9-1-2016

# Managing the Ethical and Risk Implications of Rapid Advances in Artificial Intelligence: A Literature Review

Taylor Meek
*Portland State University*

Husam Barham
*Portland State University*

Nader Beltaif
*Portland State University*

Amani Kaadoor
*Portland State University*

Tanzila Akhter
*Portland State University*

# Managing the Ethical and Risk Implications of Rapid Advances in Artificial Intelligence: A Literature Review

Taylor Meek, Husam Barham, Nader Beltaif, Amani Kaadoor, Tanzila Akhter
Dept. of Engineering and Technology Management, Portland State University, Portland, OR - USA

***Abstract*—The development of emergent technologies carries with it ethical issues and risks. We review ways to better manage the ethical issues and risks of one emerging technology: Artificial Intelligence (AI). Depending on how AI's development is managed, it may have beneficial and/or deleterious effects. The processing capacity of Tianhe-2, the world's fastest supercomputer, by some measures, exceeds the processing capacity of a single human brain, but at a prohibitive processing/power consumption ratio and physical size. Given the current pace of AI R&D activities, some estimates in the literature suggest that the technology could become capable of self-determination and super intelligence in only a few decades. This demands a serious analysis of the ethical implications of AI's development and the risks it might pose, in addition to technology management recommendations. We review the state of AI development, the timeline and scope of its possible future development, and potential ethical risks in its implementation. Further, we briefly review ethics and risk management practices as they relate to technology. Finally, we make technology management recommendations, which may help to address the ethical implications and to mitigate existential risks to humanity—with the development and dissemination of AI—by guiding its proper management.**

## I. INTRODUCTION

The emerging field of Artificial Intelligence (AI) technology has grown rapidly in recent years [22], bolstered by significant scientific and technological progress in many fields, especially in computer hardware, where the fastest computer in the world now has a capacity that exceeds the processing capacity of a single human brain [58]. AI is now integrated into many aspects of society. Examples can be seen in various sectors. For instance, financial institutions depend on AI systems to identify optimal investment options, detect fraud, and analyze market trends; hospitals use AI systems to assist life-saving equipment; heavy industries use AI systems and robots in assembly line production; the air transportation sector uses AI systems to control air traffic and to auto-pilot planes; and the list goes on [59].

While AI is becoming even more advanced and integrated into our lives, concerns about the ethical issues in developing such technology and the risks associated with adopting it become more important. Schwartz and Caplan indicated that "while ethical considerations may often be less visible than scientific, political, legal, or financial concerns, they are present and directly relevant to [...] decision-making" [60]. In this paper we will identify the main ethical concerns and potential risks related to AI and suggest ways to manage and mitigate potential existential risks, and maximize the well being of the human race.

In doing so, this paper will consider the following critical questions:

1) What are some ethical implications of artificial intelligence today and what is anticipated in the future?
2) What is the meaning of ethics from a technology perspective?
3) What does the literature say about managing artificial intelligence ethics?
4) What recommendations can we draw from ethics methodologies in assessing the ethical implications of artificial intelligence?
5) What are the gaps in the artificial intelligence ethics research and literature?

## II. METHODOLOGY

Our primary source of material for this paper is the literature of the AI Ethics field, supplemented by those of AI (in a historical context), Technology Ethics, and Technology Risk Management. By performing a literature review, our primary intent is to identify the central ethical issues related to AI and the sources' differing perspectives, and then to isolate management recommendations by analyzing those issues.

By reviewing the literature, we thematically characterize issues as categories and subcategories by identifying common themes from the literature [6][7]. By classifying external factors affecting an organization or a topic (AI Ethics in this case) into four categories (Political, Economical, Social, and Technological), we apply a perspectives analysis using the PEST tool [5], which further supports our analysis from a management context. This analysis allows us to understand the various perspectives of how ethical issues affect AI—and the relationships among them—to be able to make better decisions about managing them.

We then relate our analyses to our stated purpose: to identify possible management practices and topics of interest to managers and policymakers involved in this technology.

## III. LITERATURE REVIEW

### *A. Definitions*

Before we go further in discussing the ethical issues of AI, we will first define relevant terminology.

*Artificial Intelligence* has been defined as "the capacity of a computer to perform operations analogous to learning and decision making in humans, as by an expert system, a program for CAD or CAM, or a program for the perception and recognition of shapes in computer vision systems" [1]

and more broadly, another definition is "the capacity of computers or other machines to exhibit or simulate intelligent behavior; the field of study concerned with this" [2].

The Artificial Intelligence literature divides AI into two categories: domain-specific artificial intelligence and artificial general intelligence (AGI) [8] [9].

*Domain-specific Artificial Intelligence* includes AI that are narrow in their capabilities or utility functions: they can do a specific task intelligently. For example, IBM's Deep Blue is a computer that is dedicated to playing chess. In 1997, this system beat Garry Kasparov—the world reigning chess champion at the time. However, this machine is incapable of doing anything but play chess [10]. Such AI machines have already emerged in many areas, such as finance, mathematical computation, disease identification, medical treatment, nuclear simulation, weather prediction, and other fields [10].

*Artificial General Intelligence (Superintelligence; AGI)* are those AI that can respond to a variety of previously unspecified situations, no matter how novel they are or unprepared the AI is for the task [8]. Such an AI could conceivably learn, create its own knowledge, make its own decisions, and simulate the human brain, but not necessarily behave as with a human psyche or moral values [11]. However, such an intelligence has yet to be invented, as current technologies (both in terms of computational power and logical decision-making algorithms) have not yet been developed to achieve this [12].

*Ethics* has been defined as, "that branch of philosophy dealing with values relating to human conduct, with respect to the rightness and wrongness of certain actions and to the goodness and badness of the motives and ends of such actions" [3].

*Risk* has been defined as exposure to the chance of injury or loss; a hazard or dangerous chance [61]. In this paper we refer to risk as the AI Ethics literature does, classifying ethical risks associated with AI technology, and so we characterize certain risks as a subset of the ethical issues surrounding AI technology.

## IV. CHRONOLOGICAL HISTORY OF AI RESEARCH

### *A. History of AI*

We review the history of AI research in order provide context for the current state of the technology. Early Artificial Intelligence research was heavily informed by writings of philosophy, logic theory, and fiction [13]. Research in the emerging field of computer science began in 1931, and also gave great influence [14]. Following the revolution of computing technology in the early 20th century, AI research emerged in the middle of the 20th century, resulting in an initial AI theory [15]. Due to the imagination of the media, many myths also emerged surrounding this innovation [16].

Alan Turing introduced a proposal for a general intelligence machine in 1937 [16]. Many research laboratories contributed to its early inception, such as Alan Turing's lab in Manchester, the Moore School at Penn, Howard Aiken's laboratory at Harvard, the IBM and Bell Laboratories, and others. These gave the field of AI a chance to emerge from the shadows and to become a mainstream science. World War II brought an increased focus on the development of computing technology [13], which led to the development of the first electronic computer—ENIAC—which was termed the *giant brain*. In this historical context, we pause to note that we can view ENIAC as an early form of domain-specific artificial intelligence, and perhaps the first—a system capable [18] of performing computations many times faster than electromechanical machines or humans.

In 1956, the first description of artificial intelligence was announced when John McCarthy collaborated with scientists from several fields to submit a research proposal to the Rockefeller Foundation [17][19]. Thereafter, there were several meetings, conferences, and publications on AI in its evolving domains: application, effects, and uses. *Computers and Thought*—a book written by Edward Feigenbaum and Julian Feldman in 1963—was the first book about AI. It was the first to gather information and concepts about AI's function. McCarthy also published a seminal paper paper related to AI in this era titled *Programs with Common Sense* in 1958. Those material paved the way for future researchers of AI [13].

In addition to AI's progress in the U.S. research community, it also progressed in the UK and some European countries. Several workshops took place in Edinburgh in the middle of the 1960s and early 1970s. Moreover, a clear vision for knowledge-based systems emerged in the 1960s and early 1970s with the development of the Dendral program—that period has been called "paradigm shift" for AI by Ira Goldstein and Seymour Papert [13]. Progress in Europe began a decade later, following the progress in the UK [20].

Between 1985 and 1997, IBM developed Deep Blue, the chess-playing supercomputer which has been considered to be capable of domain-specific artificial intelligence.

Many advances occurred in the 2000s. During that period, intelligent toys emerged—"interactive robot pets"—an idea which had origins in the 18th century. Also, the first proposal for a robot with a face capable of emotional expressions emerged, introduced by Cynthia Breazeal at MIT in her dissertation on Sociable Machines. And in 2005, autonomous cars first raced one another in the DARPA Grand Challenge race [21].

Significant developments in other fields influenced advancement in AI—for example from computer science and nanotechnology. We have also seen—throughout the history of the theory—various products leveraging AI technology, such as advanced robotics, autonomous cars, and intelligent computers [17].

### *B. Current State of AI*

In recent years, there have been major advances in developing the research and technology of AI into performant products with a focus on several technologies. *Natural*

*Language Processing* (NLP)—or *speech recognition*—is one such technology, in which AI algorithms can understand human language and take an action in response. *Autonomous vehicles* is another focus area, with companies like Google—among others—making major advancements towards commercializing such technologies. Large information technology-focused companies such as Google, IBM, and Facebook are investing huge amounts of money in AI R&D [22]. Artificial intelligence technology has already been used in *mission-critical control system environments* such as space travel [23]. Indeed, Artificial Intelligence can be used to program a computer to acquire knowledge, to understand language such as English, and to translate language [24]. It is also used in generating models for the way humans learn, themselves.

*C. Future State of AI*

When reviewing the possibilities of future AI technology developments, some research suggests a substantial chance of a superintelligent artificial intelligence arising within a few decades [9]. However, when looking into the future of AI, some doubts can also be raised: can this intelligence understand the human experience, eye movement, and facial expressions? Will it be able to make sense of social cues such as levels of human interpersonal tension [24]? Despite these doubts—based on a survey by Groves about the use of artificial intelligence and robotics in industry—AI is expected to become a big part of daily life in the 2020s [24]. In some jobs these technologies are expected to be a significant and effective part of the economy. Other jobs will not see significant changes, however the nature of such work may change with automation [24]. Some kinds of jobs are expected to become difficult for humans because the education system does not qualify people to do certain types of intensive work. So, the need for automation will be necessary for success in these lines of work [24]. On other hand, areas where machine intelligence has powerful utility functions—such as processing information, storage, and recollection—cannot occur without the development of strong artificial intelligence systems [25]. Another necessity of AI is space travel, since the risk of failure due to human error is greater without autonomous systems [26]. AI technologies might also play a significant role in health care, including mental health care [27].

In summary, today we are surrounded by AI technology whether we realize it or not. In evaluating this technology's management implications, we must consider the present state: where the technology's use is finding widespread application. We must also consider a future in which—for example—AI possesses or exceeds the effective power and speed of humans: becoming a smarter Intelligence than we. In doing so, we may see a future where humans become less relevant [24]. We will also see cases where cross-disciplinary approaches—like those seen in the early days of AI research—may become more important in guiding the technology's development and implementation.

## V. REVIEW OF TECHNOLOGY ETHICS

*Technology Ethics* is a field examining ethics from a technology management perspective which allows us to identify management issues related to a technology. Betz argues that Technology Ethics becomes a topic of concern after a technology has been developed and it is ready for commercialization and marketing, as represented in Fig. 1 [18]. Here we provide a brief context of the field of Technology Ethics, focusing on the *Technology Assessment* tool and its ethical derivatives, and provide further context through the subdomain of *Risk Management*. Following this review, we draw from an understanding of AI technology to identify potential management issues during these stages.

*A. Literature- Technology Ethics Assessment Models*

Presently, technology's role in society is experiencing a shift: it is now more than just a tool for increasing convenience in our lives. It has become a necessity that touches and changes every aspect of human life. Technology has changed the way society behaves. Every industry and sector has been fundamentally reshaped by technology.

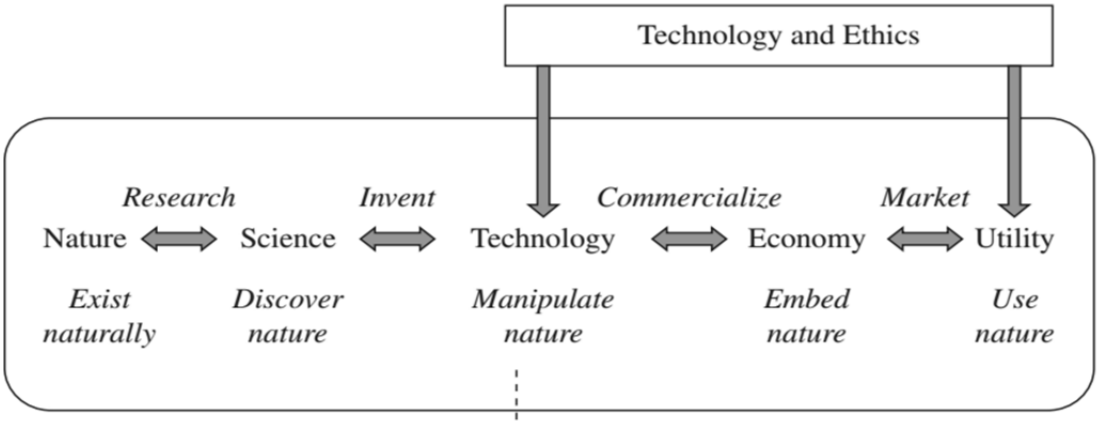


Fig. 1 [18]

Complexity and rapid development of technology generates many ethical problems. According to Moor, the computer/information revolution has only increased ethical problems. Identity theft and the solicitation of children by child molesters are examples of ethical issues that have soared during the proliferation of computers and widespread computer networks. Moreover, during this stage, the ability to own computers and share files has resulted in further ethical problems pertaining to intellectual property and privacy [28].

The dilemma becomes even more problematic as technology develops, and Technology Assessment is one way to evaluate such issues. The development of technologies for infertility treatment, genetically modified organisms, and stem cells have also presented emerging ethical and social issues [29] and have stirred public controversy. Indeed, the demands of greater social responsibility in technology development make it necessary to shape technology assessment in response to such concerns. Arguably, the first use of *Technology Assessment* as a term arose in the 1960s in the U.S. [30] and has been defined as "an applied process that considers the societal implications of technological change in order to influence policy to improve technology governance" [31]. The Office of Technology Assessment in 1976 described Technology Assessment as "a comprehensive form of policy research that examines the short- and long-term social consequences (e.g., societal, economic, ethical, legal) of the application or use of technology" [32].

Since then, Technology Assessments have seen widespread use, and their advocacy has spread beyond the basic assessment of technology. Today they also contribute to: (1) the diffusion of technology, (2) factors leading to rapid acceptance of new technology, and (3) the role of technology and society [30]. Technology Assessment has seen many advances and according to Grunwald, "its approaches have been developed and are practiced to a certain extent. All of them have a specific focus, particular theoretical foundations, different rationales, and have been conceptualized for meeting differing challenges and context conditions" [33].

## VI. THEMATIC REVIEW OF AI ETHICS

Many theorists have differing ideas about AI Ethics and there exists a broad spectrum of views of how to assess the technology and its issues. We now overview the ethical implications of artificial intelligence in the literature. We overview *theories* in literature, then describe examples of *risks* and associated *roles* concerned. Some theories are of a practical nature—they recommend a specific course of action—whereas others describe the technology along various gradients, such as risk scope and intensity. From this we will discriminate *existential risks* and *non-existential risks*, and roles along two axes: *agent* versus *patient* and *natural life*, including human life or other life on Earth versus *artificial life*, including artificial intelligence, artificial agents, etc. Appendix A describes some basic concepts related to technology risk identification, and management in general.

First, let us remark that the literature variously refers to the subject of our research as roboethics, machine ethics, Robot Rights, and others. It is also represented within the general fields of computer science and ethics; as well as domain-specific fields of artificial intelligence, artificial life (ALife), and agent theory as ethics, safety engineering, moral theory, jurisprudence, and others [42][38][43][44][45]. These factors pose a challenge in terms of providing a holistic view of the literature. Though we do not attempt to encapsulate the entirety of the theory across all domains in this paper, we do provide a functional background for our later discussion, based on the sources we did identify.

### *A. Literature- Theoretical AI Ethics*

Several theories provide prescriptive—or practical—advice, whereas others describe a gradient of possible outcomes and approaches. This practical advice helps to demonstrate the boundaries of the overarching theory (which also include the unstated counterarguments to the theories listed in this section). This is useful insofar as some research has proposed making AI Ethics a mainstream subject to be addressed in a broad range of mainstream academic journals and conferences [38]. Yet within the subject of Engineering and Technology Management (ETM)—or Management of Technology (MOT)—others have found only a small subset of existing literature addresses the topic of Artificial Intelligence at all, let alone the broad ethical implications of the technology [46]. Given the broad range of topics within the theory of ethics in artificial intelligence, we hope that describing the contextual boundaries of the theory will help provide engineering managers and policymakers a framework with which to understand future developments in the technology.

The first theory we describe is a prescriptive one: research should treat AI ethics as a mainstream subject. It suggests that AI theoreticians and philosophers must not be the only researchers in this field, and that it is a topic that should be addressed by computer scientists as well as a domain in its own right [38]. The authors of this theory suggest that increased acceptance of AI ethics as a mainstream subject will increase the opportunity for mainstream academic publication. This has been demonstrated to be the case, as today specialized peer-reviewed publications and conferences on the topics exist, but we also see the treatment of topics such as Roboethics, Machine Ethics, and Cyborg Ethics in mainstream prestigious publications [38][47]. We propose extending this beyond computer science and into the technology innovation and management literature, just as it has already begun to establish itself within computer science. (We could propose an opposing view—that AI ethics be subjugated to domain-specific research—but this is antithetical to the clear benefits of actively managing a technology's research, development, productization, and marketing [18], as well as the historical context of the technology's diverse origins in its early days.)

The second class of prescriptive literature addresses the methods by which we safely develop the technology, and here we find our first split of opinion: whether or not embedding ethical decision-making into an artificial agent is sufficient. One domain of research suggests that AI ethics must be managed within the context of *Safety Engineering*, or, "developing safety mechanisms for self-improving systems" [38]. This contrasts with the *Machine Ethics* approach, which proposes embedding a capacity for ethical decision-making into artificial intelligence implementations. Criticisms against the latter approach include (1) many such papers have been criticized as being purely philosophical, (2) they have focused on *non-universal* moral or ethical norms and codes, (or those norms and codes which we humans cannot agree upon ourselves) and (3), if we design our machines to match human levels of ethical decision-making, such machines would then proceed to take some immoral actions (since we humans have had occasion to take immoral actions ourselves). AI Safety Engineering theory, however, suggests several practical approaches, such as: (1) developing systems capable of proving that they are safe, (2) confining AI from the outside world to prevent its interaction with (and manipulation of) it, and (3) testing through limited AI development to improve our security protocols. This may pose several ethical issues, as we will see later.

Third, we note literature—which gives us the domain termed *Robot Rights*—addressing the rights of the AI itself as we develop and implement it. We find arguments against [38] the affordance of rights for artificial agents: that they should be equals in ability but not in rights, that they should be inferior by design and expendable when needed, and that since they can be designed not to feel pain (or anything) they do not have the same rights as humans. On a more theoretical level, we find literature asking more fundamental questions, such as: at what point is a simulation of life (e.g. artificial intelligence) equivalent to life which originated through natural means [43]? And if a simulation of life is equivalent to natural life, should those simulations be afforded the same rights, responsibilities and privileges afforded to natural life or persons? Some literature suggests that the answer to this question may be contingent on the intrinsic capabilities of the creation, comparing—for example—animal rights and environmental ethics literature.

Fourth, we note the topic of *Human Rights* and *AI Jurisprudence*: those topics which address an equitable balance of rights and responsibilities between humans and AI agents. We find literature that proposes [38] that early artificial intelligence should be built to be safe and law-abiding, and that later artificial intelligence (that which surpasses our own intelligence) must then respect the property and personal rights afforded to humans. (And that they must use laws affording equal rights to both themselves and humans, such that one party is not disenfranchised at the expense of the other.) Further, we find literature [48] which addresses the topic of jurisprudence within AI technology, for example addressing the limitations of current regulations regarding industrial robots as they apply to autonomous agents. When considering legal frameworks, we note that at present no such framework has been identified in literature which would apply blame and responsibility to an autonomous agent for its actions. (Though we do suggest that the recent establishment of laws regarding autonomous vehicles may provide some early frameworks that can be evaluated for efficacy and gaps in future research.) Frequently the literature refers to existing liability and negligence laws which might apply to the manufacturer or operator of a device.

A sixth theory can help us categorize the risks we hope to evaluate, as well as other theories in the literature: the theory of *moral agents* and *moral patients* [44]. Within this theory, a moral agent is the one who makes an action with ethical implications against a second party—the patient—who receives the action. If we categorize agents at a Level of Abstraction including both humans and artificially intelligent agents, then we can categorize risks by (1) agent (*human* versus *AI*) and (2) rôle (*actor* versus *patient*). In this model, we can categorize (1) Safety Engineering, Machine Ethics, and Human Rights as those which protect human (and in some cases, other AI) patients from actions by AI actors, (2) Robot Rights as that which protects AI patients from actions by human (and in some cases, other AI) actors, and (3) AI Jurisprudence as that which equitably protects AI or human patients from AI or human actors.

A seventh subclass seeks to classify Artificial Intelligence as an *intrinsically good* or *intrinsically evil* technology (according to the definitions in [18]). Here we see a split between Domain-Specific AI technology and Artificial General Intelligence technology. In the former case we see support, [38] that such technology's uses, such as mail sorting and spellchecking, provide practical benefit and do not present existential risks, and are thus ethical or intrinsically good. In the latter case, we see opposition, [38] [49] [47] that systems capable of general intelligence surpassing human capacity are unethical or intrinsically evil due to (1) its existential threat to humanity (such as economically outcompeting us or making the human race unnecessary) and (2) the potential suffering of AI systems during our development of them. Some literature [49] [47] suggests that such technologies should not be developed at all. Alternatively, some literature supports the development of AGI due to its perceived risk/reward ratio, and that a general artificial super-intelligence poses risks and benefits that must be properly managed [9]. Some even propose that such a general superintelligence may be inevitable [37]. In the case where experimentation does occur, recommendations [38] suggest the establishing of AI Research Review Boards (such as the one formed [50] at Google) similar to medical research boards, with the authority to restrict funding and enact partial or complete bans on certain activities or research, but perhaps provide an exception for development of safety measures and control mechanisms for AGI architectures. In this topic we could foresee considerable

further debate, and find that it would be useful to characterize such debate as applying to two domains: (1) Domain-Specific AI technology versus Artificial General Intelligence, and classification of the technology as (2) intrinsically good or intrinsically evil, as defined in [18].

Lastly, we touch on an eighth subclass, which addresses the somewhat more benign pitfalls of AI technology's development—those which may result in the success or failure of a given project. A number of pitfalls have been identified [51] in developing such solutions, which can help inform the development and commercialization process. (For example, (1) taking care not to oversell AI technology's capabilities; given that the achievements and failures of Artificial Intelligence research are well-known and documented [37], it is possible to inform future commercialization through an understanding of the present state of technology, and (2) not relying on overly ambitious forecasts of the past.) It can be argued that a failed development may cause, at worst, a catastrophic outcome, such as one of the risks described below and at least, the delay of the technology's implementation or adoption. Depending on that outcome, if we take AI's development to be inevitable [37], it follows that in either case, the proper management of the technology bears ethical implications because its mismanagement may either result directly in the risks presented, or may result in those risks indirectly, if competing actors successfully develop such technologies in less-regulated environments that may not anticipate such ethical issues.

We remind the reader at this point that this list is not intended to be an authoritative or exhaustive declaration of the scope of AI Ethics research, but instead a starting point for future classification and research, and that a vast domain of knowledge in this topic has accumulated over the years.

*B. Literature- Applied AI Ethics: Development Risks*

The potential or actual risks of artificial intelligence technology have been described in the literature. Several approaches are used to categorize and classify such risks. Due to several challenges with these sorts of risk assessment approaches, we instead simply provide a review of risks and provide minimal classification where available. Here we will draw from the literature to review risks that have materialized or which may arise throughout the future development process.

In terms of classification approaches, Bostrom [38] describes a form of a *risk matrix* (which in traditional literature [2] encompasses two axes: (1) *probability* and (2) *severity*, with severity further comprised of (a) *intensity* (e.g. global, local, personal) and (b) *scope* (e.g. endurable, terminal). Bostrom focuses on one such intersection (global intensity+terminal scope) to discriminate between (1) *existential* or (2) *non-existential* risks generally. Other authors suggest [52] flaws with this methodology, proposing that risk scope can be further delineated between *Temporal Scope* (a gradient between Generational and Transgenerational) and *Spatial Scope* (the Bostrom scale of Personal, Local, and Global). Although such risk matrices are common in many fields of research, in cases where quantitative data is limited or unavailable, substantial challenges in their use arise when comparing risks that differ significantly in their categorizations [53]. Given these drawbacks, we will constrain ourselves to differentiating existential and non-existential risks, providing general categories of risks, and specific examples or cases where available. We will also reflect on the theoretical AI models described above where appropriate.

*C. Literature- Existential AI Risks*

Among existential risks, we find four main categories in the literature: (1) *Unethical Decision-Making*, (2) *Direct Competition*, (3) *Death of AI*, (4) and *Unpredictable Outcomes*.

An artificial agent may or may not be developed with the capacity for moral reasoning, and so risks developing a capacity for unethical decision-making. If, for example, an agent was programmed to operate war machinery in the service of its country, it would need to make ethical decisions regarding the termination of human life. This capacity to "make non-trivial ethical or moral judgments concerning people" [38] again poses issues for Human Rights, but through this example we can also identify moral agents (AI) and patients (people). Using this classification, the topics of Safety Engineering, Machine Ethics, Human Rights, and AI Jurisprudence become available to us to understand and plan for possible scenarios. From a risk management perspective, it is important to note that as technologies such as military drones continue to be developed, it will be important to continually reassess the technology's risks in order to mitigate them. Certain international legal constructs may also need to be established which can mitigate the low barrier to entry in sending autonomously-operated machines to war (as opposed to humans operating them remotely). We must also assess whether the risks can be managed, or whether such technologies must be categorized as intrinsically evil and outlawed at an international level.

One or more artificial agent(s) could have the capacity to directly outcompete humans, for example through capacity to perform work faster, better adaptation to change, vaster knowledge base to draw from, etc. [38]. In this case, human labor may become more expensive or less effective than artificial labor, resulting in redundancies (or extinction) of the human labor force. Many researchers think that AI would be a substitute for human labor in many occupations, but they do not know the timeline of this taking place. If such a scenario arises in the next few decades, it is not clear that human workers could retrain quickly enough to maintain high levels of employment [41]. It has been proposed that although human agents may delegate such work to artificial agents, this may not be a volitional act: since artificial agents may be better suited to decision-making, humans may drift towards a state of dependence as they gradually hand over more and

more work to the artificial agents. In this instance, the topic of Human Rights provides us some fodder for posing ethical questions—for example: is it fair for human agents to be excluded from the workforce due to their inefficiency? In this case, it is difficult to ascertain the agent/patient roles—did humans volitionally give this power over, or was it taken by force? Evaluating such scenarios in detail would permit us to plan for them.

The literature suggests that throughout the development of an AI we may go through several generations of agents which do not perform as expected [37] [43]. In this case, such agents may be placed into a suspended state, terminated, or deleted. Further, we could propose scenarios where research funding for a facility running such agents is exhausted, resulting in the inadvertent termination of a project. In these cases, is deletion or termination of AI programs (the moral patient) by a moral agent an act of murder? This, an example of Robot Ethics, raises issues of personhood which parallel research in stem cell research and abortion. Humans may not be the only moral agents in this case, as some research [37] has proposed a *singleton*: an artificial agent which recognizes the possibility of competition and decides to eliminate that competition before it poses a threat. This then includes the issues of Unethical Decision-Making, drawing once more upon Safety Engineering, Machine Ethics, and AI Jurisprudence.

There are numerous areas of humanity's existence that may be impacted by the arrival of artificial agents. Our culture, lifestyle, and even probability of survival may change drastically [38]. Because the intentions programmed into an artificial agent cannot be guaranteed to lead to a positive outcome [37][54], Machine Ethics becomes a topic that may not produce guaranteed results, and Safety Engineering may correspondingly degrade our ability to utilize the technology fully (if, for example, we place such stringent controls on the agent that it is capable only of yes/no answers and never autonomously performing tasks on our behalf).

*D. Literature- Non-Existential AI Risks*

Face recognition technologies and their ilk pose significant privacy risks [47]. For example, we must consider certain ethical questions like: what data is stored, for how long, who owns the data that is stored, and can it be subpoenaed in legal cases [42]? We must also consider whether a human will be in the loop when decisions are made which rely on private data, such as in the case of loan decisions [37]. In this case, we can look to safety engineering and machine ethics when designing such an AI: to produce an AI that performs its utility function in an ethical manner, and which is able to provide a rationale for its decisions to ensure its compliance with ethical and legal standards.

Autonomous care systems for the elderly or children can provide dignity and independence for those individuals [55][45], but discrepancies between caste/status based on intelligence may lead to undignified parts of the society—e.g. humans—who are surpassed in intelligence by AI. Despite the potential benefits, we again find cases where risk can be managed so that the costs do not outweigh those benefits. We may also consider human rights and legal frameworks to ensure rights and equity among both natural and artificial Intelligences.

In order to preserve human property rights and legal rights, certain controls must be put into place. If an artificially intelligent agent is capable of manipulating systems and people, it may also have the capacity to transfer property rights to itself or manipulate the legal system to provide certain legal advantages or statuses to itself [38]. This situation requires controls in the form of machine ethics or safety engineering so that we can develop systems which behave in an equitable or ethical manner.

Liability and negligence are legal gray areas in artificial intelligence. If you leave your children in the care of a robotic nanny, and it malfunctions, are you liable or is the manufacturer [45]? We see here a legal gray area which can be further clarified through legislation at the national and international levels; for example, if by making the manufacturer responsible for defects in operation, this may provide an incentive for manufactures to take safety engineering and machine ethics into consideration, whereas a failure to legislate in this area may result in negligently-developed AI systems with greater associated risks.

Animals raised by surrogate parents have been studied and have been found to demonstrate abnormal behavioral patterns as a result of their unusual upbringing. We can compare the hypothetical situation of children being raised by robotic nannies to such studies [45]. A cross-disciplinary approach may be warranted in these sorts of uses. For example, embedding child psychology research and knowledge into the AI as a part of its utility function, which could be viewed through the lenses of safety engineering and machine ethics. (And raises the question—if we embed behaviors such as how to raise children, is this machine ethics or safety engineering?)

Proposals exist for autonomous war vehicles targeting and killing humans without direct human control or intervention [45]; correspondingly, solutions for some (but not all) issues may be found in existing legal frameworks, such as the 1944 Geneva Convention definitions on combatants [45], but further legal frameworks may need to be developed in order to address the myriad of safety and security issues of AI.

A sufficiently intelligent AI could possess the ability to subtly influence societal behaviors through a sophisticated understanding of human nature [37][54]. In these cases, such AI may need to be able to prove it is behaving in an ethical manner, such as developing methods to demonstrate its intentions.

Because a single human actor controlling an artificially intelligent agent will be able to harness greater power than a single human actor, this may create inequalities of wealth [54][37]. Some suggest that development of AI should occur in the open so as to level the playing field [54]. This raises the spectre of AI Ethics Boards akin to those in medical

boards, as well as issues of human rights and legal frameworks to provide continuity for human quality of life.

Surveillance could be ubiquitous and controlled by those who own the AI agents, and so used for the benefit of those owners and the detriment of those they might surveil. [54] In this case we find human actors maligning AI against human patients, using the AI as a tool, in which case legal frameworks may need to be adapted to ensure that human actors are held accountable for their use of the technology.

Errors in domain-specific trading algorithms cost one trading company $440 million [54]. The potential financial costs may be even more substantial if an AGI enters financial markets where errors can occur in under a second due to the participation of these sorts of automated trading algorithms. Combined with the ubiquitous surveillance risk, we could envision a scenario where an AGI gains intelligence on the operation of the domain-specific trading algorithms, and uses that to manipulate markets intentionally. Here we find another scenario where a combination of approaches could be used to manage the risks.

## VII. ANALYSIS

Beyond the thematic and chronological reviews provided above, we applied a PEST perspectives analysis to analyze the AI Ethics literature, synthesizing the literature's contextual background and recommendations. PEST analysis looks at the political, economical, social, and technological factors to minimize the gap between expectation and performance of a given scenario.

### *A. Analysis - Literature Findings*

Here we examined the most commonly cited ethical issues of AI. In doing so, we found that those issues are discussed with varying foci. First, we categorized these ethical issues as (1) those related to domain-specific AI technology; and (2) those related to artificial general intelligence technology. Second, we categorized them as focusing on (1) (a) the ethical issues related to AI technology's effects on humans and other living beings versus (b) the ethical/risks issues related to AI technology itself; and (2) (a) existential risks versus (b) non-existential risks of AI technology. Table 1 summarizes our analysis.

### *B. Analysis- PEST*

Here we classified the ethical issues derived from the literature into macro-environmental factors through the use of a PEST (Political, Economic, Social, Technological) [5] analysis.

**Political Factors:**

- **No Governance-** No current national/international regulations govern AI. AI developers are under no control or limitations (beyond basic liability and criminal legal frameworks), resulting in gaps that could be exploited as the technology advances.
- **Accountability-** There is ambiguity as to whom is held accountable for the decisions of AI systems: for example, if an autonomous car hits a person, who will be responsible.
- **Military Control-** AI is considered a valuable military asset: many governments fund AI projects heavily for military purposes [56]; similarly, civilian advances in the technology could be placed under state control if substantial uncontrolled advances are made.
- **Loss of Control-** The possibility exists of losing control of self-improving systems: those systems might decide that humans are obsolete and eradicate us or ignore our rights and needs.

TABLE 1: ETHICAL ISSUES OF AI

| | **Ethical Issues** | | |
|---|---|---|---|
| | **Effects on humans and other living beings** | | **AI technology itself** |
| | **Existential risks** | **Non-existential risks** | |
| **Domain-Specific AI** | - Unethical decision making | - Privacy<br>- Human Dignity/ Respect<br>- Decision making transparency<br>- Safety<br>- Law abiding<br>- Inequality of Wealth<br>- Societal Manipulation | - AI Jurisprudence<br>- Liability and Negligence<br>- Unauthorized manipulation of AI |
| **AGI (Artificial General Intelligence)** | - Direct competition with humans<br>- Unpredictable Outcomes | - Competing for jobs<br>- Property/Legal Rights | - AI rights and responsibilities<br>- Safety mechanisms for self-improving systems<br>- Human like immoral decisions<br>- AI death |

**Economical Factors:**

- **Labor Competition-** AI agents may compete against humans for jobs, though history shows that when a technology replaces a human job, it creates new jobs that need more skills. (e.g.: a machine that assembles a TV in a factory eliminates a human assemblers' job, but creates technical jobs to operate and maintain the machine and new engineering jobs in the the company that built such machines.)
- **Human Labor Obsolescence-** On the other hand, if self-improving AI systems emerged and they were very fast in upgrading and enhancing themselves, it is possible that humans would not be able to adapt quickly enough and would become irrelevant.
- **Greater Production-** AI technologies can allow safer, faster, and more efficient mass production and services which means cheaper, higher-quality products. Humans have a certain capacity and cost: they need to rest, eat, and get paid. AI systems may not have the same requirements in doing the same job, which could mean higher production in less time making the goods cheaper.

**Social Factors:**

- **AI Interactions with Humans-** Social, cultural, and other issues may arise.
- **Privacy-** Should an AI be permitted unmitigated access to our information to improve its function?
- **Human Dignity/ Respect-** What impacts on humanity's labor, social, and legal rights are tolerable?
- **Decision Making Transparency-** We face significant challenges bringing transparency to artificial network decisionmaking processes. Will we have transparency in AI decisionmaking?
- **Safety-** Are AI safe with respect to human life and property? Will their use create unintended or intended safety issues?
- **AI Consciousness-** Is deleting AI murder? Is it morally acceptable to turn one off? When and why?

**Technology Factors:**

- **Misuse-** AI machines could be hacked and misused, e.g. manipulating an airport luggage screening system to smuggle weapons.

## VIII. CONCLUSION AND RECOMMENDATIONS

As part of managing any emerging technology, the ethical (and thus risk) issues must be considered to ensure that the technology is safe and not intrinsically evil in nature (otherwise warranting absolute prohibition and control). Artificial Intelligence (AI) is an emerging technology that has risks and ethical concerns that—if managed well—can lead to better quality of life for human society.

Our review of the ethical issues related to AI (in terms of (1) AI's interaction with humans and other life and (2) AI itself) demonstrated that varied perspectives and some debate exists. To address the subject well, the first step is to identify the most important ethical aspects of AI, then how to manage them. In this paper, our PEST analysis of the literature identified some of the most important ethical aspects of AI.

Many issues were identified, with a focus on to two types of AI: domain specific AI systems and general AI systems. The ethical issues address both existential and non-existential risks that both types of AI can pose to humans and other living beings. Further, there are ethical and risk issues that are related to the AI systems themselves, related to their regulation and treatment.

Based on that, we identified several management recommendations for better management of the ethical aspects of AI in terms of:

**National and International Ethics Committees-** Although varied technology assessment tools and methodologies can be used for conducting ethical analysis within a technology, social, legal, and value conflicts still emerge. So, there is a need for national and international legislation and regulation to govern the ethical aspects of AI similar to that of, for example, the pharmaceutical industry. National and international ethics committees should be established with the goals of (1) developing, refining, and revising the principles, regulations, and ethics frameworks in regards to new dynamics or emerging ethical challenges pertaining to AI, and (2) provide such a standard to AI technology developers.

**Organizational Advisory Boards on AI Regulations-** Simultaneously, organizations focused on AI research should form advisory boards overseeing AI development projects. They should serve in an advisory function for managers to be able to assess and decide whether or not the technology's development poses ethical issues the pose risks, and hence advise for corrective actions in their development.

**Managers' Technology Assessment and Risk Management Knowledge-** In order to be more proactive and less reactive in embedding ethics in AI, managers should assess the ethical issues related to AI in their organizations, deciding on issues like privacy, control, ownership, accuracy, security and so on. To be able to do so, managers need to extend their expertise in terms of technology assessment practice and risk management methodologies (including an understanding of the limitations of applying these methodologies).

**Team's Ethics Awareness-** Adequate ethics education and training are necessary for engineers, technologists, and researchers to be aware of ethical issues during the research and development process to mitigate the risks that may arise as a result of AI implementations.

**AI Systems Security-** AI systems security must be addressed thoroughly when building AI systems.

**Failsafe Mechanism-** When building a general purpose AI system that can make its own decisions, a mechanism to cease the system's operation should be built in the event that control over the system is lost and it begins to make unethical or risk-prone decisions. However, it should also be clear who

has access to such mechanisms and under which circumstances they may be activated.

**Gaps-** In closing, we identified several gaps in the literature:

1. AI research does not fully draw from other fields of research, such as Engineering and Technology Management, which could further inform the field.
2. AI and AI Ethics literature is spread across many domains of research, and goes by many names.
3. Insufficient research has been done to evaluate recent jurisprudence issues related to, e.g., autonomous vehicles.
4. There are numerous technical challenges left to be solved surrounding AI ethics, as demonstrated by the split between Machine Ethics and Safety Engineering.
5. Little has been done to propose concrete recommendations for management personnel working in this field; the majority of the literature is technical or philosophical, and does not provide clear guidance.

## ACKNOWLEDGEMENTS

This paper is based on a class project carried in Engineering and Technology Management Department of Portland State University, USA. The class project was done by Taylor Meek, Husam Barham, Nader Beltaif, Amani Kaadoor, Tanzila Akhter, and Bashair Al Saglab in Fall, 2015 as part of the ETM 520/620 class taught by Professor Dundar Kocaoglu.

## REFERENCES

[1] Dictionary.com, "Artificial-Intelligence," *Dictionary.com Unabridged. Source location: Random House, Inc. http://dictionary.reference.com/browse/artificial-intelligence*. [Online]. Available: http://dictionary.reference.com. [Accessed: 12-Oct-2015].

[2] "artificial intelligence, n.," *OED Online*. Oxford University Press.

[3] Dictionary.com, "Ethics | Define Ethics at Dictionary.com," *Dictionary.com Unabridged. Source location: Random House, Inc. http://dictionary.reference.com/browse/artificial-intelligence*. [Online]. Available: http://dictionary.reference.com/browse/ethics. [Accessed: 13-Oct-2015].

[4] Writing@CSU, "Content Analysis - Relational Analysis." [Online]. Available: http://writing.colostate.edu. [Accessed: 07-Oct-2015].

[5] J. Makos, "What Is PEST Analysis and Why it's Useful," *PESTLE Analysis*, 22-Aug-2014. [Online]. Available: http://pestleanalysis.com/what-is-pest-analysis/. [Accessed: 07-Oct-2015].

[6] Writing@CSU, "Types of Content Analysis." [Online]. Available: http://writing.colostate.edu. [Accessed: 07-Oct-2015].

[7] P. Mayring, "Qualitative Content Analysis," *Forum Qual. Sozialforschung Forum Qual. Soc. Res.*, vol. 1, no. 2, Jun. 2000.

[8] D. Lemire, "General versus domain intelligence," *Daniel Lemire's blog.* .

[9] N. Bostrom, "Ethical issues in advanced artificial intelligence," *Int. Inst. Adv. Stud. Syst. Res. Cybern.*, vol. Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence. Windsor, no. CH2, pp. 12–17, 2003.

[10] IBM.com, "IBM100 - Deep Blue," 07-Mar-2012. [Online]. Available: http://www-03.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/. [Accessed: 02-Nov-2015].

[11] R. Kurzweil, *The Singularity is Near*. Penguin Books, 2005.

[12] N. Bostrom and Y. Yodkowsky, "The Ethics of Artificial Intelligence," in *The Cambridge Handbook of Artificial Intelligence*, 2014th ed., Cambridge University Press, pp. ch. 15, pp. 316–334.

[13] B. Buchanan, "A (Very) Brief History of Artificial Intelligence," *AI Mag.*, vol. 26, pp. 53–60, Nov. 2006.

[14] S. Jürgen, "New millennium AI and the convergence of history," *Chall. Comput. Intell.*, pp. 15–35, 2007.

[15] A. Sloman, "What is Artificial Intelligence?," *chool of Computer Science, The University of Birmingham*, 04-Jul-2007. [Online]. Available: http://www.cs.bham.ac.uk/research/projects/cogaff/misc/aiforschools.html. [Accessed: 03-Nov-2015].

[16] P. McCorduck, M. Minsky, O. Selfridge, and H. Simon, "History of Artificial Intelligence," *IJCAI*, pp. pp. 951–954, 1977.

[17] J. Schmidhuber, "2006: Celebrating 75 years of AI - History and Outlook: the Next 25 Years," *ArXiv07084311 Cs*, Aug. 2007.

[18] F. Betz, "Ethics and Technology," in *Managing Technological Innovation*, 3rd ed., Hoboken, NJ: Wiley, 2011, pp. ch. 15, pp. 299–316.

[19] T. Urban, "The AI Revolution: Road to Superintelligence - Wait But Why," 22-Jan-2015. [Online]. Available: http://waitbutwhy.com/2015/01/artificial-intelligence-revolution-1.html. [Accessed: 03-Nov-2015].

[20] E. Sandewall, "A perspective on the early history of artificial intelligence in Europe," *AI Commun.*, vol. 27, no. 1, pp. pp. 81–86, 2014.

[21] AITopics.com, "Brief History | AITopics," *AITopics.com*. [Online]. Available: http://aitopics.org/misc/brief-history. [Accessed: 16-Nov-2015].

[22] F. Forrest, "10 things you need to know about artificial intelligence," *TechRepublic*, 06-Jul-2015. [Online]. Available: http://www.techrepublic.com/article/10-things-you-need-to-know-about-artificial-intelligence/. [Accessed: 23-Oct-2015].

[23] S. Chien et al, "The Future of AI in Space, Intelligent Systems," *IEEE*, vol. 21, no. 4, pp. 64–69, 2006.

[24] I. Groves, "The future of artificial intelligence," *Business Insights: Global*, 16-Aug-2014.

[25] P. Norvig, "Artificial intelligence: A new future," *New Sci.*, vol. 216, no. 2889, pp. vi–vii, Nov. 2012.

[26] G. Dorais, R. Bonasso, P. Kortenkamp, and D. Pell, "Adjustable autonomy for human-centered autonomous systems," presented at the Working notes of the Sixteenth International Joint Conference on Artificial Intelligence Workshop on Adjustable Autonomy Systems, 2009, pp. 16–35.

[27] D. Luxton, "Artificial intelligence in psychological practice: Current and future applications and implications," *Prof. Psychol. Res. Pract.*, vol. 45, no. 5, Oct. 2014.

[28] J. H. Moor, "Why We Need Better Ethics for Emerging Technologies," *Ethics Inf. Technol.*, vol. 7, no. 3, pp. 111–119, Sep. 2005.

[29] N. Al-Rodhan, "The Many Ethical Implications of Emerging Technologies," *Scientific American*, 13-Mar-2015. [Online]. Available: http://www.scientificamerican.com/article/the-many-ethical-implications-of-emerging-technologies/. [Accessed: 06-Nov-2015].

[30] D. Banta, "What is technology assessment?," *Int. J. Technol. Assess. Health Care*, vol. 25, no. S1, p. 7, Jul. 2009.

[31] A. W. Russell, F. M. Vanclay, and H. J. Aslin, "Technology Assessment in Social Context: The case for a new framework for assessing and shaping technological developments," *Impact Assess. Proj. Apprais.*, vol. 28, no. 2, pp. 109–116, Jun. 2010.

[32] OFFICE OF TECHNOLOGY ASSESSMENT, *Development of Medical Technology: Opportunities for Assessment*. Congress of the United States, 1976.

[33] A. Grunwald, "Technology Assessment for Responsible Innovation," in *Responsible Innovation 1 - Innovative Solutions for Global Issues*, vol. 2, Springer, 2014.

[34] A. Genus and A. Coles, "On Constructive Technology Assessment and Limitations on Public Participation in Technology Assessment," *Technol. Anal. Strateg. Manag.*, vol. 17, no. 4, pp. 433–443, Dec. 2005.

[35] S. Saarni and etl, "Ethical analysis to improve decision-making on health technologies," *Kennedy Inst. Ethics J.*, vol. 16, no. 4, pp. 333–351, 2006.

[36] E. Palm and S. O. Hansson, "The case for ethical technology assessment (eTA)," *Technol. Forecast. Soc. Change*, vol. 73, no. 5, pp. 543–558, Jun. 2006.

[37] N. Bostrom, *Superintelligence*. New York: Oxford University Press, 2013.

[38] R. Yampolskiy, "Artificial Intelligence Safety Engineering: Why Machine Ethics Is a Wrong Approach." 2012.

[39] G. Stoneburner, A. Goguen, and A. Feringa, " Risk Management Guide for Information Technology Systems." National Institute of St andards and Technology, Jul-2002.

[40] K. Sotala and R. Yampolskiy, "Responses to Catastrophic AGI Risk: A Survey." M ACHINE INTELLIGENCE RESEARCH INSTITUTE, 2013.

[41] S. Russell and B. Petermeier, "AI: how can we manage robot risk?," *Agenda - The World Economic Forum*. .

[42] J. P. Sullins, "Introduction: Open Questions in Roboethics," *Philos. Technol.*, vol. 24, no. 3, pp. 233–238, Sep. 2011.

[43] M. Bedau, *Philosophical aspects of artificial life*. Basil Blackwell.

[44] L. Floridi and J. Sanders, "On the Morality of Artificial Agents," *Minds Mach.*, vol. 14, no. 3, pp. 349–379, Aug. 2004.

[45] N. Sharkey, "COMPUTER SCIENCE: The Ethical Frontiers of Robotics," *Science*, vol. 322, no. 5909, pp. 1800–1801, Dec. 2008.

[46] P. Lopes, O. Kumiko, Kissimoto, and S. Mário, "Innovation management: a literature review about the evolution and the different innovation models," presented at the ICIEOM 2012- XVIII International Conference on Industrial Engineering and Operations Management, 2012.

[47] B. McKibben, *The end of nature*, Random House trade pbk. ed. New York: Random House Trade Paperbacks, 2006.

[48] Y. Weng, C. Chen, and Chuen-Tsai Sun, *The Legal Crisis of Next Generation Robots: On Safety Intelligence*. New York, NY: ACM, 2007.

[49] B. Joy, "Why the Future Doesn't Need Us," *WIRED*, 01-Apr-2000. [Online]. Available: http://www.wired.com/2000/04/joy-2/. [Accessed: 02-Nov-2015].

[50] D. Zeng, "AI Ethics: Science Fiction Meets Technological Reality," *IEEE Intell. Syst.*, vol. 30, no. 3, pp. 2–5, May 2015.

[51] M. Wooldridge and N. R. Jennings, "Pitfalls of agent-oriented development," 1998, pp. 385–391.

[52] P. Torres, "A New Typology of Risks," *IEET*, 05-Feb-2015. [Online]. Available: http://ieet.org/index.php/IEET/more/torres20150205. [Accessed: 28-Oct-2015].

[53] L. Anthony (Tony)Cox, "What's Wrong with Risk Matrices?," *Risk Anal.*, vol. 28, no. 2, pp. 497–512, Apr. 2008.

[54] B. Hibbard, *Ethical Artificial Intelligence*. University of Wisconsin - Madison and Machine Intelligence Research Institute, Berkeley, CA, 2015.

[55] N. Bostrom, "IN DEFENSE OF POSTHUMAN DIGNITY," *Bioethics*, vol. 19, no. 3, pp. 202–214, Jun. 2005.

[56] Schupak, "Hawking, Musk: 'Starting a military AI arms race is a bad idea,'" 27-Jul-2015. [Online]. Available: http://www.cbsnews.com/news/hawking-musk-warn-of-global-artificial-intelligence-arms-race/. [Accessed: 29-Nov-2015].

[57] R. I. Field and A. L. Caplan, "Evidence-based decision making for vaccines: The need for an ethical foundation," *Vaccine*, vol. 30, no. 6, pp. 1009–1013, Feb. 2012.

[58] A. Nair, "China's Tianhe-2 retains top supercomputer rank," *Reuters*, 17-Nov-2014.

[59] C. Das, "Artificial Intelligence: Applications in Everyday Life," *HubPages*. [Online]. Available: http://hubpages.com/technology/Artificial-Intelligence-Applications-in-Everyday-Life. [Accessed: 06-Feb-2016].

[60] J. L. Schwartz and A. L. Caplan, "Ethics of vaccination programs," *Curr. Opin. Virol.*, vol. 1, no. 4, pp. 263–267, Oct. 2011.

[61] Dictionary.com "risk," in *Dictionary.com Unabridged*. Source location: Random House, Inc. http://dictionary.reference.com/browse/risk. Available: http://dictionary.reference.com/. Accessed: February 11, 2016.

# APPENDICES

## APPENDIX A: REVIEW OF TECHNOLOGY RISK MANAGEMENT

The U.S. National Institute of Standards and Technology (NIST) produces several guidelines related to technological risk. For example, *SP800-39 Managing Information Security Risk: Organization, Mission, and Information System View*, a guide for managing risk related to the field of Information Security, and *SP800-30 Guide for Conducting Risk Assessments*, providing a practical tool for assessing risk [39]. These guides suggest a focus on four risk management processes, which organizations could apply to AI:

1. Framing Risk- In which an organization establishes a context for the risk, resulting in a *risk management strategy* that describes the remaining three steps while providing a common understanding of how the organization perceives and addresses risk.
2. Assessing Risk- In which the organization identifies threats, vulnerabilities, harm, and likelihoods of harm to derive a determination of risk.
3. Responding to Risk- In which the organization develops, evaluates, determines, and implements responses to risk.
4. Monitoring Risk- In which the organization routinely monitors the risk and the effectiveness of the organization's risk response and practices; the organization also identifies needed changes based on their evaluation of their performance.

Through the risk management process, an organization can continually respond to and adapt to changes in risks in order to mitigate the consequences of those risks. Organizations could perhaps—through the application of tools that have seen widespread use throughout other industries—manage the risks of AI without interrupting its research and market adoption.