Learning in Autonomous and Intelligent Systems: Overview and biases from data sources

Abstract: Autonomous and Intelligent Systems (A/IS, to adhere to the terminology of the IEEE Ethically Aligned Design report) can gather their knowledge by different means and from different sources. Learning algorithms are in principle neutral, it is rather the data they are fed with during the learning period that can introduce biases or a specific ethical orientation. Human control over the learning process is more straightforward in learning from demonstration, where data sources are restricted to the choices of the demonstrator (or teacher) but even in the unsupervised versions of reinforcement learning biases are present via the definition of the reward function. In this paper we provide an overview of learning paradigms of artificial systems: supervised and unsupervised methods, with the most outstanding examples of each category, without too much technical detail. Furthermore, we describe the types of data sources that are presently available and in use by the robotics community. We do also focus on observable bias in image datasets and originated by human annotation. We point at quite recent research on bias in social robot navigation and end with a brief reflection about ambient influences on future learning robots.

KEYWORDS:
Autonomous and Intelligent Systems; automatic learning methods; bias in data sources

Resumen: Los sistemas autónomos e inteligentes (A/IS por sus siglas en inglés, en concordancia con el informe del IEEE sobre diseño alineado con la ética) pueden obtener sus conocimientos a través de diferentes procedimientos y de fuentes diversas. Los algoritmos de aprendizaje son neutros en principio, son más bien los datos con los que se alimentan durante el período de aprendizaje que pueden introducir sesgos o una orientación ética específica. El control humano sobre el proceso de aprendizaje es más directo en aprendizaje por demostración, donde las fuentes de datos están restringidas a las elecciones del demostrador (o profesor), pero incluso en las versiones no supervisadas del aprendizaje por refuerzo los sesgos están presentes a través de la definición de la función de recompensa. En este artículo proporcionamos una visión general de los paradigmas de aprendizaje de los sistemas artificiales: métodos supervisados y no supervisados, con los ejemplos más destacados de cada categoría, sin profundizar demasiado en el detalle técnico. Además describimos los tipos de fuentes de datos disponibles actualmente y su uso por la comunidad robótica. También enfatizamos el sesgo que se observa en bases de datos de imágenes y originados por anotación humana. Destacamos una investigación muy reciente sobre sesgo en navegación de robots sociales y finalizamos con una breve reflexión sobre influencia del ambiente sobre futuros robots que aprenden.
PALABRAS CLAVE:
Sistemas autónomos e inteligentes; métodos de aprendizaje automático; sesgo en fuentes de datos

Dr. Pablo Jiménez Schlegl
Institut de Robòtica i Informàtica Industrial (CSIC-UPC)
Llorens i Artigas, 4-6
Campus Sud UPC, Edifici U
08028 Barcelona

1. Introduction

Why learning? Why should an artificial intelligent system such as an autonomous robot be endowed with this cognitive function? Is careful planning not enough to cope with all the tasks it has to perform? The first answer is straightforward: it can hardly earn the appellative of intelligent if it is not able to learn. If it performs an action, be it mental or physical, which in given circumstances results in failure, we would certainly qualify the system as stupid if it performed the same action once and again under the same or similar circumstances. Autonomy is also a feature that requires

learning capacities: the way to gather knowledge without depending on a provider. Systems performing their activity in a limited and controlled world, like expert systems or industrial robots, certainly may complete their duties with just a set of rules (the knowledge base) and data introduced by their human programmers. But even in such restricted environments the trend is to widen the versatility and to enrich the interaction capabilities of the artificial systems, which includes the necessary skills to deal with unforeseen situations and to adapt their behavior to varying circumstances. This fact was already identified in the late 80's. «Learning capabilities are needed for intelligent systems that can remain useful in the face of changing environments or changing standards of expertise» (Buchanan,1989: p. 251). For this reason expert systems are being progressively implemented as neural networks, whose knowledge base has to be trained and/or refined via learning, and this is also the reason why industrial robots are shifting from the «caged dangerous machine» towards the *cobot* (*co*llaborative ro*bot*) paradigm. The need of being endowed with learning skills becomes all the more paramount in the development and deployment of field, service, domestic, and personal robots. Beyond the controlled and structured environments that constituted the limited world of most robots up to now, they have to be released into a new, open, complex, wide world they know nothing about. The human programmers of robots may certainly furnish them with a declarative representation of some of the world's features, as well as with some procedural information about the tasks they have to perform. But such a formal description is clearly incomplete, inaccurate, simplistic and hardly useful for mere survival. Stated differently, robots have to deploy their activity in a world that is partially known, (i.e., the knowledge about the world is incomplete), partially observable (it cannot be taken for granted that even the relevant observations for the robot's task will be registered), and dynamic (i.e., changing). Changes in the world are due either to ambient phenomena or to actions performed by other agents. Some of such changes can be anticipated with reasonable assumptions on expected behaviors: ambient transitions either follow physical laws (natural phenomena) or anthropogenic ways (e.g. traffic rules), whereas agent actions can be predicted from the knowledge about its motivations, goals, as well as capabilities. Even in the case that these declarative facts and procedural skills have been previously encoded in the robot's knowledge base, it may still be necessary to accommodate such knowledge to the particular circumstances affecting the deployment of the robot. Consider for example a domestic robot, which is endowed with a set of basic skills by the manufacturer, but has to adapt its behavior to the specifics of its new home and owners: from the very geometry and layout of rooms and furniture, up to the wishes and whims of the human inhabitants. In sum, non-coded as well as unpredictable knowledge, and adaptation to upcoming changes and refinement of skills render learning as an unavoidable requisite for autonomous robots to be deployed in unstructured environments. To this end, different families of machine learning techniques apply.

2. Fundamentals in Learning

Learning strategies are typically grouped into two big families, namely **supervised** and **unsupervised** methods. This classification is quite enlightening and, at the same time, intuitive, as it refers to a basic methodological question: to learn by external guidance or to learn by its own.

2.1. Supervised learning

This modality requires the involvement of an agent in the role of a teacher (generally a human), who either provides to the learning (artificial) system a controlled learning set of inputs and the corresponding responses, or gives feedback to the system about its learning performance, correcting it (if necessary) after execution of the learned action or task. Expressed otherwise, assume that the goal of learning is expressed as computing the function f that relates a given input X with an output Y, that is, $Y = f(X)$. Then, in supervised learning, the corresponding Y to certain X is always provided by the teacher, who supervises the learning process in terms of function f being

approximated increasingly better along successive iterations. The teacher does also decide when the learning system has achieved an acceptable level of performance, and in this moment learning is completed. Within supervised learning methods we may further distinguish between **classification methods** (each output *Y* is a class or category, and function *f* has to correctly assign each individual *X* to its class) and **regression algorithms** (both variables are numbers and *f* may be an analytical function such as for example in linear regression, where inputs and outputs are related by a linear law). This is straightforward if numerical variables are considered, but supervised learning can also take place at a symbolic level, like **inductive logic programming**, which aims at synthesizing a minimal logical program capable of providing the correct true or false values to the corresponding input variables. Popular classification algorithms include:

• Support Vector Machines (SVM), which compute separating hyperplanes between classes (e.g., lines in the two-dimensional case) such that data are classified as belonging to one class or another depending on which side of this hyperplane they are. These planes are computed in a way that the gap separating the data belonging to different classes is maximized (the base procedure is linear, but the so-called kernel trick allows also for non-linear separating functions).

• Statistical methods like Bayesian learning and its variants maximum likelihood and expectation maximization. This application of the Bayes rule allows to determine the probability of the input data belonging to a given class, knowing the priors of class distribution in advance (that is, the probability of occurrence of each class) and having determined, during the training phase, the likelihoods of the data for each class.

• Neural networks (NN, aka Artificial Neural Networks, ANN), which consist in combining basic computational units, called neurons, linked by weighted connections. Each such neuron computes the weighted sum of its inputs and fires (activates) if this sum is larger than a given threshold. In the learning phase, the weights associated to each neuron input are updated until the network performs a satisfactory classification. NN allow for online learning, but they provide no insight into the classifier. Deep neural learning, which is state-of-the-art in many applications nowadays, bases on relatively recent variants of NN, among which the most employed are Convolutional Neural Networks (CNN). «Deep» refers to the presence of multiple layers in the network. Each one of these layers is trained to transform its input to a more abstract representation, but most of what happens in these intermediate layers turns out to be opaque for the human trainer. This poses serious concerns about traceability and explainability of the resulting classifier, which are hot topics of current research.

The supervised learning paradigm par excellence in Robotics is **Learning from Demonstration** (LfD, aka imitation learning). The teacher performs the task next to the robot, in a way that the robot is able to perceive the execution of the task, be it by processing the images or videos captured by its cameras (more on Computer Vision in Section 3), or by registering its configurations and the forces exerted on its arm while it is pushed and pulled by the teacher along the desired trajectory (this is known as *kinesthetic teaching*). LfD can be performed at different levels of abstraction, being the lowest one the *trajectory level*, that is, to learn basic sensory-motor skills, e.g. to learn how to reach and carry a glass full of water, or to wrap a scarf around a person's neck. It is crucial to be aware that not an exact reproduction of any of the shown trajectories is sought, but a *generalization* over the demonstrations, which means to capture the relevant features that are implicit to the learned skill (to continue with the examples, these would be to hold the glass in an upright position so as not to spill the water, or to reach a final configuration of the scarf such that both ends are more or less similar in length). To this end, the teacher should provide demonstrations with a certain variability, so that only the significant parts remain constant along them. In this way, once the robot applies the learned skill, it is able to adapt its execution to the current circumstances (e.g., the actual position of the glass to be grasped). Another way to stress the relevant features of a skill is by using social cues, such as gazing, pointing, or using verbal statements[1]. Even at trajectory level learning, distinctions have to be made about whether the robot has to reproduce the articular

1 Within Robotics, these topics are studied by the field of Human-Robot Interaction (HRI).

motions of the teacher (e.g., if it has to imitate its arm gestures), or just the trajectory of the hand, or even if only the final position is meaningful (but not the way it is reached). Furthermore, the quality of the learned skill has to be quantified in some way. For example, it has to be established whether the goal is to attain the same final relative position, the same absolute position, or the same relative displacement as in the demonstration. All of these topics (meaningful parts of a skill, granularity of the imitation, the metric of imitation performance) address the so-called *what-to-imitate* problem, whereas *how-to-imitate* is concerned about the *correspondence problem*: due to to the different embodiment of teacher and learner[2], the mapping of the demonstrations from the teacher to the robot is far from trivial (Nehaniv and Dautenhahn, 2002).

LfD can also take place at a higher abstraction degree, known as *task level*. Here the task is composed by predefined atomic actions which are represented symbolically, for example via rules, that require a set of preconditions (statements about the surroundings, aka world state) to be true in order to be executed, and they have some effects in the form of new (or modified) world statements. Symbolic learning means to process the sensory input, and to segment it into meaningful world transitions, which correspond to the symbolic actions. Sequences of such actions lead the world from an initial to a goal state, this is what another cognitive function, namely planning, is aimed at. Sequencing means to learn precedence constraints between actions. Some actions have to precede others in any case, and this strict precedence is uncovered if a sufficient number of demonstrations is performed with a certain variability which allows to distinguish it from other circumstantial temporal orderings between actions.

It is obvious that the more demonstrations, the better chances the skill or the task to be learned properly. However, the more demonstrations does also mean the more time of the human teacher devoted to a task that becomes in this way tedious and dull. In the ideal case, a few demonstrations should be enough to highlight the relevant parts of the task, but except in very simple instances, such a set of selected demonstrations is quite difficult to design. This issue may be tackled via some sort of incremental learning. The idea behind this approach is to provide first a small set of demonstrations, so that the robot is able to learn a first rough approximation of the skill or the task. By observing the performance of the robot in the execution of the skill, the teacher detects where improvements are needed, and focuses the attention of the robot towards the still erroneous parts of the task. This teaching strategy is known as *scaffolding* or *moulding*, it can be also seen as a variant of *coaching*. Driving the attention to specific parts of a task requires some sort of communication mechanisms, as those developed in the field of *Human-Robot Interaction* (HRI). Social cues have investigated and adapted to this end. They include non-verbal cues such as pointing or gazing, and verbal instructions as well. Computer vision techniques of gesture recognition or gaze following, as well as natural language processing allow such interaction to take place in a natural way for the non-expert user. Alternatively, refinement or improvement of the learned task can be transferred to unsupervised learning as explained in Section 2.3.

2.2. Unsupervised learning

In this family of learning methods, no information about Y is provided to the learning system. As no teacher is present, the system has to determine the implicit structure underlying the input data X. That is, it tries to find an explicit model for such implicit distribution or structure. Learning performance can then be quantified in terms of how well new data adjust to the found structure. Unsupervised learning includes *clustering methods* (inputs are grouped in clusters by some proximity or partitioning criterion, k-means clustering being a popular such algorithm) and *association rule learning* (i.e., to discover rules describing large portions of input data).

2 They have different kinematic structure, such differences being just a question of scaling, or of different proportions, or even of different number, disposition or type of joints.

Reinforcement learning (RL) is one of the most successful and widespread learning paradigms. In its unsupervised version[3], it aims at obtaining an optimal or near-optimal *policy* (action selection as a function of the current state) without guidance of a teacher. The setting is conceived as a *Markov Decision Process*, i.e. the effect of applying an action depends just on the current state where the action is applied, without taking the previous history into account. Full observability of each state is assumed (all the relevant information can be observed), although partial observability formulations do also exist. The only feedback provided to the learner comes from the environment, where the robot's actions take place. Despite the existence of different RL techniques and continuous development of new variants, there are some common traits:

- a set of environment and robot states *S*;
- a set of actions *A* that can be executed by the robot;
- policies of transitioning from states to actions;
- rules that determine the scalar immediate *reward* of a transition;
- rules that describe what the agent observes.

The crucial feature that directs the learning process is the reward (indirectly, as will be seen in short). It expresses the degree of desirability or satisfaction associated to the last attained state (or of the action that has lead to this state), and thus obviously it is provided by state observations. However, it is the designer of the RL algorithm who decides the environmental variables that are considered in the computation of the reward (and how they are evaluated in this computation). The learning process is guided by the so-called *value functions* (o *utility functions*, in the best benthamian tradition), that express long-term desirability of a transition (and thus of an action executed in a given state), as opposed to the rewards, that represent immediate satisfaction degrees. Rewards are actually used for the computation of value functions, but a certain state (or the transition that has led to it) may have a high reward and a poor value or vice-versa, they are not necessarily equivalent in qualitative terms. Values are computed from the rewards of the estimated optimal course of actions leading to the final goal of the learning process. This means that while rewards are associated directly to a certain state observation, values need to be estimated once and again from the different action courses related to the sequences of state observations made by the robot along its entire lifetime. The relative influence of immediate versus long- term desirability can be tuned via a *discount factor* associated to future rewards, as done by certain RL algorithms. Most RL methods are not deterministic as is exact Dynamic Programming (a mathematical optimization algorithm which sweeps over the entire search space) but can be considered as stochastic approximations, which sample states according to the underlying probabilistic model. Alternatively to value estimation, evolutionary optimization methods such as genetic algorithms or simulated annealing, search directly the policy state. Such optimization methods are not suitable for online learning, as they do not allow to interact with the state space while learning (whereas value function estimation certainly does), but they can be used to compare the performance of RL as for the obtained results. Online learning with RL leads to another question, namely the tradeoff between the *exploration* and the *exploitation* phases: the former means to explore new, possibly more rewarding states, whereas the latter means to exploit the knowledge gathered so far. A common strategy to balance out the two phases is the ε-greedy method: the action currently believed to be optimal is chosen with probability $1-\varepsilon$, and another random action is chosen with probability ε. Values of ε close to 0 favor exploitation, whereas exploration is favored by values between 0.5 and 1. Finally, as has been mentioned, value estimation requires some kind of future projections about the behavior of the system. The availability of a model which mimics this behavior (i.e., allows to simulate different possible courses of actions and the corresponding evolution of the system) enables the so-called model-based RL algorithms, such as Adaptive Real-time Dynamic Programming. In opposition, model-free algorithms, which do not need any knowledge about consequences of individual actions, are represented by the Q-learning algorithm.

3 Supervised versions do also exist, for example *Inverse Reinforcement Learning* (IRL), where the optimal policy is observed from the teacher and the algorithm tries to compute the reward function, which is explained below.

2.3 Combining the two learning paradigms.

As already suggested at the end of Section 2.1, supervised and unsupervised learning may be combined in order to obtain the best of the two worlds. Learning may be viewed as a search of the best course of actions to achieve a certain goal. There are many possible combinations or sequences of such actions, and the set of all these possible combinations constitutes the search space, which may be quite huge. In LfD, the teacher clearly highlights the solution within this search space, at the cost of devoting their valuable time to this task, which may require a certain number of demonstrations to achieve generalization. In RL the robot determines the solution on its own, but the initial guess may be quite distant from the actual solution, requiring huge amounts of exploration-exploitation efforts in this search space to approach the solution. So the first obvious way is to provide one or a few demonstrations to constrain the search space to the area where the actual solution lies, and then trigger the RL process to refine the learned skill within this restricted neighborhood. Thus the teacher is not required to provide further demonstrations to enhance the performance of the learned skill. In a complementary fashion, a method presented in (Martínez, Alenyà and Torras, 2016) is focused at enhancing a relational RL approach with occasional requests to the teacher, whose demonstrations prune the search space where required, following the suggestions of the very own system.

3. Perception

Static computer-resident AI is generally fed with data provided by a human programmer. This is the case of Expert Systems, trained with financial or medical data. But in the case of a robot, due to its embodiment, embedded in physical surroundings, learning heavily relies on perception. Perception is the input stream from which descriptions about the current state of the robot and of the world can be extracted, which in turn allows to couple sensed changes in the environment to particular actions performed by the robot. Later, in Section 4, some datasets available to the robotics community will be presented, on which learning algorithms (among others) may be tested. But even these datasets have been obtained from perception at the first place. Perception can be seen as a process, an information flow traversing different phases: *acquisition* or *sensing*, where a physical magnitude is captured by a sensor, *preprocessing* (which means conditioning the resulting signal), *feature extraction* (i.e., making the relevant information explicit), and finally *interpretation*. Eventually, also *sensor fusion* may take place, that is, the combination of the informations provided by different perception channels to obtain a more complete insight on the state of the world. For simple sensors these steps are straightforward (if they exist at all), but the quite more involved computer vision systems include additional processing such as *segmenting* the image into regions before feature extraction. Two main types of perception may be distinguished, as explained next.

3.1. Proprioception

This is about perceiving the (internal) state of the robot. The most relevant information concerns the robot's pose, for example the configuration of its arm(s), and it is provided by sensors called encoders located at the robot's joints. Such encoders measure the rotation angle of each one of the robot's axes, and thus the overall configuration of the arm (or equivalently, the position and orientation of the robot's gripper) can be computed. An encoder mounted on a mobile robot's wheel of course also counts its turns, and from this internal information, an external measurement can be computed, namely the displacement of the robot on the terrain (this is called odometry), and thus it is not proprioception in a strict way. Measuring the charge state of the battery or the energy consumption of the robot's electrical motors are also proprioceptive perceptions.

3.2. Exteroception

A large number of sensors allow to detect or measure different features of the surroundings:

- Contact: microswitch, bumpers;
- Proximity, (contactless) presence: inductive, capacitive, photoelectric, or ultrasonic sensors;
- Position (with respect to a fixed reference in the world): beacons plus camera, GPS;
- Distances, volumetric shapes: LIDAR (laser scans);
- Heat: thermal imaging with infrarred sensors;
- Sound: microphones;
- Touch: contact sensors on fingers, artificial skin;
- Forces: Force/Torque sensors mounted generally on the robot's wrist;
- Vision: Monocular or stereo cameras, RGB+D cameras (which not only sense light but also depth, i.e., distances to the camera).

Any other type of sensor may potentially mounted on a robot as well, such as chemical sensors to detect the presence of specific gases, or radioactive sensors to detect radioactivity. From all the exteroceptive channels, the most informative, involved, widespread and the one to which most research efforts have been devoted is clearly Computer Vision (CV). The outcome of image processing (with a certain relevance to robotics) can be categorized into *detection* (of an object, a defect, a face, etc. within an image), *identification* (of an individual object, a specific place, a particular person, a fingerprint, iris, or the like), *recognition* (or classification, in other words, assigning a specific view of an object to its corresponding class, which has been previously specified or has been learned, simple scenes such as «person on a bike» do also enter in this category), and *tracking* (following a moving object, animal, or person along a sequence of images or a video). Where CV still encounters unsurmountable obstacles is in scene *understanding*, as this requires some sort of general knowledge, which is beyond current AI.

Most perceptive channels are far to simple to convey any ethically significant bias. That a bumper informs about a mobile robot colliding against something is an objective fact without any ideological load. The same can be said about the robot's arm pose or the position of an autonomous vehicle in world coordinates. A microphone will capture all the sounds in the audible spectrum, and a camera all the colors in the visible range, it is in the higher processing and interpretation steps where the system may be shaped to ignore high pitched voices or to consider the skin tone of a person as a revealing feature. Such a shaping of the higher perception functions may lead towards unfair situations, such as for example a service robot ignoring the verbal requirements of children or high-pitched voice adults, or even the presence of people with certain skin tones. If a robot has been trained exclusively by people wearing glasses, could it possibly fail at recognizing a non-glasses-wearing person as a valid interlocutor? No, as long as it is able to recognize a person by other facial and/or bodily features. Unfair bias may be of course intentionally introduced, but beyond the intentionality of the programmer, can it appear as an inadequate or inaccurate choice of the learning data? The answer is affirmative, but to elaborate on this first the input data sources for robot learning have to be presented.

4. Data sources

4.1 Datasets for Robotics

The data that artificial intelligent systems use as input sources for learning consist either in controlled sets provided by their human programmers, or in environmental data they have access to. The boundary between these two categories is somehow blurred, as the environment (or the means employed by the system to access the information it contains) can of course also be quite controlled by the human designers of the learning process, as are always the experimental settings in the

laboratories or in industrial environments. Furthermore, these data can be just abstract alphanumeric sets or sensory information that requires some processing to become perception. In between particular data bases[4] and direct perception of the surroundings a third source has to be mentioned, namely the world wide web. Again, the internet may be a provider of raw data, e.g. by searching images associated to a given term with your favorite browser (or by similarity with an input image, more on this in Section 4.2), or it may contain/give access to datasets specifically constructed for machine learning applications hosted in universities and research centers around the world. Such databases are generally intended for educational and research purposes, and open source regarding such finalities, often with a creative commons license. As we are focusing on potential bias in robot learning, a closer look on data repositories specifically aimed at robotics is taken in what follows. In order to provide a common testbed where the performance of different algorithms that aim to solve the same problem can be compared, many research groups provide access to the data sets they have collected and have used to test their own methods. Some of them have become standard benchmarks in specific robotic applications. Categories of datasets include:

- **Navigation data**: Data gathered by onboard sensors of a mobile robot while traveling around indoor or outdoor environments. The initial aim was to test SLAM algorithms (simultaneous localization and mapping, that is, the robot autonomously constructs a map of its surroundings and localizes itself within). Examples include MIT's Radish[5], which contains files of odometry, laser and sonar data taken from real robots and from simulated robots, as well as environment maps generated by robots or by hand (i.e., re-touched floor-plans) or the Mobile Robot Programming Toolkit[6] with odometry, LIDAR and vision (also olfaction!). Nowadays the domain has extended to include traffic conditions data for autonomous vehicles, as well as non-terrestrial media:
  - Autonomous vehicles: The KITTI Vision Benchmark suite (Geiger et al. 2013) includes datasets with stereo, optical flow, visual odometry, 3D object detection and 3D tracking, where ground truth is provided by an accurate LIDAR system and GPS, with up to 15 cars and 30 pedestrians per image in urban, rural, and highway scenarios. Also benchmarks for each visual task together with evaluation metrics are provided. Another examples is the Ford Campus Vision and LIDAR Data Set (Pandey, McBride and Eustice 2011), and there are many more. Some of these databases focus on specific actors of the traffic scenario, mainly on pedestrians, such as the FCAV M-Air Pedestrian (FMP) Dataset (FCAV, 2019).
  - Underwater robots: The Marine Robotics Datasets from the Australian Centre for Field Robotics (ACFR, 2013) are about the spatial distribution of habitats, identification of ground characteristics and of living species, etc.
  - Flying robots: Multidrone Public DataSet (Multidrone, 2020), which contains annotated audiovisual material for recognition and tracking of bicycles, football players, human crowds, etc.
- **Motion capturing data**: Data obtained while observing human motion, including walking, gestures, etc. For example, the Locomotor Control Systems Laboratory of the University of Michigan hosts, among others, a dataset containing leg joint kinematics, kinetics, and EMG activity registered with cameras and electromiografic muscular sensors of able-bodied individuals walking at different steady speeds and inclines on a treadmill (Locolab, 2019).

- **Object, place, people and scene recognition data**: Computer vision datasets of labelled images have a longstanding tradition beyond robotics, but are clearly quite helpful for this application area as well. They are huge collections of images on diverse categories of

4 The terms database, dataset (or data base/set) and repository are used in an interchangeable fashion.
5 https://dspace.mit.edu/handle/1721.1/62236
6 This repository started as an initiative of the MAchine Perception and Intelligent Robotics (MAPIR) research group of the Universidad de Málaga, with posterior contributions of other research groups. https://www.mrpt.org/

objects, living beings including humans, activities, simple scenes, etc. One of the most popular ones is ImageNet, which is organized according to the WordNet[7] hierarchy (only nouns for now), with an average of over five hundred images per node. Currently there are over 14M images, and 21841 synsets indexed (ImageNet, 2019). Since 2010 the annual ImageNet Large Scale Visual Recognition Challenge is run, where contestants compete with their algorithms for the best classification rate on ImageNet sets (currently over 95% accuracy). Other generic image datasets are the PASCAL Visual Object Classes (Everingham et al. 2010) or the SUN database (Xiao et al. 2010). More specifically oriented at Robotics, the Fukuoka Datasets for Place Categorization provide 3D depth, RGB and reflectance images for indoor and outdoor scenarios in Fukuoka, Japan. (Martínez-Mozos et al. 2019). Outdoor place categories include forest, urban area, indoor parking, outdoor parking, coast areas, and residential areas, whereas indoor place categories include corridor, office, lab, study room, kitchen, and laboratory. More constrained to indoor domestic environments, Robot@Home provide raw and processed (i.e., including 3D reconstructions and 2D geometric maps) data (87000 time-stamped observations) from RGB-D cameras and 2D lasers, with annotated ground truth about rooms and objects (Ruiz-Sarmiento, Galindo, Gonzalez-Jimenez, 2017).

- **Action recognition data**: RoboNet from Berkeley Artificial Intelligence Research contains 15M video frames of different robots (varying also grippers, camera perspective, etc.) interacting with different objects in a table-top setting, with the goal of pre-training reinforcement learning models before learning in the actual environment (BAIR, 2019). Learning with pre-trained models turns out to be faster and more successful than learning from scratch. The RoboTurk database contains 111 hours of videos of teleoperated robots demonstrations executing difficult and dexterous tasks, together with robot joint information and the human operator control stream (Mandlekar et al. 2019). The Google Brain Robotics Data (Levine, 2016) contains datasets with robot pose information as well as images for robot hand-eye coordination while performing various tasks such as grasping (650000 examples), pushing (59000 examples), pouring, and depth image encoding. The dataset of daily interactive manipulation (Huang and Sun, 2019) collects the position, orientation, force, and torque of objects manipulated in daily tasks, including 1603 trials of 32 types of daily motions as well as 1596 trials of pouring.

- **Others**: Many other datasets for various robotics-related tasks can be found, such as AprilTags Visual Fiducial System (Wang and Olson 2016) which provides fiducial tags and the CV software to identify them, as well as to compute their precise 3D localization and orientation (this has various applications in augmented reality, robotics, and camera calibration), or the MIT/Tübingen Saliency Benchmark (Kümmerer et al, 2018) for saliency[8] model evaluation.

Just a few examples have been provided for each category, as the aim is not to provide an exhaustive list, but rather an overview on what such datasets contain. Some of these references have been extracted from a recent online review (Lim, 2020), see also (Choi, 2019). Most datasets are clearly neutral, as for their potential social impact: they address strictly technical issues such as SLAM or robot hand-eye coordination for object grasping. It is in HRI (including *social navigation*, that is, how the robot should move in an environment populated with humans and in relation with individuals), when humans and robots come together and interact, where flawed learned attitudes of

7 WordNet® is a lexical database of English, hosted at Princeton University (https://wordnet.princeton.edu/), containing nouns, verbs, adjectives and adverbs, which are categorized into sets of cognitive synonyms (called synsets, currently there are 117000), each one for a concept, with conceptual-semantic links between them. It is suitable for computational linguistics and natural language processing.

8 Saliency refers to the parts of an image that for some reason are outstanding and attracting attention. It is related to understanding human eyes movements and attention mechanisms.

the robots may result in unfair behavior. This can include topics such as motion capturing systems, if they are later applied to rehabilitation robotics or to prosthetics, and gender-related differences in body constitution have not been taken into account (e.g. only male adults have taken part in the gathering of data). Action recognition data may also potentially trigger ethical concerns as soon as such actions involve some interaction with a person, although, as for today, they are basically on object manipulation, as illustrated above. People recognition data will surely be the most prone to misuse, which begins in the very categorization of persons from images (see Section 5).

4.2 Data annotation

Image classification requires a well-trained image recognition system, and this training, to be reliable and effective, has to be performed over a large number of input data, that is, labelled, tagged or annotated images. The so-called *automatic image annotation* systems, which associate *metadata* (such as keywords or captioning) to the actual image data aren't in general but previously human-trained image classifiers. The aim of textual image annotation, besides construction of databases for machine learning, is to allow annotation-based image retrieval (ABIR, which is a kind of query-by-text), as opposed to content-based image retrieval (CBIR, aka query-by-example) which tries to find equal or similar images to a given one (Inoue, 2004). It includes systems such as the one described in (Hervé and Boujemaa, 2007), where scene recognition (and eventually annotation) of day/night, urban/rural, indoor/outdoor images is based on the fusion of low-level local and global descriptors. The underlying semantic content has however been made explicit at some point. More «automatic» are those systems that infer annotation from contextual information: for example, in (Ramisa et al, 2018), the loose connections between news articles and the images that illustrate them are sought. The recognition of single objects at parts of an image and their relative positions may also lead to the recognition (and annotation) of a scene described by a simple sentence. A real automatic alternative is to resort to photorealistic Computer Graphics (CG): the objects and characters in the scene are already defined (and thus tagged), as they are CG models of static objects or dynamic agents that play some role in the CG simulation. Annotated videos and images can thus be extracted from the simulation in any required number for machine learning. This approach has been taken in (Johnson-Roberson et al 2017) in the context of self driving cars. However, there is still a long way to capture the richness and unexpectedness of the real world in simulated scenarios.

As for human annotation, we may distinguish between the cases that require expert knowledge and the ones that not. Representative of the first category are the Marine Robotic Datasets mentioned above, as they require marine biologists to identify specimens, whereas the second corresponds to the annotation of everyday objects such as done in ImageNet. A relevant distinction is made in (Lyons 2020) between *constructed datasets* and *scraped datasets*, where the former have been «carefully designed and photographed by research groups under controlled laboratory conditions», whereas the latter are just «images scraped in bulk from the internet». Although annotation is not explicitly addressed in the paper, it is clear that if annotated, images in the case of constructed datasets will be labelled more systematically and consistently. The use of human workforce to perform this type of tasks that are difficult for a computer to execute is sometimes called *human-based computation* (HBC) or *human-assisted computation* (also *human algorithms* or *distributed thinking*). In the case of image annotation, such tasks consist in providing (typing or selecting) a keyword or label for an image or for parts of it, or spotting a specific item in an image and pointing at it (e.g. drawing a bounding box around). The outsourcing to humans of certain computing steps generally involves some kind of *microwork*, which is compensated in different ways: economically (as in Amazon's Mechanical Turk), providing fun or online status (via *gamification*, that is,

converting the task in a kind of online-game, see for example (von Ahn, Liu and Blum 2006)), or self-esteem (altruistic volunteering, desire to contribute to science, etc.). Microwork in itself is a controversial issue, but in this paper we are more concerned about its outcome. Under-payed people labeling images for hours are certainly no guarantee for accurate annotating, and the system has to be very carefully designed not to allow tendentious annotations related to boredom, ideological extremism, or directly maleficence. Even the tagging of objects or places is not devoid of political content. Consider for example an image of the Highlands tagged «United Kingdom» without any reference to Scotland. Or a washing machine semantically linked to «Woman» because one or various annotators consider domestic chores as «women-work». But the most controversial tagging is related to the categorization of people. This is discussed in the following section.

5. Bias in datasets

When datasets are distinguishable from one another, even if restricted at the same object category, there is a clear sign of built-in bias, as shown by a playful experiment in (Torralba and Efros 2011). Bias may compromise the generalization capabilities of the trained recognition algorithms, and under given circumstances also originate unfair behavior of the intelligent autonomous system. The authors also point at the perils of recognition algorithms becoming too overfitting, because of their progressive adaptation to a specific dataset whose associated competition they strive to win, loosing generality. They propose cross-dataset generalization to uncover bias (training from one dataset and testing in another one) and provide ways to minimize the effects of the different sources of bias:

- selection bias, how data are retrieved: obtaining data from multiple sources;
- capture bias, e.g. photographs of a category of objects always taken from the same viewpoint: transformations including flipping, jittering or cropping the image;
- negative set bias, i.e., of objects not pertaining to the category: adding negatives from other datasets.

When it comes to annotating images of people by other people, personal sympathy and more often antipathy introduce a severe bias in the labeling of certain social groups, be it by gender, sexual preferences, race, or other attributes that supposedly can be read from an image, that is, inferred from appearance. This has been dramatically stressed in ImageNet's former categorization of «people» in (Crawford and Paglen 2019). In their essay «Excavating AI»[9] they point at the 2833 subcategories under «Person», categorized attending to «race, nationality, profession, economic status, behaviour, character, and even morality», with the most crowded ones being «gal», «grandfather», «dad» and «chief executive officer». The real problem comes when categories appear for «Bad Person», «Loser», «Wimp», or «Kleptomaniac», as well as many racist and misogynistic terms.

Other labelings are just nonsensical. In any case, the authors of the essay point at the reductionist assumption that the richness of human features can be captured in just a few disjoint categories (male/female, a few races, etc.) and that even the character of an individual can be inferred from just appearance. Furthermore they link such assumptions to the pseudoscientific physiognomists and phrenologists approaches of the XIX and early XX century. They go one step further by questioning the assumption that fixed, universal, internally consistent and transcendentally

9 Some aspects of their work have been contested in (Lyons 2020): the author was one of the coauthors of the JAFFE dataset mentioned in the essay, and strongly rejects that the original aim of the project was to constitute a dataset for machine learning (nor that it was funded by the military). He further points at the «self-contradictory stance regarding informed consent for the use of facial images», as the use of the images taken from JAFFE in the essay and associated artistic exhibits went beyond the terms in the original informed consent of the photographed subjects. Despite all, these and other inaccuracies stressed in (Lyons 2020) do not invalidate in our opinion the key findings in (Crawford and Paglen 2019), and Michael Lyons even shares their opposition to AI-based surveillance.

grounded concepts do exist at all, as well as that their underlying essence expresses itself naturally in the form of images. The essay concludes

> «The whole endeavor of collecting images, categorizing them, and labeling them is itself a form of politics, filled with questions about who gets to decide what images mean and what kinds of social and political work those representations perform.» (Crawford and Paglen 2019)

In the view of such findings, the question is pertinent about whether should or not the robot learn to distinguish social groups and to draw some behavioral expectations from this knowledge. An assistive or a service robot should obviously be aware on the physical impairments or psychological limitations of a specific individual it has to interact with, and if this interaction is circumstantial and for a short time, it may certainly rely on some appearance-related evidence such as the individual being on a wheelchair or being of short age. But anything beyond that is at least problematic and requires a careful approach. Bias and associated unfairness may arise even in apparently neutral tasks such as approaching or accompanying a person, as shown in the pioneering work of (Hurtado, Londoño and Valada, 2021) about unfairness in socially-aware robot navigation. Learning proxemics etiquette from humans, for example, may result in unfair behavior by sides of the robot, as it is gender conditioned, while the robot is (or should be) gender neutral. They propose a two step approach, with a learning phase of safe and socially-compliant behavior, and a second *relearning* phase where observed unfair biases (using e.g. evidence from clustering) are addressed. This method may be applied to other domains of HRI as well.

6. Conclusions

Machine learning applied to Robotics may result in inappropriate or unfair behavior from sides of the robot if potential bias, favoring some social groups and neglecting others, is not addressed properly. Learning algorithms are neutral *per se*, besides punctual factors as the definition of the reward function in RL, the significant source of eventual bias lies in the input, be it the demonstrations fed to a LfD system or the data in classification algorithms such as deep learning. Human annotation of images used in training classificators may convey the prejudices of the annotators, specially in people categorization. If datasets such as ImageNet have the vocation of becoming the visual database of everything, learning systems eventually accessing to them surely will perpetuate and amplify built in bias. The question is whether robots should acquire such knowledge at all. Robots that have to provide some kind of assistance (physical, psychological, or informational) just need to be aware of the physical and mental limitations of the individual they have to assist, without necessarily ascribing them to any social group.

What about future personal or domestic robots becoming something between an assistive appliance and a pet or even a family member or surrogate colleague/friend/partner? Besides whatever knowledge they may be endowed with from the manufacturer, clearly they will learn from their immediate surroundings (and the physical and cybernetic spaces they may have access to). Human ambient bias will in this way be perpetuated in a natural way (just as with children's education), unless something such as critical sense or ethical firewalls can be built in the robot's software. This is a complex, but maybe some day necessary, endeavor.

Acknowledgements

This work has been partially funded by the European Union Horizon 2020 Programme under grant agreement no. 741930 (CLOTHILDE) and by the Spanish State Research Agency through the María de Maeztu Seal of Excellence to IRI[MDM-2016-0656] and the project HuMoUR TIN2017-90086-R.

References

References

Australian Centre for Field Robotics, ACFR. *Marine Robotics Datasets*. Available at http://marine.acfr.usyd.edu.au/datasets/

Berkeley Artificial Intelligence Research, BAIR. *RoboNet: A Dataset for Large-Scale Multi-Robot Learning*. Available at: https://bair.berkeley.edu/blog/2019/11/26/robo-net/

Buchanan, Bruce G. (1989). Can Machine Learning Offer Anything to Expert Systems? *Machine-Learning*, 4: 251-254. https://doi.org/10.1007/BF00130712

Choi, Sunlok. *The Awesome Robotics Datasets*. Available at: https://github.com/sunglok/awesome-robotics-datasets

Crawford, Kate and Paglen, Trevor. *Excavating AI: The Politics of Images in Machine Learning Training Sets.* Available at: https://excavating.ai

Everingham, Mark; Van Gool, Luc; Williams, Christopher K.I.; Winn, John and Zisserman, Andrew (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*. 88 (2): 303-338. https://doi.org/10.1007/s11263-009-0275-4

Ford Center for Autonomous Vehicles (FCAV), University of Michigan. *FCAV M-Air Pedestrian (FMP) Dataset of monocular RGB images and Planar LiDAR data for pedestrian detection*. Available at: https://github.com/umautobots/FMP-dataset [updated November 10th 2019; cited January, 15th 2021]

Geiger, Andreas; Lenz, Philip; Stiller, Christoph and Urtasun, Raquel (2013). Vision meets Robotics: The KITTI Dataset. *International Journal of Robotics Research.* 32 (11): 1231-1237. https://doi.org/10.1177/0278364913491297

Hervé, Nicolas and Boujemaa, Nozha (2007). Image annotation: which approach for realistic databases? In: *CIVR '07: Proceedings of the 6th ACM international conference on Image and video*

retrieval, Amsterdam, The Netherlands, July 2007. New York, NY, USA: Association for Computing Machinery, pp. 170-177. https://doi.org/10.1145/1282280.1282310

Huang, Yongkiang and Sun, Yu (2019) A dataset of daily interactive manipulation. *The International Journal of Robotics Research*. 38(8):879-886. https://doi.org/10.1177/0278364919849091

Hurtado, Juana Valeria; Londoño, Laura, and Valada, Abhinav (2021), From Learning to Relearning: A Framework for Diminishing Bias in Social Robot Navigation. *Frontiers in Robotics and AI, 8: 650325.* https://doi.org/10.3389/frobt.2021.650325

ImageNet, Stanford Vision Lab, Stanford University, Princeton University. Available at: https://www.image-net.org/index.php

Inoue, Masashi (2004). On the need for annotation-based image retrieval. In: *Proceedings of the ACM SIGIR Workshop on Information Retrieval in Context (IRiX), Sheffield, UK, July 29th 2004.* Department of Information Studies, Royal School of Library and Information Science Copenhagen, Denmark: pp. 44-46

Johnson-Roberson, Matthew; Barto, Charles; Mehta, Rounak; Sridhar, Sarath Nittur; Rosaen, Karl and Vasudevan, Ram (2017). Driving in the matrix: Can virtual worlds replace human- generated annotations for real world tasks? In: Proceedings of the *IEEE International Conference on Robotics and Automation, 29 May-3 June 2017, Singapore.* IEEE: pp. 746-753. https://doi.org/10.1109/ICRA.2017.7989092

Kümmerer, Matthias; Bylinskii, Zoya; Judd, Tilke; Borji, Ali; Itti, Laurent; Durand, Frédo; Oliva, Aude and Torralba, Antonio. MIT/Tübingen Saliency Benchmark. Available at: https://saliency.tuebingen.ai/ [updated 2018; cited January, 16th 2021]

Levine, Sergey. Google Brain Robotics Data. Available at: https://sites.google.com/site/brainrobotdata/home, [updated August 5th 2016; cited January, 15th 2021]

Lim, Hengtee (2020) 18 Best Datasets for Machine Learning Robotics. Available at: https://lionbridge.ai/datasets/17-best-robotics-datasets-for-machine-learning/

Locolab, University of Michigan. The Effect of Walking Incline and Speed on Human Leg Kinematics, Kinetics, and EMG. Available at: https://ieee-dataport.org/open-access/effect-walking-incline-and-speed-human-leg-kinematics-kinetics-and-emg

Lyons, Michael (2020). Excavating “Excavating AI”: The Elephant in the Gallery. *Submitted:* https://arxiv.org/abs/2009.01215 https://doi.org/10.2139/ssrn.3901640

Mandlekar, Ajay; Booher, Jonathan; Spero, Max; Tung, Albert; Gupta, Anchit; Zhu, Yuke; Garg, Animesh; Savarese, Silvio and Fei-Fei, Li (2019). Scaling Robot Supervision to Hundreds of Hours with RoboTurk: Robotic Manipulation Dataset through Human Reasoning and Dexterity. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3-8 November 2019*. IEEE, pp. 1048-1055. https://doi.org/10.1109/IROS40897.2019.8968114

Martínez, David; Alenyà, Guillem and Torras, Carme (2017). *Artificial Intelligence.* 247: 295-312. https://doi.org/10.1016/j.artint.2015.02.006

Martínez Mozos, Oscar; Nakashima, Kazuto; Jung, Hojung; Iwashita, Yumi, and Kurazume, Ryo (2019). Fukuoka datasets for place categorization. *International Journal of Robotics Research*, 38 (5): 507-517. https://doi.org/10.1177/0278364919835603

MultiDrone EU Project Dataset. Available at: https://multidrone.eu/multidrone-public-dataset/ [updated January 23rd 2020; cited January, 15th 2021]

Nehaniv, Chrystopher and Dautenhahn, Kerstin (2002). The Correspondence Problem. In: K. Dautenhahn and C. L. Nehaniv (eds.) *Imitation in Animals and Artifacts.* Cambridge, MA, USA: MIT Press, pp. 41-61,

Pandey, Gaurav; McBride James R. and Eustice, Ryan M. (2011). Ford Campus vision and lidar data set. *The International Journal of Robotics Research*. 30 (13): 1543-1552. https://doi.org/10.1177/0278364911400640

Ramisa, Arnau; Yan, Fei; Moreno-Noguer Francesc and Mikolajczyk Krystian (2018). BreakingNews: Article annotation by image and text processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40 (5): 1072-1085. https://doi.org/10.1109/TPAMI.2017.2721945

Ruiz-Sarmiento, Jose Raul; Galindo, Cipriano and Gonzalez-Jimenez, Javier. (2017) Robot@Home, a robotic dataset for semantic mapping of home environments. *The International Journal of Robotics Research*, 36 (2): 131-141. https://doi.org/10.1177/0278364917695640

Torralba, Antonio and Efros, Alexei A. (2011) Unbiased look at dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, 20-25 June, 2011*. IEEE, pp. 1521-1528. https://doi.org/10.1109/CVPR.2011.5995347

von Ahn, Luis; Liu, Ruoran and Blum, Manuel (2006). Peekaboom: a game for locating objects in images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06).* New York, (USA): Association for Computing Machinery, pp. 55-64. https://doi.org/10.1145/1124772.1124782

Wang, John and Olson, Edwin (2016) AprilTag 2: Efficient and robust fiducial detection. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejon, Korea, 9-14 October, 2016.* IEEE, pp. 4193-4198. https://doi.org/10.1109/IROS.2016.7759617

Xiao, Jianxiong; Hays, James; Ehinger, Krista A; Oliva, Aude and Torralba, Antonio (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, 13-18 June, 2010*. IEEE, pp. 3485-3492. https://doi.org/10.1109/CVPR.2010.5539970