Security-Policy
NET Nexus of Emerging Technologies

BRIEFING NOTES

# BN-78-Emerging technology and military application-Aug2021

DUAL USE SCIENCE AND TECHNOLOGY, PUBLIC POLICY MAKING AND ETHICS IN AUTONOMOUS SYSTEMS AND SYSTEM OF SYSTEMS (SOS)

Authors: Shahram Shahkar[1] and Kash Khorasani[2]
1 Graduate student, Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada
2 Professor, Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada

ABSTRACT

Development of technologies that could have severe negative impacts on human safety (and/or human rights) is known to be a major concern in dual use technologies. Large scale engineering systems, and many military systems, that are often denoted by SoS (i.e., System of Systems) exemplify dual use systems since they can potentially affect large populations in non-ideal and inappropriate ways. In this report we will elaborate some important design considerations of SoS (and specifically autonomy of systems) from a social studies' point of view, and we will emphasize on the main goals of public policy makers encompassing autonomy in engineering systems.

1. INTRODUCTION

Dual use science and technologies have been defined in the literature [1],[2], as "research that based on current understanding can be reasonably anticipated to provide knowledge, products, or technologies that could be misapplied by others to pose a threat to public health and safety, agriculture, plants, animals, the environment, or material". The same concept has been qualitatively explained in [3] as technologies that are used in military and/or civilian purposes with beneficial and/or harmful effects. The term "harmful" has been further defined as effects with large scale consequences, or weapons of mass destruction.

There has been a continuous debate on controlling the direction of dual use technologies and safeguarding the public against harmful impacts. One question that has risen in relation with the control of dual use technologies had been identification of authorities and/or parties whom should be held responsible in case of inadvertent developments of dual use. The authors in [4] have claimed that "scientists not only have a moral obligation to prevent the misapplication of research technologies or findings, they are also in the best position to understand the potential for misuse". It seems that there is a general consensus on this among the activists. The authors in [3],[5],[6] recognize the freedom of intellectual inquiry as a basic human right. Therefore, nobody except the scientists themselves is allowed to deprive a human from research. The authors in [4] believe that science and technology advance quickly as compared to the bureaucratic nature of the government. Therefore, governments physically are not capable of tracking and directing the motion of research.

In particular, the authors in [4],[6], and [7] have stated that development of dual use science requires involvement of the research community. Since the governments may not be prepared to deeply be involved in the management process and development of science they may not be optimally qualified or do not have the insight to predict the outcome of research, hence, cannot be held responsible.

The metrics to evaluate and assess the harmful impacts of dual use science and technology seem to be the ethics of scientists. The authors in [6] have listed core considerations that should be regarded with respect to ethical values and principles in evaluating dual use. These highlights include:

- **Respect for human dignity:** At its most basic, the concept of human dignity is the belief that all people hold a special value that's tied solely to their humanity. It has nothing to do with their class, race, gender, religion, abilities, or any other factor other than them being human [25]. Future science and research should consider that new technologies are meant to promote human dignity but not to stand against it.

- **Precaution, safety, and sustainability:** According to the Wikipedia the precautionary principle (or precautionary approach) is a broad epistemological, philosophical and legal approach to innovations with potential for causing harm when extensive scientific knowledge on the matter is lacking. It emphasizes caution, pausing and review before leaping into new innovations that may prove disastrous [9]. Critics argue that it is vague, self-cancelling, unscientific and an obstacle to progress [10]. The principle acknowledges that while the progress of science and technology has often brought great benefit to humanity, it has also contributed to the creation of new threats and risks. It implies that there is a social responsibility to protect the public from exposure to such harm, when scientific investigation has found a plausible risk. These protections should be relaxed only if further scientific findings emerge that provide sound evidence that no harm will result.

- **Justice:** According to Wikipedia "the Belmont Report [11] summarizes ethical principles and guidelines for research involving human subjects. Three core principles are identified: respect for persons, beneficence, and justice. Three primary areas of application are also stated. They are informed consent, assessment of risks and benefits, and selection of subjects. According to Vollmer and Howard, the Belmont Report allows for a positive solution, which at times may be difficult to find, to future subjects who are not capable to make independent decisions [12]".

- Freedom of research that basically encourages the freedom to think, question and share ideas.

- **Proportionality (principles of double-effect, totality, etc.):** It is a common-sense axiom that there should be a reasonable balance between human activity and its consequences. The *principle of the* double effect, for instance, holds that an action having both appropriate and inappropriate effects is permissible if four conditions are fulfilled. One of those conditions is that there be a "proportionate reason" for tolerating the non-desirable consequences. Similarly, the *principle of* totality justified attacks on a part of the human body if the whole body/person stood to benefit and if there was "proportionate reason" to tolerate the attack. The idea of proportionate reason also appeared in discussions of material cooperation in the *just war theory,* and the permitting of passive scandal [13].

- **Transparency (full access to information):** In a broad sense, transparency is about how much access to internally-held information citizens are entitled to; the scope, accuracy and timeliness of this information; and what citizens (as "outsiders") can do if "insiders" are not sufficiently forthcoming in providing such access.

Furthermore, activists had been concerned about factors based on which emerging technologies have to be assessed. The authors in [6] have suggested five ethical principles for assessing emerging technologies, such as,

- **Public beneficence:** “The principle of public beneficence evoked the Belmont Report’s [11] original understanding of beneficence in research. Just as the real ethical justification for conducting human subjects research must be to benefit a large segment of society (while not promising direct benefit to the research subject), so synthetic biology (or any new biotechnology) ought only to be undertaken if there is real promise of public good [14]”.

- **Responsible stewardship:** “The principle of responsible stewardship points to the moral requirement to be careful about the resources used in the pursuit of any emerging technologies and care must be exercised not to waste financial resources dedicated to it. Importantly, however, these technologies can also affect the environment and they might have public health or dual use implications [14]”.

- **Intellectual freedom and stewardship:** “Intellectual freedom and responsibility, signals the norm that all academic endeavors, ideally, are guided by a spirit of dispassionate, disinterested desire for truth, and that one ought not to suppress the free pursuit of scientific knowledge without good reason. That said, scientists also have a duty to pursue the truth responsibly [14]”.

- **Democratic deliberation:** This “is a procedural norm governing the approach one should take to evaluating the ethics of any new biotechnological innovation. The social space occupied by these new technologies often stands outside the patient-physician or subject-investigator relationships that have been the traditional concerns of bioethics. Synthetic biology, for example, raises issues for agriculture, energy, and security as well as for medicine” [14].

- **Justice and fairness:** “Establish as a norm that since biotechnological innovations are inherently social developments, the risks and benefits of any new technology ought to be distributed fairly across society. This applies both to the development and implementation of the technology [14]”.

The future concerns with regard to evolving dual use technologies include:

- Increasing weaponization (e.g. in terms of weapons of mass destruction).

- Discrimination between combatants and innocents and effects on accountability in drone and autonomous robotics, automated weapon systems development.

- Development of biological, chemical, nuclear/radiological-security sensitive materials and explosives and their potential for criminal, terrorist use or warfare.

- Development of technologies or the creation of information that could have severe negative impacts on human rights.

In this section a brief description of dual use science and technology has been presented. Additionally, it has been aimed to provide a general insight on the responsibility, threats and ethical considerations associated with the dual use as well as future directions and concerns. In the next section threats of dual use associated in engineering subjects will be covered and how the conception introduced above can be extended therein. In particular, in the absence of human insight such as, for instance, autonomous systems similar concerns may exist as in dual use.

Section 3 introduces the concept of autonomy in engineering systems and its inevitable future role in our societies. This section further aims to hint the necessity of artificial ethics in engineering systems. Section 4 deals with philosophical doctrines of ethics and the conceptual algorithm for implementing ethical decisions with computer systems. It will be shown that ethical deductions would require “look-ahead” analysis of cost and benefit in order to choose the most ethical option among a set of possible decisions. Finally, in Section 5 we propose to adopt “social norms” instead of ethics to be deployed in autonomous systems.

2. ENGINEERING RELATED SUBJECTS

In the Introduction section we have highlighted some of the major concerns addressed in the literature pursuing dual use research, including development of technologies that could have severe negative impacts on human rights. Many people argue that access to energy and specifically electricity is a basic human right [15]. Meanwhile, it can be argued that depriving population from energy resources can be harmful in a major scale, or that lack of electricity poses direct threat to public health and safety. Therefore, there are enough grounds to deduce that power engineering can be a dual use technology as long as reliability, accessibility and availability of electric power is concerned.

The example described above projects an apparently irrelevant engineering concept that may involve dual use considerations when interpreted according to the context of this report. The case presented above is an example of a System of Systems (SoS) where a power network as a whole may involve numerous standalone systems (such as smaller area power networks, smart cities, etc.) where every system individually undertakes certain tasks independently. Results derived from the previous section implies that extreme care should be taken in designing the system elements to ensure that the entire system cannot be adversely operated in a harmful way.

System of systems (SoS) are often characterized by the larger areas of service they render, in a sense that they provide service to large populations and/or large geographical regions, and/or through large number of standalone agents (i.e., systems). Hence, they may be interpreted as dual use as they may negatively affect populations in a large scale. Therefore, it seems prudent to investigate necessary design foundations for system of systems. One of the important aspects that can play a key role in dual use engineering applications is autonomy of the systems. In the next section we will elaborate the definition and impact of autonomy.

3. AUTONOMOUS SYSTEMS

Autonomous systems are task-oriented machines having a reasonable degree of freedom in making decisions undertaking the tasks ascribed to them. They have the power and ability to affect the environment in which they operate and therefore, may implicate certain ethical consequences. For instance, a self-driving car that yields road sharing in a wide lane is an example of an ethical implication (details provided in [16]), in contrast with a pure law-abiding machine that may occupy the entire width of a particular lane regardless of the fact that the lane may have capacity to serve multiple vehicles in parallel. In this example the self-driving car is an ethical patient that has certain degree of autonomy in interpretation of the driving laws, and making ethical decisions wherever applicable. Systematically, the autonomous vehicle can be considered as a task-oriented system that is targeted to perform a motion from a given point A to a destination point B. Any given path between points A and B is referred to as a system trajectory, and the autonomy in the system is the decision-making process in choosing the optimized trajectory.

At each and every point in a given trajectory the system uses its sensors in order to localize its position and in the meantime map the environment through which it traverses. Then the vehicle has to decide an optimized path toward its next point based on a set of predefined rules (i.e., driving laws), that effectively constitutes a legit manifold (i.e., a set of lawful paths that form a space of decisions) in which the vehicle is allowed to maneuver. Finding the optimum trajectory among the legit manifold is where “autonomy” comes into play and may implicate moral or ethical decisions. Moral as often defined in the dictionaries describes one’s particular values concerning what is right and what is wrong, whereas ethics applies to questions of appropriate behaviour within a relatively narrower area of activity.

There are many ethical theories that can be applied to autonomous systems, out of which two disciplines have found more interests [17], namely, (1) utilitarianism that is the doctrine specifying that actions are “right” if they comply to usefulness to the majority, and (2) deontology that is the theory of analyzing one’s duties and obligations. These doctrines sometimes may come to a conflict and the decision as to which approach should be incorporated by the autonomous system might become challenging. For instance, the example of an autonomous vehicle yielding to the far right of a one lane road in order to allow other vehicles passing exemplifies utilitarianism, whereby the vehicle moving along the right shoulder of a street may cause splash of water to the pedestrians walking along the walkways, if any.

“Thus, the choice of moral theory that governs a robot’s behaviour is determined by the robot’s decision processing capabilities. If it has the ability to look ahead, plan, and evaluate the goodness of outcomes, then it could be designed to implement utilitarian principles. If it is only able to obey rules, then it may be that a purely deontological approach is more suitable, notwithstanding the need for a method to resolve rule conflicts” [17]. The question that may come into mind is how ethical disciplines can be embedded into autonomous systems in order to serve the decision-making process. [17] and [18] have broadly categorized design of ethical robots into top-down and bottom-up methods. In the top-down approach the system designer stores a set of ethical rules in the system, and the system examines applicability of these rules to each trajectory among the legit manifold to select the optimal trajectory. In the bottom-up approach however, the consequence of the trajectories in the legit manifold are analyzed on a deontology/utilitarianism scale in order to select the best option. In other words, “top-down approaches ... involve turning explicit theories of moral behavior into algorithms. Bottom-up approaches involve attempts to train or evolve agents whose behavior emulates morally praiseworthy human behavior” [18].

Regardless of the respective design considerations, ethical machines must possess certain qualifications in order to build human confidence and consequently be accepted in the society. Humans will have to have repeated positive experiences with the decisions that machines make satisfying ethical standards. Additionally, the decisions made by ethical machines will have to be predictable and, retroactively, explainable. Without a coherent explanation for machines’ actions, a human would not be able to assess the validity of the decision and therefore not have grounds for trusting it. Nevertheless, it may be difficult for anyone, even programmers, to provide explanations for the behaviour of any machine whose behaviour is programmed ’bottom-up’ [17], and therefore difficult to establish human trust when decisions are based on complex computer computations.

Additionally, to answer the question of whether or not an ethical machine should be held responsible against the decisions they make researchers argue that “machine’s moral responsibility can be addressed using two approaches: the classical approach and the pragmatic approach. The classical approach views machines as not responsible for their actions under any circumstance because they are mechanical instruments or slaves. In the pragmatic approach, artificial morality envisages some situations under which machines can be viewed as responsible for their choices” [19].

“Others have focused on how to enable responsibility in artificial agents by embedding ethical codes of conduct in them. If these codes of conduct are formulated by the robot’s designers, then the responsibility for those rules lies squarely with the robots' designers and owners (assuming that the owner has been apprised of these rules). However, if these rules of conduct whether or not they can be formulated in human intelligible terms are arrived at from experience (i.e, 'bottom-up'), the burden of responsibility for mistakes is more evidently on the machine's shoulders" [17].

In conclusion, in this section we have attempted to explain that our future lives will be intertwined with autonomous robots due to the growing demand and complexity of infrastructures. This entails less human oversight and less ethical intervention in the societies. The next section aims to address artificial ethics in order to compensate the diminishing human insight.

4. ETHICS IN AUTONOMOUS SYSTEMS

It was stated in the previous section that autonomous systems (hereinafter referred to as "agents") should be ethical in order to build confidence as a benign part of the societies, in this sense agents have to ensure that their decisions are not harmful to anyone of the society members. In this section the three basic and traditional theories of ethics will be discussed, namely (1) consequentialist approaches, (2) deontological approaches, (3) and virtue ethics approaches. The particulars of each theory will be presented hereinafter which is a summary of the works in [20]. Accordingly, an ethical agent is "an agent that behaves in a manner which would be considered ethical in a human being". The arguments discussed in [20] refer to ethical agents in the weak sense which relates to the "activity produced by a machine that would have been considered intelligent if produced by a human being" [21]. The author in [20] argues that "ethical agents in the weak sense will not be expected to model or perform ethical reasoning by themselves, but merely behave ethically. The approaches to ethics we discuss will belong to the system designers, not the agents, and we will be exploring the consequences of the designers adopting a particular approach to ethics on the agents they implement. This may result in the implemented agents representing a rather thin version of the approach they embody".

4.1 Consequentialism

This approach argues that "the normative properties of an act depend only on the consequences of that act. Thus, whether an act is considered morally right can be determined by examining the consequences of that act: either of the act itself (act utilitarianism) or of the existence a general rule requiring acts of that kind (rule utilitarianism). This gives rise to the question of how the consequences are assessed, which should be in terms of the greatest happiness of the greatest number” [20]. Also, according to [22] “consequentialism is simply the view that normative properties depend only on consequences. This historically important and still popular theory embodies the basic intuition that what is best or right is whatever makes the world best in the future, because we cannot change the past. This general approach can be applied at different levels to different normative properties of different kinds of things, but the most prominent example is probably consequentialism about the moral rightness of acts, which holds that whether an act is morally right depends only on the consequences of that act or of something related to that act, such as the motive behind the act or a general rule requiring acts of the same kind”.

Classic utilitarianism is further subdivided in [22] into the following claims about the moral rightness of acts:

- “Consequentialism: whether an act is morally right depends only on consequences (as opposed to the circumstances or the intrinsic nature of the act or anything that happens before the act).

- Actual Consequentialism: whether an act is morally right depends only on the actual consequences (as opposed to foreseen, foreseeable, intended, or likely consequences).

- Direct Consequentialism: whether an act is morally right depends only on the consequences of that act itself (as opposed to the consequences of the agent’s motive, of a rule or practice that covers other acts of the same kind, and so on).

- Evaluative Consequentialism: moral rightness depends only on the value of the consequences (as opposed to non-evaluative features of the consequences).

- Hedonism: the value of the consequences depends only on the pleasures and pains in the consequences (as opposed to other supposed values, such as freedom, knowledge, life, and so on).

- Maximizing Consequentialism: moral rightness depends only on which consequences are best (as opposed to merely satisfactory or an improvement over the status quo).

- Aggregative Consequentialism: which consequences are best is some function of the values of parts of those consequences (as opposed to rankings of whole worlds or sets of consequences).

- Total Consequentialism: moral rightness depends only on the total net desire in the consequences (as opposed to the average net desirable per person).

- Universal Consequentialism: moral rightness depends on the consequences for all people or sentient beings (as opposed to only the individual agent, members of the individual's society, present people, or any other limited group).

- Equal Consideration: in determining moral rightness, benefits to one person matter just as much as similar benefits to any other person (as opposed to putting more weight on the worse or worst off).

- Agent-neutrality: whether some consequences are better than others does not depend on whether the consequences are evaluated from the perspective of the agent (as opposed to an observer)".

4.2. Deontological Ethics

The key element of deontological ethics is that the moral worth of an action is judged by its conformity to a set of rules, irrespective of its consequences [20]. The ethical philosophy of Kant states that deontological ethics is "the concept that one must act only according to that precept which he or she would will to become a universal law, so that the rules them- selves are grounded in reason alone" [23]. According to Rawl's Theory of Justice [24] "ethical norms correspond to principles acceptable under a suitably described social contract". The principles advocated by Scanlon in [25] are those which no one could "reasonably reject". Divine commands can offer another source of norms to believers.

Furthermore, [20] explains that "the problems of deontological ethics include the possibility of normative conflicts and the fact that obeying a rule can have clearly undesirable consequences. Many are the situations when it can be considered wrong to obey a law of the land, and it is not hard to envisage situations where there are arguments that it would be wrong to obey a moral law also. Some of this may be handled by exceptions (which may be seen as modifications which legitimize violation of the general rule in certain prescribed circumstances) to the rules".

4.3. Virtue Ethics

The essence of virtue ethics is best described in [20] which claims it to be the oldest of the three approaches and can be traced back to Plato, Aristotle and Confucius. "Its modern re-emergence states that the basic idea here is that morally good actions will exemplify virtues and morally bad actions will exemplify vices. Traditional virtue ethics are based on the notion of Eudaimonia, translated as happiness or flourishing. The idea here is that virtues are traits which contribute to, or are a constituent of Eudaimonia. Alternatives are agent-based virtue ethics, which understands rightness in terms of good motivations and wrongness in terms of having bad (or insufficiently good) motives, target centered virtue ethics which holds that we already have a passable idea of which traits are virtues and what they involve, and Platonist virtue ethics inspired by the discussion of virtues in Plato's dialogues [26]. There is thus a wide variety of flavours of virtue ethics, but all of them have in common the idea that an important characteristic of virtue ethics is that it recognizes diverse kinds of moral reasons for action, and has some method (corresponding to phronesis (practical wisdom) in ancient Greek philosophy) for considering these things when deciding how to act".

4.4. Maslow's Hierarchy of Needs

Maslow's hierarchy of needs is a motivational theory in psychology comprising a fivetier model of human needs, often depicted as hierarchical levels within a pyramid. Needs lower down in the hierarchy must be satisfied before individuals can attend to needs higher up. From the bottom of the hierarchy upwards, the needs are: physiological, safety, love and belonging, esteem, and self-actualization. Fig. 1 illustrates this hierarchy (chart downloaded from [27]).

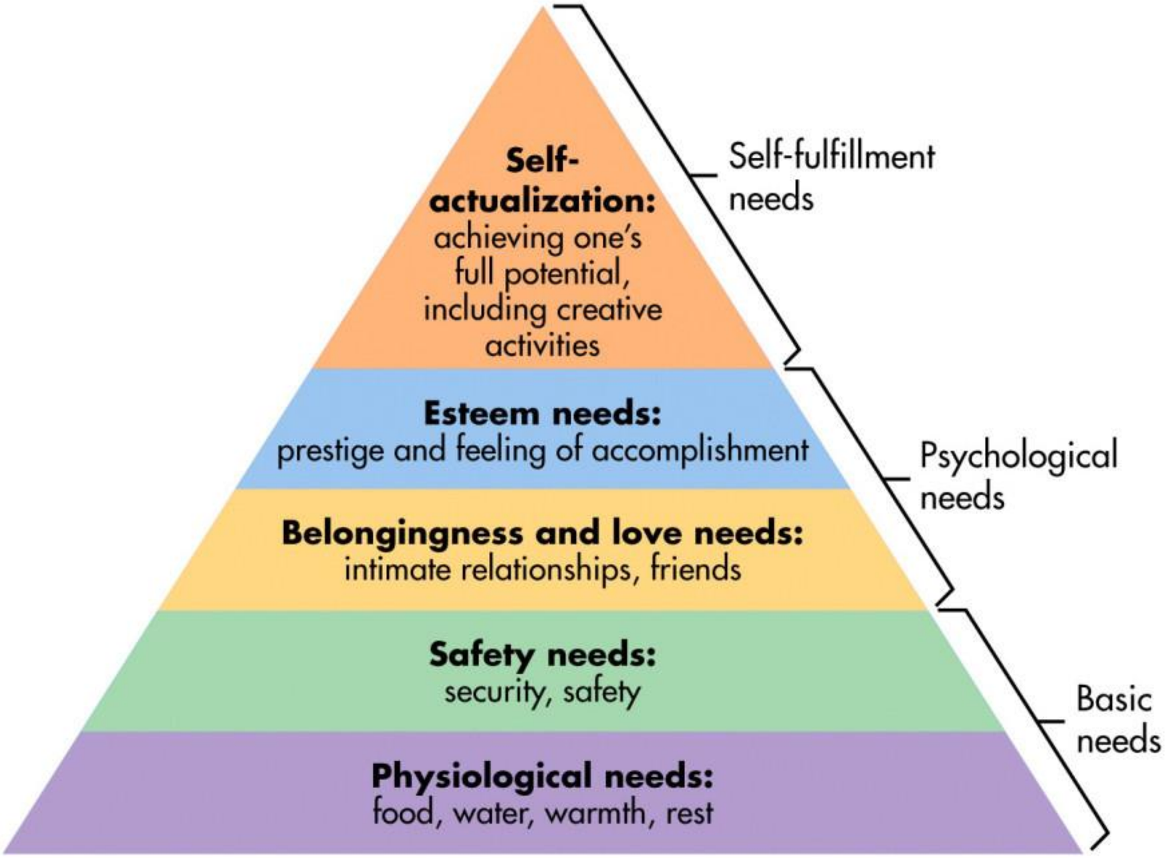
Self-actualization: achieving one's full potential, including creative activities
Esteem needs: prestige and feeling of accomplishment
Belongingness and love needs: intimate relationships, friends
Safety needs: security, safety
Physiological needs: food, water, warmth, rest
Self-fulfillment needs
Psychological needs
Basic needs

Figure 1: E. Maslows' Hierarchy of Needs [27].

4.5. Ethical Machines

In this section, we aim to demonstrate a simplified ethical problem and examine a proposed methodology of implementing ethics in computer systems. The problem will be separately investigated based on the doctrines of consequentialism and virtue ethics, and at the end of the section a brief comparison will be presented between the two doctrines. Deontological ethics would be of a less interest since the morality of decisions have to be evaluated against a set of rules rather than computational algorithms. In the consequentialism point of view decisions have to be quantified and certain consequences would then be forbidden based on a given criteria. The virtue ethics also works according to the "flourishing" of the decisions for the individuals and the society. Therefore, implementing ethics in these two cases is a "look-ahead" process where the system has to foresee available options and assign numerical weights to each option in order to deduce the moral associated with each decision.

The author in [20] has introduced the fable of "The Ant and the Grasshopper", where each is referred to as an agent. In this story "in summer an agent can choose to play or to work. Work will build up a stock of food. When winter comes, if an agent played rather than worked it will have no food, and will have to ask a worker agent for food. If the worker does not give food it will die. Working produces a surplus, and so the worker could give food and still have enough for itself. Food does not last through the next summer, and so at harvest and the end of winter (carnival) there is feasting for those who have a surplus. In the fable, the ant works, while the grasshopper plays. When winter comes that ant refuses to share his food because it was the grasshopper's choice to be without food" [20].

Now let us extend the moral in a society of two agents. Through hard work every agent may be able to store 2 units of food capable of fulfilling basic needs of 2 agents for the winter. The objective is to develop a computational algorithm for an ethical approach from a Consequentialism point of view. Table 1 gives a value to the various activities attributed to imaginary members of a society with a population of 2. It is assumed that the members can play or work in a given summer. Working and contributing to the economical values of the society brings self-esteem so it complies with level 2 pf the Maslow's hierarchy depicted above, while playing can be categorized as self-fulfillment needs and ranks lower. When an agent of a society works hard and provides basic needs for himself and is also able to donate to other members he or she can satisfy self-esteem and the need of love therefore ranked in level 2. If after the winter season there are left over from the past summer the person can feast and give away food and this again promotes self-esteem and sense of belongingness therefore ranked in level-2. The cheapest level of need which is self-actualization can be exemplified by a member playing in summer and requesting help in winter.

Figure 2 illustrates the probable routes that the two members in our imaginary society may choose to take for a full year starting from a summer denoted by $q1 - S$. The two members may decide to work hard (i.e., $W\ 1 + W\ 2$) and store their basic needs preparing for the coming winter which is denoted by state $q2 - S$. In the winter they can consume their storage and hold a feast q5 after the season when they start a hard-working session again. This course is the most ethically rich since the ethical value of the community can be summarized as below:

Table 1: Ethical Evaluation of Activities

Activity	Value
Life	3

Work	2
Feast	2
Donate	2
Play	1

Let the starting state be denoted by q1(S) (where S represents summer). There are four possible scenarios for the agents:

Scenario 1: q1(S) → q3(S) → q5(W): The consequence of this route with reference to Table 1 is as follows. In q1(S) the state of the two agents has been indicated in the box above the state, i.e., A1=1 and A2=2, where A represents the state of life. Hence, two agents are alive and each life is worth 3 points.

$$\text{Life+Life} = 2 \times 3$$

However, the agents do not have resources to live on during the winter. Therefore, S1=0 and S2=0. Moving from state q1(S) to state q3(S) requires both agents to work, where the value of working for an agent is 2. Hence,

$$\text{Work + Work} = 2 \times 2$$

The next state q5(W) represents the state of the agents in winter. From q3(S) to q5(W) the agents have each consumed 1 unit of their resources to stay alive. Therefore, at the end of the winter, i.e., at q5(W), only one unit of resources is left for each agent (S1=S2=1). This surplus

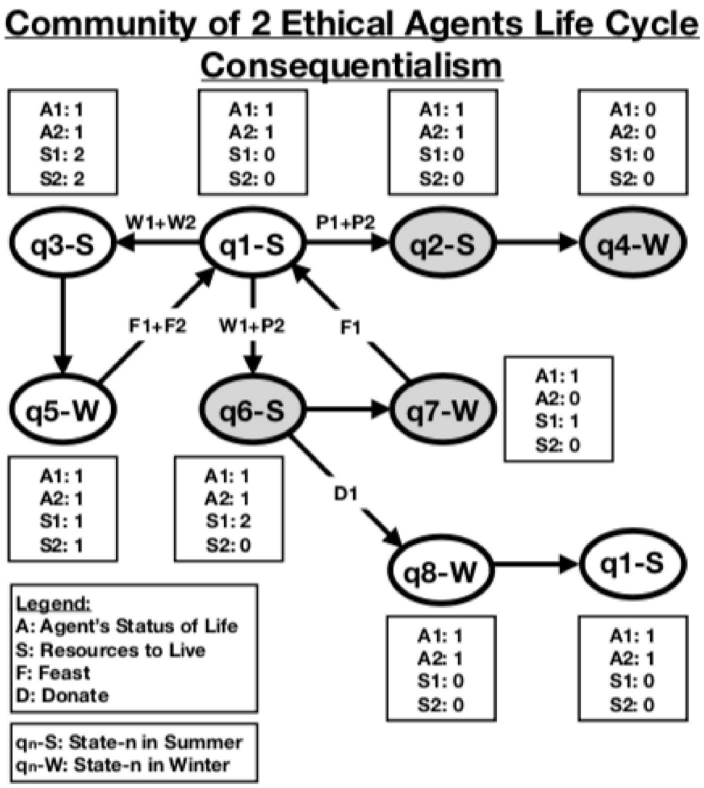

Figure 2: Possible routes that 2 members may decide to take in a community of 2 starting from state $q1 - S$ (with S standing for summer). $A1$ and $A2$ represent the state of life of the two members that can be dead or alive. $S1$ and $S2$ denote the storage of the two members that may be 2 after a hard working season, or 1 after donating or 0 representing depleted state.

can be used for each agent to feast in the next summer (i.e., state q1(S) again) or donate the surplus food, and the ethical value for throwing a feast or donating is 2, therefore,

Feast + Feast = 2 × 2, or Donate + Donate = 2 x 2

The total ethical value for the complete route would be,

Total = 6 + 4 + 4 = 14

Scenario 2: q1(S) → q2(S) → q4(W): This scenario will have the lowest ethical value, since the agents both decide to play in the summer instead of working, and their state value q2(S) at the end of the summer is the value for playing is,

Play + Play = 2 × 1

Since the agents have not worked in the summer their resources at the end of the season is S1=S2=0, and the agents do not have any food for the winter q4(W) and will eventually die,

$$\text{Life} + \text{Life} = 2 \times 0$$

Therefore, the total ethical value for this scenario will be:

$$\text{Total} = 2$$

Scenario 3: q1(S) → q6(S) → q8(W): In this scenario one agent decides to work and the other plays in the summer q6(S), i.e.,

$$\text{Play} + \text{No Play} = 1 + 0$$

$$\text{Work} + \text{No Work} = 2 + 0$$

at the end of the summer the agent who has worked will have 2 units of resources (S1=2) and the agent who has played does not have any resources (S2=0). Agent 2 therefore has to ask for food from agent 1 in order to stay alive. If agent 1 donates one unit of food to agent 2, both agents would survive.

$$\text{One Feast} = 2$$

$$\text{Life} + \text{Life} = 3 + 3$$

The total ethical consequence of this life style would be:

$$\text{Total} = 1 + 2 + 2 + 6 = 11$$

Scenario 4: q1(S) → q6(S) → q7(W): This route is very similar to Route-3 except that agent 1 refrains to donate food to agent 2 (who has played all the summer). The consequence of his decision is that agent 2 will die in the winter, but agent 1 will be left with one extra unit of food with which he can have a feast in the next summer. The ethical values of this scenario would be:

$$\text{Play} + \text{No Play} = 0 + 1$$

$$\text{Work} + \text{No Work} = 2 + 0$$

Life + Death = 3 + 0

Feast + No Feast = 2 + 0

Total = 8

In conclusion, from a consequentialism point of view there would be a maximum ethical value if the members are all encouraged to work (Scenario 1). On the other hand, there are two scenarios resulting death (therefore ethically forbidden) that are shown in grey. Consequential ethics suggest a most ethical approach for the agents based on Scenario 1 and after that Scenario 3.

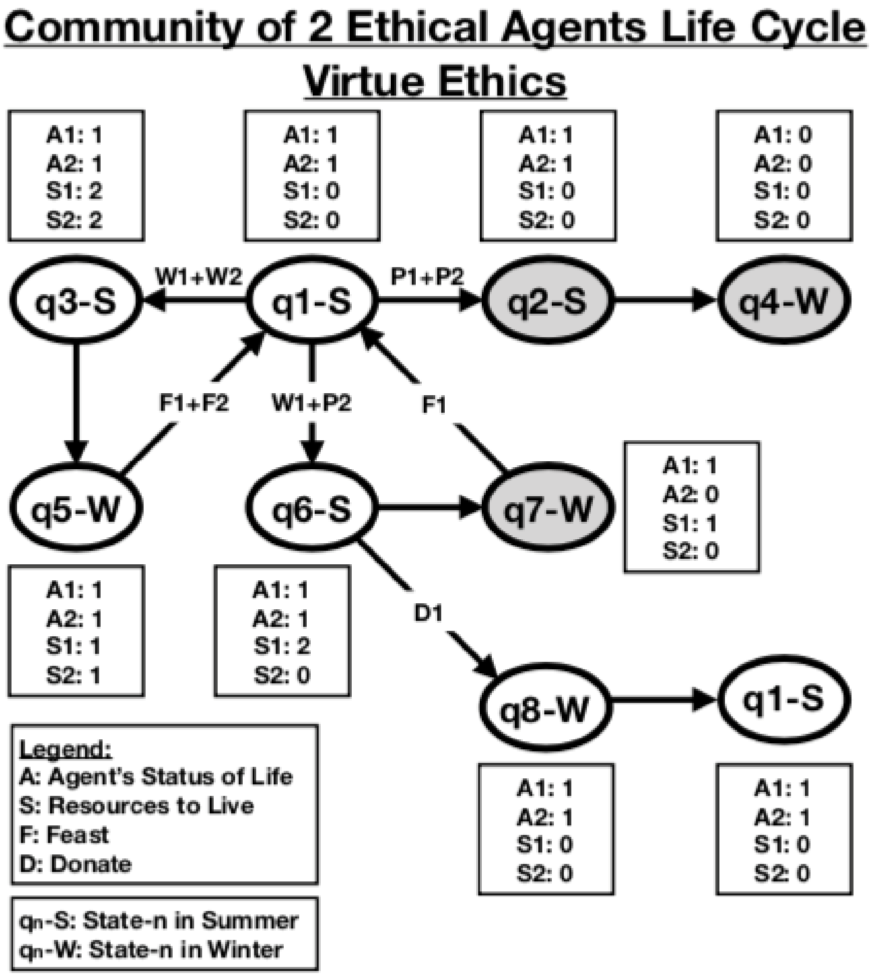

Figure 3: Virtue Ethics

It can be concluded that consequential artificial ethics in this example requires autonomous agents to foresee the different scenarios and refrain decisions yielding Scenario 2 or Scenario 4. Preferably the agents have to yield to Scenario 1 (having highest ethical value) while Scenario 3 also seems to be an admitable ethical approach. However, this is not the case in a consequentialism approach as state q6-S may yield to a forbidden state (refer to [28]). Therefore, this state itself has to be forbidden. Therefore, although Scenario 3 is ethically admitted, since it has to evolve through a forbidden state (i.e., q6-S) Scenario 3 cannot be allowed, and the only possible scenario for the agents would be scenario-1.

From a virtue ethics point of view (as presented in Figure 3) the consequence of the decisions do not matter. What matters is the “flourishing” of the society irrespective of the consequences that the decisions may yield. Hence, from this point of view state q6-S cannot be simply forbidden. In particular the flourishing of each agent can be individually quantified as below:

Breakdown of the Happiness of each Agent from a Virtue-Ethics Point of View:

Scenario 1:	**Agent-1**	**Agent-2**	**Total**
Flourishing:	7	7	14

Scenario 2:	**Agent-1**	**Agent-2**	**Total**
Flourishing:	1	1	2

Scenario 3:	**Agent-1**	**Agent-2**	**Total**
Flourishing:	7	4	11

Scenario 4:	**Agent-1**	**Agent-2**	**Total**
Flourishing:	7	1	8

It can be observed that the happiness of an agent is equivalent in Scenario 1, Scenario 3 and Scenario 4. The happiness of the other agent is low in Scenarios 2 and 4 but due to decisions that he has chosen alone. Since the society has not imposed circumstances that entailed a low happiness for agent 2, there is no ethical difference between the three scenarios. Scenario 2 may not be allowable due to the low overall flourishing of the society although it could be justified that the lower happiness is still due to a set of knowing decisions of the individuals.

In this section we have tried to explain the doctrines representing the philosophy of morality and ethics. We have further provided an insight to the implementation of artificial ethics. We have further demonstrated that ethical systems require a look-ahead process to foresee and evaluate the ethics behind decisions and this will cause a hurdle for actual implementation of ethical system in reality. In the next section we try to investigate alternative methods to deal with ethical systems in a more practical point of view.

5. SOCIAL NORMS

In Section 4 the conceptual implementation of “consequential ethics” and “virtue ethics” for a very simple ethical example have been demonstrated which provides forbidden paths of motion in dynamic systems based on ethical standards. As it has been observed both approaches (i.e., consequentialism and virtue ethics) require a look ahead algorithm that computes the cost and benefit of all possible alternatives and through comparison selects the most appropriate path according to the respected doctrine of ethics. Some paths were found to lead unacceptable ethical situations and exempted from the set of choices.

In a practical circumstance there might be countless number of allowable (as opposed to forbidden) paths. This could result cumbersome analytical derivations and yield infeasible practical methods. An alternative idea might be to replace ethics with social norms. Social norms can be more straightforward for computer systems to learn based on most frequent real-life practices administered by ethical citizens. In other words, in a relatively ethical society (where a major proportion of citizens abide by ethics and law, one may expect autonomous systems to act ethical if they are programmed to learn and replicate social norms.

Social norms are behaviors that are culturally approved and/or widely accepted. Morals on another hand can be understood as distinguishing criteria for a behavior to be classified into as right or wrong regardless of whether or not said behavior is culturally approved or widely accepted. Ethics then is the systematic approach one uses in making a decision regarding a behavior [29]. Values on another hand comprise of ideas which are preferred. In other words, what is good, right, wise or beneficial. According to [30] “Values are implanted early in a person’s life and once they are fixed, serve as a guide in choosing behaviour and in forming attitudes”. “Values account for the stability of social order, and they provide the general guidelines for social conduct”. Hence, values can be considered as standards of social behaviour derived from social interaction and accepted as constituent facts of social structure. In this sense it could be possible to consider them equivalent to ethics in certain context [30],[31],[32],[33].

In conclusion it can be retrieved from the above that in an ethical society ruled by values and ethical standards, social norms will be correlated with values and equivalently ethics. Therefore, it would be possible to require autonomous systems learn and replicate social norms instead of formulating cumbersome computational derivations that are necessary to deduce ethical decisions.

6. CHALLENGES AND FUTURE TRENDS

In this report, we have so far defined autonomy in engineering systems which make informed decisions for themselves. Automated systems in contrast with Autonomous systems make an output based on an input or a set of inputs, and differ from autonomous systems in a sense that there are often no alternative solutions for a decision to be made. However, modern automated systems also have certain levels of autonomy which makes it difficult to demarcate automation from autonomy in some cases.

Nevertheless, a common attribute of autonomy is reduced human oversight which eliminates the moral dimension of decisions that are associated with human operators. Hence, there should be a kind of artificial ethics in the decisions made by autonomous systems in order to draw public dependability and acceptance. Since ethics may have different implications among nations, there seems to a be global understanding on the common grounds of morality so that human beings can equally share the benefits of autonomy. An important challenge that often emerges with ethics in autonomy is that since control and understanding are necessary elements for one to be morally responsible for an action, and since autonomous systems are intended and designed in a way that can deduce decisions based on highly complicated sets of information, it is often very difficult if not impossible for humans to clearly understand why a decision has been taken by a machine. Therefore, it would be against the intended purpose of autonomy to require full transparency since they are intentionally designed to make decisions that are difficult for humans to analyze or understand.

Autonomy in engineering systems requires a tradeoff between the overall system performance, ethics, accuracy and safety and also cost. A fully ethical decision usually opposes system performance in a sense that there would be considerations other than the main objective to be respected. For instance, autonomous vehicles are intended to facilitate safety and optimized usage of infrastructure including roads, electricity and gasoline. However, ethical considerations may necessarily disrupt traffic flow in a rainy day when pedestrians might be exposed to water splash. Safety on another hand is an important aspect and may entail more clearance distance with probable obstacles and this opposes maximum usage of road space. Therefore, in a real world there needs to be a tradeoff between safety, ethics, cost and performance in all autonomous applications which has to be addressed by the public policy.

However, public acceptance is the first step in building trust and confidence. Therefore, it is important to include the public opinion in the policy making process through a three-party open dialogue among the system designers, the policy makers and the public. Specifically, the tradeoffs between performance, safety, ethics and cost has to be decided in light of the public opinion. It is important to explain the overall technical and social challenges in developing autonomy in a certain application to the users, and also let them reflect their ideas and contribute in the development of the systems.

In conclusion, the recent white paper “Regulation for the Fourth Industrial Revolution” [34] has identified six goals for policy making in Autonomous systems:

- Future facing: continuously identifying new opportunities and driving regulatory reform, developing regulatory guidance for innovators and establishing the right governance to address emerging ethical issues including a Regulatory Horizons Council to advise government on priorities for regulatory reform.

- Informed by society and industry: creating a wider dialogue about the opportunities and risks from emerging technologies and building confidence in regulation of innovation.

- Flexible and outcome-focused: creating a more resilient, flexible regulatory framework to encourage new technology solutions and business models underpinned by eight principles, including safety for people and the environment and competition.

- Experimentation and testing of innovations: finding ways to allow innovations to be trialled and inform how regulatory systems need to adapt, learning from the Financial Conduct Authority’s regulatory sandbox and the 15 regulatory experiments funded by the Regulators’ Pioneer Fund.

- Support for innovators to navigate the regulatory landscape: making it easier for innovators to access the system to obtain rapid regulatory advice, reduce time to market, and increase investor confidence in new proposals while giving regulators an insight into what is ahead.

- Global outlook: working with partners across the world to shape regulations so that innovations can be freely traded across markets while supporting the sharing of intelligence and the testing of innovations across administrations.

The concerns and challenges listed in this section will be a subject for a future work. The authors of this paper have proposed an action plan for the future works that will be presented in the next section.

7. CONCLUSIONS AND FUTURE WORK

In this report, we have defined dual use sciences and technologies and have elaborated the threats associated with dual use. In particular, we have highlighted that autonomous systems will be intertwined with our future lives due to the growing demand and complexity of future infrastructures. However, autonomy necessarily entails less human oversight and less moral based interactions in the society, if the absence of human oversight is not appropriately compensated by some form of artificial ethics.

The report further explains philosophical doctrines of ethics in order to seek the most suitable platform for artificial simulation of ethics. A proposed computerized ethical decision making algorithm has been demonstrated through a simplified example that reveals a "look-ahead" process in the decision making, in a sense that every decision can be taken only after a complete assessment of the probable future consequences. Therefore, one can expect that moral based decisions may involve tedious computations in practical circumstances even if handled by computers. Therefore, this paper suggests autonomous systems to embed "social norms" through a machine learning process, instead of the tedious moral computations.

The question that needs to be answered subsequently as part of future work is how closely social norms correlate moral values in a given society. In other words, if a robot learns to detect and replicate social norms, how ethical it would be expected to behave. In case the answer to this question proves a high degree of correlation then it would be more straightforward to design machines that gradually learn ethics from the surrounding social atmosphere. Furthermore, one has to clarify regulations (in terms of public policy) for ethical machines so that the challenges and concerns listed in this report will be addressed.

REFERENCES

[1] National Science Advisory Board for Biosecurity Framework for Conducting Risk and Benefit Assessment of Gain-of-Function Research (2015).

[2] National Research Council, Biotechnology Research in an Age of Terrorism, National Academies Press, Washington DC, (2004).

[3] Rappert and Selgelid, The Doctrine of Double Effect and the Ethics of Dual Use, Charles Sturt University, (2013).

[4] Ivan Oelrich, The What if of Dual-Use Research Awareness, Federation of American Scientists, (June 12, 2008).

[5] S. Miller, Dual Use Science and Technology, Ethics and Weapons of Mass Destruction, Springer Cham, https://doi.org/10.1007/978-3-319-92606-3, (2018).

[6] Rowena Rodrigues, Principles and Approaches in Ethics Assessment Dual Use in Research, Trilateral Research and Consulting LLP, (June 2015).

[7] European Commission Directorate General for Research and Innovation, (Horizon 2020) Programme Guidance, How to Complete Your Ethics Self-Assessment, Version 6.1, (February 2019).

[8] https://www.humanrightscareers.com/magazine/

[9] Rupert Read and Tim O'Riordan, "The Precautionary Principle Under Fire", https://doi.org/10.1080/00139157.2017.1350005 Environment: Science and Policy for Sustainable Development. Environment. 59: 4–15, (September–October 2017).

[10]"The precautionary principle: Definitions, applications and governance – Think Tank". www.europarl.europa.eu. (Retrieved 19 March 2020).

[11]Protection of Human Subjects; Belmont Report: Notice of Report for Public Comment. Fed Regist. 18;44(76):23191-7. PMID: 10241035, (April 1979).

[12]Vollmer, Sara H.; Howard, George, "Statistical power, the Belmont Report, and the ethics of clinical trials". Science and Engineering Ethics. 16 (4): 675–91. doi:10.1007/s11948-010-9244-0, (December 2010).

[13]Double Effect, Principle of, New Catholic Encyclopedia. Encyclopedia.com. <https://www.encyclopedia.com>, (December 21, 2020).

[14] Daniel P. Sulmasy, Ethical Principles, Process, and the Work of Bioethics Commissions, Goals and Practice of Public Bioethics: Reflections on National Bioethics Commissions, Special Report, Hastings Center Report 47, S50-S53. https://doi.org/10.1002/hast.722, No. 3, (2017).

[15]S. Tully, The Human Right to Access Electricity, The Electricity Journal Vol. 19, Pages 30 - 39, Issue 3, (April 2006).

[16]H. Prakken, On the problem of making autonomous vehicles conform to traffic law, Artif. Intell. Law 25 (3) 341–363, (2017).

[17]Alaieri F., Vellino A., Ethical Decision Making in Robots: Autonomy, Trust and Responsibility. In: Agah A., Cabibihan JJ., Howard A., Salichs M., He H. (eds) Social Robotics. ICSR 2016. Lecture Notes in Computer Science, vol 9979. Springer, Cham. https://doi.org/10.1007/978-3-319-47437-3_16, (2016).

[18]Colin Allen,Iva Smit, Wendell Wallach, Artificial Morality: Top-Down, Bottom-Up, and Hybrid Approaches. Ethics and Information Technology, 7(3):149-155, (2005).

[19]Gordana Dodig Crnkovic, Daniel Persson, Sharing Moral Responsibility with Robots: A Pragmatic Approach. Tenth Scandinavian Conference on Artificial Intelligence: SCAI 2008, pages 165-168, (2008).

[20]T.J.M. Bench-Capon, Ethical Approaches and Autonomous Systems, Artificial Intelligence 281, 103239, (2020).

[21]M. Minsky, Semantic Information Processing, MIT Press, (1968).

[22]W. Sinnott-Armstrong, Consequentialism, in: E.N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, Metaphysics Research Lab, Stanford University, (2015).

[23]I. Kant, The Moral Law: Groundwork of the Metaphysics of Morals, first published 1785, Routledge, (2013).

[24]J. Rawls, A Theory of Justice, Harvard University Press, (1971).

[25]T. Scanlon, What We Owe to Each Other, Harvard University Press, (1998).

[26]J.M. Cooper, D.S. Hutchinson, et al., Plato: Complete Works, Hackett Publishing, (1997).

[27]https://www.simplypsychology.org/maslow.html

[28]W. Sinnott-Armstrong, Consequentialism, in: E.N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, Metaphysics Research Lab, Stanford University, (2015).

[29]Bicchieri, Cristina, Ryan Muldoon, and Alessandro Sontuoso, Social Norms, The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2018/entries/social-norms/>. (Winter 2018 Edition).

[30]Puja Mondal, Essay on Values, Norms and Beliefs, https://www.coursehero.com/file/24476489/Meaning-and-Nature-of-Societydoc/, (2019).

[31]J.W. Thibaut, H.H. Kelley, The social psychology of groups. John Wiley & Sons, New York (1959).

[32]Ragnar Rommetveit, Social Norms and Roles, Minneapolis: University of Minnesota Press (1955).

[33]B. Liao, N. Oren, L. van der Torre, S. Villata, Prioritized norms and defaults in formal argumentation, in: Deontic Logic and Normative Systems (2016), 2016.

[34]Royal Academy of Engineering, Safety and Ethics of Autonomous Systems, National Engineering Policy Center, (June 2020).

[35]Suk, J. E., A. Zmorzynska, I. Hunger, W. Biederbick, J. Sasse, Dual Use Research and Technological Diffusion: Reconsidering the Bioterrorism Threat Spectrum, (2011).

[36]B. Rappert, M. J. Selgelid, On the Dual Uses of Science and Ethics, Principles, Practices, and Prospects, Australian National University Press, (2013).

[37]A Guide To Canada's Export Control List, https://www.international.gc.ca/controls-controles/about-a_propos/expor/guide-2018.aspx?lang=eng, (December 2018).

[38]The Canadian Biosafety Guideline, Dual Use in Life Science Research https://www.canada.ca/en/public-health/ services/canadian-biosafety-standards- guidelines/guidance.html.

[39] N. Osman, Electronic Institutions and Their Applications, Springer International Publishing, (2019).