty to animals, and wanton destruction of inan-

[15] See generally, I Kant, *Groundwork for the Metaphysics of Morals* (Thomas E Hill and Arnulf Zweig tr and eds, OUP 2002) 60-65.

[16] C Korsgaard, *The Sources of Normativity* (CUP 1996) 136.

[17] See eg Kant's acceptance of the dignity of wrongdoers and the need not to mistreatment or hold them in contempt, Kant, *The Metaphysics* (1996) (n 1) 105-109, 209-210.

imate objects during warfare (including infrastructure, municipal buildings, and housing).[18] Such rules derive from the categorical imperative to treat humanity, whether in your own person or in the person of any other, never simply as a means, but always at the same time as an end. Not being cruel to animals and not destroying inanimate objects upholds personal human dignity and, therefore, treats humanity as an end rather than a means to engage in personal desires. Although this is still a human-centric approach in that ethical conduct towards animals and inanimate objects is governed by the need to maintain human dignity, it does show Kant's appreciation of a wider perspective to potential beneficiaries of moral rules.

The categorical imperative explains *how* to identify rules capable of universalisation but does not explain *why* we should 'will' them to become 'universal law'. Kant develops concepts such as 'autonomy of the will', 'rational beings', and 'humanity as an end itself', to consider what motivates people to act and what may be the ultimate aim of moral action.

2.2. *Autonomy of the will*

Kant defines autonomy of the will as 'the property the will has of being a law to itself (independently of every property belonging to the object of volition)'.[19] Moral conduct and a sense of duty towards rules derive not from external factors (eg sanctions imposed by the state) but the human will acting autonomously to provide reason. Thus, autonomy of the will refers to 'the will of every rational being as a will which makes universal law'.[20] This may sound chaotic and advocating freedom for humans to do as they please but the autonomy principle is necessarily limited by the requirement that any rule chosen must be capable of universalisation. Kant explains this as 'never to choose except in such a way that in the same volition the maxims of your choice are also present as universal law'.[21] Such rule-making or universal law-making capacity is

[18] ibid 192-193.

[19] Kant, *The Moral Law* (n 1) 101 para 440.

[20] ibid 94 para 432.

[21] ibid 101 para 440.

independent from personal desires and inclinations.[22] But how is it possible to legislate without relying on desires and inclinations? Surely even an objective legislator at some point succumbs to personal inclinations? Kant considers personal motives as part of the deliberative process of a rational being's autonomy of the will; if these can be reconciled with rules that are capable of universalisation then they are permissible.[23]

Autonomy of the will is the motivator for making and abiding by moral rules that can lead to personal and wider societal freedom. Kant expresses this in terms of mutual reciprocity in rule-making and rule adherence: 'A rational being belongs to the kingdom of ends as a member, when, although he makes its universal laws, he is also himself subject to these laws'.[24] So there is an inner aspect relating to personal benefit, and an outer aspect concerning wider benefit to humankind. Personal self-worth and dignity are derived from participation in a rule-making process, and freely choosing to be bound by the rules created. Wider benefit to humankind derives from respecting the freedom of rational beings to exercise reason in making rules.

Conceptualising autonomy of the will as intrinsic to rational beings engaging in universal rule-making leads Kant to consider an ideal state of moral conduct in what he refers to as 'the kingdom of ends'. This is 'a systematic union of different rational beings under common laws'.[25] It is a misconception that Kant advocated for world government in contradiction to the autonomy of the will. He opposed the idea of world government referring to it as a 'soulless despotism' not in accord with reason or the freedom of states.[26] 'The kingdom of ends' at the individual and societal level concerns the emergence of a community of rational beings engaging in universal rule-making that ensures moral conduct. It is mirrored at the international level by Kant's idea of a 'federative union of states' agreeing to peaceful conduct of their affairs in order to secure freedom and rights.[27] It is an ideal that provides a reason for hav-

[22] ibid 96 para 434.

[23] See eg Kant's discussion of heteronomy of the will, ibid 102-106 paras 441-444.

[24] ibid 95 para 434.

[25] ibid 95 para 433.

[26] I Kant, 'Toward Perpetual Peace' in M Gregor (ed and tr), *Practical Philosophy* (CUP 1996) 8:367.

[27] ibid 8:356, 8:367, 8:385. For an excellent account of Kant's transitional approach to securing perpetual peace through a 'federative union of states' see C Corradetti,

ing morals and is intended to motivate and inspire moral conduct. This seems abstract and far removed from the reality of disagreements and differences that occur in rule-making. It does not address how differences may be resolved and assumes rational beings will adopt non-self-serving motives to create moral rules.[28] But abstraction, to an extent, is necessary to make sense of the elements that make up moral human conduct separate from purely expedient, personal, and selfish reasons. Abstraction enables forward-thinking through reasoning of what *ought* to constitute moral conduct rather than what actually is.[29] Autonomy of the will shows how humans have a capacity and disposition to choose and make moral rules, irrespective of the actual action taken, leading to a greater sense of individual and community freedom, dignity, and moral authority.[30]

2.3. *Rational beings and rational thinking capacity*

As we have already seen, the 'rational being' or 'rational agent' is the primary subject in Kant's analysis of how moral conduct emerges. 'Rational being' refers to the human capacity to understand and reason, which leads to action or conduct.[31] There are four defining features of Kantian rational agency: (i) capacity to understand and reason; (ii) capacity to set and be subject to universal moral rules; (iii) practical reasoning; and (iv) self-reflective and deliberative capacity. Rational beings have the capacity to understand and reason so they can set 'ends'; ob-

'Kant's Legacy and the Idea of a Transitional Jus Cosmopoliticum' (2016) 29 Ratio Juris 105–121.

[28] For detailed consideration of these criticisms see TE Hill, *Respect, Pluralism, and Justice: Kantian Perspectives* (OUP 2000) 33-56, 87-118 and 200-236; C Korsgaard, *Creating the Kingdom of Ends* (CUP 1999) 188-221.

[29] See Kant's distinction between the *noumenal world* (based on thinking about the world in terms of understanding, rationality, and freedom to explore possibilities), and the *phenomenal world* (based on knowing the world in so far as it is given to the senses), I Kant, *Critique of Pure Reason* (Paul Guyer and Allen Wood trs, CUP 1998) A235-260/B294-315.

[30] See A Reath, 'Autonomy of the Will as the Foundation of Morality' in A Reath, *Agency and Autonomy in Kant's Moral Theory: Selected Essays* (OUP 2006) ch 5.

[31] See A Reath, 'The Categorical Imperative and Kant's Conception of Practical Rationality' ch 3.

jectives or justifications and reasons for certain actions.[32] Reasons or justifications for moral rules can be referred to as the 'normativity' element of Kant's categorical imperative; if an act is morally wrong, then there is some genuinely normative reason not to do it.[33] Only rational beings with autonomy of the will may set universal rules and be subject to them.[34] As mentioned above, autonomy of the will relates to the capacity to freely act according to principles provided by reason. Reason is 'the faculty that provides the principles of cognition *a priori* [knowledge that is independent of any experience]'.[35] Rational beings will engage in universal rule-making independent of personal desires and inclinations. A rational being belongs to the 'intelligible world'[36] and the only way he can make sense of his own will is through the idea of freedom.[37] Practical reasoning enables setting ends that are moral commands for a person to raise himself from 'the crude state of his nature … and more and more toward humanity'.[38] Finally, rational beings have a self-reflective and deliberative capacity which enables consideration and choice of options before taking action.[39] The basis of this deliberative capacity is a sense of freedom; a rational being cannot make decisions to act without feeling they are free to make the choice:

> 'reason creates the idea of a spontaneity, which could start to act from itself, without needing to be preceded by any other cause that in turn determines it to action according to the law of causal connection'.[40]

According to Kant, judgment, the faculty of thinking the particular as contained under the universal (rule, principle, or law), is an 'inter-

[32] Kant (1969) (n 1) 90-91 paras 428-429, 99 para 438.

[33] S Darwall, 'Morality and Practical Reason: A Kantian Approach' in D Copp (ed), *The Oxford Handbook of Ethical Theory* (OUP 2007) 282.

[34] Kant, *The Moral Law* (n 1) 93-94 paras 431-432, 107-109 paras 446-448.

[35] Kant (n 29) A11/B24.

[36] See Kant's discussion of the *phenomenal world* (n 29).

[37] Kant, *The Moral Law* (n 1) 113 para 453.

[38] Kant, *The Metaphysics* (n 1) 151 para 6:387.

[39] See Korsgaard on the 'deliberative perspective' of Kantian rational agency (n 16) 100, 128-129.

[40] Kant (n 29) A533/B561.

mediary between understanding and reason'.[41] Judgment relates to human perceptual, social and interactional competencies that enable deciding whether something particular falls within a general rule.[42] Rational beings acquire knowledge by making 'analytic judgments', in which the predicate is contained in a concept of the subject, and 'synthetic judgments', in which the predicate is external to the subject and adds something new to our conception of it.[43] Both types of judgment are necessary. But synthetic judgments enable us to understand concepts such as freedom and autonomy of the will, without necessarily experiencing or having prior knowledge of these, and to formulate objective moral rules capable of universalisation.

The deliberative rational agency of humans has been referred to as a 'practical identity' which enables some normative conception of ourselves as 'something over and above all [personal] desires' who choose which desires to act on.[44] Deliberative and self-reflective qualities enable humans to make decisions, and create moral and legal norms. Such qualities form part of 'human central thinking activities' involving the ability to feel, think and evaluate, and the capacity to adhere to a value-based system in which violence is not the norm governing human relations.[45]

[41] I Kant, *Critique of the Power of Judgment* (Paul Guyer ed, Paul Guyer and Eric Matthews trs, CUP 2000) 64 para 5:177, 66 para 5:179.

[42] See eg L Suchman, *Plans and Situated Actions: The Problem of Human-Machine Communication* (Xerox Corporation 1985) studies perceptual, social and interactional competencies that are the basis for associated human activities, and how humans exercise judgment through self-direction that cannot be specified in a rule; J Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation* (WH Freeman & Company 1976) especially ch 8 refers to judgment as wisdom which only human beings possess because they have to 'confront genuine human problems in human terms'.

[43] Kant (n 29) A7/B11. Eg 'All bodies are extended' is an analytic judgment whereas 'All bodies are heavy' is a synthetic judgment because it requires additional thinking as to how 'heavy' relates to the concept of a body.

[44] Korsgaard (n 16) 100.

[45] O Ulgen, 'Human Dignity in an Age of Autonomous Weapons: Are We in Danger of Losing an 'Elementary Consideration of Humanity'?' (2016) 8(9) ESIL Conference Paper Series 1-19, 7-8 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2912002> (forthcoming in OUP edited collection – updated copy with author); O Ulgen, 'Autonomous UAV and Removal of Human Central Thinking Activities: Implications for Legitimate Targeting, Proportionality, and Unnecessary Suffering' (forthcoming) 1-45.

2.4. *Human dignity and humanity as an end in itself*

Kant's moral theory encourages a transcendent value-based ethics through the idea that rational beings should act in a way that treats humanity as an end in itself.[46] Humanity as an end in itself is expressed in Kant's maxim, 'act in such a way that you always treat humanity, whether in your own person or in the person of any other, never simply as a means but always at the same time as an end'.[47] What does it mean to treat humanity as an end in itself? It is based on Kant's human-centric approach and essentially relates to recognising and upholding the status and value of human dignity. Not treating humans as mere means to ends is the Kantian notion of human dignity. It is the basis to all moral conduct and the means by which humans represent objective rather than relative ends. Relative ends are values based on personal desires, wants, hopes, and ambitions. They are easily replaced and replaceable. Objective ends, however, cannot be replaced with an equivalent. They are reasons for morals governing human conduct which are capable of universalisation and valid for all rational beings. If humans are objective ends they cannot be replaced by relative ends, which are transitory and subjective, and based on wants, hopes, and desires. Objective ends are superior because they possess a particular moral value; dignity.[48]

Human dignity is a pervasive idea in international human rights law and many constitutions, sometimes expressed as a right in itself or as a moral value informing other substantive rights.[49] It is given expression

[46] See eg Kant, *The Moral Law* (n 1) 90-93 paras 427-430.

[47] ibid 91 para 429.

[48] ibid 90-91 paras 428-429.

[49] 1948 Universal Declaration of Human Rights, Preamble and arts 1, 22, 23(3); 1966 International Covenant on Civil and Political Rights, art 10; 1966 International Covenant on Economic, Social and Cultural Rights, art 13; 1965 Convention on the Elimination of All Forms of Racial Discrimination, 1979 Convention on the Elimination of All Forms of Discrimination Against Women; 1984 Convention Against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment, Preamble; 1989 Convention on the Rights of the Child, Preamble and arts 23, 28, 37, 39, 40; 2006 Convention on the Rights of Persons with Disabilities, Preamble and arts 1, 3(a), 8(1)(a), 16(4), 24(1)(a), 25(d); 1978 Spanish Constitution, s 10(1); 1949 German Basic Law, arts 1(1) (as a duty), 79(3) (amendment to the duty is inadmissible); 1996 South African Constitution, s 1 (as a constitutional value), s 10 (as a right); see also P Carozza,

as an elementary consideration of humanity in the Martens Clause; a fundamental principle of customary international law protecting civilians and combatants in all circumstances not regulated by international law.[50] Human dignity as a transcendent value-based ethic is an important basis for deontological thinking as it enables conceptualisation of morality beyond the individual, group or nation-state to encompass the wider world. It means that rather than devising outcome-based rules, the rule-making process is prospective and aspirational in terms of what humanity should aim for.

Treating humanity as an end in itself in the Kantian sense means recognising rational beings have intrinsic worth and a self-determining capacity to decide whether or not to do something. They are not mere objects or things to be manipulated, used or discarded on the basis of relative ends (eg personal wants, desires, hopes, and ambitions). So humanity's intrinsic value is not dependent on personal characteristics.[51] Human dignity gives a person a reason for doing or not doing something. That reason takes precedence over all others. It means setting moral and rational limits to the way we treat people in pursuit of relative ends.[52]

Having considered core elements of Kantian ethics, sections 3-7 explore how each of these apply to artificial intelligence and robotics.

3. *Moral rules capable of universalisation in relation to artificial intelligence and robotics*

How will artificial intelligence and robotics implement Kant's main formulation of the categorical imperative – 'act only on that maxim

'Human dignity in constitutional adjudication' in T Ginsberg, R Dixon (eds), *Comparative Constitutional Law* (Edward Elgar 2011).

[50] Additional Protocol I to the Geneva Conventions of 12 August 1949 (1977) ('API'), art 1(2); Hague Convention respecting the Laws and Customs of War on Land (1907) ('Hague Convention IV') Preamble; Hague Convention with respect to the Laws and Customs of War on Land (1899) ('Hague Convention II').

[51] See TE Hill, 'In Defence of Human Dignity: Comments on Kant and Rosen' in C McCrudden (ed), *Understanding Human Dignity* (OUP 2014) 316.

[52] For elaboration of Kant's humanity principle as an objective end representing human dignity see TE Hill, *Dignity and Practical Reason in Kant's Moral Theory* (Cornell UP 1992) 43-44.

through which you can at the same time will that it should become a universal law'? Artificial intelligence and robotics do not possess human rational thinking capacity or a free will to be able to understand what constitutes a rule that is inherently desirable, doable, and valuable for it to be capable of universalisation. But there is human agency in the design, development, testing, and deployment of such technology so that responsibility for implementing the categorical imperative resides with humans. Humans determine which rules are programmed into the technology to ensure ethical use and moral conduct. For these rules to be capable of universalisation they must be 'public and shareable'. In the civilian sphere, for example, there is much debate about open access and use of artificial intelligence to gather personal data, potentially compromising privacy.[53] In the military sphere, discussions on lethal autonomous weapons under the auspices of the UN Convention on Certain Conventional Weapons represent a process for universalisation of rules which may regulate or ban such weapons. Indeed, there is emerging *opinio juris* among some states for a preventative prohibition rule, and a majority of states recognise that any rules regulating lethal autonomous weapons must take account of ethical, legal, and humanitarian considerations.[54] The potentially broad purposes and uses of artificial intelligence and robotics technology may lead to competing rules emerging which may or may not be capable of universalisation. Some preliminary issues related to the nature and type of rules are considered here.

How will rules be generated to regulate ethical use and operation of the technology? This depends on whether the technology is intended to completely replace human functions and rational thinking or to complement and supplement such human characteristics. Fully autonomous technology refers to artificial intelligence and robotics replacing human rational thinking capacity and free will so that rules emerge from the technology itself rather than humans. Human-machine integrated technology, on the other hand, refers to technology that supports and assists

[53] See Ipsos Mori Report, 'Public views of machine learning: findings from public research and engagement conducted on behalf of the Royal Society' (April 2017) 48-51.

[54] For discussion of emerging *opinio juris* on lethal autonomous weapons see O Ulgen, '"World Community Interest" approach to interim measures on "robot weapons": revisiting the *Nuclear Test Cases*' (2016) 14 New Zealand YB Intl L (forthcoming) s III.A.

humans in certain circumstances so that rules are created, influenced, controlled, and tailored by a combination of human and machine interaction and intervention. Both kinds of rule-generating approaches have ethical implications.

3.1. *Fully autonomous rule-generating approach*

A fully autonomous rule-generating approach would mean the technology produces its own rules and conduct without reference to or intervention from humans. After the initial design and programming by humans, the technology makes its own decisions. This is 'machine learning' or 'dynamic learning systems' whereby the machine relies on its own databank and experiences to generate future rules and conduct.[55] Fully autonomous weapons systems, for example, would have independent thinking capacity as regards acquiring, tracking, selecting, and attacking human targets in warfare based on previous experience of military scenarios.[56] Such an approach presents challenges. There is uncertainty and unpredictability in the rules that a fully autonomous weapons system would generate beyond what it has been designed to do, so that it would not comply with international humanitarian law or Kantian ethics. In the civilian sphere, fully autonomous technology may generate rules that adversely impact on human self-worth and progress by causing human redundancies, unemployment, and income instability and inequality. Adverse impact on human self-worth and progress, and uncertainty and unpredictability in the rule-generating process are contrary to what is fundamentally beneficial to humankind; such a process cannot produce rules that are inherently desirable, doable, valuable, and capable of universalisation. A perverse 'machine subjectivity' or

[55] See P Asaro, 'Roberto Cordeschi on Cybernetics and Autonomous Weapons: Reflections and Responses' (2015) 3 Paradigmi. Rivista di critica filosofica 83-107, 96-98; MJ Embrechts, F Rossi, F-M Schleif, JA Lee, 'Advances in artificial neural networks, machine learning,and computational intelligence' (2014) 141 Neurocomputing 1-2.

[56] See 'Report of the ICRC Expert Meeting, Autonomous Weapon Systems: Technical, Military, Legal and Humanitarian Aspects' (9 May 2014) ('2014 ICRC Report'); 'Report of the ICRC Expert Meeting, Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons' (15-16 March 2016) ('2016 ICRC Report'); Ulgen (forthcoming) (n 45).

'machine free will' would exist without any constraints, similar to Kant's 'hypothetical imperatives' formed by human subjective desires.

3.2. *Human-machine rule-generating approach*

A human-machine rule-generating approach currently exists in both the civilian and military spheres. IBM, for example, prefers the term 'augmented intelligence' rather than artificial intelligence because this better reflects their aim to build systems that enhance and scale human expertise and skills rather than replace them.[57] The technology is focused on practical applications that assist people in performing well-defined tasks (eg robots that clean houses; robots working with humans in production chains; warehouse robots that take care of the tasks of an entire warehouse; companion robots that entertain, talk, and help elderly people maintain contact with friends, relatives, and doctors). In the military sphere, remotely controlled and semi-autonomous weapons combine human action with weapons technology. Human intervention is necessary to determine when it is appropriate to carry out an attack command or to activate an abort mechanism. This kind of rule-generating approach keeps the human at the centre of decision-making. But what happens if there are interface problems between the human and machine (eg errors; performance failures; breakdown of communication; loss of communication link; mis-coordination)?[58] This may prove fatal in human-weapon integrated systems reliant on communication and co-ordination, and a back-up system would need to be in place to suspend or abort operations. What happens if the technology is hacked to produce alternative or random rules that cause malfunction, non-performance, or harmful effects? The same problem applies to fully autonomous technology and seems a good reason for restricting use and performance capability to set tasks, controlled scenarios or environments where any potential harm is containable.

[57] F Rossi, 'Artificial Intelligence: Potential Benefits and Ethical Considerations', Briefing Paper to the European Union Parliament Policy Department C: Citizens' Rights and Constitutional Affairs European Parliament (October 2016) <www.europarl.europa.eu/RegData/etudes/BRIE/2016/571380/IPOL_BRI(2016)571380_EN.pdf>.

[58] Asaro (n 55) 90-91.

The potential exclusion of non-human animals and inanimate objects from Kant's human-centric approach to the categorical imperative may directly apply to formulating ethical conduct in artificial intelligence and robots. If there is concern about machine unpredictability and uncertainty in generating its own rules, human intervention can set the categorical imperative as 'the technology must always prioritise human life over damage to property or non-human animal life'. This human-centric approach is already being trialled in self-driving cars with the German Government recently approving ethical guidelines for autonomous vehicles requiring that:

> 'the protection of human life enjoys top priority in a balancing of legally protected interests. Thus, within the constraints of what is technologically feasible, the systems must be programmed to accept damage to animals or property in a conflict if this means that personal injury can be prevented.'[59]

4. *Difference between 'human will' and 'machine will'*

Kant's autonomy of the will is hard to transpose into technology because it is reliant on concepts such as self-worth, dignity, freedom, rule-making capacity, and interaction. A machine would not have a sense of these concepts or be able to attach value to them. "Human will" develops through character and experience to inform moral conduct. 'Machine learning' or 'dynamic learning systems' that generate rules and conduct based on a databank of previous experiences may resemble a form of 'machine will' that makes ethical choices based on internally learned rules of behaviour.[60] But the human will is much more dynamic, elusive, and able to cope with spontaneity in reaction to novel situations

[59] Rule 7, see Federal Ministry of Transport and Digital Infrastructure, 'Ethics Commission: Automated and Connected Driving' (June 2017) 11 <www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission.pdf?__blob=publicationFile>.

[60] M O Riedl, 'Computational Narrative Intelligence: A Human-Centered Goal for Artificial Intelligence' (2016) CHI'16 Workshop on Human-Centered Machine Learning, 8 May 2016, San Jose, California, USA; M O Riedl, B Harrison, 'Using Stories to Teach Human Values to Artificial Agents' (2015) Association for the Advancement of Artificial Intelligence.

which sit outside rule-based behavioural action and derive from human experience and intuition.

Autonomy of the will requires inner and outer development of the person to reach a state of moral standing and be able to engage in moral conduct. This is suggestive of an innate sense of right and wrong.[61] Can machines emulate this sort of 'will'? Artificial intelligence in autonomous weapons may allow machine logic to develop over time to identify correct and incorrect action, showing a limited sense of autonomy. But the machine does not possess a 'will' of its own nor does it understand what freedom is and how to go about attaining it by adopting principles that will develop inner and outer autonomy of the will. It has no self-determining capacity that can make choices between varying degrees of right and wrong. The human can decide to question or go against the rules but the machine cannot, except in circumstances of malfunction and mis-programming. It has no conception of freedom and how this could be enhanced for itself as well as humans. The machine will not be burdened by moral dilemmas so the deliberative and reflective part of decision-making (vital for understanding consequences of actions and ensuring proportionate responses) is completely absent. There is a limited sense in which artificial intelligence and robotics may mimic the outer aspect of Kant's autonomy of the will. Robots may have a common code of interaction to promote cooperation and avoid conflict among themselves. Autonomous weapons operating in swarms may develop principles that govern how they interact and coordinate action to avoid collision and errors. But these are examples of functional, machine-to-machine interaction that do not extend to human interaction, and so do not represent a form of autonomy of the will that is capable of universalisation.

When we talk about trust in the context of using artificial intelligence and robotics what we actually mean is reliability. Trust relates to claims and actions people make and is not an abstract thing.[62] Machines without autonomy of the will, in the Kantian sense, and without an ability to make claims cannot be attributed with trust. Algorithms cannot determine whether something is trustworthy or not. So trust is used

[61] See commentary on Kantian human will as related to a capacity to make things happen, intentionally and for reasons, unlike robots (n 15) 94.

[62] O O'Neill, *Autonomy and Trust in Bioethics* (CUP 2002)

metaphorically to denote functional reliability; that the machine performs tasks for the set purpose without error or minimal error that is acceptable. But there is also an extension of this notion of trust connected to human agency in the development and uses to which artificial intelligence and robotics are put. Can we trust humans involved in developing such technologies that they will do so with ethical considerations in mind – ie limiting unnecessary suffering and harm to humans, not violating fundamental human rights? Once the technology is developed, can we trust those who will make use of it to do so for benevolent rather than malevolent purposes? These questions often surface in debates on data protection and the right to privacy in relation to personal data trawling activities of technologies. Again, this goes back to what values will be installed that reflect ethical conduct and allow the technology to distinguish right from wrong.

5. *Kantian notion of rational beings and artificial intelligence*

Kant's focus on the rational thinking capacity of humans relates to potential rather than actual possession of rationality, taking account of deficient rationality, immoral conduct, and situations where humans may deliberately act irrationally in order to gain some advantage over an opponent. Technology may be deemed to have rational thinking capacity if it engages in a pattern of logical thinking from which it rationalises and takes action. Although Kant's concept is specifically reserved for humans who can set up a system of rules governing moral conduct (a purely human endeavour and not one that can be mechanically produced), the capacity aspect may be fulfilled by artificial intelligence and robotics' *potential* rather than *actual* rational thinking. But this seems a low threshold raising concerns about predictability and certainty of the technology in real-life scenarios. So there would need to be much greater clarity and certainty about what sort of rationality the technology would possess and how it would apply in human scenarios.

When we compare machines to humans there is a clear difference between the logic of a calculating machine and the wisdom of human judgment.[63] Machines perform cost effective and speedy peripheral processing activities based on quantitative analysis, repetitive actions, and sorting data (eg mine clearance; and detection of improvised explosive devices). They are good at automatic reasoning and can outperform humans in such activities. But they lack the deliberative and sentient aspects of human reasoning necessary in human scenarios where artificial intelligence may be used. They do not possess complex cognitive ability to appraise a given situation, exercise judgment, and refrain from taking action or limit harm. Unlike humans who can pull back at the last minute or choose a workable alternative, robots have no instinctive or intuitive ability to do the same. For example, during warfare the use of discretion is important to implementing rules on preventing unnecessary suffering, taking precautionary measures, and assessing proportionality. Such discretion is absent in robots.[64]

6. *Universal and particular moral reasoning in artificial intelligence and robotics*

How will artificial intelligence and robotics engage in moral reasoning in order to act ethically? Should the technology possess universal or particular moral reasoning? In ethical theory 'universality' of moral reasoning means in any situation where an agent morally ought to do something, there is a reason for doing so.[65] Kant's categorical imperative

[63] Weizenbaum (n 42), critically discusses the limitations of computer-based logical thinking after he developed the ELIZA computer programme to mimic the behaviour of a psychoanalyst; argues that computer intelligence is 'alien to genuine human problems and concerns' at 213, and that 'there is an aspect to the human mind, the unconscious, that cannot be explained by the information-processing primitives, the elementary information processes, which we associate with formal thinking, calculation, and systematic rationality' at 223.

[64] E Lieblich and E Benvenisti, 'The obligation to exercise discretion in warfare: why autonomous weapons systems are unlawful' in N Bhuta, S Beck, R Geiss, H-Y Liu, C Kress (eds), *Autonomous Weapons Systems Law, Ethics, Policy* (CUP 2016) argue that autonomous weapons systems violate the duty to exercise discretion under international humanitarian law because they have pre-determined decision-making capability which does not respect the individual by considering their case/position carefully and exercising discretion where necessary.

[65] Darwall (n 33) 286.

makes it clear that it is a specific type of reason; one based on a rule capable of universalisation. In contrast, ‘particular’ moral reasoning does not rely on universal rules to justify moral obligations and reasons for actions, instead looking for analogous situations from which rules emerge.[66] Would universal moral reasoning in artificial intelligence include reference to all particular instances requiring particular moral reasoning?

Ongoing developments in the civilian and military spheres highlight moral dilemmas and the importance of human moral reasoning to mediate between competing societal interests and values. Companion robots may need to be mindful of privacy and security issues (eg protection and disclosure of personal data; strangers who may pose a threat to the person’s property, physical and mental integrity) related to assisting their human companion and interacting with third parties (eg hospitals; banks; public authorities). Companion robots may need to be designed so that they do not have complete control over their human companion’s life which undermines human dignity, autonomy, and privacy. Robots in general may need to lack the ability to deceive and manipulate humans so that human rational thinking and free will remain. Then there is the issue of whether fully autonomous weapons should be developed to replace human combatants in the lethal force decision-making process to kill another human being. Is there a universal moral reasoning that the technology could possess to solve such dilemmas? Or would it have to possess a particular moral reasoning, specific to the technology or scenario?

Human moral reasoning involves a combination of comprehension, judgment, experience, and emotions. It may also be dependent on societal, cultural, political, and religious factors. Arguably, the 1948 Universal Declaration of Human Rights provides a common standard of universal moral reasoning in setting out general human rights that are deemed universal, indivisible, and inviolable.[67] Particular moral reasoning may seek to limit factors relevant to reasoning based on the technology’s capability or the scenario in which it is used. For example, an autonomous weapon that is capable only of targeting and destroying

[66] See eg J Dancy, *Moral Reasons* (Blackwell 1993); B Hooker, MO Little (eds), *Moral Particularism* (OUP 2000).

[67] Universal Declaration of Human Rights (United Nations) UN Doc A/810, 71, UN Doc A/RES/217(III) A, GAOR 3rd Session Part I, 71.

buildings will not have to consider factors relating to the location, appearance, intentions, or activities of a human combatant. On the other hand, if the weapon is employed in uncomplicated and non-mixed areas and is capable of human targeting, it would have to engage in moral reasoning that complies with the principles of distinction, proportionality, and unnecessary suffering.[68]

Machine moral reasoning, however, may or may not be able to interpret the relative significance and value of certain human rights which could lead to arbitrary and inconsistent application. It may be designed to use bias or cultural preferences in reaching moral decisions (eg favouring or not favouring different categories of rights such as 'first generation' rights to liberty and security, fair trial, privacy, freedom of political association and assembly; 'second generation' rights to housing, water and sanitation, education, and economic development; or the 'third generation' right to protection of the environment). One way to overcome this is to design the technology to be value-neutral in identifying human lives so that it is not based on cultural, racial, gender, or religious biases. An example is the German government's new ethical guidelines for autonomous vehicles which states that 'in the event of unavoidable accident situations, any distinction based on personal features (age, gender, physical or mental constitution) is strictly prohibited'.[69] Or could the universal moral reasoning be identified from *jus cogens* norms and obligations *erga omnes*? Perhaps there is universal moral reasoning under 'third generation' rights, subject to some modification to take account of recent developments under the principle of humanity to broaden the content of such rights to include prevention of war, prevention of harm and violence, and protection against unnecessary suffering.[70]

7. *Can artificial intelligence and robotics respect human dignity and humanity as an end in itself?*

Human dignity is accorded by recognising the rational capacity and free will of individuals to be bound by moral rules, as well as through

[68] Arts 35(2), 48, 51, 57 API (n 50).

[69] Rule 9 (n 59).

[70] See generally Ulgen (n 54).

notions of accountability and responsibility for wrongdoing.[71] We accept that when wrongdoing is committed someone needs to be held accountable and responsible. How can artificial intelligence express person-to-person accountability and fulfil this aspect of human dignity (ie accountability for wrongdoing means respecting moral agents as equal members of the moral community)? Could we ever accept artificial intelligence as equal members? There is also the matter of whether artificial intelligence and robotics will be able to treat humanity as an end in the Kantian sense.

In the military sphere the use of lethal autonomous weapons are arguably used for a relative end (ie the *desire* to eliminate a human target in the *hope* of preventing harm to others). For Kant, relative ends are lesser values capable of being replaced by an equivalent. Killing a human being in the *hope* that it will prevent further harm is insufficiently morally grounded to override human dignity and may be reckless if alternatives and consequences are not considered. Utilitarians may counter that balancing interests involves consideration of the greater good, which in this instance is to prevent harm to others.[72] Consequentialist thinking and the utilitarian calculus are reflected in the proportionality principle under art 51 API, requiring assessment of whether an attack is expected to cause excessive incidental loss of civilian life in relation to the concrete and direct military advantage anticipated. But utilitarianism cannot overcome the problem of applying a quantitative assessment of life for prospective greater good that treats the humans sacrificed as mere objects, and creates a hierarchy of human dignity. Unless autonomous weapons can only be used to track and identify rather than eliminate a human target, they would extinguish a priceless and irreplaceable objective end possessed by all rational beings; human dignity.[73]

Using autonomous weapons to extinguish life removes the reason for having morals in the first place; human dignity of rational beings with autonomy of the will. In doing so a relative end is given priority over an objective end.[74] Lack of face-to-face killing creates a hierarchy

[71] See Darwall (n 33) 291 discussing morality as mutual accountability.

[72] Bentham (n 11).

[73] See generally Ulgen (2016) (n 45).

[74] C Heyns, 'Autonomous weapons systems and human rights law' (Presentation made at the informal expert meeting organised by the state parties to the Convention on

of human dignity. Military personnel, remote pilots, commanders, programmers, and engineers are immune from rational and ethical decision-making to kill another human being and do not witness the consequences. By replacing the human combatant with a machine the combatant's human dignity is not only preserved but elevated above the human target. This can also be seen as a relative end in that it selfishly protects your own combatants from harm at all costs including violating the fundamental principle of humanity as an objective end.[75]

8. *Conclusion*

Kantian ethics provide a human-centric approach to formulating moral rules. The central elements of Kantian ethics lead towards a focus on human self-determining capacity for rule-making and rule adherence. These elements illustrate the essential ways in which human attributes and capabilities, such as practical reasoning, exercising judgment, self-reflection and deliberation allow for the formation of moral rules that are capable of universalisation. Such human attributes and capabilities are non-existent in artificial intelligence and robotics so that human agency must be at the forefront of designing and taking responsibility for their ultimate conduct and action. A limited sense of rational thinking capacity can be programmed in the machine but it will not have the self-reflective and deliberative human capacities, as developed under the Kantian notion of rational beings, so that the machine will not be able to assess a given situation and exercise discretion in choosing a particular action or not. In closed scenarios where the technology is used for defined tasks, as seen in the civilian sphere, this limited rational thinking capacity may be appropriate as it will not be necessary to exercise discretion.

Whether rules can be created to meet the Kantian categorical imperative standard depends on whether there is a fully autonomous rule-

Certain Conventional Weapons 13–16 May 2014, Geneva, Switzerland) 8; C Heyns, 'Autonomous weapons systems: living a dignified life and dying a dignified death' in Bhuta (et al) (n 64).

[75] See Hill (n 52) ch 10 considering whether Kantian human dignity allows for this sort of hierarchy in relation to terrorists and hostage situations.

generating or human-machine rule-generating approach. Both raise ethical concerns in terms of who ultimately decides on the rules that will govern ethical conduct and whether this is sufficiently controllable and alterable in case of malfunction or detrimental harm. More complicated scenarios involving open-ended tasks with machine learning or dynamic learning systems used to generate rules raise concerns about uncertainty and unpredictability. Such a process would not be fundamentally beneficial to humankind as it cannot produce rules that are inherently desirable, doable, valuable, and capable of universalisation. There is also a limited sense in which the technology can actually be deemed to have a 'will' of its own; certainly not in the Kantian sense of autonomy of the will but perhaps a 'machine will' that has the capacity to set rules and abide by them. This limits the rule-making capacity to machine-to-machine interaction to the exclusion of human ethical concerns.