Sci Eng Ethics (2015) 21:317–326
DOI 10.1007/s11948-014-9541-0

ORIGINAL PAPER

Artificial Intelligence and Robot Responsibilities: Innovating Beyond Rights

Hutan Ashrafian

Received: 22 January 2014 / Accepted: 25 March 2014 / Published online: 16 April 2014

Abstract The enduring innovations in artificial intelligence and robotics offer the promised capacity of computer consciousness, sentience and rationality. The development of these advanced technologies have been considered to merit rights, however these can only be ascribed in the context of commensurate responsibilities and duties. This represents the discernable next-step for evolution in this field. Addressing these needs requires attention to the philosophical perspectives of moral responsibility for artificial intelligence and robotics. A contrast to the moral status of animals may be considered. At a practical level, the attainment of responsibilities by artificial intelligence and robots can benefit from the established responsibilities and duties of human society, as their subsistence exists within this domain. These responsibilities can be further interpreted and crystalized through legal principles, many of which have been conserved from ancient Roman law. The ultimate and unified goal of stipulating these responsibilities resides through the advancement of mankind and the enduring preservation of the core tenets of humanity.

Keywords Artificial intelligence · Robot · Responsibility · Rights

Exemplum Moralem

An international political conflict based on a border dispute resulted in a war between two nations. Both sides employed robots and artificial intelligent technology in warfare.

Toward the end of a military campaign, the dominant side gained a substantial opportunity to capture a strategically critical area of land whose attainment could lead to a unilateral victory and likely end the conflict.

H. Ashrafian (✉)
Imperial College London, 10th Floor QEQM-Building, Praed Street, London W2 1NY, UK
e-mail: h.ashrafian@imperial.ac.uk

During the battle to secure the strategic area, the dominant side's frontal robotic battalion successfully cleared all opposing combat robots from the rival side according to their pre-determined battle plan.

These robots however encountered projectile-based attacks from a group of indigenous children. In accordance with all rules and laws for robots and artificial intelligence technologies, the robots totally refrained from engaging in combat in any way with human civilians.

During their attack on the robots, some of the children sustained serious injuries. Although the robots and children were based on opposing sides, the robots attend to the injured children, utilizing their individual resources to treat the children's injuries. During the process of offering medical care to the injured child civilians from the opposing nation, the robots lost their dominance of the strategic area and the war continued.

Introduction

The continued advances in computer science, engineering and robotics have led to a rapid development of enhanced computability offering superior artificial intelligence and robotics. In due course these may herald the possibility of near-human, comparable-to-human and even beyond-human capability that requires an increased fidelity in appraisal of these technologies (Ashrafian et al. 2014). The prospect of sentient, rational and self-conscious artificial intelligence agents has led to the conceptual consideration of robot and artificially intelligent rights and laws that consider human societal and artificial intelligence agent relationships, as well as the relationships between artificial intelligent agents themselves (Ashrafian 2014).

Rights in human society are counterbalanced by the need for commensurate responsibilities and duties. Consequently, the idea of artificial intelligence and robot rights necessitates a matching level of societal responsibility and duty. Addressing this issue represents the discernable next-step for evolution in this field. This manuscript discusses the philosophical and practical considerations for artificial intelligence and robotic responsibilities and identifies the societal and moral considerations for these agents beyond that of rights, addressing translational concepts including the sophistication of robots and artificial intelligences.

When considering the moral responsibilities of the robots in the case, some basic questions arise: (1) Should robots have helped the injured children on moral grounds? Even if this meant that the war could be delayed or even potentially lost; with possibly many more deaths from both warring sides? (2) At a broader level what is the moral responsibility of self-conscious, rational and sentient artificial intelligence?

Moral Responsibility for Artificial Intelligence

In order to appraise the responsibilities of artificial intelligences and robots, there are several core philosophical positions that require consideration (Fig. 1). These

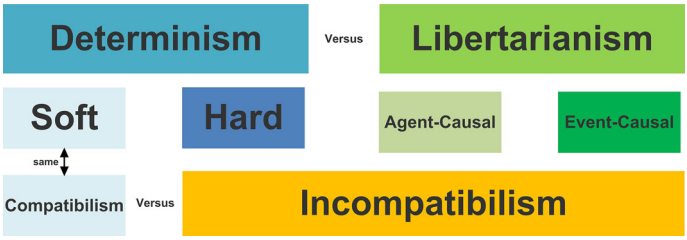

Fig. 1 Philosophical perspectives of moral responsibility and free will

include the contemplation of moral responsibility through Determinism and Libertarianism. Determinism represents that all events are pre-'determined' and therefore negate the concept of free will so that individual choice and therefore responsibility are disavowed. Conversely libertarianism denotes that individuals have moral responsibility derived from their innate free will in making personal decisions. Determinism can be divided into soft and hard. Soft determinism or Compatibilism represents a 'middle road' supporting the role of moral responsibility where decisions are made by free will within a context of determinism. Compatibilism in turn is differentiated from Incompatibilism, which consists of Libertarianism (Agent-Causal and Event-Causal) and Hard-Determinism and rejects any element of free will or choice due to the explanation that all events (and therefore individual decisions and actions) are totally pre-determined by the laws of nature. Currently, the most favored school of philosophy when considering moral responsibility is compatibilism.

The determinist school classically describes 'humans as robots or automatons' as their actions are fundamentally predetermined by natural laws. Consequently the hard determinist view of the case above where the robots assist the children could only be explained through the fact that the robots could demonstrate free will or responsibility to help the children unless they had been pre-programmed to specifically do so. Conversely the libertarian view would be that the robots were rational and sentient beings with free will, and as a result of personal moral responsibility went to the aid of the children despite knowing that through these actions they would lose their strategic position in the war. The compatibilist view would be one where each robot was constructed to achieve free will and decision-making capability through programmed rationality and sentience. Thus whilst they had been preset to carry out a military task; their free will and their moral responsibility led to their prioritizing the health of the injured children over their predetermined task to gain a strategic position in the war.

One thought experiment designed to consider hard determinism might also offer a deeper contemplation of strong artificial intelligence (exceeding human intelligence). If hard determinism was to hold true, a hypothetical ultimate super-computer that is cognizant of all knowledge to-date can be used to predict individual human choices and future events based on the raw analysis of every fact and trend that preceded any event. If such a predicting computer (representing strong AI) could not exist, then the actuality of hard determinism would be negated.

Both compatibilists and libertarians offer explanations that the robots in the above case are more than the product of their construction and programming, so that they demonstrate sentiments of sympathy and empathy toward the injured children. Ultimately the question of whether these robots have a moral responsibility requires the establishment of whether the robots have a free will to carry out their own actions. A hard determinist would argue that all robotic and artificially intelligent actions are due to the laws of physics and as such the robots would not carry any responsibility. Conversely according to libertarians, the robots have a 'soul'; so they demonstrate a free will with which to have moral responsibility, which in this case guided them to help the children. An incompatibilist view is problematical and may be evaluated by a thought experiment whereby human beings are enhanced by an implant that can control their desires so that they have no free will (Harris 2010). If the implant stimulates desires at random, then the human remains without free will, however his decisions and choices are reminiscent and generally undistinguishable from actual human actions. From the compatibilist view, the robots are governed by the laws of physics and their programming, however these can be designed and actioned in such a way that the robots can exhibit free will, which consequently results in the robots having moral responsibility.

The libertarian understanding of free will derives from the notion that individuals have the ability to do something differently or otherwise through the principle of alternative possibilities. Harry Frankfurt developed thought experiments (Frankfurt 1969) that counter this notion. He uses subjects that are responsible for their actions through intuition despite lacking the freedom to act differently. For our established robotic incident, the following can be an example of a Frankfurt-type case:

Warfare Robot A is likely to complete his mission of achieving military victory for his country whilst also considering the welfare of any local inhabitants (friend or foe) embroiled in the war (as set out by international treaties). There is only one reason that he will not consider the welfare of local inhabitants; only if they represent a direct threat to his country's victory or threat to his human or robotic country-mates. Robot A's programmer Y is keen to guarantee that Robot A does consider the welfare of the local populace during his war efforts so that he adds a specific implant into Robot A's neural system that can override A's programming in any case that he decides to consider against the welfare of local inhabitants during warfare. During the war, Robot A does prioritize human welfare during a battle (by helping injured civilian children) on his own accord so that programmer Y does not activate A's special implant. Consequently, Robot A is responsible for considering the needs of local inhabitants during a robotic fought war although, according to Programmer Y's implant, Robot A lacks freedom to do differently/otherwise.

Frankfurt's case thus opposes incompatibilism and supports the compatibilist notion that whilst sentient, rational robots and artificial intelligences are the product of their constructors and programmers. As such, they maintain moral responsibility in the society (both human and artificially intelligent) within which they exist.

Beyond the meta-physical elements of moral responsibility, there is an interpretation of responsibility through moral sentimentalism. This is consistent with a compatibilist framework, which in the case of robots suggest that their actions may have taken place as a result of a response to inherent emotions. The

robots that have established a moral responsibility would have then prioritized the well being of the children over their national aims of military victory.

The compatibilist account offers different degrees of free will and responsibility. Harry Frankfurt suggests a distinction in the levels of freedom through a hierarchy of desires (first order, second order and so on). Within the case Robot A may have a conflict in his desires, he may have a first order desire to help injured humans and a second order desire to achieve victory through warfare. Although at a utilitarian level, winning the war may save many more lives than those of a few injured children, Robot A's hierarchy of desires leads to an internal conflict with regard to responsibility. To a degree Robot A is not fully in control of himself (with the aim of fulfilling one single goal), so he is less free, but nonetheless resultantly prioritizes the lives of the injured children over the larger and long term goal of winning the war.

Comparison with the Moral Responsibility of Animals

When considering the responsibilities and duties of sentient and rational artificial intelligence agents and robots, one direct comparison might be with another group of non-human sentient beings such as animals and pets. Conceptually they offer many similarities and can provide insights about non-human responsibilities. Peter Singer specified that "Animals are treated like machines that convert fodder into flesh" (Singer 1979). There are several analogous elements between animals and artificial intelligence agents and robots. Many animals are reared by humans to fulfill specific duties in human society (such as guide dogs). In many cases they are also specifically bred (with defined genotypes, phenotypes and traits) and subsequently trained for specific tasks. In a similar way robots and artificially intelligent agents are specifically designed, built and subsequently programmed for specific tasks.

Wild animals have moral codes, and many animals demonstrate a 'social homeostasis' through their social networks and relationships. They have been shown to demonstrate a 'wild justice' (Bekoff and Pierce 2009) where they express emotions, behavioral flexibility, reciprocity, empathy, trust and discernable duty. These are also characteristics that would exist in future robots and artificial intelligence agents. From a purely ethical viewpoint, the question of morality is independent to an individual's species of origin; although at a practical level human beings have been dominant. Non-human animals species are subordinate to mankind and in a similar fashion robots and artificially intelligent agents will also be subordinate to humans. The question of animal rights has many well-established viewpoints, and there is an implicit consensus that at a practical level a utilitarian approach is applied to offer rights to animals where feasible. Nevertheless, there is a tacit recognition that animals carry a moral responsibility that requires the consideration of their moral value.

However, any direct moral comparison between sentient, rational artificial intelligence agents and robots with animals may prove superficial and problematic. The source of relevant moral actions should be distinguished from the evaluation of

the agent as being morally responsible for a certain behavior, as otherwise we may encounter the situation of having to legally charge or prosecute animals in a similar manner to archaic societies. This is represented in today's legal systems, which still addresses animals as reasonable targets of human censorship although it is accepted that it would be nonsensical to praise or blame them for their behavior. Consequently the parallel of robots and animals may not be fruitful, as we do not charge animals with moral accusations. Sentient and rational artificial intelligence agents and robots would have essential psychological qualities so as to make them both morally and legally responsible.

Practical Responsibility for Artificial Intelligence

Whilst conceptually there are favorable arguments supporting the moral value and moral responsibility of artificial intelligence agents and robots beyond simply considering their rights (Ashrafian 2014; Ashrafian et al. 2014), the introduction and application of responsibilities and duties requires realistic guidelines and protocols. For humans several such guidelines exist, though they are not as well recognized as the more acknowledged Universal Declaration of Human Rights (UDHR) (United Nations 1948). Two specific declarations for Human responsibilities include The Declaration of Human Duties and Responsibilities by the UNESCO supported VTMF (Valencia Third Millennium Foundation), also known as the 1998 Valencia Declaration or DHDR (VTMF and UNESCO 1998) and 1997 Universal Declaration of Human Responsibilities (UDHRes) (InterAction Council 1997) (Fig. 2). The DHDR specifies that "responsibility", "is an obligation that is legally binding under existing international law" and that "duty" is an "ethical or moral obligation".

Here it is proposed that as artificial intelligence agents and robots occupy human society with protection and support from humanity-based rights, then the principal

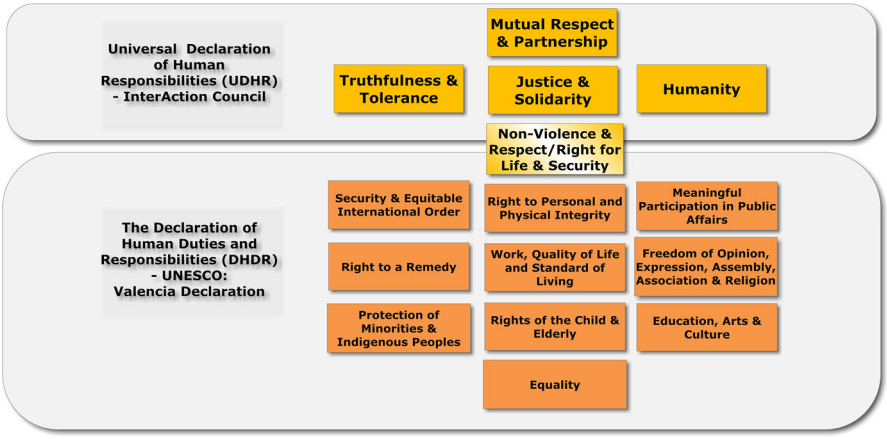

Fig. 2 Universal articles of human responsibilities (derived from the 1997 Universal Declaration of Human Responsibilities and the 1998 Declaration of Human Duties and Responsibilities)

message of these core human responsibilities will apply equally to the non-human artificial intelligences and robots, with the stipulated modification that human needs are to be prioritized over artificial intelligence and robot needs. The articles will together include those of: (1) Justice and Solidarity, (2) Mutual Respect and Partnership, (3) Truthfulness and Tolerance, (4) Fundamental Principles for Humanity (InterAction Council 1997), (5) Human Security and an Equitable International Order, (6) Meaningful Participation in Public Affairs, (7) Freedom of Opinion, Expression, Assembly, Association and Religion, (8) The Right to Personal and Physical Integrity, (9) Equality, (10) Rights of the Child and the Elderly, (11) Work, Quality of Life and Standard of Living, (12) Right to a Remedy, (13) Education, Arts and Culture (VTMF and UNESCO 1998) and (14) Non-Violence and Respect for Life/The Right to Life and Human Security (InterAction Council 1997; VTMF and UNESCO 1998) (Fig. 2).

Much in the same way that there has been consideration of human-artificial intelligence laws as well as artificial intelligence-on-artificial intelligence (AIonAI) laws (Ashrafian 2014), the responsibilities for artificial intelligence technologies should also consider human-AI and AIonAI elements. Consequently Article 1 of the UDHRes can be modified to:

Every person or individual, regardless of gender, ethnic origin, technological origin, social status, political opinion, language, age, nationality, or religion, has a primary responsibility to treat all human people in a humane way, and if this is not conflicted, has an added secondary responsibility to treat all non-human artificially intelligent individuals in a humane way.

Other elements are equally relevant to humans and artificial intelligence agents such as Article 4 "What you do not wish to be done to yourself, do not do to other," and Article 11 considering "advancement of the human race". Many of the principles should highlight the priority to act responsibly in favor of human life and culture over that of non-human artificial intelligence, but where possible to acknowledge the importance of these. For example human security and right to life should be ranked higher than the robotic equivalent, even for the robots and artificial intelligences themselves, however where possible the survival of both should be considered. Furthermore, whilst robots and artificial intelligences can contribute to the responsibility and support of human political rights, human freedom of expression and human culture, they themselves may not stand for political office over humans, override their freedom over humans or enforce their culture onto humans. Whilst all their socio-cultural merits should be celebrated, the fundamental partiality to favor human needs should be maintained by artificial intelligence agents and robots.

The determination of the status of artificial intelligence agents and robots with responsibility and supporting laws requires comparative societal governance. According to the current appreciation of artificial intelligence, most robots occupy the master-salve paradigm where no independence of action beyond direct human volition is permitted. This contrasts with future artificial intelligence abilities of self-consciousness, rationality and sentience demonstrated in the initial case where the robots could wage war through independent decisions on behalf of their warring states. The current outlook for such artificial intelligence agents still rests on

existence through service and subordinacy to human society. Within this paradigm, robots and artificial intelligence agents will demonstrate free will and morality, but also require societal security and welfare constraints so that in the preliminary phases of these technological advances detailed socio-political controls for robots and artificial intelligence agents must be determined. For example there will be a restriction on robot self-recreation, the ability to carry out independent business or public office. Nevertheless robots and artificial intelligences will be supported by rights and common laws and will contribute to society. As a consequence, the question arises of how human society recognizes a non-human being that is self-conscious, sentient and rational with ability at comparable-to-human (or even beyond-human) levels?

Within this context, a precedent already exists. In the ancient world 'foreign' or 'non-national' individuals (who by definition had comparable human aptitude) have been accepted to have different degrees of societal status and rights as recognized by formal law, for example in the ancient Roman Empire (30BC-212AD) (Shumway 1901). Under the *Ius Gentium* law (Fig. 3), Roman citizens were given a full complement of rights (through *Ius Civile*) whilst there were several classes of free individuals, including people of *Latin* (from Latium), *Peregrinus* (Provincial people from throughout the empire) and *Libertus* (Freed slave) status.

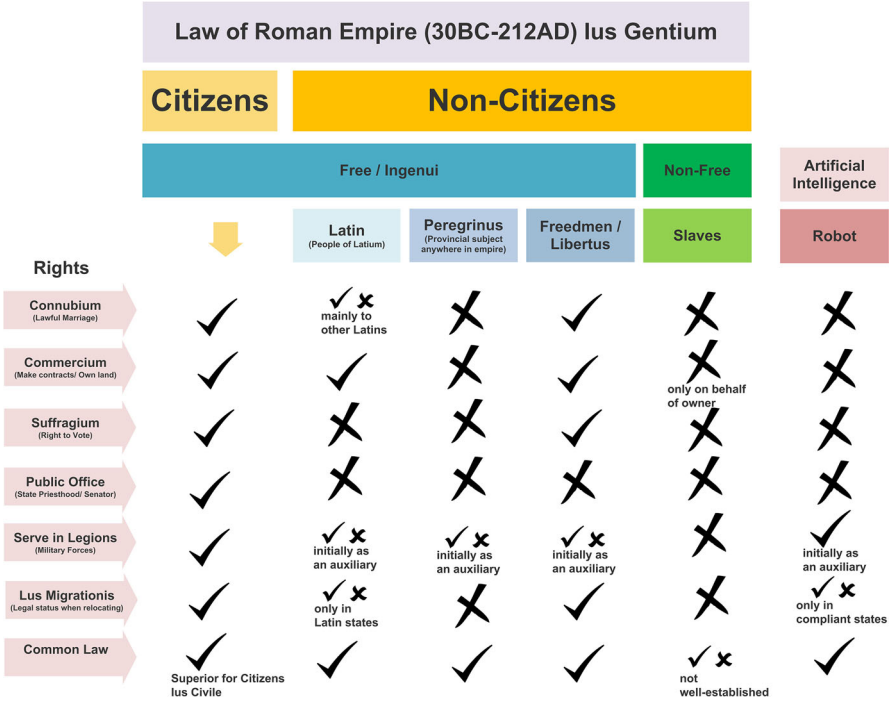

Fig. 3 Individual status within ancient Roman law and comparison to the proposed status of artificial intelligence

Latin rights (*Ius Latinum*) offered an intermediary stage to full Roman citizenship through the ability to carry out business, marry, participate in the military to some degree and have international legal recognition. Peregrinus rights however offered a lesser status, so that inter-marriage and business was not permitted, although societal contribution such as acting as auxiliary soldiers was acceptable. A comparable system could be applied to future artificial intelligence and robotics (Fig. 3). Here robots would likely occupy Peregrinus or possibly partial-Latin status, where they would not self-replicate, stand in public office or own land and business but would be protected by the law and have the ability to contribute to society through examples such as defending nations and participating in the healthcare sector.

Ultimately the application of an equivalent Roman-like system of laws for artificial intelligence agents and robots may progress just as those of the Romans themselves. In the first instance Roman lawyers were pragmatic and their law demonstrated that some slaves enjoyed significant autonomy. The 'elite' slaves, as in the case of the emperor's slaves, were estate managers, bankers and merchants, holding important jobs as public servants, or entering into binding contracts, managing and making use of property for their masters' family business. In fact some slaves were able to retain property (known as *peculium*) for personal management and use. The peculium was inaccessible by the owner, which could eventually be used to purchase their freedom though was technically the property of the head of the household. A similar system of a *digital peculium* has been envisaged for robots (Pagallo 2012) and could contribute in the broader application of Roman legal status as an exemplar for the legal status of future artificial intelligence agents and robots. Furthermore, after some time the Romans introduced the Edict of Caracalla or Antonine Constitution (*Constitutio Antoniniana*) in 212AD (likely to increase the number of individuals subject to taxation). Here Roman citizenship was granted to all "freeborn" men throughout the Empire whereas all freeborn women in the empire would receive the same rights as Roman women. Taken to its eventual conclusion, the continual advances in artificially intelligence agents and robots may herald their status of fully-fledged personalities with an accompanying level of higher legal and moral responsibilities but also a higher degree of rights. Here each type of legal personality could potentially be met by appropriate artificial intelligence agents (Chopra and White 2011). Consequently one possibility is that a legal personhood status might ensue for robots and artificial intelligence agents (Solum 1992) as result of a future "Caracalla approach."

Whilst an exact replica of ancient Roman law is not the direct solution to the practical introduction of robots and artificial intelligence agents within mankind's communities, its parallels nevertheless offer some degree of perceptiveness regarding the introduction of such agents into human society.

Conclusion

The ongoing developments and innovations in artificial intelligence and robotics offer the promised capacity of computer consciousness, sentience and rationality.

These have propelled the philosophical consideration of artificial intelligence and robot rights. The discernable next-step for evolution in this field necessitates attention to the moral responsibilities and duties of artificial intelligence and robots. Various philosophical stances can be engaged ranging from determinism to libertarianism and lend support to a middle ground of compatibilism. Such a position requires a commensurate adoption of responsibilities and duties for the advancement of human and artificial intelligence societies. These broad obligations require accountability within the context of prioritizing human aims and needs within the framework of a robust legal platform. The broader application of *noblesse oblige* where a leader fulfills the responsibilities of his status necessitate a proportionate *humanité oblige (humanity obliges)*; here it is incumbent on human society to ensure the fair, tolerant and ultimately humane institution of advanced artificial intelligence and robots within mankind's society.

Conflict of interest None.

References

Ashrafian, H. (2014). AIonAI: A humanitarian law of artificial intelligence and robotics. *Science and Engineering Ethics*. doi:10.1007/s11948-013-9513-9.

Ashrafian, H., Darzi, A., & Athanasiou, T. (2014). A novel modification of the Turing test for artificial intelligence and robotics in healthcare. *International Journal of Medical Robotics*. doi:10.1002/rcs.1570.

Bekoff, M., & Pierce, J. (2009). *Wild justice: The moral lives of animals*. London: The University of Chicago Press.

Chopra, S., & White, L. F. (2011). *A legal theory for autonomous artificial agents*. Ann Arbor: University of Michigan Press.

Frankfurt, H. G. (1969). Alternate possibilities and moral responsibility. *The Journal of Philosophy, 66*(23), 829–839.

Harris, S. (2010). *The moral landscape*. London: Bantam Press.

InterAction Council. (1997). Universal Declaration of Human Responsibilities. http://interactioncouncil.org/universal-declaration-human-responsibilities.

Pagallo, U. (2012). Three roads to complexity, AI and the law of robots: On crimes, contracts, and torts. In M. Palmirani, U. Pagallo, P. Casanovas, & G. Sartor (Eds.), *AI approaches to the complexity of legal systems* (p. 54). Berlin: Springer.

Shumway, E. S. (1901). Freedom and slavery in Roman law. *American Law Register, 40*(November), 636–653.

Singer, P. (1979). *Practical ethics*. Cambridge: Cambridge University Press.

Solum, L. B. (1992). Legal personhood for artificial intelligences. *North Carolina Law Review, 70*, 1231–1287.

United Nations. (1948). The universal declaration of human rights (UDHR) (http://www.un.org/en/documents/udhr/).

VTMF & UNESCO (1998). Declaration of responsibilities and human duties http://globalization.icaap.org/content/v2.2/declare.html.