Explorative Experiments: A Paradigm Shift to Deal with Severe Uncertainty in Autonomous Robotics

Viola Schiaffonati
AI & Robotics Lab,
Department of Electronics,
Information and Bioengineering,
Politecnico di Milano

This paper presents a case of severe uncertainty in the development of autonomous and intelligent systems in Artificial Intelligence and autonomous robotics. After discussing how uncertainty emerges from the complexity of the systems and their interaction with unknown environments, the paper describes the novel framework of explorative experiments. This framework presents a suitable context in which many of the issues relative to uncertainty, both at the epistemological level and at the ethical one, in this field should be reframed. The case of autonomous robot systems for search and rescue is used to make the discussion more concrete.

1. Introduction

If uncertainty is incompleteness of knowledge, severe uncertainty is that form of uncertainty difficult to be evaluated also in probabilistic term. This paper presents a case of severe uncertainty in the context of technological development, and in particular in the development of autonomous and intelligent systems in Artificial Intelligence (AI) and autonomous robotics. In this context severe uncertainty ~~is related to~~ two different but related aspects: the first one is the complexity of the technical ~~artefact~~, the second one is the complex interaction of it with an unknown environment. Moreover, the current evolution of AI tools from model-based top-down approaches to data-based bottom-up solutions has further increased the degree of uncertainty in these fields. One of the main issues is relative to the poor capacity of the designers to predict the behavior of the systems under development. ~~The d~~esigners have only a loose control on the models

Perspectives on Science 2022, vol. 30, no. 2

https://doi.org/10.1162/posc_a_00415

used to calculate the solutions and, thus, have reduced possibilities to understand what the different parts of the models represent and to foresee their behavior in situations different from those considered during the development of the systems.

We argue that to deal with this uncertainty is not only a matter of developing appropriate techniques, but rather of adopting new conceptual categories, suggesting that some problems have to be addressed with methods having a philosophical nature. To this end, we introduce the notion of explorative experiment which gives reason of part of the experimental practice currently adopted in AI and autonomous robotics.

To make the discussion more concrete we ground it on a relevant class of AI systems: autonomous robots. In autonomous robotics AI systems are able to operate in unpredictable environments without a continuous human supervision (Siciliano and Khatib 2008). In particular we focus on the specific scenario of search and rescue settings, where robot systems operate in environments that have just endured disastrous events (e.g., earthquakes, nuclear accidents, or gas leaks) with the general goal of supporting human rescuers by assessing the state of the environment and detecting and rescuing victims (Tadokoro 2009). The scope of the purposes of robot systems in this context is wide and ranges from reporting means of access within the environment to finding and transporting victims to safe areas. Robot systems in search and rescue situations can have different degrees of autonomy, from robots teleoperated by human operators to fully autonomous robots that are only supervised by operators. In contrast to industrial settings in which the sources of uncertainty are controlled, in search and rescue settings robots operate in harsh and largely unpredictable environments. The search and rescue environment is representative of several other cases in which the increasing use of autonomous robots has a progressively large impact on our life, such as in the case of autonomous vehicles or assistance robots. In these cases it is clear how to predict the behavior of autonomous robots is essential to avoid dramatic consequences.

The goal of this article is not to provide an analysis of the different ways in which the notion of uncertainty can be defined ~~and its different relations with close concepts, as for example risk~~, but to address the problem of the predictability of the behavior of autonomous and intelligent systems as a case of severe uncertainty, that is incompleteness of knowledge difficult to be probabilistically quantified. Our focus here is on prediction, and in particular on the challenges human designers face in predicting the behavior of the systems under development, that is to predict the performance of the robot systems in the real settings in which they have to operate. Rather than the trial and error approaches currently adopted in autonomous robotics, we argue that to deal with the problem of how autonomous systems will perform when operating in unknown settings new forms of experimentation ~~for investigating the behavior of these systems~~ are required. Finally, we discuss the adoption of an existing ethical framework for experimenting with autonomous robots to overcome possible ethical issues.

Consider a hypothetical scenario to evidence the difficulty in predicting the behavior of a robot system in a search and rescue scenario. A multi-robot system is used to assess the presence and location of possible victims in a building where a gas leakage occurred. The system is composed of different mobile robots that can communicate with each other's. These robots have to explore the damaged building and report to the search and rescue team a map of it with the precise locations of the possible victims. This map can be used by human rescuers to get into the building and reach the victims very quickly, while reducing the risks they can face. Suppose that this multi-robot system can have two different configurations: configuration 1 and configuration 2. In configuration 1, the robots have long-range sensors, so they have an extended communication range, but they are very slow. In configuration 2, robots have short-range sensors, so they have a limited communication range, but they are faster than in configuration 1. What is interesting for the current discussion on uncertainty is that, at the moment, it is impossible to choose a priori (i.e., before actually deploying the robots in the damaged building) the best configuration between 1 and 2 because a prediction framework for this does not exist.

The structure of this article is as follows. In section 2, the shift from the classical model-based approach to the data-driven one in AI is presented. Section 3 discusses the issues in predicting the behavior of autonomous and intelligent systems and the importance of experimentation by focusing on the case of autonomous robot systems. In section 4 the concept of explorative experimentation is presented and related to the discussion about the issues of prediction in autonomous robotics. This section emphasizes that a different epistemological framework is needed when dealing with experimentation in autonomous robotics and in the engineering sciences in general. Section 5 presents the application of an ethical framework for experimental technologies to the case of robot systems for search and rescue. It ~~argues that also new ethical frameworks are required and~~ emphasizes how the explorative experimental approach offers both an epistemological and ethical framework to deal with severe uncertainty in the prediction of robots behavior within complex environments.

2. AI: From Models to Data

According to Nils Nilsson (1998) AI can be defined as the discipline concerning the development of intelligent behavior in ~~artefacts,~~ thus involving perception, reasoning, learning, communication, and acting in complex environments. Three main phases can be recognized in the process of developing an AI system according to the so-called classical approach (Russell and Norvig 2009). In the first phase (representation) the human designer starts from a problem and provides a model of that part of the reality that is relevant to solve this problem. The model computationally represents the main features of that part of reality, such as ~~their~~ properties and ~~their~~ interactions, and is expressed in an appropriate programming language. In the second phase the model is processed by a computer program to automatically derive from the represented knowledge the new knowledge representing the solution to the problem (reasoning). In the third phase the human designer evaluates in a critical way the solution that has been obtained: if she is not satisfied with it, she will iterate from the first phase in order to improve the solution.

Notwithstanding its simplicity, this conceptualization can give reason of many of the processes adopted in the development of AI systems over the years. For example, if we consider a computer program to play chess, the role of conceptualization based on representation and reasoning, together with their critical evaluation, is clear. In the first phase the designer represents, by using a programming language, the relevant elements of the game, including the state of the game and the legal moves. This model usually includes both declarative (explicitly represented) and procedural (implicitly represented in the program code) knowledge. Declarative knowledge is used to represent, for instance, which pieces are initially present on the chessboard, while procedural knowledge to represent how pieces move on the chessboard. In other words, declarative knowledge represents the state of the game and procedural knowledge allows the computer program to generate possible sequences of legal moves. In the second phase the computer program plays the game by operating on the computational model for generating possible sequences of moves and selecting the one that brings to a victory or to an advantageous non-terminal state. During these operations, the AI systems interact with the external world: the computer program to play chess performs the first move, then evaluates the move of the opponent, and generates possible sequences of moves from the new state of the game. In this case, the external world is considered deterministic and largely known. In the third phase the designer evaluates the performances of the computer program over a number of games and modifies the program accordingly, such as for instance the way in which the program evaluates non-terminal states.

The increasing adoption of Machine Learning (ML) techniques has challenged the above approach for the development of AI systems. According to Tom Mitchell (1997) ML is concerned with the issue of how to construct artificial systems able to automatically improve from their interaction with experience. It is worth noticing that learning in this context always refers to a specific task and a performance measure so that ML systems are always bounded by these elements. Also in this case, the designer starts from a problem, but the modeling activity of the first phase is different from the classical approach: she identifies a family of models, then the computer program builds the actual model starting from data. The family of models to start with can be, for example, neural networks and the data labelled examples (Haykin 2008). It is clear that in this case the designer loses control over the details of the built model (e.g., the correct weights are not set up by the designer but are modified by an algorithm) that will be used later to solve the problem. In other terms, this model is a black box and not even the designer can fully understand the very complex mechanisms of its functioning and the reasons why some solutions are achieved instead of others. Accordingly, the nature of the iterative repetitions of the three phases previously described changes: if in the ~~classic~~ approach to AI these repetitions can be seen as informed adjustments of the initial model based on the observed outcomes, in the ML approach these are mostly trial-and-error attempts to change the family of models initially identified by the human designer. It is worth stressing that, when adopting ML techniques, the number of activities performed by the computer program increases. While in the first phase the effort performed by the human designer is reduced, in the third phase her effort to understand and interpret the functioning of the systems is increased.

3. Predicting the Behavior of Autonomous and Intelligent Systems

The increasing adoption of ML techniques in the field of autonomous robotics is coupled with the intrinsic complexity of robot systems. Uncertainty is connected here to prediction, that is, the difficulties faced by human designers when dealing with predicting the performance of robot systems in the real settings in which they have to operate. This complexity refers to two different but related aspects. The first one is the complexity of the technical ~~artefact,~~ such as autonomous robot systems; the second one is the complex interaction of it with an environment that is not known a priori.

Unpredictability is not unique to AI systems: to some extent all complex computer programs are unpredictable. The difference, however, is that unpredictability of AI systems is somehow intrinsic. These systems can be more complex than usual programs; moreover, they are designed to operate in largely unknown environments. Such elements make it difficult, also for designers, to predict the behavior of these systems. In other words, this unpredictability is evident already at the level of design and is not only a problem related to the use. Predicting the outcomes of these systems before deploying them in contexts of use is very difficult.[1] Usually in autonomous robotics, to cope with the difficulty to predict how robot systems will perform when facing unknown settings, trial-and-error approaches are adopted. Unfortunately, trial-and-error approaches are not suitable to be employed in cases in which fragile and critical systems are involved, such as for instance in the case of flying robots that cannot experience too many collisions. A solution could be to use simulations to test the functioning of these systems and, only after a proper training, to use them in the real world. However, the gap between simulation and reality is very seriously perceived in autonomous robotics, and quite often the reasons why some results obtained in simulation can be generalized to the real world are not clear (see, e.g., Sadeghi and Levine 2017). A similar problem is that related to the passage from the controlled environment of a laboratory, where robot systems are developed and tested, to the external world, where these systems are supposed to be deployed at a later stage. For example, the case of the perception system of an autonomous robot developed by using offline data-sets evidences that the data used for training a system for robots and the data the system encounters during its operation are very different (Tzeng et al. 2016). One of the reasons for this mismatch is that training data are usually collected in environments more controlled (such as laboratories) than the unstructured environments in which robots actually operate. Several cases reinforce the idea that, even for their designers, is very difficult to foresee the performance of robot systems in settings that are also only slightly different from those in which they have been preliminary tested. Predicting what can happen in new settings without actually experimenting in them is almost impossible.

Although several techniques are under development to address these issues (see for example the methods developed to predict the value of a performance measure for a setting in which the robot has not yet operated (Amigoni et al. 2018)), these attempts are only at the beginning. In many cases, notwithstanding the careful design and testing of robotic systems, the exact knowledge about their behavior can be achieved only after performing experiments in the field.

The case of autonomous robotics can be considered representative of the difficulties met in predicting the behavior of autonomous and intelligent systems and their interactions with largely unpredictable environments, making them very similar to black box systems, whose intricate internal mechanisms are not completely clear. Moreover, systems that include

[1] It is interesting to consider this issue also in the light of the idea that every technical design is a form of creating determinacy (Fritzsche 2009). AI technologies challenge this understanding of technology as determinacy resulting from the process of engineering.

components built according to the ML approach makes things even more complex and the prediction of their behavior more difficult. For example, the case of some robot systems using convolutional neural network for place recognition and expected to be used onboard of autonomous vehicles clearly illustrates this aspect (Sünderhauf et al. 2015). In this case it is only after conducting experiments that an interesting result has been discovered: some neural networks trained for scene categorization perform better than neural networks trained for object recognition. This surprising result was not predictable before actually running the experiments. This case highlights that prediction is extremely important not only for scientific reasons, but also for guaranteeing safety to the human users.

Experimental procedures to understand the systems under development and be able to predict their behavior play an important role in this context. Rigorous experimentation is essential when trying to understand the behavior of robot systems and AI systems in general, also when ML techniques are adopted. Rather than looking for an a priori (e.g., theoretical) understanding of the behavior of a complex system, the focus is on iteratively refining this understanding through a form of experimentation that progressively discovers and conceptualizes how a system behaves in given circumstances.

4. Explorative Experiments

In this section we focus on how experimentation can play an important role for the problem of prediction described in the last section and we propose the framework of explorative experimentation. Our aim is not to provide a technical solution to this problem, but to suggest that the approaches allowing for the prediction of the behavior of AI systems should be developed within the framework of explorative experimentation, taking into account some epistemic peculiarities of experimentation in AI and robotics. The goal is to get a more solid empirical understanding of how such systems work and to better design and develop them.

4.1. Experimentation in Autonomous Robotics

Until few years ago experiments in robotics have been mainly devoted to assessing that a robot system works properly or, in the best cases, that it works better than comparable systems. Recently, a trend toward more rigorous experimentation has emerged and experimentation in autonomous robotics has developed in the direction of forms of experimentation at least in principle similar to those of the natural sciences. The traditional experimental principles of the natural sciences (reproducibility, repeatability, comparison, generalization, …) have inspired robotics in developing its experimental ~~approach~~. For example, the increasing tendency to publicly distribute code and datasets is a sign of how reproducibility is considered important (Amigoni et al. 2009). However, a systematic review of how experiments are conducted in this field show that some of above mentioned principles are far to be concretely adopted in the everyday practice of experimentation (Amigoni et al. 2014). For example, the testing of robot systems in different environments is still limited and, thus, it is rarely possible to generalize experimental results obtained in a particular environment to similar ones.

Within this context it is reasonable to ask whether it does make sense to apply the same standards (e.g., experimental principles) of the natural sciences to the engineering ones, such as robotics and AI. There are important differences in the experimental practices of the engineering sciences that justify the introduction of a different epistemological framework based on the notion of "explorative experiment" (Schiaffonati 2016, 2020). Explorative experiments are ~~a~~ technological ~~form~~ of experimentation with the goal of testing whether the designed technical ~~artefacts~~ meet their desired specifications, rather than to test a theory or to refute theoretical hypotheses. To better understand explorative experiments we consider now their main features.

4.2. Technical Artifacts

In explorative experiments technical artifacts (in the case of autonomous robotics: programs, software systems, computers) are evaluated. Experiments in the engineering and in the natural sciences have different goals: to test artifacts in the former case, to test theories or models in the latter one. Technical artifacts are physical objects with a technical function and a use plan: they are intentionally built by humans to fulfill a given function (Vermaas et al. 2011). The fact that a technical artifact is the result of an intentional human action plays an important role in the characterization of the way experiments are conducted. The goal in testing technical artifacts is to evaluate whether and how they are able to fulfil the technical function for which they have been designed. It is, thus, clear that a normative judgment plays a role in this experimental practice: a robot system for example can be evaluated as better or worse with respect to a given function that works as a reference model. On the contrary, a natural phenomenon (which is usually what is investigated in an experiment in the natural sciences), such as an electron, cannot be good or bad: the electron in the experiment is evaluated without any reference to its supposed technical function, that is, without any normative constraint with respect to its correct functioning.

Consider again search and rescue robotics depicted in Section 1. In this context robot systems support human rescuers to locate possible victims in a disaster environment. These robot systems are developed to cover a technical function, for instance to explore in the minimum amount of time the most extended portion of an unknown area (e.g., the building in which a gas leakage has occurred). During this exploration, robots can collect information about possible victims, safe paths to be used by human rescuers, and the structural safety of the damaged building. More precisely, the technical function of these robots can be defined according to two different dimensions: the task they have to perform, such as for instance to detect the highest number of victims in the shortest period of time; the environment in which this task occurs, such as the damaged building after the gas leakage. In this case the task is composed of both an activity (detecting possible victims) and a way to quantitatively measure the performance of the robots during this activity (the time constraints that are introduced). Because these robots are designed to implement a technical function, so their physical composition depends on this. More precisely, the physical components of the robot system are selected according to the task and the environment involved. For example, these robots can move on wheels only if the environment in which they have to operate is smooth enough and without big obstacles. If this is not the case, locomotion needs to be reconsidered, for example by using legged or aerial robots. In this example the experimental evaluation of robot systems as technological artifacts is given in the process of assessing their performance with respect to a reference model (for instance to locate all the victims in a given environment in a given time) and a metrics (for instance the difference between the real number of the victims and the number of victims located by the robots in the given time).

4.3. Designers are Experimenters

Another important feature of explorative experiments is that in many cases those who design technical artifacts are also those who test them.[2] This is an important difference with respect to the traditional experimental protocol of the natural sciences in which the independence of the experimenter from the investigated phenomenon is—at least ideally—prescribed. This prescription plays a role in trying to prevent or mitigate the effects of the experimenter's prejudice. On the contrary, in ~~the engineering sciences~~ it is not clear how the experimenter can be independent from the artifact that she has created (Tedre 2011). The need for the designer to test an artifact she

[2] Exceptions are, of course, possible such as, for instance, the following examples: those experiments aimed at testing the computer/robot-human interaction and those initiatives promoting open source libraries to share source code and datasets to improve the experimental practices of a larger community.

has designed and implemented resides in the complexity of the artifacts themselves and of their interaction with their environment that ~~makes~~ the prediction of their behavior highly uncertain (see Section 3).

In the case of search and rescue robots, for example, the experimental evaluation requires a degree of dynamicity that means to take into account the features both of the robot system and of the experimental context. The attempt to anticipate during the design phase the issues arising at the implementation level is one of the tenets of the engineering practice, given that the distance between the expected behavior and the effective one is always greater than expected (Vincenti 1990). However, these issues are amplified in the case of autonomous robots employed in critical situations, such as search and rescue contexts, given that it is impossible to model during the design phase all the issues that the robot system could meet in practice. This form of severe uncertainty cannot be probabilistically quantified and concerns the "unknown unknown" that the robot systems can meet especially when interacting with unpredictable environments, such as an environment while and after a natural disaster. To discover the actual behavior of these technical artifacts in their environments a partially different experimental framework is needed such as that of explorative experimentation.

4.4. A Different Notion of Experimental Control

So far we have seen how explorative experiments in the engineering sciences differ from controlled experiments in the natural sciences because the objects of their investigation are technical artifacts that are evaluated with respect to their technical function. Here we focus on the differences in the experimental methodology, and especially on the notion of control. Experimental control is at the core of experimental activity. It consists of the selection of the factors to be investigated in the experimental process and in their manipulation and is an essential part in the design of a successful experiment. When considering experiments with autonomous robots, this form of control is not possible to be managed in the design phase but is partly accomplished only when these robots are inserted in the environment in which they have to operate.

Consider an autonomous robot moving from location A to location B when obstacles are present but are not explicitly considered and modeled in the design phase. It almost impossible to predict the behavior of this robot in the case of non-polygonal obstacles, such as round ones that are very often present in the real world. This difficulty is due to several elements: errors occurring when deciding how to deal with a perceived obstacle, errors in the locomotion to avoid these obstacles, errors in the software used to decide how the robot should behave after an obstacle has been detected, errors occurring when a planned action is not executed in the expected way. Some of the errors can be solved by adopting the traditional tools provided by software testing. However, not all of them can be avoided in this way, since the environments in which the robots operate are poorly controlled. To model and predict all these aspects is not only a problem due to a lack of technological knowledge, but is intrinsic to the context. To try to anticipate which elements of the environments are not adequately represented in the model is not possible before the robot system is deployed and employed in a real environment. The term "explorative" is precisely used to emphasize this element: explorative experiments are performed to explore different possibilities, to investigate possible interventions and to collect information, that are later used to improve the technical artifacts under investigation. As a consequence, ~~the~~ experimental control cannot operate already from the beginning of the experimental process, and in particular in the design phase. The traditional experimental protocol, in which the experimenter acts from a centre of command and control outside the experimental scenario, progressively disappears (Kroes 2016).

4.5. Directly Action-Guiding Experiments

The difference between epistemic experiments and directly action-guiding experiments, ~~(~~discussed in Hansson 2015, 2016), can further clarify the notion of explorative experiment~~s~~ and its differences with the traditional one~~s~~. Historical and philosophical accounts are not always capable of giving reason of the multifaceted nature of the notion of experiment and tend to present a uniform notion of it as emerged during the Scientific Revolution of the XVII century.[3] Looking back to the pre-scientific and non-academic forms of experimentation, (Hansson 2015, 2016) individuates technological forms of experimentation taking place in ancient Syria already from the XIII century. They are labelled directly action-guiding experiments to evidence the strong connection between knowledge and action. In other words, it is by means of an action, or better of those people performing that action, that the efficiency of the knowledge achieved after the experimental process is made manifest. This is, for instance, the case of a physician who wants to know more about a disease to prevent or to treat it, and not only to obtain knowledge in general.

Direct action-guiding experiments thus are different from epistemic ones: if an epistemic experiment has the goal to improve knowledge about the way in which some phenomena occur in the natural world, a directly action-guiding experiment ~~has to~~ satisfy two criteria. The first one is that

[3] Werret (2019) represents an interesting exception.

the expected result should consist in the attainment of a desirable goal for human action: in a clinical trial of an analgesic, for example, the desired result is an efficient reduction of pain with minimal collateral effect. The second criterion is that the planned intervention should be implemented in non-experimental contexts to obtain the desired goal: in the case of a clinical trial of an analgesic the experimental intervention is the correct treatment to be administered in order to reduce pain of patients outside the clinical trial. The systematic test of an autonomous robot to be employed to assist an elderly person in his home can be seen as another example of a directly action-guiding experiment: the expected result is the correct functioning of the robot and its proper interaction with the elderly person in his home; the experimental intervention consists in the fine-tuning of the abilities of the robot to achieve this result. Direct action-guiding experiments are independent from theories, or better, they have a higher degree of independency with respect to epistemic ones. Being devoted to action, they are focused on effects: if one wants to know whether it is possible to obtain A by means of B, one tries to implement B and checks if A has been obtained.

Explorative experiments can be seen as directly action-guiding experiments. Their goal is not epistemic, but concretely devoted to an action for intervening in the reality.[4] The idea of experiments partly independent from theories is not new. For example, in the philosophy of biology the label exploratory experiment has been used to describe those experimental processes that are not guided by theories. Steinle (1997) uses the label "exploratory" to describe experiments performed to obtain empirical regularities when theories or theoretical frameworks are not available (for further similar uses see Burian 1997; Elliot 2007). However, it is not by chance that we use the term explorative, rather than exploratory, to stress a form of investigation not guided by a theory and addressed to exploring the functioning of a technical artifact (and not to investigate a natural object).

Moreover, explorative experiments share some features with experimentation in the social sciences, and in particular with the notion of field experiment often related to the very idea of exploration. For instance, similarly to explorative experiments, design experiments, as proposed~~,~~ for instance~~,~~ in the political sciences (Ansell 2012), show the problem of isolating the effect of single variables in field experiments and the fact that their focus is not to test a theory, but to refine an intervention in the world.

[4] In general, all types of experiments have an epistemic component, also the explorative ones. However, in explorative experiments the epistemic component is not primary such as in other cases.

Given the many similarities with field experiments, however, explorative experiments are a technological form of experimentation on technical artifacts where the normative component (how much the tested technical artifacts conform to its design) plays a central role. This makes them similar to some methodological reflections developed in the field of Design Science Research, where the iterative and evolutionary nature of design improvements through exploration is emphasized (Gill and Hevner 2013).

In conclusion an explorative experiment is a form of experimental investigation devoted to test a technical artifact without the control boundaries typical of an epistemic controlled experiment. Its goal is to investigate the possibilities and limits of the technical artifact and its interaction with the surrounding environment. The design of this investigation is not conducted on the basis of a well-formed theory or a systematic theoretical background. Accordingly, the initial hypotheses cannot always be formulated in a clear way, and the type of knowledge which is the goal of this experimentation is oriented to evaluate the performances of the technical artifact with respect to ~~their~~ technical function. The experimenter is often the designer of the artifact, thus her independence from the experimental context is not guaranteed.

5. An Ethical Framework for Explorative Experiments

The uncertainty in predicting the behavior of autonomous robots operating in unknown environments has an impact not only at the epistemological level, but also at the ethical one. Severe uncertainty in autonomous and intelligent systems can have dramatic effects on our lives, such as in the case of autonomous vehicles, household robots for assistance, or robot systems in search and rescue scenarios. So far, we have emphasized the need of an epistemological shift to deal with this type of uncertainty and introduced the notion of explorative experimentation. In this section we discuss how these new epistemological categories have an impact on ~~new ethical frameworks~~. Like the traditional epistemological and methodological categories of the natural sciences cannot be directly applied to the engineering sciences, in the same way the moral principles developed for experimentation in the natural sciences need to be reconsidered in the case of the explorative experimentation with new technologies, such as autonomous and intelligent systems.

5.1. Experimental Technologies

Autonomous robots can be considered as experimental technology, meaning that the operational experience relative to their effective behavior is limited and, therefore, the attempts to precisely assess their societal risks and benefits are uncertain: this means that their impact on humans and societies ~~are~~ mostly unknown and difficult to predict (van de Poel 2016). This focus on the experimental nature of autonomous robot systems ~~further~~ makes evident how the inherent uncertainty that ~~characterize~~ them cannot be solved with traditional tools (such as statistical judgments) but need to be managed in the context of a framework which is novel from both an epistemological and ethical point of view. In general, any robot system can be considered an experimental technology; however, in the case of autonomous robot systems this uncertainty is more evident as the robots operate in highly unstructured environments.

On the one hand, anticipation of the societal consequences of technologies has been employed to establish moral and regulatory frameworks; in particular, to try to overcome uncertainty, traditional predictive methods (e.g., cost-benefit analysis or risk assessment) have been adopted. Recently, for example, a vast effort has been put in place by IEEE (*Institute of Electrical and Electronics Engineers*) with the Ethically Aligned Design document (IEEE 2018) formulated by the *IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems*. Its goal is to solve, also by means of anticipation, potential ethical issues in the design of intelligent and autonomous systems. Anticipation, however, is not always enough to deal with these issues: as seen, these are experimental technologies characterized by a limited operational experience which is insufficient to clearly anticipate their risks and benefits. Therefore, their impacts are largely unknown and difficult to predict: a form of severe uncertainty is inherent in the introduction of these new technologies into society. On the other hand, adopting the precautionary principle can bring several disadvantages, first of all to stop any technological development for which the knowledge about its impact is uncertain.

To acknowledge the uncertainty in the development of experimental technologies means to recognize that unexpected events can always occur, and that a different approach is required. This approach is a form of incrementalism, where experimental technologies are gradually introduced into society to constantly monitor the societal effects that emerge and to iteratively improve their design accordingly. In the case of autonomous robotics, incrementalism means that experiments are concerned with the gradual introduction of robot systems in their real context of use. Explorative experiments, devoted to acquiring knowledge on the behavior of robot systems in the real world, are therefore also crucial to address ethical issues related to the impact on such robots on society.

5.2. An Ethical Framework for Autonomous Robots

The conceptualization of autonomous robots as experimental technologies presents an interesting advantage, that is to adopt an ethical framework

~~already~~ developed for experiments concerning experimental technologies in general (van de Poel 2016).[5] This ethical framework is composed of sixteen different conditions that can be organized around the four moral principles of bioethics: non-maleficence, the attempt to avoid any damage; beneficence, trying to prevent harm and promote well-being; respect for autonomy, ensuring that people are free to act according to their personal values and beliefs; respect for justice, the attempt to fairly distributing costs and benefits. It is worth noticing that these principles are not guidelines for solving any moral problem but form an ethical framework that constantly evolves and be used in different contexts. The conditions of this framework are provisional, meaning that under some circumstances they cannot be applied. In other words, they are constantly open to be improved on the basis of the experience acquired after their concrete application that offers a test bench to critically revise them.

Once again, to make the discussion more concrete, robots for search and rescue scenarios are considered (for further details see (Amigoni and Schiaffonati 2018)). In this context we do not suggest that the problems related to uncertainty can be solved only by an appropriate design of robot systems. Neither that extended testing can overcome this uncertainty. Rather we claim that to concretely minimize the risks associated to experimental technologies the first step is to understand what it means to experiment on them. The framework of explorative experiments has thus an impact not only at the methodological level, but also at the ethical one, where the traditional moral categories need to be revised to deal with such experimental technologies.

We now consider the conditions applied to robot systems as experimental technologies. These are the conditions relative to the principle of non-maleficence:

1) *Absence of other reasonable means for gaining knowledge about risks and benefits*. This condition requires that, before adopting a robot system in a search and rescue scenario, any other means to acquire knowledge about the robot and its risks have been investigated. To adopt an incremental perspective means that, before deploying robots in real contexts of use, it is necessary to experiment with them in laboratories or in controlled environments in order to evaluate their performances under different conditions. The final goal is to

[5] Another interesting proposal for the assessment of ethical requirements, specifically developed for AI technologies, is provided by the European efforts implemented in the Ethics Guidelines for Trustworthy AI and the associated Assessment List: https://ec.europa.eu/futurium/en/ai-alliance-consultation.

develop a robot system able to avoid harm or, more realistically, to minimize it after its adoption in a search and rescue scenario.

2) *Monitoring of data and risks while addressing privacy concerns.* This condition requires the continuous monitoring of the operations of robot systems for search and rescue in order to avoid damage to both the victims and the operators and to respect their privacy. For example, the human operator must constantly monitor robot's operations to avoid possible damages to the victims. Although risks related to privacy are not immediately perceived in a search and rescue context, nonetheless any measure to protect the privacy of the victims has to be implemented. Here it is clear that traditional tools to deal with privacy issues, such as informed consent, cannot be applied due both to the emergency situation and to the use of autonomous robots ~~in~~ it. Thus, new solutions need to be developed.
3) *Possibility and willingness to adapt or stop the experiment*. This condition is connected to the previous one and requires that any activity of the robot system could stop or be modified in the case of possible damages to the humans involved in the scenario. For instance, a robot for search and rescue should be designed to be equipped with a large red button to be pushed in the case of danger for the victims.
4) *Containment of risks as far as reasonably possible*. This condition prescribes that, at least in principle, risk should be minimized. However, a situation without any risk is impossible in those contexts in which robots are employed for searching and rescuing victims of a disaster. In a more realistic way, this condition requires that possible measures to contain risk should be adopted, such as the use of soft materials for the external parts of the robot devoted to the rescue of victims.
5) *Consciously scaling up to avoid large-scale harm and to improve learning*. This condition refers to the gradual introduction of the robot systems in the search and rescue environment in which they operate. A possibility could be to use them in smaller buildings without victims at first, and then to progressively use them in larger areas with different victims under different conditions. This is a way to gradually improve the knowledge about the working of the robot systems to avoid large scale harm.
6) *Flexible setup of the experiment and avoidance of lock-in of the technology.* This condition requires different levels of flexibility in setting up the experiment to progressively increase its complexity. Without such flexibility a possible error is to lock-in a specific technology without considering other possibilities. This is rather common as, for instance, when a wheeled robot is deployed in a search and

rescue scenario, while a flying robot would work better, but to switch from one to another is very expensive.

7) *Avoid experiments that undermine resilience*. This condition expresses the idea that minimization of risk cannot be measured in absolute terms but depends on the resilience of each of the humans involved in an unexpected risk. It makes a substantial difference to minimize the risk for robots operating in environments without human victims with respect to environments in which victims are to be located and transported. Moreover, there exists a difference in the resilience of operators trained to work in critical situations and of victims without any training who are particularly vulnerable because of their condition.

These are the conditions relative to the principle of beneficence:

1) *Reasonable to expect social benefits from the experiment*. This condition expresses that the introduction of robots in search and rescue scenarios should offer benefits for society. Given that the use of robots is not free of risks, the advantages for society should compensate for these risks, for example, increasing the possibility to save human lives or to better protect human rescuers.
2) *Clear distribution of responsibilities for setting up, carrying out, monitoring, evaluating, adapting, and stopping of the experiment*. This condition is critical for robotics: it requires a sharp distinction between scientifically trained experimenters and the subjects of experiments. Responsibilities have to be clearly distributed, such as in the different experimental steps and in the respective involvement of experimental subjects.

These are the conditions relative to the principles of respect for autonomy and justice:

1) *Experimental subjects are informed*. This condition applies every time human subjects are involved in the experiments and can be fulfilled by adopting informed consent. However, informed consent cannot be used in emergency situations. To partly avoid this lack of information robots could be equipped with sounds and lights to make their presence visible to the victims.
2) *The experiment is approved by democratically legitimized bodies*. This condition applies when the consent is not individually provided but derives from a democratically legitimized organ. In the case of search and rescue robots it could be easier to obtain this type of consent rather than the informed consent of the single subjects.

3) *Experimental subjects can influence the setting up, carrying out, monitoring, evaluating, adapting, and stopping the experiment.* This condition guarantees that the previous condition cannot be abused by a group deciding to impose sacrifice on an individual. Every involved experimental subject should have a role in the setting up of the experiments. This is particularly critical in the case of search and rescue robots as the victims are not in the position to be involved in this process. A possibility could be to involve former victims in the process of defining the most appropriate experiments.
4) *Experimental subjects can withdraw from the experiment.* This condition is in line with the previous one. For example, a conscious victim in a search and rescue setting should always have the possibility to ask to not be transported by a robot.
5) *Vulnerable experimental subjects are either not subject to the experiment or are additionally protected or particularly profit from the experimental technology (or a combination).* This is a condition requiring that vulnerable subjects should be protected, or even not involved, in experiments with experimental technologies. This condition is critical with search and rescue robots because the victims are particularly vulnerable while, at the same time, are the subjects for whom these technologies are developed.
6) *Fair distribution of potential hazards and benefits.* This condition requires a fair distribution of risks and benefits but, again, is particularly critical in the case of search and rescue robots. For example, in the case in which a robot decides a priority in the order for which the victims are reported to the human rescuers.
7) *Reversibility of harm or, if impossible, compensation of harm.* This condition requires that, in case of unavoidable damage, some compensation mechanisms are put in place. A form of compensation could be a monetary one, similarly to what happens with insurance, in the case of robots potentially able to cause damages to the victims that they have to rescue.

6. Conclusions

This paper presents a case of severe uncertainty in technological development. We have discussed how uncertainty raises from the difficulty of predicting the behavior of autonomous robots in their real context of application. This uncertainty is further amplified with the shift in AI from the classical approach to ML techniques, where the modeling activity can be seen as a black box also for its designers. To overcome this opacity a novel approach labelled as "Machine Behavior" has been recently proposed

(Rahwan et al. 2019). The proponents suggest that we need new tools to study the behavior of increasingly complex machines in connection with the environments in which they operate. The invitation, in this case, is to use the tools of the social sciences, given for granted the analogy between animals and machines. In particular, it is proposed that the current methods of the behavioral sciences should be adopted to deal with the difficulties related to the unpredictability of machines. In this paper we have suggested, instead, that a specific experimental methodology must be developed for the engineering sciences. Explorative experimentation is adopted to stress the peculiarities of experimental processes devoted to test technical artifacts, such as autonomous robots, and the uncertainty in predicting their behavior. We have argued that this different framework based on the idea of explorative experimentation can help in the conceptualization of the issues related to uncertainty and in dealing with the ethical issues of experimenting and adopting these robots. This paper is an attempt to show how some problems related to uncertainty have to be addressed with methods having a philosophical nature. Further work is needed both in the refinement of the concept of explorative experiment, with the support of other fields of research, and in its application in different technological domains.

References

Amigoni, F., V. Castelli, F. Bonsignorio, and M. Luperto. 2018. "Predicting Robot Performance: Why and How." *IJCAI-ECAI/ICML/AAMAS Federated AI for Robotics Workshop (FAIR).*

Amigoni, F., M. Reggiani, and V. Schiaffonati. 2009. "An Insightful Comparison between Experiments in Mobile Robotics and in Science." *Autonomous Robots* 27(4): 313–325. https://doi.org/10.1007/s10514-009-9137-8

Amigoni, F., V. Schiaffonati, and M. Verdicchio. 2014. "Good Experimental Methodologies for Autonomous Robotics: From Theory to Practice." Pp. 37–53 in *Methods and Experimental Techniques in Computer Engineering*. Edited by F. Amigoni and V. Schiaffonati. Cham: Springer. https://doi.org/10.1007/978-3-319-00272-9_3

Amigoni, F., and V. Schiaffonati. 2018. "Ethics for Robots as Experimental Technologies: Pairing Anticipation with Exploration to Evaluate the Social Impact of Robotics." *IEEE Robotics and Automation Magazine* 25(1): 30–36. https://doi.org/10.1109/MRA.2017.2781543

Ansell, C. 2012. "What is a 'democratic experiment'?" *Contemporary Pragmatism* 9(2): 159–180. https://doi.org/10.1163/18758185-90000235

Burian, R. M. 1997. "Exploratory Experimentation and the Role of Histochemical Techniques in the Work of Jean Brachet, 1938–1952." *History and Philosophy of the Life Sciences* 19: 27–45. PubMed: 9284641

Elliot, K. C. 2007. "Varieties of Exploratory Experimentation in Nanotoxicology." *History and Philosophy of the Life Sciences* 23(3): 313–336.

Fritzsche, A. 2009. "Engineering Determinacy: The Exclusiveness of Technology and the Presence of the Indeterminate." Pp. 305–312 in *Philosophy and Engineering*. Edited by I. van de Poel and D. E. Goldberg. Dordrecht/Heidelberg/London/New York: Springer. https://doi.org/10.1007/978-90-481-2804-4_26

Gill, T. G., and A. N. Hevner. 2013. "A Fitness-Utility Model for Design Science Research." *ACM Transactions on Management Information Systems* 4(2): 5–24. https://doi.org/10.1145/2499962.2499963

Hansson, S. O. 2015. "Experiments before Science? – What Science Learned from Technological Experiments." Pp. 81–110 in *The Role of Technology in Science: Philosophical Perspectives*. Edited by S. O. Hansson. Dordrecht: Springer. https://doi.org/10.1007/978-94-017-9762-7_5

Hansson, S. O. 2016. "Experiments: Why and How?" *Science and Engineering Ethics* 22: 613–632. https://doi.org/10.1007/s11948-015-9635-3, PubMed: 25721443

Haykin, S. 2008. *Neural Networks: A Comprehensive Foundation*. New York: Prentice Hall.

IEEE Standards Association. 2018. *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems*. IEEE.org. https://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

Kroes, P. 2016. "Experiments on Socio-Technical Systems: The Problem of Control." *Science and Engineering Ethics* 22: 633–645. https://doi.org/10.1007/s11948-015-9634-4, PubMed: 25702146

Mitchell, T. 1997. *Machine Learning*. New York: McGraw-Hill.

Nilsson, N. 1998. *Artificial Intelligence: A New Synthesis*. San Francisco: Morgan Kaufmann.

Rahwan, I. et al. 2019. "Machine Behaviour." *Nature* 568: 477–486. https://doi.org/10.1038/s41586-019-1138-y, PubMed: 31019318

Russell, S., & Norvig, P. 2009. *Artificial Intelligence: A Modern Approach* 3rd ed. New York: Prentice Hall.

Schiaffonati, V. 2016. "Stretching the Traditional Notion of Experiment in Computing: Explorative Experiments." *Science and Engineering Ethics* 22(3): 647–665. https://doi.org/10.1007/s11948-015-9655-z, PubMed: 26018042

Schiaffonati, V. 2020. *Computer, Robot Ed Esperimenti*. Milano: Meltemi.

Sadeghi, F., and S. Levine. 2017. "CAD2RL: Real Single-Image Flight without a Single Real Image." *Proceedings of Robotics: Science and Systems*.

https://arxiv.org/pdf/1611.04201.pdf. https://doi.org/10.15607/RSS.2017.XIII.034

Siciliano, B., and O. I. Khatib. (Eds.). 2008. *Springer Handbook of Robotics*. Heidelberg: Springer. https://doi.org/10.1007/978-3-540-30301-5

Steinle, F. 1997. "Entering New Fields: Exploratory Uses of Experimentation." *Philosophy of Science* 64: S65–S67. https://doi.org/10.1086/392587

Sünderhauf, N., S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford. 2015. "On the Performance on Convnet Features for Place Recognition." *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems* (IROS): 4297–4304. https://doi.org/10.1109/IROS.2015.7353986

Tadokoro, S. 2009. *Rescue Robotics*. London: Springer-Verlag. https://doi.org/10.1007/978-1-84882-474-4

Tzeng, E., C. Devin, J. Hoffman, C. Finn, P. Abbeel, S. Levine, S. Saenko, and T. Darrell. 2016. "Adapting Deep Visuomotor Representations with Weak Pairwise Constraints." *Proceedings of the Workshop on the Algorithmic Foundations of Robotics (WAFR).*

Tedre, M. 2011. "Computing as a Science: A Survey of Competing Viewpoints." *Minds & Machines* 21(3): 361–387. https://doi.org/10.1007/s11023-011-9240-4

van de Poel, I. 2016. "An Ethical Framework for Evaluating Experimental Technology." *Science and Engineering Ethics* 22: 667–686. https://doi.org/10.1007/s11948-015-9724-3, PubMed: 26573302

Vermaas, P., P. Kroes, I. van de Poel, M. Franssen, and W. Houkes. 2011. *A Philosophy of Technology: From Technical Artefacts to Sociotechnical Systems*. San Rafael, CA: Morgan & Claypool Publishers. https://doi.org/10.2200/S00321ED1V01Y201012ETS014

Vincenti, W. 1990. *What Engineers Know and How They Know It*. Baltimore and London: The John Hopkins University.

Werret, S. 2019. *Thrifty Science*. Chicago: The University of Chicago Press. https://doi.org/10.7208/chicago/9780226610399.001.0001

AUTHOR QUERY

No query.