New Ideas in Psychology 54 (2019) 107–111

Contents lists available at ScienceDirect

New Ideas in Psychology

journal homepage: www.elsevier.com/locate/newideapsych

Is morality the last frontier for machines?

Bipin Indurkhya

Institute of Philosophy, Jagiellonian University, Cracow, Poland

ARTICLE INFO

Keywords:
Machine ethics
Machine morality
Autonomous decision making

ABSTRACT

This paper examines some ethical and cognitive aspects of machines making moral decisions in difficult situations. We compare the situations when humans have to make tough moral choices with those in which machines make such decisions. We argue that in situations where machines make tough moral choices, it is important to produce justification for those decisions that are psychologically compelling and acceptable by people.

1. Introduction to morality in machines

As autonomous decision-making systems are becoming more and more commonplace — from self-driving cars, to military robots, including drones, to social companion robots including sex robots — they are rapidly entering into domains where moral and ethical values play a key role. For example, a self-driven car, on sensing a mechanical failure, may have to decide whether to hit some pedestrians, or drive into a ditch thereby risking the lives of its occupants. A military robot may have to decide whether to fire a shell at a house where a terrorist and also five other possibly innocent people are hiding. (This was the theme of a recent movie *Eye in the Sky.)* A companion robot may have to decide whether to lie to the human companion with a terminal disease about whether she or he will recover. These issues have ignited an intense interdisciplinary discussion on how machines should be designed to handle such decisions, and if machines should handle such decisions at all (Arkin, Ulam & Wagner 2012; Levy 2007; Lin, Abney & Bekey 2011; Pandey et al., 2017).

My goal in this paper is to identify issues in the design of machines that make human-like moral decisions. One key requirement for such machines, which we will refer to as *moral machines,* is that humans should find their decisions reasonable and acceptable. My approach is to analyze, with a few examples, how humans make complex moral choices and, more importantly, how they justify them, and then apply these insights to the design of moral machines. In this regard, we regard both humans and AI machines as cognitive systems, and focus on their input-output behavior. Some researchers find this assumption problematic, and maintain that humans and AI/Robotic systems are of different kinds, and their decision-making processes are also entirely different. For instance, it has been argued that the domain for ethical decisions is essentially a human forte, and a machine ought not to venture there (Bryson 2016; Bryson et al., 2017). It has been argued that machines should be deliberately designed to make it obvious that they are machines, so that no moral agency is attributed to them. Though this is a complex issue requiring a detailed discussion, which I would like to sidestep in this article, I would like to raise two issues that raise doubts about this position, at least with regard to the advent of moral machines.

1.1. Emergence of human-robot blends

The current debate on whether a machine can be moral agent or not is based on assuming a clear separation between robots and humans: robots are machines designed by humans using mechanical and electronic components; humans are biological beings who are born with some genetic dispositions inherited from the parents, and who develop their cognitive functionalities over time. However, this boundary is being blurred slowly. On one hand, people are incorporating robotic components in their bodies and brains to increase their physical and cognitive abilities (Schwartzman 2011; Warwick 2003, 2014). On the other hand, researchers are designing machines and robots using biological material (Ben-Ary & Ben-Ary 2016; Warwick 2010). At the moment, the state-of-the-art is still far from generating a human-robot blend that would be hard to classify as a robot or a human, but it might soon become a reality. For example, consider a person X who has artificial arms, legs and heart; uses cognitive enhancement devices for sight, hearing and smell; and has brain implants for enhanced memory functions and for Bluetooth-based communication with other beings with similar implants; and so on. X meets another man Y at a bar and recognizes him (albeit mistakenly) through the enhanced memory function as the man with whom X had a fight ten years ago and Y has beaten up X. X lightly slaps Y across his face, but the extra-strong titanium arm breaks Y's jaw. Who is responsible for Y's injury? X, the manufacturer of the titanium arm, the manufacturer of the memory

E-mail address: bipin.indurkhya@uj.edu.pl.

https://doi.org/10.1016/j.newideapsych.2018.12.001
Received 26 June 2017; Received in revised form 22 January 2018; Accepted 25 December 2018
Available online 11 January 2019
0732-118X/

implant, or all of the above? In this situation, perhaps many people will maintain that X is still the moral agent. But as more and more cognitive functions are delegated to devices, it is not clear at what point the person turns into an android and loses moral consciousness.[1]

1.2. Machines are susceptible for hacking

It is sometimes argued that robots, especially military robots, should not be completely autonomous because they can be hacked (Lin 2011). This, however, is a problem with humans as well. Ever since the dawn of history, there have been many examples where some human was bribed or blackmailed into turning against their own side, or changed their moral stance. (Consider Judas, Brutus, Alfred Redl, Harold Cole, Mir Jafar, Aldrich Ames, and so on.) Therefore, this cannot be a basis for denying machines moral agency. I should emphasize again that we are taking a cognitive systems approach to comparing humans with AI systems/robots. Hacking and bribing/blackmailing are similar with respect to the techniques used and the effect they produce. In hacking, one looks for a vulnerable, insecure piece of code to gain entry into the system to subvert it from its original goal. In bribing/blackmailing, one looks at a vulnerable personal behavior trait or past indiscretion in order to coerce an agent to betray his or her own side.

In the rest of this paper, I would like to focus on a more pragmatic issue. Assuming that the development of technology cannot be stopped by creating legal barriers — indeed, there are already companion robots that interact with people in a human-like way and try to fulfill their social needs — the issue I will address is how to make their decisions acceptable to humans from an ethical point of view. First, I will discuss one approach that has been generally accepted when humans have to make tough moral choices, and then I will examine some problematic issues that arise when this approach is applied to machines.

2. Humans facing tough moral choices: *Sophie's choice* effect

What does a human do when confronted with two choices that are both horrifying? I refer to this as *Sophie's Choice* effect based on the film with this title, where a Nazi officer forces a Polish mother (played by Meryl Streep) to choose one of her two children, whose life would be spared. One can find many similar real-life cases, especially during natural disasters like earthquakes and floods, or during man-made disasters like wars. No matter what one chooses, such decisions usually leave a deep psychological scar and can traumatize the person for the rest of her or his life.

For example, one real-life Sophie's choice situation was forced on Aneta Gadleva, when Chechen militants took more than 1000 people hostages at School No 1 in Beslan, North Ossetia, on 1 September 2004 (Gracheva 2016). Aneta Gadleva was among the hostages with her two daughters: nine-year-old Alana and one-year-old Milena. On the second day of the hostage crisis, the militants allowed women of young children to be released but they could only take with them one breast-feeding infant. It was a panic situation, with the militants shouting orders and no time to think. The risk was for the mothers and all the children to be left behind. In the end, 11 mothers and 15 babies were let go, with many of the mothers, including Aneta Gadleva, being forced to leave older siblings behind, many of whom died on the third day as the crisis ended in a violent confrontation between the Russian security forces and the militants.

This issue has been explored extensively in recent years as what is known as the trolley problem and its variants (Bruers & Braeckman 2014; Navarette et al., 2012; Nucci 2013). The basic trolley problem is a hypothetical situation set up to present the participant with an ethical dilemma: do nothing and five people may die, or take an action killing a different person and saving the five people. These experiments, however, do not reveal what a person would actually do in such a situation, what psychological trauma they will face as a result of it, and how they justify their choices. Sometimes the participants provide some justification, but then varying the experimental conditions show that they do not necessarily act according to their own justification (Bauman et al., 2014). Nonetheless, these experiments provide fodder for how autonomous machines like self-driving cars might be programmed with moral rules (Brogan 2016). I should emphasize that the trolley problem and its variants seem far removed from the real-life examples presented above, but this is precisely the point here. It is not clear what significance do the results from trolley problem experiments have for real life situations. On the contrary, we can get a much better insight into the complexity of moral decision-making in humans by focusing on real-world scenarios.

Such dilemmas are also faced by governments and social groups, often in war campaigns, in starting big construction projects like dams, in social projects like relocating slums, and so on. In such situations, some public justification is often provided, though, in almost all such cases it is not accepted by everyone. Perhaps the most well-known case may be the justification put forth by the US Government for dropping nuclear bombs on Hiroshima and Nagasaki, namely that it saved lives of American as well as Japanese people (Morton 1960). We can learn from such explanations as to what is acceptable or not acceptable by social groups.

A very useful case in point is the development of triage systems to determine the priority of medical treatment of patients, which is widely used (Iserson & Moskop 2007; Moskop & Iserson 2007; Robertson-Steel 2006). The triage system was originally applied during Napoleonic wars, and evolved through the First World War, in the military hospitals on the battlefields, where the resources in terms of the medical staff, transport to the medical facilities, available beds, operating facilities etc. are limited, but the number of wounded soldiers needing treatment is very large. The goal of the triage system is to identify, through a quick assessment, the people likely to benefit the most from immediate medical care, as opposed to people who are unlikely to live regardless of the medical care, and those who are likely to survive without getting immediate medical care. It has been extended and applied to the emergency departments in modern hospitals (Christ et al., 2010). These days many versions of triage are widely used: for example, the reverse triage system determines which patients can be discharged from the hospital.

What is relevant for the discussion here is that the process of making triage decisions has been formalized into algorithms so that lesser-trained healthcare professionals can also make triage decisions speedily and effectively (Larson et al., 1973). Similarly, the decision-making process to determine when a person experiencing chest pains might be having a heart attack was formalized by Lee Goldman. He devised an algorithmic way to combine the ECG data with three risk parameters — is the pain unstable angina; is there fluid in the patient's lungs; and is the systolic blood pressure below 100 — to quickly determine the optimal treatment (Browner 2006; see also Gladwell 2005, Chap. 4.). When such decision procedures are accepted by the medical community, healthcare professionals are trained in it and they do not have to agonize over moral complexities of each individual decision: how can I deny medical treatment to this seriously wounded soldier crying for help, even though I realize that nothing much can be done for him and he does not have long to live anyway? This suggests that when machines are concerned, a similar approach could be effective. This issue is examined in the following section.

3. Machines making moral decisions?

Humans generally show a reluctance to accept machine superiority. Almost always, humans have challenged the machine when it made a foray into some human domain. There is the legend of John Henry, who

[1] Some might argue that such scenarios are impossible or too far in the future, in which case the issues of this paper will not be relevant to them.

competed with a steam-powered hammer, won the contest, but then died immediately after as his heart gave out. Deep Blue defeated the reigning world champion Garry Kasparov in 1997, but some claim that the computer does not 'understand' chess because it does not play as humans do (Linhares 2014). More recently, in 2011, the computer system Watson won the quiz game Jeopardy against the best humans, but many scholars deny that it has any 'understanding' of the questions or related concepts (Searle 2011). Following this trend, it may not be surprising that moral machines, where they have yet to demonstrate their superiority in some way, is considered a taboo topic by some researchers.

Examples of chess and Jeopardy show that when we extract the cognitive mechanism underlying some human behavior, make an algorithmic version of it, and *have it run on a machine,* people generally do not accept it. This is exactly opposite of what we saw above when humans are trained to implement the algorithmic version, as in the case of triage, they readily make tough moral choices. So why is that when an algorithm making moral choices is run on a machine, people are reluctant to accept it? The example of actuarial tables in making parole decisions may provide some answers, and we discuss this below.

Actuarial tables contain statistical data from the past to show the likelihood of some event happening. Insurance companies use them extensively to determine the premium levels. So, for example, insurance premiums for teen drivers are higher because past history shows that a teen driver is more like to be involved in a traffic accident. Evidence has been put forth to show that actuarial tables are more reliable than human experts, but their role in legal decision-making is hotly disputed (Dawes, Faust & Meehl 1989; Krauss & Sales 2001; Litwack 2001; Starr 2014) for various reasons. For instance, consider some of the arguments raised by Starr (2014) against evidence-based sentencing. She argued that the group characteristics predicted by the regression model do not directly translate into the behavior prediction of an individual: "… [T]he instruments [such as actuarial tables] provide nothing close to precise predictions of individual recidivism risk. The underlying regression models estimate average recidivism rates for offenders sharing the defendant's characteristics. While some models have reasonably narrow confidence intervals for this predicted average, the uncertainty about what an *individual* offender will do is *much* greater." (Starr 2014, p. 3, emphasis original; see also the discussion in pp. 26–32.)

Moreover, Starr points out that if, based on the past history of a group (as reflected in an actuarial table), an individual is judged to be more likely to return to crime, it violates the constitutional rights of that individual in getting a fair and unbiased judgment: "… sentencing based on such instruments [as actuarial table] amounts to overt discrimination based on demographics and socioeconomic status." (p.66.) This is because any statistical or algorithmic method based on the past behavior patterns of a population reflects its past biases and prejudices. So, in this respect, such methods do not model the Kuhnian revolutions of social norms (Indurkhya 2016).

It should be emphasized here that the human decision-making process goes through such revolutions. Consider the application of the legal concept terra nullius by Australian courts. According to it, if a land is unoccupied, then a settler can claim property rights to it. Until 1992, the courts maintained that Australia was unoccupied before the settlers arrived, and made decisions accordingly. However, in Mabo v Queensland (No.2) [(1992) 175 CLR 1], the Australian High Court held that previous decisions — holding that Australia was terra nullius at settlement — were legally flawed, because the aborigines were already living and settled in Australia when the settlers arrived (Hunter and Indurkhya 1998). Computer algorithms based on past judgments cannot make this leap, or gestalt shift, which becomes a crucial limitation of such systems.

Another example is provided by Martindale's (1990) sociological theory of creativity in arts. In this model, the society encourages some amount of novelty to keep things interesting. But there is also habituation, so what was novel once gradually becomes less and less interesting. When a new style is introduced, initially it offers novelty so even relatively simple artifacts are considered interesting. Over time, more and more ideas are explored in that style, but eventually the potential of that style is used up, and a new style is called for. This model predicts a periodic change of style. Martindale provided empirical support for his theory by historiometric analysis of French, British, and American poetry, European and American painting, Japanese prints, and other genres. The aspect of the model relevant for the discussion here is simply that algorithmic methods based on past history may able to predict evolution within a style, but not the emergence of a new style. Perhaps they can predict that a new style will emerge, based on the periodicity of the style change, but they cannot predict the content of the style. Consider, for instance, the work of Ni et al. (2011), who trained their program with the official UK top-40 singles chart over the past 50 years to learn as to what makes a song popular. A program like this might successfully predict the winner of the future Eurovision competitions, but it cannot predict drastic changes in the aesthetic values and tastes like atonal music or abstract art. The implication of this for moral machines is that a machine-learning approach based on past data cannot learn, for example, that slavery is immoral, for it requires modeling something akin to a Kuhnian revolution.

Another limitation of the algorithmic methods to predict the future based on the past is that they do not take into account what the effect of the action based on the prediction will have on the future. Starr (2014, p. 3) points this out as well: "What judges need to know is not just how 'risky' the defendant is in some absolute sense, but rather how the sentencing decision will *affect* his recidivism risk. For example, if a judge is deciding between a one-year and a two-year prison sentence for a minor drug dealer, it is not very helpful to know that the defendant's characteristics predict a 'high' recidivism risk, absent additional information that tells the judge how much the additional year in prison will reduce (or increase) that risk. Current risk prediction instruments do not provide that additional information." (Emphasis original.)

Moreover, once the algorithmic methods are known, people alter their behavior in order to achieve the desired result. This can be considered as the *Minority Report* effect, for it was the basis of a short story with this title by Philip K. Dick. Some examples illustrating this phenomenon are the manipulation of electoral district boundaries in the US by individual parties in order to give them a demographic advantage, which is known as *Gerrymandering* (Mann 2006), and various efforts people make to increase the rank of their webpage in Google search queries (Hamilton 2009).

Thus, we see that when a decision process is formalized based on rational principles, and an algorithmic version is devised, people accept the results of this process. But a similar approach applied to decision making by machines is not so easily accepted by people. We can notice an analogous trend in robot behavior. Humans practice for long hours over weeks, months and years, to be able to juggle blindfolded, or do synchronized swimming or dancing with a partner with perfect precision. In sports competitions like Olympics, smallest imperfections with respect to the timing may cost crucial tenths of a point. For robots, on the other hand, such tasks as juggling blindfolded or doing synchronized actions (swimming or dancing) are relatively easy, but it also makes them more machine-like (which they are). For example, starting from a bounce-juggling robot designed by Claude Shannon in the 1970s, there have been many other robots that juggle by predicting the trajectory of the bounced or tossed ball. This is because it easy to predict the trajectory of a tossed ball and place the robot arm at the expected place. However, to juggle based on visual perception, as in juggling with a partner, is a much harder task. It is only recently that robots have been developed which can play catch based on visual feedback. Similarly, robots can be programmed easily to move in synchrony. But when a robot engineer programs a group of robots to dance, they are purposely programmed to dance a little bit off so that it is more human like (See for instance, Peng et al., 2016.).

The point here is that in order to make the machine behavior (in this

case moral decision making) acceptable to people, we need to generate explanations behind decision making that are psychologically compelling. So even though a machine may make a decision based on some calculated probabilities based on logic, it is important to explain it from a psychological point of view. This is illustrated by the experience of the designers of one of the first AI backward-chaining expert systems, *Mycin* (Shortliffe 1976), which was found to be lacking in explanations, and this feature was added later in *Emycin* (Ulug 1986).

A related psychological issue that needs to be considered is that generally people apply different moral criteria for themselves than to others. This is illustrated by the research of Bonnefon, Shariff, and Rahwan (2017), where they found that when it comes to Autonomous Vehicles (AVs), even though people approve of utilitarian AVs, which may sacrifice their passengers for the greater good, they themselves would prefer buying an AV that prioritizes protecting its passengers.

4. Conclusions: Is morality the last frontier?

Assuming that the autonomous decision-making systems are here to stay, and that there will be situations in which they will be making moral and ethical decisions, in order that these decisions are accepted by many (if not all) humans, it is crucial to generate explanations underlying those decisions that are psychologically convincing. To emphasize, a rational or logical explanation is not always psychologically compelling. On the other hand, there is much research to show that people's behavior is determined by factors that seem irrational but are nonetheless predictable (See, for instance, Ariely 2009.). We need to incorporate these factors in an explanation module that generates psychological justifications for the decisions of any autonomous system.

Some AI researchers are fully aware of this challenge. Ian Sample (2017), in a recent news article, interviewed Alan Winfield, a professor of robot ethics: "… [Alan Winfield] agreed that tech firms might struggle to explain their AI's decisions. Algorithms, especially those based on deep learning techniques, can be so opaque that it is practically impossible to explain how they reach decisions. 'My challenge to the likes of Google's DeepMind is to invent a deep learning system that can explain itself,' Winfield said. 'It could be hard, but for heaven's sake, there are some pretty bright people working on these systems.'"

The same concern has been echoed by the head of Google's self-driving car project Dmitri Dolgov: "Over the last year, we've learned that being a good driver is more than just knowing how to safely navigate around people, [it's also about] knowing how to interact with them." (Quoted in Wall 2016). BBC Technology Editor, Matthew Wall notes: "Driving isn't just about technology and engineering, it's about human interactions and psychology." The same can be said about moral decision-making: it is not just about rationality and logic. To make moral decisions that can be supported by psychologically acceptable explanations, it is important to research how humans reason and what arguments they find persuasive, and incorporate this ability in robots and other autonomous systems.

Precisely this point is echoed in a recent article by Kuang (2017): "In many arenas, A.I. methods have advanced with startling speed; deep neural networks can now detect certain kinds of cancer as accurately as a human. But human doctors still have to make the decisions – and they won't trust an A.I. unless it can explain itself. This isn't merely a theoretical concern. In 2018, the European Union will begin enforcing a law requiring that any decision made by a machine be readily explainable, on penalty of fines that could cost companies like Google and Facebook billions of dollars." All this has stimulated a flurry of research activities on *Explainable Artificial Intelligence* in recent years (Aha et al., 2017; Gunning 2016; Indurkhya and Misztal-Radecka 2016; Olah et al., 2017; Schwiep 2017).

References

Aha, D., (Ed.). (2017). *IJCAI-97 workshop on explainable AI (XAI). 20 August 2017. Melbourne, Australia*In: *http://www.intelligentrobots.org/files/IJCAI2017/IJCAI-17_XAI_WS_Proceedings.pdf*, Accessed date: 13 January 2018.

Ariely, D. (2009). *Predictably irrational: The hidden forces that shape our decisions.* New York: Harper Perennial (exp. rev. edition.

Arkin, R. C., Ulam, P., & Wagner, A. R. (2012). Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust and deception. *Proceedings of the IEEE, 100*(3), 577–589.

Bauman, C. W., McGraw, A. P., Bartels, D. M., & Warren, C. (2014). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Social and Personality Psychology Compass, 8*, 536–554.

Ben-Ary, G., & Ben-Ary, G. (2016). Bio-engineered brains and robotic bodies: From embodiment to self-portraiture. In D. Herath, C. Kroos, & Stelarc (Eds.). *Robots and art* (pp. 307–326). Singapore: Springer.

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2017). The social dilemma of autonomous vehicles. *Science, 352*(6293), 1573–1576.

Brogan, J. (2016). *Should a self-driving car kill two jaywalkers or one law-abiding citizen?* http://www.slate.com/blogs/future_tense/2016/08/11/moral_machine_from_mit_poses_self_driving_car_thought_experiments.html/ Accessed 5.9.16.

Browner, W. S. (2006). Decision analysis. In R. M. Wachter, L. Goldman, & H. Hollander (Eds.). *Hospital medicine* (pp. 43–50). (2nd ed.). Philadelphia: Lippincott, Williams & Wilkins.

Bruers, S., & Braeckman, J. (2014). A review and systematization of the trolley problem. *Philosophia, 42*(2) 251 –169.

Bryson, J. J. (2016). Patiency is not a virtue: AI and the design of ethical systems. *Proceedings of the AAAI spring symposium on ethical and moral considerations in non-human agents* (pp. 202–207). Palo Alto, CA: AAAI Press.

Bryson, J. J., Diamantis, M. E., & Grant, T. D. (2017). Of, for, and by the people: The legal lacuna of synthetic persons. *Artificial Intelligence and Law, 25*, 273–291.

Christ, M., Grossmann, F., Winter, D., Bingisser, R., & Platz, E. (2010). Modern triage in the emergency department. *Deutsches Ärzteblatt International, 107*(50), 892–898. http://doi.org/10.3238/arztebl.2010.0892.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*(4899), 1668–1674.

Gladwell, M. (2005). *Blink: The power of thinking without thinking.* New York: Little Brown & Co.

Gracheva, A. (2016). *The Beslan mum who could only save one of her children.* 12 June 2016BBC News (Magazine)http://www.bbc.com/news/magazine-36378981, Accessed date: 12 April 2017.

Gunning, D. (2016). *Explainable artificial intelligence (XAI).* https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf, Accessed date: 15 January 2018.

Hamilton, P. (2009). *Google-bombing - manipulating the PageRank algorithm. Class white paper.* http://userpages.umbc.edu/~pete5/ir_paper.pdf/, Accessed date: 11 January 2017.

Hunter, D., & Indurkhya, B. (1998). 'Don't think, but look' a gestalt interactionist approach to legal thinking. In K. J. Holyoak, D. Gentner, & B. Kokinov (Eds.). *Advances in Analogy research: Integration of theory and data from the cognitive, computational, and neural sciences* (pp. 345–353). Sofia (Bulgaria): NBU Press.

Indurkhya, B. (2016). A cognitive perspective on norms. In J. Stelmach, B. Brożek, & Ł. Kwiatek (Eds.). *The normative mind* (pp. 35–63). Kraków (Poland): Copernicus Center Press.

Indurkhya, B., & Misztal-Radecka, J. (2016). Incorporating human dimension in autonomous decision-making on moral and ethical issues. *Proceedings of the AAAI spring symposium on ethical and moral considerations in nonhuman agents* (pp. 226–230). Palo Alto, CA: AAAI Press.

Iserson, K. V., & Moskop, J. C. (2007). Triage in medicine, part I: Concept, history and types. *Annals of Emergency Medicine, 49*(3), 275–281.

Krauss, D. A., & Sales, B. D. (2001). The effects of clinical and scientific expert testimony on juror decision making in capital sentencing. *Psychology, Public Policy, and Law, 7*(2), 267–310.

Kuang, K. (2017). Can A.I. Be taught to explain itself? The New York times magazine. Nov. 21, 2017 https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html, Accessed date: 25 November 2017.

Larsen, K. T., Jr., Vickery, D. M., Collis, P. B., & Folland, E. D. (1973). Triage: A logical algorithmic alternative to a non-system. *Journal of the American College of Emergency Physicians, 2*(3), 183–187.

Levy, D. (2007). *Love and sex with robots: The evolution of human-robot relationships.* New York: HarperCollins.

Lin, P. (2011). Introduction to robot ethics. In P. Lin, K. Abney, & G. A. Bekey (Eds.). *Robot ethics: The ethical and social implications of robotics* (pp. 3–16). Cambridge (Mass: MIT Press.

Lin, P., Abney, K., & Bekey, G. A. (Eds.). (2011). *Robot ethics: The ethical and social implications of robotics.* Cambridge (Mass: MIT Press.

Linhares, A. (2014). The emergence of choice: Decision-making and strategic thinking through analogies. *Information Sciences, 259*, 36–56.

Litwack, T. R. (2001). Actuarial versus clinical assessments of dangerousness. *Psychology, Public Policy, and Law, 7*(2), 409–443.

Mann, T. E. (2006). Polarizing the house of representatives: How much does gerrymandering matter? In P. S. Nivola, & D. W. Brady (Eds.). *Red and blue nation? Characteristics and causes of America's polarized politics* (pp. 263–300). Baltimore: Brookings Institution Press.

Martindale, C. (1990). *The clockwork muse: The predictability of artistic change.* New York: Basic Books.

Morton, L. (1960). The decision to use the atomic bomb. In L. Morton (Ed.). *Command Decisions. Office of the chief of the military history of the army. Washington D.C* (pp. 493–518). . http://www.dod.mil/pubs/foi/Reading_Room/NCB/361.pdf, Accessed date: 5

November 2016.
Moskop, J. C., & Iserson, K. V. (2007). Triage in medicine, part II: Underlying values and principles. *Annals of Emergency Medicine, 49*(3), 282–287.
Navarrete, C. D., McDonald, M. M., Mott, M. L., & Asher, B. (2012). Virtual morality: Emotion and action in a simulated three-dimensional "trolley problem". *Emotion, 12*(2), 364–370.
Ni, Y., Santos-Rodriguez, R., Mcvicar, M., & De Bie, T. (2011). Hit song science once again a science? *Fourth International Workshop on Machine Learning and Music: Learning from Musical Structure, Sierra Nevada, Spain*.
Nucci, E. D. (2013). Self-sacrifice and the trolley problem. *Philosophical Psychology, 26*(5), 662–672.
Olah, C., Mordvinstev, A., & Schubert, L. (2017). *Feature visualization: How neural networks build up their understanding of images.* https://doi.org/10.23915/distill.00007.
Pandey, A. K., Gelin, R., Ruocco, M., Monforte, M., & Siciliano, B. (2017). When a social robot might learn to support potentially immoral behaviors in the name of privacy: The dilemma of privacy vs. Ethics for a socially intelligent robot. *Proceedings of the workshop on privacy-sensitive robotics, HRI 2017*.
Peng, H., Hu, H., Chao, F., Zhou, C., & Li, J. (2016). Autonomous robotic choreography creation via semi-interactive evolutionary computation. *Int J of Soc Robotics, 8*, 649–661.
Robertson-Steel, I. (2006). Evolution of triage system. *Emergency Medical Journal, 23*(2), 154–155.
Sample, I. (2017). *AI watchdog needed to regulate automated decision-making, say experts.* The Guardian. 27 January 2017 https://www.theguardian.com/technology/2017/jan/27/ai-artificial-intelligence-watchdog-needed-to-prevent-discriminatory-automated-decisions/, Accessed date: 28 January 2017.
Schwartzman, M. (2011). *See yourself sensing: Redefining human perception.* London: Black Dog Publishing.
Schwiep, J. (2017). *The state of explainable AI.* https://medium.com/@jschwiep/the-state-of-explainable-ai-e252207dc46b, Accessed date: 15 December 2017.
Searle, J. (2011). Watson doesn't know it won on "Jeopardy". *Wall Street Journal* 23 February 2011.
Shortliffe, E. H. (1976). *Computer-based medical consultations: MYCIN.* New York: Elsevier/North Holland.
Starr, S. B. (2014). Evidence-based sentencing and the scientific rationalization of discrimination. *Stanford Law Review, 66*.
Ulug, F. (1986). *Emycin-Prolog expert system shell*Master's Thesis. Monterey, California: Naval Postgraduate School.
Wall, M. (2016). *Would you bully a driverless car or show it respect?* http://www.bbc.com/news/business-37706666/, Accessed date: 21 October 2016.
Warwick, K. (2003). Cyborg morals, cyborg values, cyborg ethics. *Ethics and Information Technology, 5*(3), 131–137.
Warwick, K. (2010). Implications and consequences of robots with biological brains. *Ethics and Information Technology, 12*(3), 223–234.
Warwick, K. (2014). The cyborg revolution. *Nanoethics, 8*(3), 263–273.