London South Bank University

EST 1892

Algorithms of Vision. Human and machine learning in computational visual culture

Nicolas Malevé

https://orcid.org/0000-0001-7041-8742

A thesis submitted in partial fulfilment of the requirements of London South Bank University for the Degree of Doctor of Philosophy

This research programme was carried out in collaboration with The Photographers' Gallery

January 2021

84087 words (not including references and appendices)

Abstract

Current computer vision algorithms largely depend on the availability of images labelled by human annotators at very high speed. The mode of production of these annotations strongly resonates with an early experiment conducted in 2007 at Caltech by Fei Fei Li, initiator of ImageNet, one of the most popular visual datasets. In a laboratory, the subjects were asked to describe photographs shown for a few milliseconds and to filter them through a taxonomy. The Caltech experiment is used, in the thesis, to engage with the photographic elaboration of computer vision: the model of vision, the photographic alignments and the micro-temporal rhythm that subtend the modes of production of labelled data and the labour behind it.

The written and practice components of the submission elaborate a novel method and document the path towards it. The method has developed in the context of practice-led research in collaboration with The Photographers' Gallery and crystallised into a project, *Variations on a Glance*, a series of re-enactments based on the Caltech experiment. The original experimental protocol is submitted to several variations, called re-experiments, exploring its potential to produce a time-critical model of vision and collective visual interpretations. The experimental protocol is re-designed iteratively to explore specific configurations of micro-temporal vision and different configurations of collectives of human and non-human participants. The thesis examines the dynamics of these collectives, in particular how they reach consensual interpretation, and how the taxonomic practices of the lab interfere in this process.

The contribution of this research is a mapping of the entanglement of computer vision and photography and a method embedded in practice that does not attempt to resolve the differences and tensions between photography and computer vision but provides a device to explore the texture of their relation. The research complements and complicates the recent critiques related to bias and discrimination in machine learning and the exploitative work conditions it relies on. Finally it offers to the photographic institution and its public a mode of intervention into the making of computer vision.

Acknowledgements

This research would not have been possible without the financial support of London South Bank University, School of Arts and Creative Industries, and The Photographers' Gallery.

My gratitude goes to my supervisors Andrew Dewdney and Katrina Sluis for their guidance and their generosity. I wish our conversation never ends.

I would like to thank the team of The Photographers' Gallery and in particular Sam Mercer, Jon Uriarte, Janice McLaren, Clare Grafik, and Brett Rogers for their trust and their complicity in adventures such as a Robot Vision Geekender, workshops of psychophysics experiments or datasets birthday parties.

I am grateful for the rich dialogue with the researchers of the Centre for Study of the Networked Image, in particular Ioanna Zouli, Geoff Cox, Annet Dekker, Lozana Rossenova, Raquel Campos Valverde, Nicola Baird, Magda Tyżlik-Carver, Elena Marchevska, Theresa Kneppers, and Marloes de Valk. And Gaia Tedone with whom I still have the chance to collaborate closely. At the School of Arts and Creative Industries, Daniel Alexander and Adam Brown made me feel welcome and offered me the opportunity to make the first pilot of this study. Thanks to Wolfgang Brückle at Hochschule Luzern to give me the opportunity to develop further this research.

In addition to London South Bank University and The Photographers' Gallery, various cultural institutions and organisations gave me the opportunity to test my ideas and develop my research practice. Thanks to Erg (Brussels), the course of Computational Aesthetics at Aarhus University (Denmark), the Algolit collective (Brussels), the Mediawissenschaft MA at Potsdam University (Germany), Hangar (Spain), FotoColectania (Spain) and the Master fotografia at the Universitat Politècnica de València (Spain). In England, the Cambridge Digital Humanities Network, Cultures of the Digital Economy Research Institute (CoDE) and Cambridge Big Data.

Thanks to all the people who participated in the re-experiments: Esben Thamning Dahl, Michael H.W. Jensen, Sari Pallesen, Caglar Araz, Tugba Akbulut,Martin Paaske, Laura Krog, Lada Halounova, Simon Troldborg Rafn, Alex Nielsen and all the others who have chosen to remain anonymous. I am grateful, inspired and impressed by your engagement. I hope this work does justice to it.

To Laurence Rassel for her inspiring thinking, her love, her patience and support.

This work has been nurtured by many collaborations, friendships and conversations. Thanks to my accomplices at the Institute for Computational Vandalism, Michael Murtaugh and Ellef Prestsæter. This research would not have been possible without the years long dialogue with the members of the Constant collective: Femke Snelting, Peter Westenberg, An Mertens, Wendy Van Wynsberghe, Élodie Mugrefya and Martino Morandi. As well as with the fantastic people collaborating with Constant: among others, Jara Rocha, Seda Gurses, the team of Open Source Publishing, Peggy Pierrot, Michel Cleempoel, Denis Devos. Ruben Van de Ven, Cristina Cochior, Cornelia Sollfrank, Sebastian Schmieg have been of great help.

A special thanks to Didier Demorcy who taught me how to look at scientists and experiments.

For this research, I travelled many times from Brussels to London. Thanks to all who made my stay in London a marvellous experience: Yaiza Hernández Velázquez and Wesley Phillips, Olga Goriunova and Matthew Fuller, Elvira Dyangani Ose.

Moral support: Bettina Knaup and Anne Quirynen, Rafaella Houlstan-Hasaerts, Xavier Le Roy and Scarlet Yu, and Karin Nygård. My parents and sister, Monique, José and Marie-Laure Malevé. As well as Ghislaine and Raymond Rassel.

Finally I would like to acknowledge Laurence Pilfold who provided me with his invaluable assistance in proof-reading this thesis.

Table of Contents

Illustration Index

Index of Tables

Chapter 1. Introduction

1.1. Photography, ubiquity at scale

In *A Sack in the Sand*, Martin Lister (2007) makes the argument that after a decade of anxiety about its announced obsolescence, photography is more than ever at the core of contemporary visual experience. In opposition to early accounts that construed digitization as a factor threatening its very existence, Lister argues that a relative consensus has been reached among a host of media theorists over an interpretation that sees the digital as contributing substantially to the medium's pervasive presence. The vast increase in photographic production and circulation over the decade seems to support this claim. Photographs are published on a massive scale via the Internet. Reportedly, 52 million photos are uploaded daily on Instagram (Statisticbrain, 2017) and 9000 snaps sent per second (Smith, 2017). These are only a fraction of all the photographs that are produced and archived on hard drives, memory sticks or in the cloud.

Yet if photographs are taken and circulated in abundance, their technological entanglement has undergone important changes which the sole criterion of quantitative growth cannot explain alone. For Martin Hand (2012), photography's ubiquity does not simply refer to the medium's increased presence, it also signals its reconfiguration. The field of photography is undergoing a process of differentiation. The forms and functions of the key apparatuses of photography have multiplied. The camera, the iconic photographic device, is still marketed as a device in its own right, a high-end product targeted at professionals always in need of higher resolution images. It is also bundled as a component with other devices like phones, webcam equipped laptops, spectacles, Google Glasses, implants etc. The ubiquity of the camera has led in two decades to a gradual softening of the difference between professional and amateur equipment (Cruz and Meyer, 2012). And it has opened up new contexts to photographic capture. If photography has always been a diverse field encompassing a variety of practices[1], the range of these practices has extended. This diversity of practices reflects a diversity of actors (i.e., more people in rich countries have a camera regardless of their sex or age). Importantly an increasing variety of non-human agents partake in the creation and interpretation of images (Zylinska, 2017). With the increasing mechanisation and algorithmic control of society, the range of applications of photography exceeds the human sensorium. Much

[1] Photography theorist John Tagg suggests the use of the term photographies (Tagg, 2009, pp14-15)

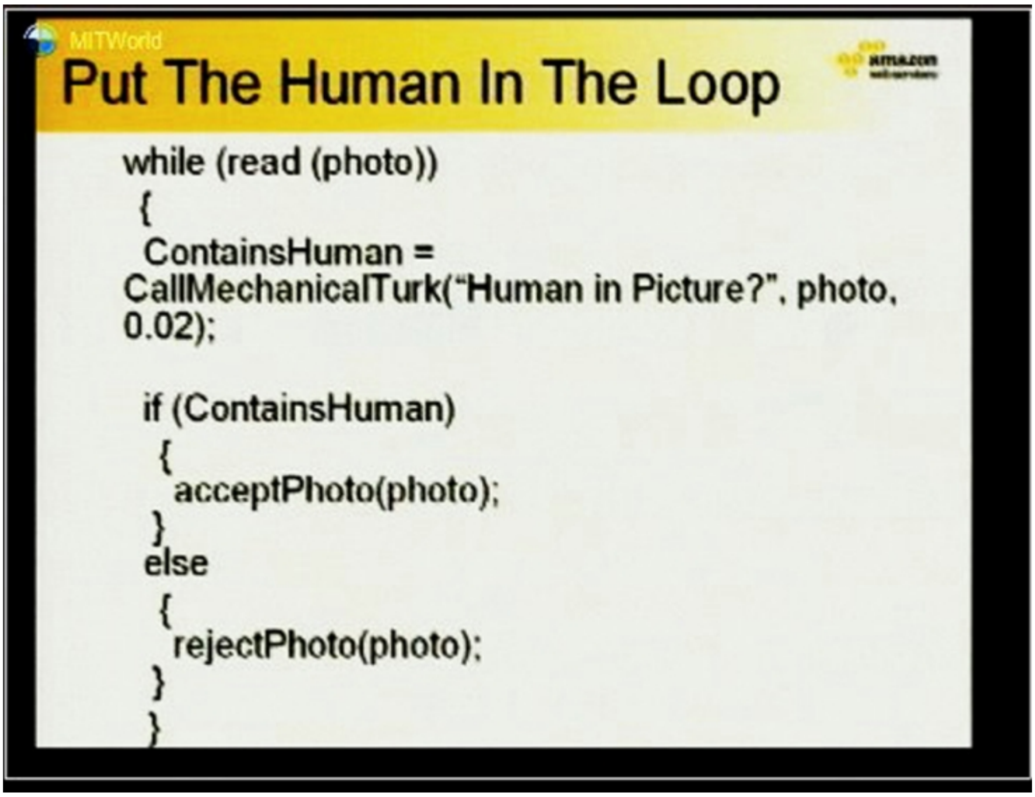
photographic production is now automated. Photography has increasingly become a technique to capture data to be processed by machines (Toister, 2019), in a world where most photographs are not necessarily consumed by human eyes anymore (Hoelzl and Marie, 2017).

1.2. Machine vision

Photography's ubiquity has another consequence. As we are confronted on a daily basis with millions of images on the Internet[2], the task of organizing them in a coherent manner seems increasingly crucial. The translation and articulation of the different registers of photography between devices, and between human as well as non-human actors, become equally crucial. Over the years, algorithmic techniques developed in the field of computer vision have evolved creating a new articulation of photography, vision, information and knowledge. They offer sophisticated solutions relevant to the scale of photography's ubiquity: to be able to parse, classify, interpret and make sense of photographs both in terms of their quantity and their diversity. The algorithms used in computer vision, as they become increasingly capable of capturing the structuring patterns of visual content, themselves become pervasive. They detect cancerous cells (Silverio, 2020), they help regulate traffic (Krishna *et al.*, 2016). Their use encompasses border checking (Sanchez del Rio *et al.*, 2016), home automation (Hasnain R. *et al.*, 2019), quality control in the assembly line (Nerakae *et al.*, 2016) or the analysis of the imagery of large aperture telescopes (Kremer *et al.*, 2017).

The increased technological grip on objects such as photographs has generated a growing sense of suspicion and anxiety. In many countries, facial recognition becomes a default routine that runs in the background of citizens' lives: as of today, 117 million Americans have their faces indexed by the police (Buolamwini, 2016). Examples of racism (Kasperkevic, 2015) or sexism (Calyskan *et al.*, 2017) in decisions taken by algorithms have come up, as their role in filtering, ranking, moderating or labeling online visual content becomes prominent. Along with the accelerated development of algorithmic techniques comes a growing demand to open up software to scrutiny, and to understand critically how algorithms, the new black boxes, operate (Kitchin, 2017). And the questions of identifying the knowledge relevant to this task and of finding ways to intervene have become matters of concern for a constellation of people that includes among others critical technologists, media artists and political activists.

[2] Reportedly, 52 million photos are uploaded daily on Instagram (Statisticbrain, 2017) and 9000 snaps sent per second (Smith, 2017)



Put The Human In The Loop

```
while (read (photo))
 {
 ContainsHuman =
CallMechanicalTurk("Human in Picture?", photo,
0.02);

  if (ContainsHuman)
   {
    acceptPhoto(photo);
   }
  else
   {
    rejectPhoto(photo);
   }
  }
```

Illustration 1: Pseudo-code for a detector by Jeff Bezos (Irani, 2015)

Yet the object of concern, what is commonly called the algorithm, proves to be elusive in surprising ways. A vignette featuring Amazon's CEO Jeff Bezos will help clarify how one of the problems attached to this term will be posed more particularly in this thesis. In her article *The Cultural Work of Micro-work*, Lily Irani (2015) evokes how Bezos, in a keynote given at the Massachusetts Institute of Technology, introduces his platform Amazon Mechanical Turk (AMT) and the work performed by the Turkers, the workers performing micro-tasks on the platform. AMT, in the words of its creator, gives the programmer the possibility to extend her code by integrating the output of external sources. This paradigm, known as *software as service*, takes a particular form on the AMT platform. The remote function called on by the software is not performed by a silicon-based processor but by humans. To illustrate this idea, Bezos shows a slide containing a simplified computer programme (see illustration 1). The programme receives a photograph as input. It evaluates if the photo represents a person and stores a binary value, yes or no, in a variable named ContainsHuman. If ContainsHuman is positive, the photo is accepted. If it isn't, it is rejected by the programme. The message Bezos wants to get across is that the most complicated part of the computation is not performed within the programme itself, but through a call to his platform. To evaluate whether a person is pictured in the photograph, the programme doesn't run through a complex set of rules. Instead, it uses the function CallMechanicalTurk, which asks this question to a worker, and then integrates the answer and goes on to the next instruction. What AMT provides is an abstract procedure to treat the worker as a software component, an algorithm that can be plugged in transparently. AMT extends the software-as-service paradigm. It offers, in Bezos' terms, the human-as-service.

This slide and its crude message condense the interconnected problems which motivated me to start this thesis. To begin, Bezos shows us a pseudo-code fragment, a sketch of the procedure through which a machine interprets a photograph. Yet the code is hollow. This piece of software is, as far as its most important element is concerned, a form of delegation. We are not facing a complex algorithm whose elaboration demonstrates the abstract capabilities of numbers. What we see is code marketed as a template for the social organisation of labour.

A second problem entwined with the first can be better understood if we consider the context in which the code fragment is displayed. Bezos chooses to strike a chord with his audience. In 2006, to detect the presence of a person in a photograph represents a daunting task for an algorithm. For a human viewer, however, the task is considered trivial. As Amazon Mechanical Turk manages to recruit a cheap labour force, it can provide an alternative to developing such a complex algorithm: to tap into the cognitive abilities of a population of Turkers[3] paid a few cents per task. Ironically, the example that highlights Bezos's selling point is the identification of a human in a photo whilst the programme obfuscates human labour behind a façade of code. Today, a programme performing such detection has become rather trivial. Coders do not require humans behind the scenes providing answers in real time. Silicon-based processors are able to perform such tasks autonomously. Or so we are told. As we will see in the next chapters, the entanglement of labour, photography, representation and algorithm does not end when the task can be performed by a silicon-based processor. The delegation is being played out in larger alignments. Algorithms are said to learn from human viewers and do not need to perform direct queries to human surrogates. Yet, as we will see, the division of labour and the mechanisms of delegation sketched in this slide continue to operate in more resilient and subtle ways.

A third problem emerges from the assumptions made in such code. When Bezos inserts a call to a remote function performed by a human operator, he assumes that the response will come back to the programme without disrupting its execution. For this, the programme must receive a response without delay. To achieve this, AMT must operate at a certain scale. Its workforce must be large enough to be always available to answer remote calls. It is an empire that must, in Irani's term, "follow the sun" (Irani, 2015, p.7). Such instantaneous responses also require that the detection of a

[3] A name given to workers on the AMT platform.

human presence in a photograph should be a task that can be performed without stalling the programme's execution. For a human viewer, detection is then assumed to be immediate. Immediacy of perception is as infrastructural to this code fragment as labour management.

The delegation at work in Bezos' code marks the beginning of the research journey. To give a sense of how delegation and the questions that it raises are approached in this study, I will now introduce another vignette that assonates and dissonates with the first. This time human viewers, unlike in the previous example – where they were hidden behind a code façade, take centre stage.

A student of psychology is sitting alone in front of a computer in a dark room. A marker is displayed at the centre of the screen and a photograph appears for a brief moment. The display time varies between 27 milliseconds to half a second. She has no means to control the duration of the stimulus. She is asked to write a description of what she recalls. And make herself ready for the next stimulus. The photographs, culled from the Internet, represent a large variety of scenes. Urban or natural landscapes, people engaged in sporting activities, running animals or sleeping children, cluttered scenes or single objects. There is no constraint on the style or length of the descriptions she produces. She spends several hours concentrating on this task. Sometimes, she barely distinguishes a shape or a patch of colour. Sometimes, she is able to decide whether the scene is indoors or outdoors. When the display time is slightly longer, she can even identify the different entities populating the scene and enumerate them. In other rooms, several students are performing the same task. While they participate to the experiment, the students are isolated from each other. They have no means to compare notes or discuss their experience together. The subjects are not explained the rationale for the experiment. They are, in the vocabulary of the researchers, *naïve* to the experiment. Their competences are taken for granted. To parse a visual scene at high speed is treated as a natural skill that can be actioned on demand. Later, a second team of students assay the results and map them onto a taxonomy. Going through the hundreds of descriptions, the assayers identify the common objects perceived by the subjects and measure the correspondence between the granularity of the descriptions and the duration of the stimuli.

These students are taking part in an experiment conducted by Fei Fei Li in 2007 at Caltech. Li, a lead scientist of the field of computer vision, conducted the experiment in the frame of her doctoral research where she studied human perception and its relevance to the design of machine vision algorithms. The stated goal of the experiment was to study scene understanding by human subjects inside the walls of the computer vision lab (Fei Fei, 2007). As a result, the experiment establishes the taxonomy and the technical conditions under which subjects are able to produce a relatively stable description of the contents of a photograph under a heavy time constraint.

The thesis postulates that a particular social ontology pervades the experiment. The first principle of this ontology is that the social is composed of bare individuals understood as isolated atoms. Individual subjects are treated as agents who perceive independently from each other and whose differences of interpretation are resolved without their involvement. Consensus is not reached through collective deliberation. Instead, the researchers average individual responses and their mean is treated as consensus. More, perception itself is also dissected in self-contained units that can be apprehended in isolation. To see is to respond to discrete stimuli that can be triggered at will and are independent from each other. The subject does not learn from the process and makes use of already acquired knowledge in an instant. To see is not to encounter the world, it means to apply filters.

This social ontology is inseparable from the apparatus which enacts and stabilises the subjects, photographs and other entities emerging through the experiment. This apparatus enables what I am calling in this study photographic alignments: chains of actions that affect the nature of a photograph. These chains of actions comprise the acquisition of online images, their transformation into stimuli, or their enactment as mental images. Core to this apparatus is the notion of micro-time that can be measured in precise temporal units, milliseconds, and their perceptual correlate, the fixation (the focus of the gaze on a single point in space).

The combination of this social ontology and the apparatus makes possible a perception at scale where scale is defined quantitatively, rhythmically and temporally. As subjects are responding on-demand, applying filters at a micro-temporal rhythm, they are made ontologically compatible with the demands of the nascent AI industry. This is where the two vignettes begin to resonate: a few years after the Caltech experiment, when Fei Fei Li initiates the ImageNet project, a dataset of 14 millions of images which will strongly impact computer vision, she will be for a semester the principal client of Bezos' Amazon Mechanical Turk.

By beginning with these vignettes, I want to introduce how computer vision will be approached in this study and the kind of relation I will seek to elucidate. The thesis is starting from the deviation expressed in the first vignette. If code is hollow, if its mode is delegation rather than source, how can we engage with it? How can we follow the delegation? In the study, I will argue that this requires an inquiry into the detours of such technology rather than an epistemology of code. To be able to open up photography to code, to detect if a photograph contains the representation of a person, what is presented is an apparatus that organises a deviation, an apparatus that dis-places a problem. This displacement is not only spatial but temporal. The software instruction needs to travel in time and remain synchronised. Synchronisation, rhythm and micro-temporality will be understood as articulators of the spatio-temporal scale that subtends machine vision.

The model of vision implied in Bezos' code requires a particular construction of subjects and objects. To detect a person in a photograph is understood as an act that does not require any qualification from the worker. It is a task that can be executed interchangeably and cheaply. Furthermore, it is construed as immediate, as a mere response to a stimulus. How to account for the kinetic energy that needs to be flowing for such an immediate response? To understand the worker as a viewer producing an immediate response raises many questions ranging from automatism and incipient perception to the resolution of the photograph and the apparatus holding in place the viewer and what she sees. How can the viewing of a photograph be considered immediate? And how does photography operate in such context?

I am arguing in these pages that a productive way to answer these questions is by enquiring into the environment introduced in the second vignette: the computer vision lab where engineers collaborate with cognitive psychologists. The experiment is a component of the practice of computer vision scientists who are using humans subjects to figure out how seeing could be modelled in a way that makes it algorithmically tractable. My proposition is that the experiment does more than provide a formal account of perception to inform the design of mathematical formulas. I contend that it does provide with a social ontology that is compatible with the demands of the computer industry that requires high volumes of annotations and armies of piecemeal workers looking at images at full speed. Additionally the experiment provides with a series of readily available objects and subjects, an apparatus and a managerial template to enact and stabilise the entities and relations that are necessary for such scale to be embodied by subjects and workers. The questions of the delegation and the model of vision are not merely technical or epistemological. They are inseparable from a form of labour. The Turker and the experimental subject are looking at photographs on behalf of somebody else. Vision is delegated and supervised. By relating the computer vision lab and the crowdsourcing platform, the thesis asks how the division of labour is effectuated and travels from the former to the latter. The texture of relation explored in this study will intrinsically weave together knowledge, vision and work. The "algorithms of vision", in this thesis, refer to the distributed processes involving cognitive labour and experimental devices, not to inscrutable mathematical formulas.

The experiment is seen here as going beyond its stated goal and what needs to be grasped is how it exceeds the account of the experimentalists. Therefore the thesis cannot follow a path that is already drawn by computer scientists or cognitive psychologists. The thesis is an exploration of a relation that includes other actors and agents and is not causally linear. How the relation between computer vision and its experimental practice can be expressed, probed, tested when it is evanescent and multiple is the key methodological question this study confronts. How can the relation hold the distance, not disappear, resist dispersion? How to make it available to experience without fixing it too much? This thesis elected to follow not straight but broken lines. It is the exploration of delegation and detour, indirect relation and iterative genealogies. And an interrogation about the kind of knowledge that is relevant to this relational texture.

1.3. Overview of the thesis

The evocation of these two vignettes gives me a series of elements that allow me to formulate the research's core questions. These questions are closely related to the modes of investigation I am using to formulate and answer them. These modes of investigation are: a mapping of computer vision's objects and agents, a methodological reflection and a theorization, the development and exploration of a practice.

1.3.1. Modalities

The research develops through several modalities. The first consists in a mapping of the field of computer vision that traces a pathway that leads from the environments of annotation, like Amazon Mechanical Turk, to the lab of psychophysics. The mapping spreads over two chapters: chapter two explores the network of operation of computer vision and one of its obligatory passage points, the dataset, and chapter three studies in detail the experimental construction of computer vision's model of vision through a thorough analysis of the Caltech experiment's protocol.

The second modality consists in a theorization and a methodological reflection that lays the ground for the practical research. Chapter five proposes a form of re-enactment of the Caltech experiment, the re-experiment. The re-experiment is a means to engage with the relational fabric wherein the experimental practice of computer vision engineers, the environment of annotation of machine vision and photography are woven together. Although the method chapter concentrates more explicitly on theoretical questions, theorization nevertheless runs throughout the whole manuscript. In chapter three, a rethinking of the entities uncovered during the mapping of computer vision is offered. In chapters six and seven, the theoretical considerations that have evolved organically are brought together to reflect on the relevance of the work of re-experimentation. And the final chapter recapitulates the core concepts of the thesis and underscores their relevance for computer vision and the photographic institution.

The third modality consists in a practical engagement with the Caltech experiment. Chapters six and seven are dedicated to an account of the practice and the reflection that accompanies it. They provide an overview of the development of the practice, an analysis of its evolution and how vision, scale, photography and taxonomy are engaged through a collaborative process. In the conclusion, I discuss the relevance of the experience accumulated through the practice and its theorization for the understanding of the elaboration of computer vision and the photographic institution.

1.3.2. The mapping

The mapping follows the delegation mechanism presented in Bezos' pseudo code, its detours and its consequences. Chapter two begins with the exploration of the network of operation of computer vision and one of its obligatory passage points, the dataset. Even if current machine learning algorithms do not require humans behind the scenes providing answers in real time, they nevertheless rely on a series of operations in which human operators are crucially involved before being ready for production. A significant number of workers are recruited to label large volumes of images. This is called the process of annotation whose ramifications are explored in the second chapter. This process includes among other tasks filtering erroneous samples, tagging images, classifying or describing them. The annotation process is a pre-condition for the training of machines and involves a huge amount of manual labour. The annotated images are gathered in databases called datasets. When the dataset is complete, the algorithm enters a stage of training during which it extracts the patterns relevant to its task from the dataset's images. Through this sequence of operations, the algorithm is said to learn from human viewers. And when the training is completed, it may perform autonomously and interpret images it has never been shown before.

Central to this process is the notion of scale. The core argument running through the chapter is that computer vision undergoes a transformation that leads to a new resolution of the scale of its problem. Taking the example of ImageNet, a dataset of 14 million photographs used to train visual algorithms, the chapter examines how the new dimensionality of the field modifies it practically and conceptually. Opening up images to semantics in computer vision requires a specific apparatus: the environment of annotation wherein billions of photos are paired with labels. To invent vision as a process susceptible to translation, computer scientists require retinas in large quantities. The eyeballs of thousands of annotators are directed and accelerated at a pace dictated by tight economic constraints. In the environments of annotation, the norm is the glance. At this point a question becomes unavoidable: how can engineers establish that, by barely glancing at a photograph, a worker can extract enough meaningful information to be able to classify or label it? Realising how much the industrial process of annotation depends on what can be glimpsed of a photograph stresses the importance of the model of vision which underlies the annotation environment.

To understand how this model of vision is constructed, chapter four moves to the lab of psychophysics to examine an experiment conducted in Caltech in 2007 by Li and her colleagues, where subjects are exposed to micro-temporal stimuli and their answers filtered through a taxonomy. The experiment is used to examine the relations of speed, distance, semantics and bodily positions translated in the settings of machine annotation platforms. A close reading of the experiment's protocol helps understand how the photographic elaboration moves from experiment to production and how the resolution of computer vision's scale is distributed across various sites: the AMT and the lab of psychophysics. The Caltech experiment opens a window into the mechanisms that simultaneously enable and bracket the epistemic contribution of seeing subjects in the environments of annotation. I end the journey through computer vision's detours by making the argument that the experiment not only offers a model of vision but also probes the relational matrix of the photographic elaboration of computer vision and provides a managerial template for the annotation environment.

1.3.3. Methodology and theorization

Central to the method, the concept of *re-experiment* builds on the outcomes of the mapping and the theoretical dialogue sections in order to propose a mode of practical engagement that takes in charge the photographic elaboration of computer vision and the role played by the experimental practice of computer vision scientists. The reason that motivated the development of the method is the fact that the relations between the experiment and the annotation environment are never stated explicitly by the researchers. A common social ontology permeates the experiment and the annotation platform, the experimental apparatus resonates within the Amazon Mechanical Turk environment. And workers and subjects emerge within a specific temporality. To apprehend these relations as well as the ontologies, the devices and the rhythms they weave together, this study identifies a need for a specific mode of enquiry that make all these elements available to experience and analysis.

Chapter five introduces the method: the re-experiment, a form of re-enactment of an earlier experiment. The original experimental protocol, in this case the protocol of the Caltech experiment, is submitted to a series of variations, testing its construction and its potential to produce time-critical models of vision and collective visual interpretations. These variations mainly address two interconnected problems: the problem of the social ontology of the experiment and the problem of its experimental apparatus. The social ontology (briefly introduced above) is challenged through variations bearing on the mode of relation the re-experiment proposes to the participants. Instead of being hailed as independent subjects, the participants are invited to respond to stimuli in groups of different sizes and engage in dialogue. The division of labour and the mode of supervision are altered and the participants are invited to contribute to all the stages of the experiment. The division of labour, the distribution of roles and tasks and the elaboration of consensus are reconfigured. As a consequence, the re-experiment seeks a favourable outcome that requires the collaboration of its subjects rather than a foreseeable success where all variables are under control.

The method of re-experimentation is informed by a theoretical dialogue with various strands of philosophy of science, science and technology studies (STS) and agential realism that runs throughout the thesis. In line with the performative approach of Karen Barad (1996), particular attention is given to the role of the experimental apparatus. In line with Barad, the experimental apparatus is not seen as a passive instrument of observation. Subjects, objects, photographs are considered indeterminate entities. The work of the apparatus is to enact them through a cut that momentarily stabilises them. This agential cut resolves the indeterminacy of objects and subjects temporally. This resolution involves the creation of alignments: specific configurations where the relations between entities and apparatus give way to singularity and temporary stability. The focus on alignment responds to a need to conceive of the role of photography in the experiment as a process rather than to limit it to a set of already defined objects such as the photograph or the camera. In the re-experiment, the Caltech experiment is approached as a photographic alignment as it relies fundamentally on a sequence of resolutions of indeterminate entities named photographs. Photographs are resolved in turn as search results, dataset items, stimuli and controls. To each of these resolutions corresponds the enactment of a mode of vision and a subject of vision as well as a state of the apparatus. The resolution of indeterminate entities is performative and thereby nondeterministic. By stressing the performative dimension of the apparatus, what is sought is to probe the ability of the experiment to change and open up different ways of enacting subjects, objects, photographs and their relations. To explore the performative dimension of the experimental apparatus is to probe the horizon of change of the experiment.

The objective of re-experimenting is not limited to the goal of learning about the Caltech experiment in itself. The re-experiment emphasizes the dimension of translation existing in experimental practice and approaches the location (the where and when) of an experiment as a problem to be researched rather than as a given – which implies that what is probed in a re-experiment is not confined to the limit of the experiment it re-enacts, but also interrogates the environments in which the experiment reverberates. To re-enact the Caltech experiment is a means to interrogate the annotation environment. To engage with the apparatus of the Caltech experiment is a means to question the resolution of the annotation environment. The re-experiment is conceived as a device to probe relations that are indirect, operating in multiple registers rather than two instances linked by linear causality.

1.3.4. The practice

The 5th chapter proceeds with the account of practice and considers the design of a re-experiment and the questions raised by the collaboration. It discusses the difficult balance between a series of decisions taken by the experimenter to give a structure to the re-experiment whilst remaining open to the changes the participants bring to the process. Through a thick description, it details the development of the work of re-experimentation. The chapter concentrates in particular on the experience of early vision and the consequences of changing the parameters of the experiment to foster the collective dimension of annotation. The micro-temporal presence of the stimulus prevents the viewer from fully parsing the image hitting her retina. To stabilise the fleeting sensation, the participants are tempted to freeze it and compensate with prior knowledge for its apparent lack. Over the course of the session, however, the participants also manage to suspend the urge to freeze the sensation and learn to develop a sensitivity to the evanescent nature of the stimulus. The plasticity of their perception is tested during sessions of collective description where their sensations can resonate and be discordant, where the fleeting aspects of what has not yet coalesced into a stabilised percept can be attended to. I name this process the *composite echo* as it is made from the oral descriptions of the group and generates a conjunctive mental image that gradually takes a life of its own.

Besides the composite echo, the account of practice identifies different forms of attunement that arise from the involvement of the participants and their entanglement with the experimental device. I am using the concept of attunement from affect theory (McKim, 2008) in this context to emphasize the prominence of the rhythmical relation the participants are engaged in. Their mode of making sense of the re-experiment activates the sensitivity of their embodied faculties. The participants are constantly articulating and synchronising discordant temporal scales. An attention to the various rhythms and durations rather than the sole micro-time of the stimulus brings about another participant's body. What the participant experiences is not limited to a mechanical opening and closing of her retina, it is also the uttering of words and a complex articulation of breathing, waiting and listening that rhythmically synchronises attention.

Attunement finally is also at the core of the method used with the participants to analyse the contents of the sessions through a series of listening sessions. Active listening reveals what the participants call the “game of making sense”, a series of implicit rules underlying the selection of statements. The participants follow another protocol interwoven within the stated protocol of the re-experiment that gives value to the descriptions with a higher degree of precision and discards the contribution of those who “do not see well”. It is only through an affective listening that this protocol is felt and discussed. It is through the affective listening that this part of the re-experiment is made account-able.

If the previous chapter was seen through the lens of the emerging subjects and their affect, in the seventh chapter the wider horizon of the alignment takes centre stage. How change is effected in a re-experiment is the core question this chapter attempts to answer. Using examples from various sessions, I show that the changes occurring through the involvement of the participants with the experimental device bear on more than an interpretation of the contents of the stimulus. To resolve a photograph as a stimulus is an achievement that requires the synchronisation and alignment of a wide series of agents of different natures. An alignment offers continuity across sessions and the potential to experience coherence across varying sites. Yet there is a brittleness to an alignment. Its mode is performative, its resolution is never given. The dance between stability and instability is studied through different cases where students mumbling in the background, a participant spotting a comic book character or a couple of participants repeating instructions provoke striking reconfigurations of the re-experiment. Such changes cut across the different layers of the re-experiment. The photograph is resolved differently, the re-experiment begins and ends in different places. And its course of recognition changes accordingly: depending on what she recognizes in the stimulus, the participant is recognized in turn as a subject among subjects, as a student in a classroom or as a gallery visitor.

If there is a dance of stabilisation and instability in the re-experiment, there is a site where stabilisation reaches its maximal intensity: the taxonomy. With the classification, we enter the last stage of the Caltech experiment. After the descriptions have been made, they are scored and mapped onto a taxonomy. In the re-experiment, the participants re-use the Caltech taxonomy to filter the descriptions produced in earlier sessions. The chapter follows the ramifications of the process of classification. The taxonomy cuts into more than words. It re-orients the epistemic compass of the re-experiment. It also changes how the consensus is reached and cancels the implicit rules by which the participants had set up their collaboration. To treat the descriptions as mere lists of words to be mapped onto the taxonomic grid produces what a participant calls a “dumbing down” effect. The participants are affectively dis-engaged from the process. By resisting the various modes of reduction at work in the taxonomic process, the participants indicate more than a concern for which words are being selected in the end. They stress the solidarity between the contents of the descriptions, the alignments in which bodies are taken and they raise the larger question of how the knowledge is produced within the re-experiment and what is made with it.

In the last chapter, I take the experience of re-experimentation and the theorizing that accompanies it to revisit the environment of annotation as introduced in the second chapter and the critique it is subjected to by various strands of activism. Mobilising the notions of embodiment, resolution of scale, attunement, collaboration and photographic alignment, I engage in a differential dialogue with those critiques to enrich the perspective on the environment annotation and its photographic elaboration. By doing so, I also raise the stakes for a potential transformation of the field, arguing that the resolution of computer vision concerns primarily its resolution of scale and not local problems that can be solved in isolation. Finally I conclude by discussing what my research means for an institution of photography and invite such institutions to acknowledge their potential to play an active role in the elaboration of computer vision.

1.4. Context note on The Photographers' Gallery

Before ending this introduction, it is worth insisting that the research does not develop in a void. It is nurtured by the institution that collaborates in the research and its public. Therefore, to end this introduction, let's open the door of the collaborating institution.

The research is grounded in a specific context, the Digital Programme (DP) of The Photographers' Gallery (TPG). To introduce the programme and the institution collaborating with the research, it is useful to explain how the evolving relations of photography and technology are understood inside the gallery. As a photographic institution, TPG faces the challenge of understanding afresh the changing articulation of the medium it champions. TPG, a pioneer institution, “the first public gallery in the UK dedicated to the medium” (The Photographers' Gallery, no date a) founded in 1971, in London's Covent Garden was created with the aim of promoting photography as an art form and as an instrument to raise societal debate.[4] TPG's raison d'être is the existence of the medium, and its changing definition affects the institution's definition in turn. In comparison with a fine art institution like the Museum of Modern Art in New York (Bolton, 1992), which played a crucial role in framing the medium as an art form, TPG positions itself as a platform for diverse photographic practices. Its programme covers photographic heritage, as well as thematic, documentary and contemporary art exhibitions. The Digital Programme was created in 2012 to address the transformation of photography and its shifting relation with technology. Since its inception, the programme has been nomadic in the institution. It has a strong web presence with an active publication platform, Unthinking Photography.[5] It also comprises lectures and workshops. Furthermore, the Digital Programme has a material anchor in the institution: the Media Wall, a large eight-panel video wall located at the ground floor between the welcome desk at the entrance and the café. During the first years of its existence, the DP has occupied the physical threshold of the building with the Media Wall, the associated web platforms of the Gallery (sometimes even parasiting the main website[6]), and occasionally mobilising the whole building for exceptional events like the Geekenders[7], intense moments of workshops, parties and discussion. *All I Know is What’s on the Internet*[8], curated by Katrina Sluis, was the first show initiated by the Digital Programme to take place on an exhibition floor in October 2018. It indicated a potential change in the institutional presence of the DP. Metaphorically, and more literally as well, the programme has followed in a few years an ascending trajectory leading from the Media Wall (main floor) to the main exhibition space

[4] I.e. the Gallery opened with *The Concerned Photographer* (14 January - 14 February 1971), an exhibition curated by Cornell Capa whose stated objective was to use photography to educate and change the world (The Photographers' Gallery, 2018).

[5] See http://unthinking.photography

[6] Sebastian Schmieg's project, *Decision Space* (07 October 2016 - 05 February 2017), commissioned by the Digital Programme of The Photographers' Gallery, turned the gallery's website users into annotators of the site's photos (Schmieg, 2016).

[7] See for example, *Towards a Feminist Internet* (16-18 March 2018), a weekend of live events and workshops exploring issues of gender and technology curated by the Digital Programme in collaboration with UAL Futures’ Feminist Internet Project. (The Photographers’ Gallery, 2018b)

[8] *All I Know Is What’s On The Internet (*26 October 2018 - 24 February 2019) is an exhibition curated by Katrina Sluis whose aim was to “investigate the systems through which today’s photographic images multiply online and asks what new forms of value, knowledge, meaning and labour arise from this endless (re)circulation of content”. (The Photographers’ Gallery, 2018)

(third floor). The spatial distribution of the DP resonates with its positioning inside the institution. By “moving up”, the DP gradually becomes part of the long term perspective of the institution. At the time of writing, its double articulation, offline and online, gives it even more importance as, due to the current circumstances of the pandemics, the institution relies more than ever on programmes accessible on the internet[9].

The Digital Programme's mission is to address the current mutation of photography into a digital networked medium. This entails more than a change in equipment and technique or simply refreshing the institution's catalogue of artists. It means actively contributing to the definition of the changing limits of the institution by expanding the modes of interaction and opening the institution to the network's dynamics (Sluis, 2018). There are different reasons for the institution wanting to rethink its limits. One is the relevance of an institution that needs to find ways to adequately contribute to the understanding and appreciation of the current developments of the medium. There is equally a willingness to address a new audience, to reach out to a more diverse public and a conviction that the digital can be instrumental in reaching these audiences. The challenge therefore lies in the ways the terms of this opening will be negotiated and its consequences explored. If there is a sense of crisis associated with the digital, it must be understood more as a tension about the very re-definition of the institution and its publics rather than a blind resistance to letting the digital medium enter the institution's white walls. Even among traditional photography's champions, there is an awareness of the necessity of addressing seriously the mutation of the medium and the exhaustion of its traditional model. As Fotomuseum Winterthur's founder Urs Stahel (n.d.) observes, there is an urgent need to find a better balance between an interest in the traditional objects of photographic institutions, like negatives and print collections, and an interest in the mediation of the photographic image which includes its networked dimension.

Whilst there is a desire to open up the institution to a new digital photographic environment, there is simultaneously a resistance within institutions of photography such as TPG against the digital, which is blamed for creating environmental noise. If Stahel recommends paying more attention to the photographs that are made today, it is in equal part for their sociological importance and for the threat digital production represents to the viewer's relation to the world and her critical judgement. Stahel sees in digital photography a culture that privileges distraction and superficial viewing. The photographic institution is presented as a shield against the repeated assaults of the networked image. On the TPG website, the director Brett Rogers (2018) concurs:

[9] See for instance, the *Screenwalks* series of live-streamed artist/researcher-led explorations of online spaces (Screenwalks, 2020)

> Given the speed and force with which an image can be shared today, it is vital that TPG offers a place to slow it down; to create an arena both onsite and online where we can reflect, process, interact, produce and be inspired.

Speed reduction marks a qualitative shift. Sustained attention, concentration, slowness are endowed with positive values. Many activities of TPG's education programme are designed to teach the participants to pay attention, dedicate the “proper” time to engage with the works on the gallery's walls. *Look Again*[10] (The Photographers' Gallery, no date b) or the *Slow Art Day*[11] (Slow Art, 2017) propose group discussions of one hour or more in front of a single or a few selected photographs. A slow visual literacy that invites the viewer to take the time to analyse a work extracted from the image flow is offered as a means to restore the visitor's powers of deliberation. The networked image and its consumer are dialectically opposed to the exhibited photograph and its spectator. In the slow reading of a photograph, we find the entrenched notion that “a photograph is a ciphered message that needs to be unpacked with the tools of semiology and structuralism” as Daniel Rubinstein (2009, p.139) puts it. The emphasis is given to the interpretation of a visual surface carrying an author's intentions. In a context where technology is perceived as disruptive, the photographic institution clings on to its role of arbiter of what should be valued as photography in terms of objects and competences. And less attention is given to the ways mechanical reproduction, circulation, algorithmic manipulation are affecting photographic objects (Sluis, 2013).

TPG sees the digital as an external constraint to which it has to adapt and simultaneously as an opening to a new audience and a new interaction. It is seen as an opportunity to reach out. At the same time, it triggers a reaction of defence. One solution to such a dilemma is to “thematise” the problem. To address the questions raised by the new technological context of photography, the institution is tempted to elect a new set of artists and objects to collect, archive or exhibit and to commission the DP to do this work. Inside the DP however, the mission is understood differently. Here, the engagement with the ubiquitous dimension of photography is expected to change the institution's rule book. A lot of work goes into the reformulation of the problem the programme inherits from the institution. The question of technology is understood as a means of introducing new questions and new practices of mediation of photography, to move away from the author-centred framework, and enlarge the spectrum of agencies that intervene in what photography is becoming. These differences are negotiated during meetings, conversations in the kitchen or public presentations. Some objects or activities also provide contexts to re-articulate the terms of the

[10] Sessions of approximately one hour conducted by visual artist and neuropsychologist Janneke Van Leeuwen where participants discuss a single work.(The Photographers' Gallery, no date b)

[11] For instance, on the 14th of April 2018, the gallery's audience was invited to look at five works intensively over two hours. (The Photographers’ Gallery, 2018c)

question. The Media Wall is an example of a territory where the terms of this negotiation are explored (Dewdney, 1995). The practice at the core of this thesis, the re-experiment, proposes another one. This thesis contributes to the effort of reformulation going on in the Digital Programme of The Photographers' Gallery. It brings a new inflection to the way the question of photography's relation to technology is being raised. Within the DP, technology is mostly perceived as a vector of change in the mediation of photography and a necessary disruption of its canonical objects. My contribution stresses that these changes affect the elaboration of the technology in return. The present thesis intends to help make the institution realise its potential to bear on the elaboration of computer vision rather than to respond to its effects. The research is thought of as a device to intervene in how the institution thinks about its relation to technology. To do that, it engages in activities in situ – re-experiments to make manifest the importance of photographic mediation that is at work in computer vision.

Chapter 2. The elaboration of computer vision – resolving the scale

As the discussion of Jeff Bezos' code in the previous chapter has already suggested, computer vision may not be understood satisfactorily in terms of code and pure mathematics, as a self-standing abstraction. Problems are not solved internally, they follow a course of delegation. Taking delegation seriously, I will approach computer vision as a networked practice whose elucidation requires large detours and the resolution of a scale. In order to find ways to engage with such networked practice, a preliminary mapping is in order. In this chapter, the mapping consists of a thick description and a critical reflection on the cast of agents and apparatuses involved in producing machine vision.

To begin the mapping exercise, I will use the dataset, an obligatory passage point in computer vision (Jaton, 2017, p.823), as my point of entry.

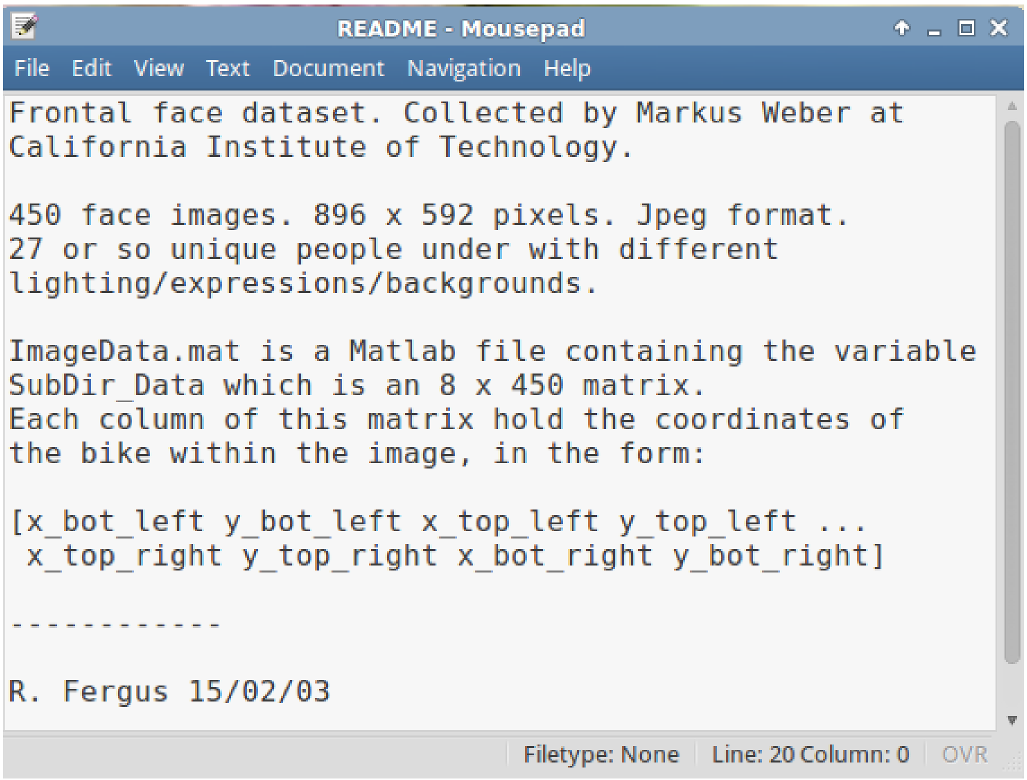
> Datasets and computer vision have always been intimately related. In fact one way of looking at the discipline of computer vision is … it's driven by datasets. Without datasets computer vision can hardly advance. ... To solve the problem of vision you have to understand your data, it's not just a feature vector (GoogleTechTalks, 2011).

These are the words pronounced by Fei-Fei Li, director of the Stanford AI Lab, when she presented her recent projects to a group of engineers at Google headquarters. Her message was at the same time simple and provoking: collections of photographs are the engine that drives the development of machine vision and this requires computer scientists to turn their attention to their visual data rather than considering them as interchangeable vectors of numbers. The methods Li invented to understand visual data – as well as how she understands vision and datasets – lie at the core of the present study.

Visual datasets are hybrid entities enabling and complicating the relations between computer science and photography. This first chapter uses the dataset as a starting point to inquire into the details of these relations. It introduces the key concepts of computer vision and the current evolution of AI and focuses on the work of annotation, the set of tasks needed to clean and label the silos of data that power machine learning. Observing the importance of the large collections of visual data for machine vision, the following pages offer a map of the entanglement of computer vision and photography.

2.1. Faces1999, photographic practice in the computer vision lab

A dataset in computer vision is a collection of digital photographs that developers use to test, train and evaluate the performance of their algorithms. Once assembled and packaged, a common set of photographs is shared among computer scientists. Using the same dataset gives the possibility for different developers to compare their work.



```
Frontal face dataset. Collected by Markus Weber at
California Institute of Technology.

450 face images. 896 x 592 pixels. Jpeg format.
27 or so unique people under with different
lighting/expressions/backgrounds.

ImageData.mat is a Matlab file containing the variable
SubDir_Data which is an 8 x 450 matrix.
Each column of this matrix hold the coordinates of
the bike within the image, in the form:

[x_bot_left y_bot_left x_top_left y_top_left ...
 x_top_right y_top_right x_bot_right y_bot_right]

------------

R. Fergus 15/02/03
```

Illustration 2: The Faces1999 readme file

By convention, a text file, the readme, is added to the collection of photographs to describe their content. Illustration 2 shows the readme file for a training set named Faces1999[12]. As the first line announces, Faces1999 contains pictures of people photographed frontally. The document states the institutional affiliation, credits, dates and provides approximative indications about the image's contents ("27 *or so* unique people", "people under with different lighting / expressions / backgrounds"). The word "bike" instead of "face" in the sentence "Each column of this matrix hold the coordinates of the *bike* within the image" is a residue from another file: the line has been copied from the readme of a training set for bicycle detection and the maintainer forgot to modify the

[12] Faces1999 is available for download from http://www.vision.caltech.edu/html-files/archive.html

reference to fit the new context. Whilst a key component of computer vision, the dataset has for a long time been treated as marginal. The mix of precision and approximation in the readme gives an idea of the informal circulation of the document. A dataset such as Faces1999 is not meant to face public scrutiny, but the habitual acceptance of scientists sharing a set of practices.

The people photographed in Faces1999 are computer scientists working at the Caltech Vision lab and their colleagues. The dataset is the lab's self-portrait, representing itself at work. The photographer, Markus Werber, is a member of the community. It is a dataset at his scale. We observe his surroundings through his eyes. It is a space where he can easily move and recruit people. He has a close relationship with his “subjects”. He can ask them to pose for him and provide him with different expressions to increase the variations in the portraits. Faces1999 is not strictly limited to the lab. Occasionally the dataset includes faces from the researchers' family circle. Then, the photographs depict house interiors, and the researcher's relatives. It is the world the photographer has access to, his universe.

There is a sense of conversation happening in the backgrounds of the photos. The backgrounds are densely covered with texts of different sorts. They are also overlaid with commentaries testifying to the mixed nature of research activity. Work regulation documents (a summary of the Employee Polygraph Protection Act), staff emails, address directories, a map of the building, invitations to conferences and parties, job ads, administrative announcements, a calendar page. All these signs suggest that more than code and mathematics is happening in this environment. A mix of bureaucratic injunctions punctuate the images. The languages of administrative interpellation and advertising respond to each other, interspersed with informal invitations and suggestions. On a door, we read “Please do not disturb” and a note signed Jean-Yves states: “Please do NOT put your fingers on the screen. Thanks.” Networks of colleagues in and outside the lab leave their marks on various surfaces captured inadvertently by the photographer. The intertwining of the social dimension and the code is felt nowhere more than in a picture of a whiteboard where complex mathematical equations cohabit with a note partially masked by the subject's head. The same surface of inscription is used for sketching the outline of an algorithm and to remind a colleague named Sony to call someone about his car.

For the photographer, there is a strong sense of familiarity with those depicted in the photos, and he can be present at all stages in the creation[13]: he selects the subjects, the settings, presses the shutter,

[13] Strictly speaking two people “authored” the dataset. However nothing prevents a dataset of this scale being curated, photographed and maintained by one person.

assembles and renames the pictures, annotates them, writes the readme, compresses the files and uploads them on the lab's website. 10 years later when Fei Fei Li, one of the researchers photographed in Faces1999, initiates ImageNet, a project with 14 million of images, it is logistically impossible for one person alone to do the work. The closed loop between lab, engineers, photographs and dataset is broken, provoking the reinvention of a long chain of epistemological assumptions and practices.

2.2. Bridging the semantic gap, from analytical decomposition to function finding

Computer vision algorithms are not limited to the confined space of the laboratory or specialized applications anymore. The field of computer vision has expanded considerably. It plays a central role in the management of image traffic on networks as well as in the preparation of diagnoses in medicine or in analysing the never-ending footage of surveillance imagery. Machine vision[14] algorithms function in digital systems that use the full spectrum of statistics to rank, filter, predict, decide. Machine vision has become a de facto software infrastructure powering the development of a large array of end products (robots, drones, autonomous cars, etc.) and the industrial production line itself (stock management, product assembly, quality testing etc.). As the field of application of computer vision is wide, my point of departure will be a specific problem that has absorbed the energy of the discipline for years. This problem, named by computer scientists the semantic gap, lies at the core of what has been branded as a revolution in computer vision: the new ability of algorithms to engage with the semantic content of an image.

To understand the terms of the semantic gap problem, let's consider the definition given by the official manual of one of the most popular software libraries of the field, OpenCV.[15] In the book, Gary Bradsky, engineer in robotics, writes: “Computer vision is the transformation of data from a still or a camera into either a decision or a new representation” (Bradski and Kaehler, 2008). What Bradski calls a decision is a description such as “there are 14 tumour cells on this slide” and what he calls a new representation is what a layman would describe as image processing (e.g., to transform an coloured image into greyscale values). In this thesis, computer vision will refer

[14] In this study, computer vision and machine vision are mostly used interchangeably. A distinction will be made when relevant.

[15] To trace out the semantic gap problem, I am drawing on a variety of sources. They include the research literature, manuals, blog posts, YouTube videos, etc. The computer vision community uses these different outlets to express different aspects of the problem and their activity. More about the source in the discussion of the emic representation of computer scientists in the next chapter.

primarily to the first part of Bradski's definition: the transformation of visual data into a description and the decision process it entails.

Making sense of a grid of numbers coming from a visual sensor is a daunting task. The indexical link that connects the image, typically a photograph, and the scene it depicts is, to say the least, tenuous and problematic. Bradsky describes what he terms the ill-posed nature of vision in these words:

> In fact, the problem, as we have posed it thus far, is worse than hard; it is formally impossible to solve. Given a two-dimensional (2D) view of a 3D world, there is no unique way to reconstruct the 3D signal. Formally, such an ill-posed problem has no unique or definitive solution. The same 2D image could represent any of an infinite combination of 3D scenes, even if the data were perfect (Bradski and Kaehler, 2008, p19).

As any 2D signal can lead to many descriptions, there is an inherent uncertainty that computer vision needs to accommodate. At the core of the problem lies a difficulty in creating a reliable “visual vocabulary” (Shah, 2012), as there is no formal model to correlate irrefutably a statement to a configuration of pixels. This problem is what computer vision developers name the semantic gap. Despite these difficulties, over the last decade, the gap has been considerably reduced. In a blog post titled *Improving Photo Search: A Step Across the Semantic Gap*, Chuck Rosenberg[16] triumphantly writes:

> we showed a major upgrade to the photos experience: you can now easily search your own photos without having to manually label each and every one of them. This is powered by computer vision and machine learning technology, which uses the visual content of an image to generate searchable tags for photos ... (Rosenberg, 2013).

This announcement features among many publications claiming significant improvements in techniques of image retrieval, automatic annotation and machine tagging (Huang *et al.*, 2015, Kastner *et al.*, 2019). What happened in ten years has often been narrated in terms of technological evolution, as the result of the increasing development of algorithmic techniques and mathematical models (Kurenkov, 2015). To understand the point of departure of my research, it is however important to narrate a different story and give agency to other agents in the process.

Traditional AI and early computer vision relied on explicit modelling. In a framework where

[16] Google I/O is an annual conference held by Google where state of the art software projects are presented and debated.

modelling is explicit, developers try to design themselves a model that matches the complexity of the problem domain. For instance, a cat's face is decomposed into simple shapes like a circle for the face, two triangles for the ears and two circles for the eyes. This approach is only efficient for a limited amount of cases. Its advantage is to produce a legible model: circle + 2 triangles + 2 circles = a cat face. Such a model may function in a strictly controlled environment, but leads to overwhelming problems in real-world scenarios. Under this paradigm, developers have to stack many local algorithms together to try to cope with the complexity of the visual world. As NVIDIA's technical expert Michael Copeland explains, to create a stop sign detector:

> People would go in and write hand-coded classifiers like edge detection filters so the program could identify where an object started and stopped; shape detection to determine if it had eight sides; a classifier to recognize the letters “S-T-O-P.” From all those hand-coded classifiers they would develop algorithms to make sense of the image and “learn” to determine whether it was a stop sign (Copeland, 2016).

When they do so, programmers quickly encounter difficulties. Stop signs in photographs are rarely perfectly centred and illuminated. Fog, partial occlusions, unexpected perspectives often change the organisation of the patterns. For each of these changes, the algorithm needs to be updated and optimised. Such an approach lacks the power to generalize as the algorithm must not only discriminate a pattern adequately but it also needs to make sense of the other patterns interfering with the pattern it attempts to detect (Deng *et al.*, 2009). Furthermore, the developer needs to acquire precise knowledge about the target object and to analytically decompose it before writing the code. To write a food classifier, a programmer needs to become a food expert.

In contrast, current techniques of machine learning based on neural networks do not rely on a previous analytical decomposition. The developer assembles a dataset reflecting the variations of the domain under study and utilizes automated means to calculate an optimal function that treats the features of the data as parameters. In computer vision, this technique, at its most simple level, uses large visual databases in which discrete units such as pixels can be considered as data points and their colour values represent the different dimensions of the data. Common techniques of machine learning in computer vision are said to be “supervised”, which means that the data is curated[17] (Beheshti *et al.*, 2016) in order to provide examples from which the machine learning algorithm extracts regularities: the software “learns by example”[18]. To come back to the example of the cat, in

[17] Beheshti et al. (2016) describe data curation as an activity that “involves identifying relevant data sources, extracting data and knowledge, cleaning, maintaining, merging, enriching and linking data and knowledge”.

[18] There are other models used in machine learning such as unsupervised learning, or reinforcement learning as well as a whole gamut of hybrid options. This study concentrates on the semantic gap and therefore crucially on the problem of classification for which the supervised learning method is optimised. Or to put it in technical terms,

the data-oriented paradigm, the developer will not try to decompose the animal into distinct shapes and explicitly summarise their relations. She will curate a large series of photographs where the cat is displayed in various positions, and the algorithm will detect the regularities traversing the dataset. Through this phase of "learning", the algorithm produces the model.

The change in algorithmic design does not merely signal an evolution of the technical process. It also provokes more broadly a reconfiguration of the positions of the actors in the field. To understand this, it may be useful to go through the general principles of calculation that undergird machine learning algorithms and observe the division of labour that corresponds to them. In its essence, machine learning can be summarised as a process of function-finding. Machine learning draws on a series of mathematical techniques to find the optimal function that maps an input to an output (i.e. to map a set of photographs to a set of labels). As there are many functions that can connect a given input to a given output, the algorithm has to additionally be able to compare and discriminate between the different functions that approximate the expected results. An operation of machine learning is therefore defined by the relation between two types of functions: the operational functions that traverse the vector of data to provide predictions and an observational function that compares the predictions of an operational function to the expected values. The relation between operational and observational functions is orchestrated in a non-linear and heavily iterative fashion which offers little analytical intelligibility. The question is approached as a problem of optimisation. As Adrian Mackenzie (2017) remarks, machine learning is not a matter of mathematical analysis, and therefore doesn't offer the same legibility as the derivation of a formula. The measure of success doesn't result from an assessment of the consistency of the computational process. It results from a constant comparison with the values the process is expected to yield. These values of reference are what the dataset offers. The gist of these general principles is encapsulated by machine learning pioneer Vladimir Vapnik's definition of learning: "learning is a problem of function estimation on the basis of empirical data" (Vapnik 1999, p.291). A choice needs to be made among approximations as there is no unique path from input to output. The process is defined by what counts as empirical data. The data is not just a passive element the machine extracts structure from. It is an instrument of supervision of the function-finding process.

Under the current machine learning paradigm, the domain under study cannot be expected to lend itself to a decomposition into neat logical steps anymore (Kirchdoerfer, 2018). The site where the modelling happens is changing. The analytical decomposition of the problem is now replaced by the production of a dataset that exhibits the regularities defining the problem. The engineers are said to

because "classification is a subcategory of supervised learning" (Roman, 2019).

be programming by example rather than by rule and let the program “discover” the rule inherent to the data. This represents a move beyond the modelling paradigm (Kirchdoerfer, 2018, p.14) to operate directly on the supplied datasets. But these datasets do not appear by magic, they need to be curated, assembled, maintained and annotated. Concretely this means that the modelling is outsourced to those who curate and annotate the dataset. The engineers say the model is learned end to end. This means in fact that it doesn't learn from them anymore. The paradigm change in machine learning has externalised the modelling process and, by doing so, has brought up a new division of labour. To substantiate this claim, I will analyse ImageNet, a dataset of reference for computer vision and from there turn to the platforms where the work of annotation happens.

2.3. What is a dataset? The example of ImageNet

As machine learning is said to approximate an empirical basis while iteratively adjusting to a set of samples, the processes of creation of the datasets have dramatically changed. The most obvious sign of their evolution is the increase of the datasets' sizes. As I wrote previously, we are now far from the dataset of a few hundred photos taken in the lab and the houses of the researcher's colleagues.[19] Current machine learning models require the availability of huge amounts of data. For example, the neural networks achieving state-of-the-art performance are trained using datasets with millions of labelled faces: Facebook’s DeepFace and Google’s FaceNet were trained using 4 million and 200 million training samples, respectively (Hu *et al.*, 2016).[20] The intricate relation between algorithm and data presupposes an intensive production of images. The training set must additionally provide variations in exposure, focus, colour, content, size reflecting the variety of the images the algorithm will be treating in production. The importance of quantity and diversity can be explained by referring again to the process of supervised learning that guides a large amount of vision algorithms. In supervised learning, the learning algorithm is given a dataset containing samples correlating input values (e.g. images) and output values (e.g. labels like cat, dog, horse, etc.). Supervised learning begins with a phase wherein the algorithm attempts to map the input to the correct given outputs. When it has optimised the function, it is tested on a dataset with other values where the trainers can measure its ability to generalise building upon what it learned in the first phase. This second stage is essential in order to control its degree of fitness. During this phase, the engineers

[19] Large efforts in the field like the FERET or MNIST datasets already existed at the time of creation of Faces1999. They represent more systematic efforts of collection than Faces1999. They nevertheless pale in comparison to the millions of photos of ImageNet and its distributed mechanism of annotation.

[20] This scale gradually becomes the default and part of the training routines in the research labs of high tech companies. Facebook researchers claim to train their networks in one hour on ImageNet (Goyal *et al.*, 2018)

check if the obtained function and its parameters are not too closely defined by its training set and are able to match new variations. If the obtained function is said to overfit (the algorithm is unable to generalise) or to underfit (the algorithm doesn't learn enough from the provided data), the programmer can tweak the learning algorithm's parameters and start the process again. The balance between overfitting and underfitting is what computer scientists call algorithmic bias. This is the first definition of bias in the present study. Here, controlling the algorithm's bias requires the availability of a training set that displays enough variability. If the sample diversity is too limited, the algorithm will not be able to generalize well (Hu *et al.*, 2016) . If the diversity is too important, it will struggle to find significant regularities. This relation between diversity and regularity is solved, in theory, by providing large datasets carefully edited and annotated to meet these requirements. We will see in the next chapters the limitations of a definition of bias in terms of underfitting and overfitting, but before that, it is crucial to understand what the creation of a large dataset concretely entails and which kind of material resources it mobilises.

One of the largest databases of human annotated visual content to date is ImageNet (Deng *et al.*, 2009). The ImageNet project is a collection of images for visual research that offers tens of millions of images manually annotated, sorted and organized according to a taxonomy. ImageNet aims to serve the needs of computer vision researchers and developers for training data. Due to its size[21] and its extensive annotations, ImageNet has become de facto the most used knowledge base in the world of computer vision (Fei-Fei, 2010). In the words of Fei Fei Li, its creator, ImageNet responds to the need for a training set at the scale of the web, a “treasure trove of images” (GoogleTechTalks, 2011). ImageNet is exclusively composed of digital photographs.[22] The dataset functions as a large cache for transient networked images. The photos are culled from the internet and are formatted to circulate on the web. They are generally low definition photographs, sometimes overlaid with timestamps or watermarks. The dataset represents a moment of online photography where Flickr enjoyed a dominant position for the sharing of photos. One half of the 14,197,122 URLs included in the dataset point to an address on the site's domain. The other half come from various sources: amateur sites, blogs, stock agencies, news sites, forums, etc. Background information as well as copyright notices are absent from the dataset. ImageNet offers the images for download to educators and researchers from its site and the list of original images URLs. As the URLs point to the image files and not the page in which they are embedded, it is generally difficult to reconstruct the context from which the photo originates. Once copied into the dataset, the link to the environment where the image was liked, shared, commented and tagged is severed. Over time, a huge amount of these

[21] The average number per category is 10.5k (Fei-Fei, 2010).

[22] As an instruction given to the dataset's cleaners makes it clear, other visual formats are not welcome: PHOTOS ONLY, NO PAINTINGS, DRAWINGS etc. (Fei-Fei, 2010).

images have been taken offline or have moved from their original location. They have become exclusively publicly available through ImageNet.[23]

ImageNet aims to provide a database that comprehensively covers the image world (Deng *et al.*, 2009). ImageNet's ambition dramatically contrasts with its contemporaneous datasets. Where other training sets were providing at most a thousand categories, Li's project offers 21,841. It includes a large botanical section featuring a dizzying array of plant varieties and a detailed catalogue of geological formations. The category “natural objects” stretches from “rock” to “extraterrestrial object”. A place of honour in the top categories is given to “sport activities” and “fungus”. The core of the database is constituted by the sets “person”, “animal” and “artefact” offering an overview of the wide variety of living bodies and manufactured objects. Finally, the rather poetic “Misc” brings together disparate entries such as “tabernacle”, “meuniere”, “shit” and “puce”. The need for a diverse set of data requires not only variation in content but also in shape. ImageNet exhibits a wide range of photographic genres. Sometimes the genre coincides with a category. For instance, the “Adenovirus” category[24] like other categories related to disease contains mostly photomicroscopic images while the “Cardioid microphone” category, as other categories related to technical equipment, consists of images from online catalogues. But more often in the same category, different photographic practices are represented. A fish is represented swimming by an underwater photography enthusiast, as a trophy by a proud weekend fisherman, or on a plate for an ad for a restaurant.

2.4. ImageNet, resolving the scale of the problem

ImageNet is conceived as a project which demonstrates that images are not interchangeable vectors, that they exhibit a large amount of variation and unexpected regularities. Large scale classification confronts the programmers with use cases that they had not anticipated and challenges the conventional wisdom derived from working on a couple of hundred of image categories (Deng *et al.*, 2010, p1). With the dataset, Li intents to provoke a change in machine vision that is not only quantitative. For Li, ImageNet is an attempt to “resolve the true scale of the problem” (GoogleTechTalks, 2011), one which, in so doing, redefines the parameters of the problem of recognition.

[23] As I will explain in the last chapter, the availability of the photographs of ImageNet has become a controversial issue and access to the complete dataset is currently prohibited by ImageNet's maintainers.

[24] Technically, in ImageNet parlance, a category is named a synset which stands set of synonyms. See below the WordNet section for more details.

To understand how much the dataset has been instrumental in the transformation of the field, one has to turn to the ImageNet challenge, organized yearly by the Standford AI lab. Through the challenge, universities and companies compete to demonstrate the quality of their algorithms. In the world of computer vision, where everything is changing at breakneck speed, ImageNet has established itself as a stable referent (Russakovsky *et al.*, 2014). Every year, the progress of state-of-the-art algorithms can be measured with a common set of indicators and agreed upon standards. The challenge consists of labelling correctly 1000 different types of objects. In 2012, Geoffrey Hinton with two of his students from the University of Toronto, won the competition. Not only did Hinton's algorithm win, but it did so by performing 10% better than the second-best competitor (Krizhevsky *et al.*, 2012). This victory re-established the scientific reputation of a whole area of machine learning. Hinton and colleagues[25] dramatically improved a computational method that had been severely criticized during decades of research on machine learning. The academic environment was overtly hostile to what was considered as a sterile field of investigation. Papers describing the nascent deep learning techniques were systematically rejected by the reviewers of computer vision conferences[26], even if the reported results showed a superiority over their competitors in terms of speed and accuracy (Kurenkov, 2015b). If the attempts to publicize deep learning's achievements through the academic channels had led to frustration, winning the ImageNet contest changed the whole perception of what was possible to do with neural networks, and made longstanding problems like the semantic gap tractable. By providing a scale for the problem, ImageNet contributed to a bifurcation in machine learning and validated a strand of research rejected by the academy. ImageNet is therefore more than a supplier of data, it is a benchmark that transforms computer vision with images.

There is another aspect of ImageNet that goes beyond its role as a supplier of images for research and testing. ImageNet provides a common visual vocabulary for many algorithms in production. As Hentschel *et al.* observe, “assembling large amounts of reliably annotated ground truth data is a costly (sometimes even prohibitively costly) and time-consuming process” (Hentschel *et al.*, 2016), and few developers and companies can afford to produce their own. To cope with this problem, many programmers resort to a technique called transfer learning. As explained in a machine learning primer written for the new users of Tensorflow, a renowned open source software for artificial intelligence:

[25] Famously, Yann LeCun, Yoshua Bengio and Geoffrey Hinton teamed together to form the “deep learning conspiracy”, a group of researchers committed to defend the cause of the machine learning after it fell into academic disgrace. For more details see Bergen and Wagner (2015)

[26] Read this thread to see the heated discussion caused by the submission's withdrawal of LeCun from the computer visionPR conference 2012: https://plus.google.com/+SergeBelongie/posts/StLxWpZ33L9

> Transfer learning is a technique that shortcuts much of this [the long process of training] by taking a piece of a model that has already been trained on a related task and reusing it in a new model (Tensorflow, n.d.).

In practice, developers reuse neural networks already trained on a large corpus and fine-tune the model using a moderate amount of target data (Hentschel *et al.*, 2016). A neural net trained on ImageNet learns the general regularities of visual data and can be specialized using a more focused and cheaper corpus. Although it performs less accurately than a network trained end to end with a dedicated dataset, the trade-off is in many cases deemed acceptable. In the words of Tensorflow developers:

> it turns out the kind of information needed to distinguish between all the 1,000 classes in ImageNet is often also useful to distinguish between new kinds of objects (Tensorflow, n.d.).

This means that ImageNet, because of its availability and the amount of networks already trained with it, becomes infrastructural and therefore, as Bowker and Starr (Bowker and Star, 2000) point out, invisible and "in the background". As developers constantly reuse existing code and build upon it, a large number of detectors are performing their work by calculating the distance between the image they analyse and their summary of the ImageNet dataset.[27] The grid of intelligibility of many deep learning algorithms is furnished by ImageNet as other options simply are not available for financial reasons.

Both examples show the combination of scientific relevance and strategic importance of the notion of scale. By changing the magnitude of reference data, ImageNet changes the nature of the computational problem. By doing so, it empowers new actors and new techniques. It also establishes itself as a referent and, due to the economic cost of the scale it imposes, becomes an infrastructure embedded in a large variety of applications. These examples have stressed the ramifications of ImageNet in machine vision. I will now examine how ImageNet itself builds upon an existing infrastructure.

2.5. How to "image" categories? WordNet, a semantic backbone

ImageNet is built on multiple technological layers and draws from various sources. For the dataset's semantic organisation, Li has drawn on WordNet, a lexical database created at Princeton in 1985 to

[27] In this context, Deselaers and Ferrari (2011) talk about the *ImageNet distance* between two images.

support the development of Artificial Intelligence. The importance of WordNet for the project is such that Li chose its name in homage to the lexical database.[28] Whilst the main feature of ImageNet is its collection of millions of images, it would be of no use if its photographs were not classified. As a dataset providing material for the training of visual recognition, the semantic organisation of the visual data is of the uttermost importance.

Like ImageNet, WordNet has gained its reputation by investing in scale. For WordNet´s creator George Miller, computational linguistics should have “a store of lexical knowledge as extensive as people have” (Fellbaum, 1998). WordNet responded to the need to have “a larger vocabulary than the toy illustrations of the day” to test the hypothesis of relational semantics (Miller in Fellbaum, 1998, p.xvii). This larger vocabulary has expanded into a database of 117,000 concepts. Over the years, the objective of comprehensiveness of lexical knowledge for the English language had to be revised. But if the number of concepts has stabilised today, the project is still growing in many directions as, for example, new translations and localisations continue to expand its reach internationally.

In a cognitivist fashion, the hierarchical organisation of concepts is essentially guided by logical relations of inclusion and inheritance. WordNet can be conceived first and foremost as a tree of concepts. The lexicon is organised around word meanings rather than word form and nouns are its first-class citizens. WordNet's concepts are defined by synonym sets of nouns, called synsets, instead of an explicit definition. As its creator George Miller explains:

> The synonym sets, {board, plank} and {board, committee} can serve as unambiguous designators of these two meanings of board. These synonym sets (synsets) do not explain what the concepts are; they merely signify that the concepts exist (Miller, 1995).

The synsets are organised hierarchically from the most abstract to the most concrete. To continue with the example of board, the synset {dining table, board} is located in the following hierarchy:

{entity} → {physical entity} → {object, physical object} → {whole unit} → {artifact, artefact} → {instrumentality, instrumentation} → {furnishing} → {furniture, piece of furniture, article of furniture} → {dining table, board} [29]

[28] In a recent public presentation of ImageNet at The Photographers' Gallery, Li (2019) explains that she met Christiane Fellbaum, WordNet's project lead, who told her about an extension of WordNet, ImageNet, that never came to fruition. Li adopted the name for her dataset.

[29] For comparison, other “board” synsets are classified in the following locations:{entity} → {abstraction, abstract entity} → {group, grouping} → {social group} → {organization, organisation} → {unit, social unit } → {administrative unit, administrative body} → {committee, commission} → {board} {entity} → {physical entity} → {matter} → {substance} → {food, nutrient} → {fare} → {board, table}

The database relies essentially on the double structure of synonymy and meronymy to express concepts and their differences[30]. In WordNet, synsets represent meanings. There is little interest in the pragmatics or the relationality of language. As Murphy observes, “Relations represented in WordNet are, for the most part, paradigmatic rather than syntagmatic” (Murphy, 2003). The paradigmatic structure of WordNet also implies the taxonomy's scale. As Monique Slodzian (2000), points out, in a paradigmatic framework every new meaning requires a distinct new synset, therefore the only way for the vocabulary to cope with polysemy is by multiplying senses and synsets. WordNet's scale therefore is not only an external indicator of its coverage of the English language. It is also an internal necessity that corresponds to the concept of language that underlies its architecture.

Even if the importance of WordNet for Li cannot be overstated, she does not import the lexical database without alterations. ImageNet is built on a selection of 21,841 synsets[31]. This is due in part to economic constraints. It is also due to the fact that the process of translation of the lexicon into visual concepts raises multiple questions. For example, one of the criteria for the selection of a synset is what Li calls its “imageability”. This term refers to the availability of images to represent the concept. To understand how the assessment of imageability of a category is made, it is useful to explain the process of acquisition of images. WordNet has been instrumental in the enterprise of curating photographs not just because it provided a structure for ImageNet's contents. WordNet synsets were key instruments to formulate the search queries the ImageNet's team used to download the dataset's photographs. When searching for images on the internet, WordNet helped Li and her team refine their dialogue with search engines. For instance, when searching for “German shepherd”, the synset can be exploited to expand the query into “German shepherd, German police dog, German shepherd dog, Alsatian” and improve the search results. Also, making use of the multiple languages provided by the Global WordNet[32], Li and colleagues were able to translate their queries and have results from pages from different countries. But, if WordNet provides a powerful semantic matrix for queries, there is no guarantee that the search engines interpret them relevantly. The researchers do not trust search engine results. They hire annotators to evaluate the results. When there is a low consensus among the annotators about the search results for a category, the

[30] Even if over the years, WordNet has added rudimentary definitions to the synsets, meronymy and synonymy are the central features used in computational linguistics to perform tasks such as search augmentation, word disambiguation or text filtering. And the example sentences included in the glosses are optional illustrations that do not lend themselves easily to computation.

[31] The hierarchical layout is also modified. Some lower level categories have become top level entries: fungus, plant/flora and natural objects are at the same level.

[32] Global WordNet refers to the translations and internationalisation of the project in multiple languages.

category is simply deemed “unimageable” (Yang *et al.*, 2019, p.5). WordNet is utilized as a backbone and at the same time it is being revised by the annotators and entire branches of the tree are discarded if they do not reach consensus.

For a dataset recognised as a benchmarking tool to evaluate competing algorithms, a hierarchical structure presents an advantage. To detect a label correctly may be conceived as a binary task: either the detector finds the right label or misses it. But when labels are organised hierarchically, the degree of control over the evaluation may be more nuanced. As Li explains, classifying a “dog” as “cat” is probably not as bad as classifying it as “microwave” (Fei-Fei, 2010). To provide training data organised in a tree-like structure makes it easier to evaluate the performance of an algorithm. It helps measuring its ability to generalise and it helps troubleshooting and re-training. This feature comes with a few strings attached, however. First, it means that the conceptual ordering of the world as organised by language can fit into a unique hierarchy that descends from abstract entities to material ones. It also means that photographs can faithfully “illustrate” these concepts and find their place in the hierarchy. Finally, there is the sense that this construction may transparently represent the concepts. These assumptions will be examined and challenged in the next chapters. But for now, I want to underline that several signs indicate that for Li herself the articulation of the lexicon and the photographic material has been uneasy. There is a notable ambivalence in Li's public assessment or defence of WordNet. As ImageNet has become the object of mounting controversies both from media activists (Crawford and Paglen, 2019) and from inside the tech community (i.e. Dulhanty and Wong, 2019; Shankar *et al.*, 2017; Recht *et al.*, 2019), Li has had to respond to a criticism mainly targeting the dataset's cultural, racial and gender bias[33]. For a large part, WordNet takes the blame for these accusations. In a recent article where they document their effort towards a fairer dataset, the team of ImageNet's maintainers refer to the taxonomy as a stagnant vocabulary (Yang *et al.*, 2019). In a previous discussion with Google engineers, Li downplays the vocabulary's authority and refers to it as a provider of a list of freely available labels. She recognizes its imperfection and argues that, even if it is imperfect, it captures enough of the linguistic structure to match broadly the regularities of the visual world (GoogleTechTalks, 2011). The difficulty is also apparent in the fluctuating relation of authority between the taxonomy and the visual data. Whilst WordNet is used as the conceptual skeleton of ImageNet, and the dubious politics of some of its categories is imported uncritically[34], it is altered in many ways. As we have already seen, when the annotators do not reach consensus, entire categories are discarded. More abruptly, the researchers may simply decide unilaterally that they don't want to bother with certain

[33] The details of this controversy are discussed in the next chapter and in the conclusion.

[34] For example the synset “Transexual” is filed under under “anomaly, unusual person”. Crawford and Paglen (2019) offer a detailed overview of what ImageNet's offensive categories.

concepts (disease for instance has been discarded with several other "fluffy"[35] concepts). They also move certain categories up, threatening the coherence of the structure. WordNet deeply structures ImageNet and yet, it is translated in ways that abruptly disrupt its rigorous ordering. The lexical database is at the same time intimately interwoven with the visual collection and treated as a movable part. Its hierarchy is valued for its use in the evaluation of the detection error and yet can be presented by Li as a mere practical list of nouns, a contingent choice waiting for a better alternative.

At the level of its components, ImageNet can be seen as a mashup between a subset of WordNet and a subset of Flickr, a mapping of the photographic platform onto the grid of the lexical database. The dataset makers call this alignment of heterogeneous apparatuses the *data collection pipeline* (Yang *et al.*, 2019). The mapping performed through the collection pipeline is the result of two years of intense work. To classify millions of images has required to set in motion a long chain of translation. This chain correlates search engine queries, sharing platforms tagging systems, the intuition of crowdsourcing platforms' workers and 20,000 categories from WordNet. It is to the annotator, a crucial actor in this chain of translation and the figure I introduced in the first vignette of the previous chapter, that I now turn.

2.6. Outsourcing the modelling of vision, Amazon Mechanical Turk and computer vision's division of labour

Computer vision datasets depend on the availability of large volumes of photographs. They rely as well on some form of manual annotation. Earlier I quoted a Google I/O developer proudly announcing the end of manual tagging. This tedious work would be executed automatically from now on thanks to the advances in computer vision. Yet to automate this task, large volumes of annotations are necessary. While the users of the new Google software may not be obliged to tag their photos anymore, thousands of hands on keyboards and retinas glued to screens are producing an unprecedented amount of labelling and descriptions to train the algorithms that automate the user's tagging. In fact, the amount of annotation work involved in the production of datasets is even more impressive than the number of photographs included in ImageNet. After all, 14 million images is only a fraction of the monthly 575 million public uploads of photos on a platform like Flickr (Smith, 2019). The work of manually cross-referencing and labelling the photos is what makes

[35] Like the synset "patriot" (GoogleTechTalks, 2011)

datasets like ImageNet so unique.[36] In fact, there have rarely in history been so many people paid to look at images and report what they see in them (Vijayanarasimhan and Grauman, 2009). The automation of vision has not reduced the number of eyeballs looking at images, of hands typing descriptions, of taggers and annotators. On the contrary, it has increased their numbers. Yet what has changed is the context in which the activity of seeing is taking place, how retinas are entangled in heavily technical environments and how vision is driven by a certain speed.

Due to the pressure of generating annotations at the scale of the web, the process needs to be massively "parallelized". In the informatics jargon, to parallelize a process means to design a workflow where two processes can take place without interference. A parallel architecture optimises the calculation and makes maximal use of all the computational resources of a machine. On the Amazon Mechanical Turk (AMT) platform where a large part of the annotation effort is produced, the workers are treated as computational processes. The workers, the "Turkers", are abstracted away. The advantage of having parallel processes is that the execution time of a given task can be divided among the available workers/processes. Li calculated the amount of human labour required for ImageNet this way: estimating that a person can annotate two images per second[37], 19 human years are necessary to verify the tens of millions of images in her dataset. Amazon Mechanical Turk allows her to compress the 19 years into two by providing the necessary workforce to divide the workload[38] (GoogleTechTalks, 2011).

Apart from speed, there is another reason, less "technical", to adopt this parallel architecture. The parallel architecture isolates the workers and makes it difficult for them to create bonds, a collective identity and to unionize. The Turkers have no name, only an anonymous identifier and the platform doesn't provide any means of communication between workers (Irani and Silberman, 2013). There are strategic reasons for crowdsourcing platforms to fear workers' unions: their working conditions are exploitative. AMT offers people with large datasets the possibility to outsource the annotation work using their massive workforce. The tasks on AMT are rewarded with micro-payments. An annotation is estimated between 1 and 4 cents and the estimated hourly revenue for a Turker is approximately two dollars (Hara *et al.*, 2018). The annotator must find an optimal trade-off between the precision and attention required to make a good annotation and a speed that allows her to "maximise" her financial gain. The hourly rate depends on the task and access to the most lucrative tasks depends on the age, the origin and the class of the workers as one needs to be "culturally

[36] And hard to replicate for specific domains. Different research projects are attempting to produce artificially the image datasets rather than collect the images (Masi *et al.*, 2016).

[37] Li gives the following formula: "Speed of human labeling: 2 images/sec (one fixation: ~200msec)" (Fei-Fei, 2010)

[38] Li (Fei Fei, 2010) estimates to have hired more than 25,000 workers in two years' time (2008-2009)

competent” to handle the request (Irani, 2015). Annotation tasks in particular require a certain familiarity with American culture. A task like “US Recognition Game: Identify the masked image as quickly and accurately as possible!” is difficult to complete successfully without an acquaintance with the cultural context.

With the Amazon Mechanical Turk (AMT), informatics revives the tradition of massive human calculation. This concept is associated with the name of Gaspard de Prony who famously produced the logarithmic tables for the French Cadastre in 1791. Inspired by Adam Smith, de Prony had devised a system of distributed calculus that could take advantage of a workforce with low mathematical skills by dividing the effort into simple additions and divisions that were executed at a rate of thousands operations a day (Daston, 1994). De Prony's principles were as mathematical as they were economical as his goal was to optimise the cost of computation. The stratification and division of labour conjoined with the miniaturisation of work into micro tasks have persisted throughout the history of mass computation. In the 1950s, when the name computer still referred to persons (women), as Wendy Chun points out, “software, as a service, was initially priced in terms of labor cost per instruction” (Chun, 2011). In *Programmed Visions*, Chun examines the details of the work executed by the women computers and the mechanisms of control exerted over them to make sure they remained focused on the repetitive operations they were given. Today, crowd workers do not perform manual additions and subtractions. They produce what the machine learning jargon calls the “ground truth”. This expression, of military origin, designates data collected on the battleground as opposed to simulations used in the headquarters (Simpson, 2012). The expression ground truth emphasizes that ultimately the data in situ has the power to contradict the representations used by strategists in a remote location. In line with the history of mass human computation, the expression ground truth also refers to a division of labour and a chain of command. If ultimately the data from the battlefield prevails, those who give orders remain in the headquarters. From de Prony to the ENIAC girls, the workers' supposed lack of skill has been a key factor in producing large scale computation at low cost.

Who can today produce the ground truth? What is the qualification of a Turker? In a nutshell, the crowd worker is a specialist at being generic. To explain this apparently contradictory description, it is helpful to have an idea of the kind of tasks that are listed on the AMT interface. These hits rarely require the Turker to be highly trained in a domain. They are the kind of tasks which are said to be easy for a human and hard for a computer. But this is only half the story. The Turkers describe the colour of objects, trace the contours of shapes, they also answer endless surveys and participate in a large number of tests ranging from psychology to interface testing. They make use of their generic skills, like the ability to describe an image, and these same skills make them the subject of choice for data collection and statistics. They are professional generic subjects. Professional because their generic character is heavily constructed. Every time they log in, they are required to perform tasks that require basic logic and reasoning. And they become faster and better subjects. They internalize the recurrent expectations of their requesters. But they must be careful not to become too good as the requesters choose them for their lack of specialisation. For instance, the annotators are asked if they have expertise on the current synset they are cleaning up. If so, they are invited to signal it to the requesters as their input may vary from other Turkers. Expertise is the exception and the worker is asked to fill in more information for her contribution to be accepted because it may likely diverge from consensus. The fact that the expert needs to take more time to fill this information in corresponds to a penalty: by losing time, she also loses money.

The workers, if they want to make a living, need to work at a pace that barely allows them to see the thumbnails appearing on their screens. For the annotators, structurally, the glance is the norm, not the gaze. Speed is built into the platform economically. Training sets must be produced fast. Lots of workers are mobilised intensely for a short period of time. Through the AMT interface, the requesters manage the cadence of the annotation work. They want to ensure the workers go fast enough to match production deadlines. And, at the same time, they attempt to preclude them from overlooking their task in order to avoid drops in quality. The interfaces of annotation are designed to control workers' productivity, to find the optimal trade-off between speed and precision. The time estimation for labelling tasks is especially difficult as it varies according to the annotator's experience and the nature of the photographs. Andrej Karpathy, now Tesla's director of AI, reported that when he annotated an ImageNet recognition challenge dataset, the labelling started at a rate of one image per minute and decreased with the accumulated experience (Karpathy, 2014). But, even if progress was noticeable, the rate was not constant, as some images like those depicting some particular breeds of animals required a longer time and additional research. To cope with the variability of the annotation process, various techniques are deployed to streamline it. To begin with, if manual labelling is costly, the priority must be given to the annotations producing the highest information gain. As Vijayanarasimhan and Grauman put it, unlabelled images must be ranked “according to their expected ‘net worth’ to an object recognition system” (Vijayanarasimhan and Grauman, 2009). The cost of the labelling effort leads computer vision researchers to approach visual content in the form of informational currency and attention scarcity. This approach informs the architecture of annotation workflows wherein a pre-labelling attempt is made by an algorithmic detector and corrected by human annotators (Papadopoulos *et al.*, 2016). Pattern recognition techniques are used to absorb the bulk of the “easy” detection work, spot the potential targets and redirect the “ambiguous” cases to the human annotators (Vijayanarasimhan and Grauman, 2009). Whilst these techniques of semi-automation are in their early stages of development, their existence indicates the anxiety of dataset makers to control the annotator's attention and prevent her distraction. The requesters invest in the annotator's attention and treat attention as an asset that needs to be protected. Besides guiding the annotator's eyes to specific regions of interest and regulating their attention, researchers are exploring how to augment the annotators' ability to absorb visual content. For instance, Krishna *et al.* (2016, p.2) claim to have devised a technique of rapid serial visual presentation that allows workers to produce one HIT (*Human Intelligence Task,* in AMT Jargon)[39] in 100 milliseconds by immersing them in an uninterrupted visual flow. As the volume of requests augments, such experiments indicate that the unit of measurement for HITs is moving towards the millisecond. Finally, others are concentrating on evaluating workers performances over large periods to identify the annotators able to sustain the rhythm over time without decline in submission quality and sifting out the “satisficers” (Hata *et al.*, 2017) who strive to do the minimal amount of work to meet the acceptance threshold. In response, the workers themselves try to figure out what is a good ratio between speed and accuracy. On social media platforms like Reddit, AMT workers exchange tips about “good paying tasks” and compare their performances. To be able to estimate the amount of work required for a task is a matter of survival in the crowdsourcing environment.

2.7. The elaboration of vision, a transepistemic relation

At this point, it is worth reflecting on key elements that have been uncovered through the journey. Following the detours of computer vision through its datasets, taxonomies and platforms, a certain notion of scale gains importance. Li stated that a crucial step to advance her discipline was to resolve the scale of the problem. I have already stressed how much ImageNet as a benchmark and as an instrument of pre-training had redefined the contours of machine vision. From this perspective, ImageNet's scale doesn't represent a mere quantitative evolution, it is the establishment of new sets of relations and infrastructures. A scale, although manifested in logistics and industrial infrastructure, is also something more. As media scholar Matthew Fuller writes, a scale provides “a certain perspectival optic by which dimensions of relationality and other scales may be 'read' ” (Fuller, 2005, p.132). From a newly sensible scale, it becomes possible to “read new dimensions of potentiality” (Fuller, 2005, p.132). As a benchmark instrument, ImageNet does more than provide an empirical reference to show the superiority of a neural network technique over another. Geoffrey

[39] For instance, an annotation or an answer to a binary choice.

Hinton had already published the performance of his algorithm but his contribution was not acknowledged by his peers. What ImageNet did was to make the potential of Hinton's technique apparent. The making of ImageNet is the making of computer vision by other means. Li's work is an articulation of scales. It is the articulation of Flickr and WordNet, the articulation of vision with the decomposition of work into micro-tasks and micro-payments. The problem scale is an articulation of the macro and the micro. 14 million images need a model of vision where the millisecond is the unit of measure for perception. When thinking about the problem scale of computer vision, we need the opposite poles of the problem's dimension: on one hand, the infinitesimally small units of perception (the eye saccades), the miniaturisation of the work process in discrete HITs that can be performed at full speed, and on the other hand, the massive amounts of photographs available on digital platforms and a vast population of precarious workers ready to annotate on demand.

This articulation of scale brings about a new distribution of labour crucially affecting the task of modelling. A task that was previously in the hands of the engineers is distributed among the annotators. The engineers want to delegate a series of crucial decisions, but they want to keep control over the way the problem is formulated and its interpretation. Through the interfaces and contracts of AMT, the engineers design the frame of transepistemic relations. Transepistemic is an adjective coined by sociologist of science Karin Knorr-Cetina (Knorr-Cetina and Malkay, 1983) to qualify a production of knowledge that unsettles the borders of a discipline. For Knorr-Cetina, an example of a transepistemic relation can be found in the interaction between a funding agency and the researchers they support. Often the funding agency participates in the framing of the problem the researchers will attempt to solve. Research happens in and ex situ, yet the contribution of the agency remains unacknowledged in the reports and research outputs. With this concept, Knorr-Cetina emphasizes the active role of external agents in the production of science and its relative invisibility. Transepistemy should not be understood however as a denunciation of an external control that determines the research process. The concept stresses the importance of the circulation of knowledge and its distributed nature. Although in a dramatically different balance of power, the engineers and the annotators contribute together to the production of the knowledge relevant to computer vision. AMT can be characterised as a transepistemic device that enables collaboration whilst masking one party's contribution. It is a key apparatus in the elaboration of computer vision. It elaborates computer vision as it provides the modelling for the training process at the appropriate scale. And it elaborates it (this time with the emphasis on 'laborare', the latin root of the word) as it articulates computer vision's division of labour.

2.8. Conclusion / transition

The photographic elaboration of computer vision refers to the effort to make vision amenable to machines as a distributed process which requires the collaboration of human and non-human agents. What becomes apparent now is that an algorithm is not an abstract entity determining human affairs from a supposed outside. Here, an algorithm is considered as part of a network of operation, not as a self-standing agent. Its elucidation doesn't require the deciphering of an inscrutable code but a detour through networks and apparatuses.

The dataset is an obligatory passage point in this network. It is a collection of photographs used to train algorithms to detect regularities in order to perform tasks such as detecting objects or faces. It is also a device to probe a scale and to articulate dimensions. This chapter worked out a notion of scale that goes beyond quantity. Understanding the scale here requires an analysis attentive to the minute details of the process as well as its breadth: the scale articulates the micro and the macro. The scale I have analysed requires the 117,000 concepts of WordNet, the billions of photos of Flickr and the army of workers of AMT, the micro-tasks executed below the second and the micro-time of eye saccades. Cutting through these myriads of entities, this scale provides a temporal resolution of a multiplicity of dimensions and apparatuses.

I ended this chapter focusing more particularly on the Amazon Mechanical Turk platform where thousands of annotators are recruited to view, label and classify photographs around the clock. When they label and classify, the workers perform indirectly a task previously under the responsibility of engineers – modelling. With the shift from explicit to implicit modelling, the Turkers have become core contributors to computer vision whilst their contribution remains unacknowledged. They do not merely belong to the algorithmic supply chain, they are engaged in a trans-epistemic relation. I am calling this process of unacknowledged delegation the elaboration of computer vision, to emphasize the intimate relation between the development of computer vision's models and the division of labour it entails.

This first chapter leaves several questions open. A first set of questions pertains to the model of vision where vision is understood as automatic and immediate. What makes computer vision engineers think investing in such scale is relevant? How can they establish that, barely glancing at a photograph, a worker can extract enough meaningful information to be able to classify or label? Realising how much the industrial process of annotation depends on what can be glimpsed of a photograph stresses the importance of the model of vision which underlies the annotation environment. Where does the model of vision come from? And how does such a model of vision provide a series of relations that can be made operational? This strongly indicates that our journey into the detours of computer vision is not over and reinforces the need to enquire into how such a model comes to being. These questions will be developed in chapter three.

The second set of questions relates to the photographic dimension of the process. I have shown a stark contrast between the photographic practices underlying the production of Faces1999 where the scientists were taking portraits of their colleagues and families and the more recent acquisition pipeline where they download photos massively from the internet. The change from explicit to implicit modelling corresponds to a change in the division of labour. It also corresponds to a change in photographic practice. The engineers, once the photographers who were controlling the process from beginning to end, become the managers of large scale annotation processes where photographs are reviewed, labelled and classified by people paid to do so. Furthermore, it corresponds to a change of apparatuses: the camera is replaced by an alignment of platforms. Flickr, the search engine, the AMT interface take prominence. How to conceive of the relation between the elaboration of computer vision and its photographic mediation? The next chapter will concentrate on the conceptual framework that will allow me to address the question of the role of photography in the process. And to move from an elaboration of computer vision to its *photographic* elaboration.

Chapter 3. Photographic elaboration

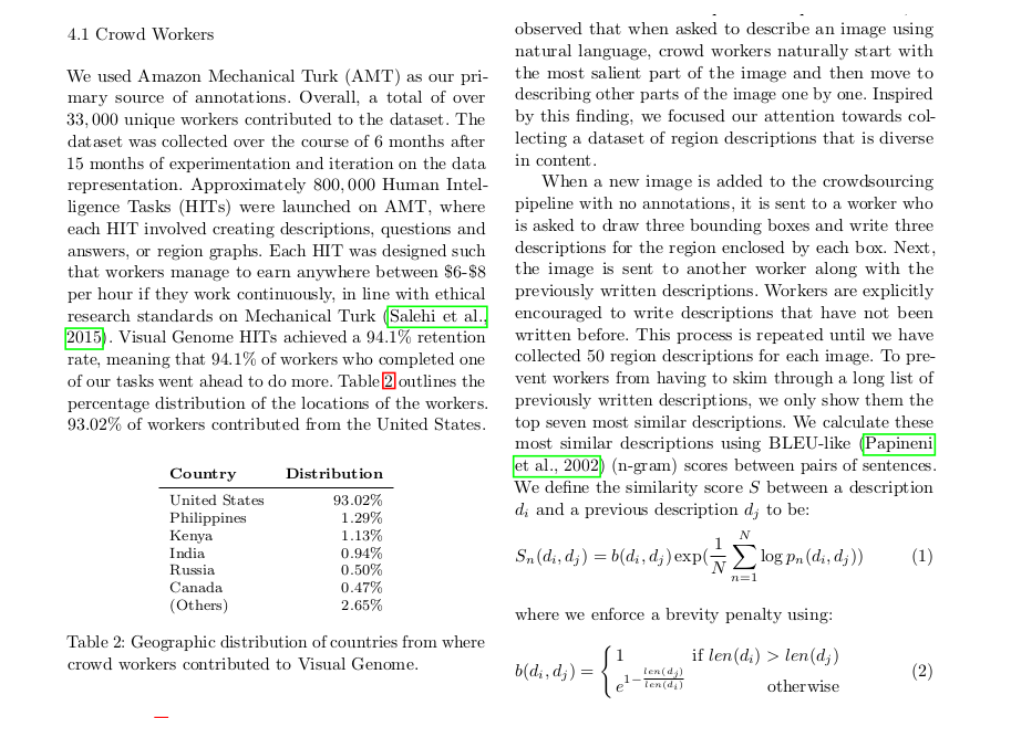
This chapter examines how re-connecting algorithms to their networks of operation calls for a specific understanding of photography. It represents the second stage of the journey of photographic elaboration. However, to get to the question of photography, a preliminary reflection about the mapping done in the previous chapter is in order. To follow the network of operation of computer vision has consequences for the understanding of algorithm, vision, scale. It has also consequences for how photography will be conceived in this thesis.

As the chapter moves on to another level of analysis, it is important to note the particular relation to theory that is cultivated through the research. The theorization in this chapter represents a flexible set of initial orientations. The argument about the elaboration of computer vision builds through the thesis and each chapter is designed to discuss what up to that point has been implicit and takes a more definite form through practical engagement in the second half of the manuscript. The discussion below is meant to open up the problem and discard pre-conceptions. It aims to revisit a series of concepts that were latent in the mapping and think about how to prepare the ground for the practice. It is not meant to dictate the terms of the practice in the sense that practice would apply theoretical principles, but to increase sensitivity to what the words and concepts used earlier bring with them, their disciplinary affiliations and to reflect critically on how they may orient the research.

3.1. Making algorithms relative

As I explained in the previous chapter, there is a burgeoning market of workers paid to look at photos, classify them and describe them. As Antonio Casilli (2019) notes in his essay on the invisible manual work in AI, we tend to easily forget that automation itself is work. The previous chapter's mapping has shown that ubiquitous photography and immersive computation require an “ambient workforce” (Scholz, 2017). Automation cannot be separated from a large network of maintainers and investors. When conceiving of the machinic agency, we must pay attention to the labour that is made invisible. An approach that doesn't pay attention to the labour that is invested to train, annotate and maintain runs the risk of fetishizing the algorithm and missing the hybrid nature of so-called machine intelligence. It risks endowing the algorithm with an autonomy that negates the material conditions that are not only necessary for its existence and maintenance but also crucial to its epistemic relevance. Such an axiomatic perspective, which understands algorithms, as Florian Jaton (2017, p.811) puts it, as “sets of instructions designed to solve given problems computationally”, concentrates the attention on the mathematical process and renders one oblivious to the modelling work outsourced to the annotators.

Dataset papers, it is true, contain various mathematical formulas. They also contain long paragraphs describing the scraping of online photographs and the management of Turkers (see illustration 3). The trajectory into the AI networks and collectives of the previous chapter shows the importance of finding an approach that remains sensitive to the annotator's contribution to the scaling operation of computer vision. And attentive to the distribution of agency over large alignments.



4.1 Crowd Workers

We used Amazon Mechanical Turk (AMT) as our primary source of annotations. Overall, a total of over 33,000 unique workers contributed to the dataset. The dataset was collected over the course of 6 months after 15 months of experimentation and iteration on the data representation. Approximately 800,000 Human Intelligence Tasks (HITs) were launched on AMT, where each HIT involved creating descriptions, questions and answers, or region graphs. Each HIT was designed such that workers manage to earn anywhere between $6-$8 per hour if they work continuously, in line with ethical research standards on Mechanical Turk (Salehi et al., 2015). Visual Genome HITs achieved a 94.1% retention rate, meaning that 94.1% of workers who completed one of our tasks went ahead to do more. Table 2 outlines the percentage distribution of the locations of the workers. 93.02% of workers contributed from the United States.

Country	Distribution
United States	93.02%
Philippines	1.29%
Kenya	1.13%
India	0.94%
Russia	0.50%
Canada	0.47%
(Others)	2.65%

Table 2: Geographic distribution of countries from where crowd workers contributed to Visual Genome.

observed that when asked to describe an image using natural language, crowd workers naturally start with the most salient part of the image and then move to describing other parts of the image one by one. Inspired by this finding, we focused our attention towards collecting a dataset of region descriptions that is diverse in content.

When a new image is added to the crowdsourcing pipeline with no annotations, it is sent to a worker who is asked to draw three bounding boxes and write three descriptions for the region enclosed by each box. Next, the image is sent to another worker along with the previously written descriptions. Workers are explicitly encouraged to write descriptions that have not been written before. This process is repeated until we have collected 50 region descriptions for each image. To prevent workers from having to skim through a long list of previously written descriptions, we only show them the top seven most similar descriptions. We calculate these most similar descriptions using BLEU-like (Papineni et al., 2002) (n-gram) scores between pairs of sentences. We define the similarity score S between a description d_i and a previous description d_j to be:

$$S_n(d_i, d_j) = b(d_i, d_j)\exp(\frac{1}{N}\sum_{n=1}^{N}\log p_n(d_i, d_j)) \qquad (1)$$

where we enforce a brevity penalty using:

$$b(d_i, d_j) = \begin{cases} 1 & \text{if } len(d_i) > len(d_j) \\ e^{1-\frac{len(d_j)}{len(d_i)}} & \text{otherwise} \end{cases} \qquad (2)$$

Illustration 3: Side by side, workers management and algorithmic similarity scoring (Krishna et al., 2016)

Whilst the map drawn in the previous chapter has been made from readings of research articles, textbooks, interviews with practitioners, and my own experience as a programmer, the notion of elaboration nevertheless differs profoundly from the axiomatic representation of the algorithm widely shared among computer vision scientists. If, in practice, computer scientists recognise the agency of data and algorithms as distributed, the axiomatic perspective continues to have currency in the profession. It continues to inform how computer vision science is being taught and to provide a rationale for the hierarchies in the profession. If, as the crowdsourcing platform CrowdFlower reports, data scientists spend the majority of their time cleansing data, they consider it the least favourite part of their job (Venkatesan, 2018) and a necessary burden taking time away from their real work.

To construe computer vision as a process of elaboration clearly contrasts with an understanding of the discipline centred on the algorithm as defined by Donald Knuth (1997, p. 4) as a “finite set of rules that gives a sequence of operations for solving a specific type of problem”. Construed as a finite set of rules, the algorithm can only be elucidated through the mastering of mathematical formulas. In this view, the algorithm is considered as independent from the language it is written in and the importance of its network of operation is dismissed as mere “implementation details” (Goffey, 2008, p.15). As Les Goldschlager and Andrew Lister (in Goffey, 2008, p.15) observe, the algorithm “is the unifying concept for all the activities which computer scientists engage in.” Therefore the stakes are high when it comes to accepting these scientists’ representations of their own practice. The emic representation, the manner in which a field represents to itself and to others what its work consists of, doesn't offer a mere definition of its activity. This definition engages the field's actors. The actors make themselves accountable according to this definition. For instance, if a machine vision scientist says she is creating algorithms, she will assume that she designs a sequence of instructions to solve a given problem. She makes herself accountable for the formal coherence of her work, and perhaps for the elegance of the solution. If her work is formally inconsistent or includes redundant instructions, she will make herself accountable and consider it as her responsibility to remedy these shortcomings or errors. But as this definition excludes the implementation details, how the inputs are constructed, assembled and regularised and how the outputs are used or further processed, she will refuse to be held accountable for any errors or problems arising from these sources. For this reason, Paul Dourish (2016), in *Algorithms and its Others*, insists on the importance of respecting the distinctions made by engineers between a notion like algorithm and similar notions like procedure, or algorithmic systems. Any theory whose purpose is to effectuate concrete change needs to be formulated in the terms used by the engineers, affirms Dourish (2016, p.9). Being able to understand technology in emic terms is crucial for pragmatic reasons as much as with respect to epistemic concerns.

However, the emic definition hardly bears scrutiny. Drawing on his fieldwork in a music startup which offers algorithmically generated playlists to its users, the ethnographer Nick Seaver (2013) emphasizes that nothing secures the stability of the emic definition. Seaver shows that the meaning given by the programmers to the notion of algorithm keeps changing, and ends in different places. The contours of the notion are not set once and for all even for those who claim authority over the concept. The axiomatic definition is therefore mobilised when an agent defines her field for an outsider even if inside the field the definition remains in flux. Furthermore, there is a hierarchy embedded in the emic representation. The more abstract the work, the more value it gets in the profession. Recently created computer languages continue to implement age old algorithms. Ffmpeg, a piece of video-processing software, is named after the Fast-Fourier transform algorithm (Arar, 2017), which is itself named in reference to the French mathematician Joseph Fourier. A renowned line detector is called Harris after the name of its creator (Harris and Stephens, 1988). These examples concur with what Jarkko Hietaniemi and his colleagues note in their programming book *Algorithms with Perl*:

> What elevates some techniques to the hallowed status of algorithm is that they embody a general, reusable method that solves an entire class of problems. Programs are created; algorithms are invented. Programs eventually become obsolete; algorithms are permanent (Hietaniemi *et al.*, 1999, p.1).

Yet such a notion of the algorithm has been at the centre of disagreements among computer scientist themselves and the emic representation should not be taken as universally shared in the field. If the canonical essays and textbooks such as Donald Knuth's *The Art of Software Programming* (1997) or Robert Kowalski's *Algorithm = Logic + Control* (1979) rarely address the hierarchies that subtend the axiomatic perspective, the manuals and other more goal-oriented publications aimed at a less academic audience tend to be more vocal on the issue of the pecking order implied in the distinctions of various levels of abstraction. As I said earlier, an algorithm operates at another level of abstraction than a program. A program is always dependent on a computer language and languages are considered to have a shorter life span. This different temporal quality corresponds to a difference of status. The hierarchy that places the algorithm on top of other entities in computing is contested in the “secondary” literature, the how-to's, the programming books, the literature of hands-on informatics. The authors of *Algorithms with Perl* gently mock the difference of status granted to the algorithm designers:

> The perspective suggested in many algorithms textbooks and university courses is that an algorithm is like a magic incantation, a spell created by a wizardly sage and passed down through us humble chroniclers to you, the willing apprentice (Hietaniemi *et al.*, 1999, p.22).

This hierarchy operates in an academic context, and in practice, the relation between algorithm and the data-structures that bridge with the real-world data is given more importance:

> ... algorithms receive more credit than they deserve. … The most important problem-solving ability is the capacity to reformulate the problem – to choose an alternative representation that facilitates a solution. … Data structures – the representations for your data – don't have the status of algorithms. They aren't typically named after their inventors; the phrase “well-designed” is far more likely to precede “algorithm” than “data-structure”. Nevertheless they are just as important as the algorithms themselves ... (Hietaniemi *et al.*, 1999, pp.22-23).

Hietaniemi *et al*. emphasize that the capacity to reformulate the problem doesn't lie necessarily in the ordered sequences described in the algorithm. They are distributed over the data-structures and data structures contribute to the re-invention of the problem.

The lesson of ImageNet could have been to demonstrate this for the field of computer vision and machine learning in general. ImageNet could have contributed to a divergent understanding of the hierarchies in computing, it could have given fresh arguments to contest an established order of what is elevated and what is not. However it stopped short of doing that. Instead, if it demonstrated the importance of a scale in quantitative terms, it did so with the use of the AMT. And therefore it left the elaboration of the scale behind the curtain. It showed how the process of modelling was crucial to the evolution of computer vision but it did not result in the celebration of the complexity of the work the process involves. It did not elevate it to the level of mastery given to algorithm's creators. It demonstrated the importance of the elaboration of computer vision's scale by dissolving the modelling process into a cloud of micro-tasks performed by anonymous workers. The modelling process has never been so central: the software is said to observe and learn from data. Yet, it has lost its symbolic value as it has been automatised. The tour de force of ImageNet is not just to have changed the scale of computer vision but to have made it without disturbing (too much) the hierarchies in place. What is at stake with ImageNet is indeed a change of scale and of paradigm. It is at the same time an attempt to provoke this change of scale whilst maintaining the status of those entitled to claim ownership over the invention in the process.

With these remarks, I hope to make clear how much the emic representation must be taken with caution. As a member's definition, it is already challenged from the inside. This definition construes the field in epistemic terms, but the axiomatic definition is also in line with a distribution of roles. There are strong reasons for the profession to keep this view in order to maintain established hierarchies, even if in her practice no computer scientist can really ignore the network of operation her work is part of.

This is reminiscent of the debates surrounding the notion of code a decade ago. Reflecting the emic understanding of computer scientists, many authors in software studies understood code as a source, a cause, and the execution of the programme as its consequence. In a culture increasingly immersed in computation, source code was considered the ciphered key to understanding the dynamics of software and its impact on power and society. Coding literacies (Vee, 2017) were seen as the answer to the problem; “view source” was the slogan. In the words of David Berry (2011, p.34) code was absolute. In the same vein today, substituting code by algorithm, the solution would be an algorithmic literacy, a patient deciphering of mathematical formulas and grammars that would offer an understanding of the hidden mechanism that controls a technology. But the arguments that were opposed then remain valid now. In *Programmed Visions*, Wendy Hui Kyong Chun (2011) formulated an articulate rebuttal of those ideas. Source code, Hui Kyong Chun insists, is only a source after the fact, a “re-source” (Chun, 2011, p.24). Coding pertains to a larger horizon of programmability.[40] To focus on source code obscures the larger network of operation in which computational effects are in play. Code, then, acts as a fetish. In a register closer to STS, Adrian Mackenzie in *Cutting Code*, gives another version of the argument, through extensive case studies. “Every abstraction is relative to a concrete framing”, writes Mackenzie (2006, p.64), and code's importance can only be appraised when following the various chains through which it is made operational. In the same vein, algorithms have to be made down to earth, insists Ian Bogost (2015). Another form of algorithmic literacy is then called for, Geoff Cox (2017, p.9) argues, where “algorithms do not act alone or with magical (totalising) power but rather exist as part of larger infrastructures and ideologies”. De-centring the understanding of computer science from code is a pre-requisite to understand it outside of the limits assigned by epistemology.

3.2. From intermediaries to mediators

At this point, I have downplayed the importance of the axiomatic perspective that limits algorithms to a set of instructions. Where does this move lead me? What is the consequence of making the algorithm contingent and relative? Does it mean simply that the modelling, once a core issue of algorithmic design, has simply evaporated in the cloud and the annotation platforms? By making the algorithm relative don't we lose the possibility to attend to what is being invented in the modelling? Having deflated the magical power of a mathematical procedure, how should we attend to the specificity of what is being invented without flattening it to an accumulation of meaningless

[40] “[understanding software] means engaging its odd materializations and visualizations closely and refusing to reduce software to codes and algorithms — **readily readable objects** — by grappling with its simultaneous ambiguity and specificity” (Chun, 2011 p.11).

tasks performed by Turkers behaving like automata? Answering this question will require a rethinking of the nature of what is invented and the agents taking part in that invention. And for that I will need to move the focus from how problems are solved to how problems are created.

To make code relative, to expand the horizon of programmability, to make the algorithm down to earth are not meant to be reductive operations but to locate agency and generativity following longer detours. The journey into computer vision's elaboration must be read as an enquiry into how the problem of vision is created in order to be resolved rather than as a problem that comes already given and is subsequently solved by engineers. So, what does it mean that the invention of the problem of computer vision is performed by a series of networks and agents rather than solved by scientists?

Science and technology studies have contributed significantly to the understanding of science as a distributed process and such understanding has been intimately related to a revision of the concept of mediation. In *The invention of modern sciences*, the philosopher of science Isabelle Stengers (2000, p.98) comments on Bruno Latour's decision to abandon the term intermediary for the term mediator. The intermediary implies the pre-existence of two objects between which the intermediary mediates. As Stengers (2000, p.98) notes, such a conception carries with it “a problematic of purity, fidelity, or distortion in relation to something that is always already present”. With the mediator, on the contrary, the emphasis is given to the invention of what needs to be translated as “susceptible to translation” (Stengers, p.98).[41] At that level, mediation is ontologically active and goes beyond the translation of a mere programme of action. The nature of all entities involved in mediation is at stake. With the mediator there is an insistence on the fact that what is translated is not untouched by the translation. It is never a mere linking but it always involves modulation or transformation. To concentrate on the mediators means to give primacy to the processes wherein the work of invention, translation and stabilisation of the entities of science is performed. Performed in this sentence is the key term. Scientific facts and objects are not given, they are constantly made and remade through processes of mediation. Science is seen not as a project of observation of the world but of intervention. With the mediator, the emphasis is given to the inventive character of mediation.

With this performative framework, we can avoid treating science as a pure world of ideas, as epistemology would have it, and also avoid reducing it to a monolithic project of instrumental reason as with the hard denunciations of technoscience (Habermas, 1968). Furthermore, as

[41] The book's English translation is “capable of translation” for the French “susceptible de traduction”. Susceptibility refers more to a sensibility, an irritability, a certain restlessness. To invent something as “susceptible to translation” means to be able to find the ways in which a certain indeterminacy in an object or a process render it translatable.

generativity is the main concern of a performative framework, it firmly opposes reducing what is being invented to the regularities of all social practices. In the context of computer vision, understanding mediation this way contrasts strongly with a classical notion of an intermediary which would focus on the algorithm that solves the problem of vision and produces a machine “capable of seeing”. To consider the activity of mediation as the work of mediators leads us instead to emphasize how vision is invented as susceptible to translation. This requires a change in how we understand invention. To clarify the notion of invention I will use in the thesis, it is worth making the distinction between invention as resolution and invention as problem solving. Invention understood as resolution implies that a problem's invention involves a temporary stabilisation of heterogeneous entities. It presupposes that the world exists in a state of indeterminacy, that the world is diverging. To invent a problem is to temporarily resolve this indeterminacy. To invent is to act upon the world to enact a version of it. It is opposed to invention understood as solution where invention is seen as the implementation of an intellectual solution to an external problem. Invention as solution implies the separability of observer and the world. In a resolution, what counts as the problem is reinvented iteratively whereas in a solution, the problem, already given out there, awaits an answer. In a resolution, the source of invention cannot be tracked to an inventor but to a distributed array of actors and actants, to a specific form of stabilisation of indeterminate entities that gains singularity and stability.

To formulate this in less general terms, let's come back to the question that initiated this reflection of the concept of mediation. What difference does it make to attend to mediators rather than intermediaries? What does it make me do differently? The first consequence is to change the focus from the readily legible objects that tender themselves as privileged sources of explanation. It forces me to ask the question: what are the relevant mediators? Who and what is relevant to the invention of vision as susceptible to translation? The dataset for instance is one good candidate as it is part of a large set of processes and objects involved in the invention of the problem of vision. As a mediator it cannot be reduced to a mere implementation detail. The question of the relevant mediators confirms the importance of the mapping of the previous chapter and how crucial it is to move away from the emic narrative. However, the question of identifying the relevant mediators is not simply answered by changing one set of intermediaries for another. Mediators differ in kind from intermediaries. They have a different consistency, they are weaved into a different relational texture. The mediators are active in a world where the contours of the objects of vision and computation are less fixed and determinate. It requires thinking about ways to engage with such indeterminacy, to attend to it. Mapping is a necessary but preliminary exercise. The challenge awaiting me is to find a way to relate to the way these mediators contribute to the invention of the problem and how their indeterminacy is resolved.

Now that I have made a distinction between solution and resolution, intermediary and mediator, it is time to clarify the use of the term mediation. Mediation becomes a confusing term as it can name a configuration of pre-existing objects and subjects, the “intermediary” version, or a configuration of agents where the relation is primary, the “mediator” version. A constellation of concepts has been proposed over the years from various sources to insist on the primacy of relation and the processual character of mediation. Intra-action (Barad, 1996, p.128) becomes a replacement for interaction as interaction still suggests that entities pre-exist the relation whereas intra-action recognizes the ontological inseparability of entities. Radical mediation (Grusin, 2015), middling (Manning, 2016, p.47) or immediation (Massumi, 2019) are candidates to replace mediation. All insist on the generativity of the process. To clarify my terminology, in the following pages I will be using the term mediation to refer to the activity of mediators, insisting on the primacy of the relation over the entities and as a term that gives importance to the strength of inscriptions (that account for the reproduction of behaviours and the delegation of action) and the force of invention, the transformative character of translation, the active role of the mediation process[42].

3.3. Photography and mediation

At this point I have sketched the contours of a framework of mediation sensitive to the specific work of invention performed by the mediators of techno-science (as opposed to intermediaries). Mediators are engaged in a process of resolution rather than problem-solving. They are not involved in a purely mental enterprise but in a material project that mobilises a large cast of human and non-human mediators. I can ask now how photography fits in this cast of agents and apparatuses? To answer this question, I must align technology and photography along a shared mediation framework. With photography, we touch upon a specific aspect of mediation, the one that concerns the objects that are commonly called media. Whilst these concepts challenging mediation have been presented here in the fields of STS and radical empiricism, they are not only relevant to the domains of science and technology. Sarah Kember and Joanna Zylinska's (2012) work on photomediation contributes crucially to a rethinking of photography in the light of a performative framework. In *Life after New Media: Mediation as a Vital Process* (Kember and Zylinska, 2012), the authors use Barad's agential realism in conjunction with other philosophical lineages (among others Henri

[42] When I will refer to the more traditional sense of mediation (intermediary between fixed entities), I will warn the reader.

Bergson, Gilles Deleuze and Bernard Stiegler) to challenge the idea of media as a static entity mediating between pre-existing objects and subjects. Their relational approach to mediation intends to make room for technology without re-introducing the causal determinism that handicapped earlier attempts to characterise technology's role. Instead of seeing the photographic medium as an intermediary, Kember and Zylinska concern themselves with the mediation processes that make media objects and actors emerge. There is no exteriority between media users and their media, no given separation at the onset. There is a constitutive entanglement. Kember and Zylinska insist on the difference between media and mediation. What we commonly call media are temporary fixings. What matters is how photography becomes defined as a certain form of relation stabilised temporarily into objects. Working with the process of mediation rather than media objects requires an approach attentive to the cuts that enact objects rather than one that takes their contours at face value. Building on Karen Barad's critique of representationalism, the authors invite us to move away from a tripartite arrangement that suggests “that there are forms of representation (images and knowledge), objects of representation (things that are known), and representers (or knowers)” (Kember and Zylinska, p.103).

Kember and Zylinska's argument also contributes to understanding media change in a wider perspective. For the authors, mediation is a flow in which cuts are performed. Therefore media are never merely replaced, and updated. The analogue camera is not replaced by the digital camera. Thinking media in terms of discrete objects leads to what they call a replacement trap, the illusion that media can be replaced denies the fact that media changes affect the whole field of relationality. The authors insist we question the given state of media objects and always engage with them as provisional within a flow of mediation. There is a sense in which a media is never fully stabilized once and for all as it “carries within itself both the memory of mediation and the loss of mediations never to be actualized” (Kember and Zylinska, p.21). This interest in the flow of mediation is not intended to simply relativise all media forms or celebrate the dynamic essence of media but to increase our sense of responsibility towards the cuts that are performed in the media flow and differentiate between them. They insist on a crucial question that comes as a consequence of the overturning of the representationalist framework: how to evaluate, how to be responsible for the cuts that are operated? Kember and Zylinska warn the reader that the proliferation of divergence, multiplicity, is not an end in itself. Leaving representation behind forces us to think anew the question of the world that emerges from the processes we engage in. If science is generative and performative, if it enacts new worlds rather than new world views, what does it mean to cut well?

To bring these theoretical considerations closer to my immediate concerns, attending to the processes of mediation of photography would not mean, in this perspective, to study a scenario where a tri-partite arrangement, the photographer, the photograph and its content, are involved in providing representations to computer vision. Instead, it would invite one to study how the processes of mediation of photography partake in the generation of computer vision's world. It would invite inquiry into how the resolution of the inherent indeterminacy of photography's objects would contribute to the resolution of the inherent indeterminacy of computer vision's objects. In such a framework, mediation flows both ways and objects and subjects of both photography and computer vision find themselves temporarily resolved through specific alignments.

One such alignment can be sketched using the description of the dataset pipeline that was introduced in the previous chapter. What matters primarily are not the representations conveyed by the photographs but how photographs, as indeterminate entities, are resolved as data. A lot more happens in the translation of a selection of photos from a series of collections (Flickr, news agencies, Justice Department databases, blogs) to another (the dataset) than a spatial transfer. The speeds and synchronisations of retinas and thumbnails, the specific cuts enacted by various apparatuses (the search engine or AMT), their different temporal regimes, their changes of pace and the resulting alignments of disparate agents across different scales take centre stage in an approach sensitive to mediation processes. The production of the domain knowledge, I claimed, is obtained through unacknowledged collaboration. We now sense better how this collaboration happens intimately through the operation of photographic mediation. Turkers, adjusting their pace to the cadence of the platform, are enacted both as subjects and workers by looking at photographs. And photographs are stabilised as data through the workers' eye saccades and clicks. I have called the annotation process an elaboration of computer vision to stress the hidden epistemic contribution and the division of labour at play in the environments of annotation. The resolution of computer vision is intimately bound to large alignments in which the stabilisation of photographs invent vision as susceptible to translation. From here on, I will speak of a *photographic elaboration* to stress that the stabilisation of the photograph as data is what is at stake in the worker's epistemic contribution and the platform's division of labour. And to engage with it, what becomes crucial to identify is not so much the contents in the photographs as the apparatuses, rhythms, scales, automatisms, embodiments that it involves.

Now that I am reaching the end of the discussion of the notion of photographic elaboration and its relation to a thick concept of mediation, I have the necessary elements to clarify the meaning of the title of this research. Algorithms of vision here is understood with an implicit criticism. The algorithms of vision this thesis refers to are the distributed processes inventing vision as susceptible to translation for machines, not the inscrutable formulas determining the fate of visual content in society. And at the core of this process of invention we find the active question of the resolution of the photograph.

3.3. The problem with datasets, representation, bias and labour

At this stage, the discussion of a performative mediation framework, if it helps consolidate the analysis of the previous chapter, remains indicative and in the next chapters, I will come back to it to enrich it and question it. However even in its present stage, it can already be usefully mobilised to address some of the concrete issues weighing on the question of mediation in the particular context of computer vision and AI. To conclude this chapter, I would like to show how the issues I have discussed here theoretically are intimately related to urgent questions that are addressed to computer vision and AI and how the understanding of mediation has consequences for the course of action to be taken. Understanding objects, subjects, photographs, scales, work as pre-existing their mediation implies a certain critique and enables a certain kind of response. The concept of mediation subtends how the current tensions are being framed and how computer vision technology is concretely being challenged. This can be illustrated with the examples of critical approaches to contemporary datasets.

The fact that computer scientists are beginning to realise the various prejudices stemming from the classification operated by the datasets is the result of a combined effort of different strands of activism and a better understanding of the discriminatory practices of AI and computer vision by tech workers, journalists and the general public. An important part of the current criticism addressed to computer vision revolves around the notion of bias as a vector of discrimination based on race and gender. To examine this criticism I will concentrate on the arguments produced by the members of the nascent but influential community of researchers investing issues of fairness, accountability and transparency (named *Fairness community* from here on) in the emergent field of machine learning meeting in international gatherings such as the FaccT[43] conference sponsored by the Association for Computing Machinery. A central figure of this community, the programmer and artist, Joy Buolamwini, promotes a counter-strategy called incoding that stands for creating intentionally inclusive code (Buolamwini, 2016). Buolamwini defines incoding as “a mindset that asks: who is missing?” The dataset is understood as a world view (Davis, 2019), a curated set of images from which a machine learns the boundaries of its world. Datasets are increasingly

[43] See the archives of the conference at https://facctconference.org/

understood as key articulators in technological networks that produce social and racial discrimination. Other technologists (e.g.Dulhanty and Wong[44], Shankar *et al.*[45], Recht *et al.*[46]) and artists (e.g. Trevor Paglen[47], Adam Harvey[48]) have launched similar attacks on datasets' bias. Many of them have formulated their critiques in terms of mis/representation. Computer scientists are held accountable for producing tools enabling social sorting. The dataset is contested as a representation of the world that simultaneously includes and excludes; it is the result of a partial choice obfuscated behind a façade of technological objectivity.

Dataset makers are responding to the critique on the same terms. The vision community, as Li and others name a loose set of disciplines including optics, psychophysics, cognitive science, signal processing and computer vision (Borji, 2017; Lapedriza *et al.*, 2013), intends to answer this criticism by "pluralising" the datasets, curating the collections of photographs in a way that ensures gender or racial diversity. Reducing bias means harmonising, protecting image attributes, re-balancing, keeping safe categories and removing offensive ones (Yang *et al.*, 2013). The response takes the form of a process of dataset mainstreaming.

Before going further, I want to acknowledge the importance of such criticism and the relevance of an activism that concerns itself with the discriminatory effects of algorithmic control. There are many documented instances of shocking injustice where the stigmatisation of segments of a population are submitted to unequal treatment or brutal prejudice directly related to the algorithmic operations and the encoding of bias in datasets (Apprich *et al.*, 2018). My position here is not to disregard such activism, but to reflect on its reliance on a framework of mediation articulated around the notion of representation. I realise how much a critique of representation, of bias, is pragmatically useful. It helps to articulate a claim that is legible to many and can lead to transformations having valuable and important effects. For instance, several large companies like IBM have reviewed their datasets in response to Buolamwini's critiques (Hardesty, 2018) and Google hired her colleague Timnit Gebru as part of an effort to increase fairness and transparency in

[44] Dulhanty and Wong (2019) conducted a demographic audit of ImageNet which revealed that the representation of males aged 15 to 29 account significantly for the largest subgroup

[45] Shankar et al. (2017) examined the geo-diversity of datasets and underline the Amerocentrism and Eurocentrism in the selected representations.

[46] Recht et al. (2019) studied the ability of algorithms trained on ImageNet to generalize on other datasets.

[47] Various works by Paglen develop a critique of datasets in general and ImageNet in particular. For instance, his work *From 'Apple' to 'Anomaly'* exhibits a selection of photographs from ImageNet to expose the "different judgements against humankind" (Barbican, 2019) the dataset enables.

[48] The critique launched by Adam Harvey against the Microsoft's dataset Msceleb resulted in the removal of the dataset's website. The critique was directed against the fact that an important number of photos had been acquired without the knowledge of the people represented in them. It also questioned the model of the human face centred on the faces of Hollywood celebrities (Megapixels, n.d.).

their systems[49].

Yet the question of mediation as I approached it earlier raises several questions about a critique of bias understood as a critique of representation. Such a critique locating representation in discrete objects, here photographs, may very well end up in what Kember and Zylinska called the replacement trap. If entities are understood as independent from their mediation, it becomes possible that they can be replaced independently. With better images come fairer algorithms. A selection of non-controversial samples would reduce algorithmic discrimination.

To be more concrete, let's take the argument of Buolamwini and Gebru's paper, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*. (Buolamwini and Gebru, 2018) Herein, the authors present bias as a means to measure the fairness of an application. As they note, face detection systems perform poorly on women and even more poorly on female subjects with darker skin. This intersectional prejudice confirms and reinforces discrimination at work in society at large. What the authors propose is a benchmark dataset, the Pilot Parliaments Benchmark, to evaluate the fairness of an algorithm. For this purpose, they curate a collection of photographs of 1,270 individuals taken from government websites of three African countries and three European countries selected for gender parity in the parliaments[50]. Predictably, major commercial face detectors performance degraded lamentably for photographs of female subjects with darker complexion. The article served as a wake-up call for many programmers of computer vision software. Its success was largely due to the fact that it repurposed age-old techniques used to evaluate other forms of errors in programming. In a gesture close to Li's, the dataset as a benchmark was utilized to define the criteria a new generation of machine learning applications would have to respect. Where Li utilized ImageNet to change the scale of the problem of image classification, Buolamwini and Gebru utilized the Pilot Parliament Benchmark to introduce a measure of fairness in the algorithmic process. Where Li changed the dimensions of a problem where bias was understood as a question of over or underfitting, the Gender Shades authors focused on the necessity to control attributes of data that have social relevance.

The authors were able to frame the problem of discrimination technically at the condition of relying heavily on a representationalist approach that bracketed the thickness of the photographic mediation lying at the core of the problem. They selected photographs of faces representing named individuals because they were representing through an electoral process a larger population of voters. The

[49] This relation however was short-lived. I will come back to it in the last chapter.

[50] Respectively Rwanda, Senegal, South Africa and Iceland, Finland, Sweden.

politicians were treated as the facial prototype of a nation. The transparent mediation of photography hand in hand with democratic representation served as the rationale to select the reference images from which measurable attributes could be extracted to remedy algorithmic bias.

If framing the problem this way had the positive effect of raising awareness to an issue that was barely addressed in the field of computer vision, it also strongly influenced the current response given to the question of discrimination. Remedies take the form of attribute-protection in which image attributes such as gender, race or age are considered and either lead to a re-balancing of the dataset (include a more equal distribution of images with different values of the same attribute like a more balanced age distribution) or a removal from the dataset (images of children below a certain age or naked people are deleted) (Yang *et al.*, 2013).

Several problems arise here. The first problem is, as mentioned, the relation of the photograph to representation. Is there anything like a photograph with fair attributes? How does the photograph either gain or lose such properties? Can a photograph travel as a pre-existing object with set attributes? Does it exist independently from the alignments that stabilise it? The second difficulty is by narrowing the problem to the choice of discrete images with given attributes, a threshold of acceptability is set that doesn't take into account the whole chain to which the selection of these photographs pertain. Since it formulates its criticism using the frame of those who can effectuate change, does such a critique not ultimately enable what it intends to contest? And finally the labour that goes in the production and stabilisation of photographic alignments, the photographic elaboration of the dataset, is elided.

Another strand of activism struggles against the exploitative dimension of platform work (Miceli *et al.*, 2020). Here the dataset is not addressed as a provider of discriminatory representations but as the end-product of a chain of production where workers are hired in precarious conditions. As I already mentioned, the production of a dataset relies on the ability of micro-labour platforms to provide a work force available around the clock for a meagre salary. Here again, different groups of workers, activists, unionists, academics have managed to give visibility to the worker's struggles. Lily Irani with projects such as the Turkopticon (Irani and Silberman, 2013) or Jeffrey Bigham with TurkScanner (Saito *et al.*, 2019) are examples among many of the imaginative interventions made on platforms such as AMT to mitigate the asymmetry between requesters and Turkers. The struggle for fairer, as well as the fight for non-discriminatory datasets, is clearly to be acknowledged as an important contribution to a progressive politics and important gains have been obtained[51].

[51] An example is the raise of the remuneration per hit in academic projects (Salehi *et al.*, 2015). More on this in the

But here too, problems come with the response. The emphasis is on economic exploitation and bargaining to obtain fairer wages. While this is an important objective, there is an underlying assumption that the work performed by the Turkers is repetitive and that it doesn't involve any valuable epistemic contribution. There is no reason to contest the hardship of the work, but a lot remains to be done to uncover what this work consists of and what its value is. In a sense, such activism often lacks the tools to be sensitive to the different practices that are in play in the different tasks performed by the workers. There is a form of internalisation of the view that they simply “respond” to the request. They are seen as automata and their automatism is conflated with mechanical repetition. Can the tasks of looking at images at speed and describing them be reduced to work in general? Such activism does little to tell us about the workers' contribution and cognitive involvement in the AI industry. As these interventions concentrate on rebalancing wages and working conditions, they pay little attention to the epistemic contribution distributed across the workers, requesters and the apparatus.

In these two examples, we find an echo of the question raised by Dourish over the pragmatics of framing the problem in the terms of those who can effectuate the changes. To discuss racism and sexism, activists and technologists agree on a representationalist framework wherein the attributes-equipped photographic objects and the negotiation will bear on the selection of these attributes. To negotiate the price of a micro-task, requesters and Turkers must agree on their work being repetitive and the conditions strenuous. The frame of negotiation, the mediation between workers and employers or activists and scientists is congruent with the framework of mediation that considers the objects and subjects of negotiation as discrete and replaceable without alteration to the relation. Image attributes are banned or privileged, remuneration per hit increases (ever so slightly). In this sense, the advantage of using such a framework of mediation is to potentially help to effectuate concrete change. But this advantage turns into a strong limitation as its consequence is to validate the terms of the relation.

Through this argument, I am trying to achieve two things. One is to acknowledge and value the work of many people who realised early the deep problems coming with the introduction of machine learning techniques in industrial, military and commercial contexts. Without such activism, my research would simply not have started. The second is to situate the intervention the research aims to perform and how it relates to such activism. My research is situated in differential resonance towards a criticism of bias and a criticism of exploitation. My aim is not to bridge this gap, but to

last chapter.

create assonances and dissonances with these two approaches. It is to complicate and enrich their objects by providing a performative reading of their objects.

With respect to the criticism of bias, I agree that photographs play a crucial role in machine vision. But photography, bias, vision will not be here considered as given. Rather they will be considered as questions. And questions that can only be studied through an engagement with machine learning middling, rather than a mediation of given objects and subjects. This raises key questions about the objects that are at the centre of the controversy. How do they acquire properties through temporal fixings? How can photography's entanglement with computer vision be approached through a practice sensitive to stabilisations and alignments? How can concepts such as bias be re-appraised by taking into account the radical entwining of viewer, apparatus and visual data?

As for labour activism, I agree the process of machine learning is one of exploitation. But the nature of the work performed and the entanglement with the apparatus of production cannot be studied just as the reproduction of a script of an already given social relation, but as a site of invention and translation to which workers contribute decisively. This will require a careful consideration of what goes on under the name of automatism and repetition. Additionally, this raises the question of understanding where the work takes place. As we have already seen, a productive way to approach machine vision is through the detour of the annotation environment. We need to consider computer vision's location as one in flux and characterised by displacement. This will require me to continue following the detours and account for the spatial enfolding of the annotation environment with the computer vision lab where techniques of translation are experimented with.

The problem with a representation based framework is that it is ill-equipped to probe and sense where difference works within the object it intends to critique. By reducing mediation to a relation between a collection of given entities through which messages are encoded and decoded, it misses the fundamental indeterminacy of these objects and the enormous work of invention and stabilisation that is required to enact them. An approach centred on re-balancing workers' rights tends to treat the photographic elaboration as any other social practice. This research establishes a dialogue with agential realism and radical empiricism to meet the process of invention and stabilisation of computer vision objects and subjects. It is used as an attempt to meet computer vision at the point where the “fixing” has not taken hold and not locked the agents into position. It asks with Mackenzie and Stengers how could actively diverging worlds appear amid the “quasi-iterative pursuit of optimization and convergence” (Mackenzie, 2017, p.102) that characterizes AI and computer vision in particular. And with Kember and Zylinska (2012, p.71), it asks how attending to the potential of change in computer vision can be done with consideration for the cuts that will inevitably be made along the way.

3.4. Conclusion / transition

This chapter presented me with the difficult task of questioning the terms I used for the mapping of computer vision's terrain and reflecting on the direction they were taking me in. By doing so, my analysis has been gradually moving from one level of analysis to another. In this chapter, I revisited the notion of elaboration of computer vision and questioned the nature of the knowledge that was produced. This has led me to examine different versions of the concept of mediation and the distribution of agency they implied. Following a host of scholars who criticize a static framework of mediation for presupposing the existence of given subjects, objects and messages before their engagement in a relation, I have turned to an understanding of mediation as a middling that gives room for the generativity and performativity of the relation. I have described the elaboration of computer vision as the resolution of a problem, its invention through the material alignments needed to stabilise its actors and processes. To resolve a problem here differs from the emic representation of solving exterior problems. To understand algorithms as networks of operation rather than sequences of symbolic instructions shifts the focus from the computer scientist to a cast of actors materially inventing the problem.

This chapter has tried to find a distance from the understanding of a practice from within, to sense its ambiguity and the tensions inhabiting it. It has aimed to establish a cautious relation with the emic representations of the field and to move away from its readily legible objects (Chun, 2011), such as the algorithms presented as immaterial centres from which computation radiates. When we move from the heads of the scientists to the world of devices and agents that make up computer vision, something happens to our understanding of invention: a different understanding of the nature of what goes by the name of invention. To invent here is to invent something as susceptible to translation, a process that is ontologically active, performed by actors of different forms stabilised by apparatuses.

It is in this context that the role of photography is considered. To approach the objects commonly called media requires the same effort regarding the dynamics of mediation as I have given to the objects of science and technology. The objects of photography have the same degree of indeterminacy and cannot be taken for granted. Attending to the processes of mediation of photography does not mean studying a scenario where a tri-partite arrangement, the photographer, the photograph and its content, are involved in providing representations to computer vision. Instead, it is an invitation to study how the processes of mediation of photography partake in the generation of computer vision's world. The central question becomes the resolution of the inherent indeterminacy of photography's objects and how their resolution contributes to the resolution of the inherent indeterminacy of computer vision's objects. In such a framework, mediation is a thick middle that flows both ways and objects and subjects of both photography and computer vision find themselves temporarily resolved through specific alignments.

With this in mind, the dataset pipeline that introduced the previous chapter can be considered as an alignment rather than a collection of discrete elements. The dataset pipeline is more than the algorithm's supply chain. And its photographic dimension exceeds the sum of the discrete objects named photographs circulating inside it. What matters primarily are not representations conveyed by the photographs but how photographs, as indeterminate entities, are resolved as data. Crucial to this resolution is the notion of scale which cuts across time, space and bodies when it articulates eye saccades, cadence, micro-labour and thumbnails. With the annotation process, we have entered into the elaboration of computer vision that relies on the hidden epistemic contribution and the division of labour at play in the environments of annotation. This elaboration is photographic in the sense that the stabilisation of the photograph as data is what is at stake in the worker's epistemic contribution and the platform's division of labour.

The notion of photographic elaboration of computer vision exists in dialogue with existing critical approaches to computer vision. It differentiates itself from a criticism of bias by questioning the centrality of representation in the effort to make datasets fairer and insisting on the importance of the alignments through which photographs are stabilised as data. It differentiates itself from a criticism of exploitation by questioning what gets created, what is specific to the workers' contributions, how they are engaged in an epistemic production – the scale being here more than quantitative and exploitative, a perspectival optics that the workers contribute to resolving through articulating rhythms and levels.

These reflections have given theoretical coordinates and reasons for engagement. They open the question of the methods and means through which the research will be conducted. They emphasize the need to gain knowledge about the photographic elaboration of computer vision, and interrogate the entanglement of photography and computer vision. They show the need for a method that will help me devise a mode of engagement that gives prominence to the processes that stabilise and differentiate the respective objects of photography and machine vision, and that can interfere with their definition. What is to be investigated is the kind of environment that reflects how computer vision “interiorises” the viewing of photographs. And what the model of vision that travels across the human-machine network of computer vision and photography can tell us about their imbrication. What is at stake is the mutual definition of the scales, the subjects of vision and the photograph that makes vision and the photograph amenable to computer vision.

Chapter 4. Experimenting early vision

The previous chapter ended on a problem: how the question of photographic elaboration could be studied and which space reflects how computer vision establishes and intensifies its relation with photography? We found such a space when we moved from an understanding of computer vision in terms of code to the concrete platforms where the annotations are being made. This detour revealed several critical elements about computer vision's photographic entanglement. I have shown how the process of machine vision learning depends on the availability of annotated photographs. I have used the expression photographic elaboration to emphasize the importance of photographic alignments for the modelling. Photographic alignments require specific practices. I have described the working conditions in which these annotations were produced and insisted on the speed at which the annotators were required to deliver the labels. To resolve photographs as data requires a population of workers involved in a specific practice of viewing at speed. The whole articulation of scales that define the infrastructure of annotation implements a model of vision and attention defined by the glance more than the sustained gaze. The importance of the glance for the process of annotation raised the question on which chapter two ended. If an infrastructure such as ImageNet is mobilised to transfer human visual knowledge to computational processes, what makes scientists confident that the regularities of human perception can be translated to machine logic? And why is the glance called to play such a significant role? The question of the provenance of a model of vision that privileges the glance leads me, in this chapter, to a dimension of computer vision I only touched upon previously: its experimental dimension. The intimate relation between the glance and the resolution of the photograph as data did not appear out of thin air in the annotation environment. As I move from problem solving to problem creation, I need to extend my journey where the problem's invention intensifies and where vision is invented at scale and at speed through the composition of photographic alignments.

To investigate computer vision's experimental dimension and its ties to the environment of annotation, I will move to a space closely related to the annotation environment, the computer vision lab. The path I will follow through the present chapter starts, as I already hinted in the introduction, in Caltech where Fei Fei Li, director of the AI Stanford Lab and initiator of ImageNet, studied human vision. Before she became a renowned figure in AI and computer vision, Li conducted extensive research in the psychophysics of vision, a branch of psychology that studies the relation between the quantitative aspect of a stimulus and the probability of a particular judgement (Read, 2015). Like many computer vision scientists, Li is interested in a very specific aspect of human vision: the glance or the ability to make sense of a scene in a split second. This mode of perception enables a perceiving subject to take immediate decisions. The glance is the near reflex perception allowing, for example, a car driver to avoid, at the last moment, a sudden obstacle. The ambition of computer vision researchers is, according to Li, to replicate with algorithms the ability to instantaneously identify and summarily describe a visual stimulus (Fei Fei, 2012). This model, which understands vision as an event happening in a split second triggering an immediate interpretation, becomes the reference for scene detection programs. But this model of immediate perception is not only connected at a particular speed, it is also related to a specific type of photographic alignment and a particular model of description. To get a better understanding of this model and the alignments it is tied to, I will now turn to its experimental construction.

4.1. What do we perceive in a glance of a real-world scene?

In her teaching activities and public talks, Li uses the findings of an experiment she conducted in 2007 to describe the model of vision needed in computer vision. Documented in a paper titled *What do we perceive in a glance of a real-world scene?* (Fei-Fei *et al.*, 2007), the experiment focused on the process occurring during the first half-second of sensory perception. Her working hypothesis was that the human subject is passing from a global interpretation of shapes and light to an increasing identification of the details of a scene in the blink of an eye. Or, in other words, that the first half-second of vision is a busy sequence of events rather than a uniform block. On this basis, she and her colleague researchers established that, after having been exposed to as little as 60 milliseconds of stimulus, a subject was able to make a description that would contain the stimulus' main objects and figures[52]. As we will see, the results of this experiment provided Li with more than a fact about early vision but with a method to think the material relations between image, screen, subject and micro-time in order to calibrate the level of precision necessary to successfully identify the main components of a scene.

The protocol described in *What do we perceive ...* can be summarized the following way. Twenty-two students from the California Institute of Technology proficient in English are recruited. The subjects are naive about the purpose of the experiment. They are cued before an image for a very short time, from 27 to 500 milliseconds. They are given unlimited time to write the description of

[52] Here, the term basic level refers to the level of specificity at which a subject tends to categorize an element (Rosch, 1978). For example, a subject will describe an animal as a cat rather than a feline or a tabby cat.

what they just perceived. Following this procedure, 1,980 descriptions are produced. The research team constructs a standardized and hierarchical list of 105 attributes based on the descriptions. The scorers evaluate the descriptions using the list of attributes. For each response, the scorer checks which attribute is referred to and if the attribute is accurately used in this description. The results are analysed according to the perception times.

The experiment has been conducted in the framework of Li's doctoral thesis *Visual Recognition: Computational Models and Human Psychophysics* (Fei Fei, 2005). Her written submission consists of four articles and straddles two disciplinary areas: computer science and cognitive psychology. In her opening statement, she writes:

> To understand how humans see and to build machines to see are two important goals in science and engineering. The purpose of this thesis is two-fold. On the human vision side, we explore properties of natural scene recognition through psychophysics experiments. On the computer vision side, we propose two algorithms that learn and recognize objects and natural scenes. (Fei Fei, 2005, p.2)

This distribution is typical of a gap that computational vision has to negotiate. On the “human side”, one finds psychophysics, where computational techniques are tools to study quantitatively human perception. While on the computer vision side, machine learning and recognition derive from theories about human perception. The way in which the knowledge is transferred between these two disciplines remains tacit. Li does not try to reconcile theoretically the two sides. The thesis' conclusion does not attempt to create a dialogue between what she learned from each discipline. She does, however, constantly import and export from one field to another. As we will see, the movement from the “human vision side” to the “computer vision side” has multiple effects. But this movement is never explicitly articulated. Her research domain is described as two juxtaposed areas and their relation is treated as evidence. It is in her work, and especially in her experiments, that this relation can be apprehended.

4.2. The researchers

Li did not conduct the Caltech experiment alone. The affiliations of the other authors correspond to the thesis' disciplinary distribution:[53] Pietro Perona, scientist and director of the Computation and

[53] There is a third author. Asha Iyer is a PhD student in the division of Biological Sciences in Caltech. At that time, she is writing her dissertation *Planning Goal-Directed Actions: fMRI Correlates in Humans and Monkeys* (Iyer, 2008) where she studies the activity of fronto-parietal regions when humans and monkeys perform goal-directed actions. Unlike the two other co-writers, Iyer will not continue to work closely with Li and her research will never coincide with theirs again. For this reason, I will take more time to elaborate on the profiles of Perona and Koch

Neural Systems Ph.D. program and Christof Koch neuroscientist and professor in the division of Biological Sciences, both at Caltech. The experiment is a temporary site of convergence for two different lines of research[54], and sometimes heterogeneous lines of inheritance and matters of concerns. Perona and Koch carry with them urgencies, epistemic approaches, methods and historical contexts that help understand the experiment's design as well as different experiences with devices and techniques.

Common to the contributors is a strong investment in computational vision. Perona, Koch and Li have already collaborated and share many research interests. Perona and Koch have worked together on modelling vision to elaborate a biologically plausible[55] computational theory of attention (Harel *et al.*, 2006) on several occasions.[56] These collaborations essentially result in improvements of algorithmic techniques drawing from insights from biology and cognitive sciences. The results of their collaborative work are published mainly in computer science outlets. Prior to the experiment, Li had already joined the duo and published, in 2002, *Rapid natural scene categorization in the near absence of attention* (Fei Fei *et al.* 2002), the report of an experiment in psychophysics wherein they observe the mechanisms of distraction and their effects on visual perception.

The three of them bring different senses of urgency to the collaboration. With Perona and Li comes a sense of momentum concerning the perception of a scene, its summarization, or the conversion of a visual stimulus into semantics. They share, in Perona's words, an impetus to transform the “digital dark matter of the internet into information” (Perona, 2010). They understand the timeliness of a project to advance the understanding of human perception of photographs in the perspective of large-scale annotation projects. Perona was engaged early on in the production of visual datasets. Just after the experiment, Perona and Li will concentrate on their own image collection projects: Perona with Visipedia, a hybrid human-machine extension to improve the treatment of visuals in Wikipedia (Perona, 2010, p.1526) and ImageNet for Li. The concomitant development of computer vision and the search engine interface for visual content has created a demand for “a repository of human understanding of visual imagery” (Perona, 2010, p.1526).

who in different ways will contribute to Li's development and integrate her work into their research.

[54] In this, the experiment follows a tradition that interweaves cognitive psychology and computer science and that can be exemplified by the work of David Marr summarised in his book *Vision* (1982).

[55] The mutual validation of models between biology and AI is a recurrent thread in machine learning theory. See recently a discussion of backpropagation in a biologically plausible deep learning framework by Illing et al (2019)

[56] For instance in *Selective visual attention enables learning and recognition of multiple objects in cluttered scenes* (Walther *et al.*, 2005), or *Graph-Based Visual Saliency* (Harel *et al.*, 2007)

For Koch the sense of urgency is the development of knowledge about early vision as the site where the (dis)entanglement of consciousness and awareness can be observed (Tsuchiya and Koch, 2005). The central theme of Koch's research is the neural correlate of consciousness, the physical state of the brain that corresponds to a certain subjective condition (Koch, 2004). As consciousness is a vast domain of investigation, Koch has strategically chosen an area that is circumscribed enough to be tractable and exemplary enough to account for one of consciousness' most remarkable traits: the phenomenon of early vision. According to Koch, vision is an easily manipulable sense and therefore lends itself to the artificiality of the experimental method. Therefore, to study consciousness through vision means to manipulate experimentally the visual stimuli and to analyse how visual information makes its way through consciousness. The glimpse plays a central role in Koch's work as the study of this phenomenon makes it possible to untangle the tight relationship of consciousness, awareness and attention (Koch and Tsuchiya, 2006). For the neuroscientist, the relations between attention, consciousness and early vision only form a continuum in exceptional circumstances. The dissociation between these states is the norm. Therefore, his experimental work studies and exploits the discontinuities between attention and vision. By doing so, he sheds light on the selective role of early vision that actively filters the information that makes its way to consciousness.

The experiment, an interdisciplinary collaboration, is primarily framed as a contribution to psychology. The paper is submitted to the *Journal of Vision* and is filed under the categories Visual Psychophysics and Physiological Optics, Crowding and Spatial Vision. There are rare mentions of computer vision or engineering in the paper. Again, the relation between the two branches is kept implicit. It is therefore Koch's strand that prevails in the context where the paper is published. But references and urgencies coming from computer vision traverse the experiment from end to end.

4.3. The experiment's after-lives

The Caltech experiment has become slowly but steadily a reference in domains as varied as the study of visual search, crowd analysis, attention, selection, change detection, autism, AI techniques, aesthetic perception, and perception of narrative through images. The experiment in these publications is seen essentially as adding to the literature on early vision and supporting the fact that early vision is a key element in how human subjects make sense of the world. In two cases, remarkably, researchers went further and replicated the experiment. In the domain of experimental aesthetics, in *Style follows content: On the microgenesis of art perception*, Augustin *et al.* (2008)

use Li's technique to explore the hypothesis that viewers are able to identify the content of an artwork before its style. A decade later, Jahanian *et al.* (2018) wrote *Web pages: What can you see in a single fixation?* where they document a variation on the experiment in which they port Li's technique to the field of usability testing. Instead of showing photographs, subjects are shown web pages with the objective of obtaining quantitative observations to improve web interface design. Li herself recently revisited the experiment when she accepted a chief scientist position at Google Cloud and launched a crowdfunding service that proposes image annotation among other tasks. In *Embracing error to enable rapid crowdsourcing* (Krishna *et al.*, 2016), an experiment conducted with Amazon Mechanical Turk workers, she tested various techniques to augment the productivity of the annotators by accelerating the display of the images they have to annotate. In *Embracing error* ..., which will be unpacked more fully in the last chapter, Li connects explicitly the method and the psychophysical findings from *What do we perceive* ... in the context of industrial annotation for machine learning.

Furthermore, the experiment has become a pedagogical tool. In various videos posted on social media platforms, Li presents her research and explains how the experiment plays a role as an introduction to computer vision's goals. In a lecture given at the eDay of Stanford's department of engineering, she explains that her students have to take part in a re-enactment of the Caltech experiment as an introduction to computer vision (Stanford University School of Engineering, 2014). Being a subject of the experiment is part of the induction to her teaching programme. The experiment is used as a pedagogical device to show what in Li's words computer vision “is after”. This "cool experiment" helps her define the cognitive task needed for image classification to her students and to a larger audience. The fact that she has conducted the experiment for fifteen years shows its symbolic importance in the academic context, but it is also a useful instrument for the various recorded presentations circulating on YouTube and other sharing platforms. Additionally, mention of the experiment is found scattered among her slides and conferences (Fei-Fei, 2010), where they help ground computer vision in human vision. Strengthening the connection between AI and human intelligence helps her consolidate the image of what she terms a human-centered AI (Simonite, 2019). The experiment is shown in her various presentations as a proof of the continuity between the human race and its technology. Lastly, her view of a “good for all AI” has become the brand image of the new services she helped to launch during the two years during which she led the AI research and development at Google Cloud (Fei Fei, 2018).

4.4. Close reading of the protocol

Having established the surrounding context, I will examine how the production of the model of vision is narrated by Li and colleagues in the *What do we perceive* ... article. For this purpose, I will engage in a close reading of the paper. The reading is an attempt to collect elements in order to answer several interlocking questions: how do the scientists frame their experiment, what is the experiment doing, and what does it tell us about the model of vision it produces?

The scientific literature in general, and experimental reports in particular, present a series of challenges for their readers. Essentially, a paper documenting an experiment is a means for the researchers to create an account of their work. The paper is not an integral and transparent catalogue of all the details of the experimental journey. It is an attempt to make the research intelligible to the researchers themselves and a community of colleagues. This entails a series of choices regarding what needs to be explained, justified, what can be left implicit or simply overlooked and the objections the authors anticipate (Lynch *et al.*, in Knorr-Cetina and Mulkay, 1983, pp. 207-208). Making an account is additionally a means for the authors to make themselves accountable and to establish the parameters of the process of accountability. Here again the emic representations must be questioned as they trace a demarcation line between what engages the scientists and what they feel free to ignore. Measurements, vocabularies and references are mobilised to demonstrate the authors' competences and at the same time to define the competences others must demonstrate to be able to question the value of their research. The chapter will therefore have to navigate different levels of meaning in the report and engage with the methods the authors use in the reporting. Through such a reading, the report becomes an object of inquiry (a theme) as much as a resource. This close reading, however, is not intended to illuminate the researchers' methods as an end in itself. A careful reading of the scientists’ words will come together with an analysis sensitive to the performative dimension of the experimental practice. The answer to the question related to the framing of the account (what are the scientists saying?) is only meaningful if it helps answer another: what is the experiment doing? Through the analysis of the account, I will look for the entities that are produced through the experiment. Key to this second level of reading is the notion of apparatus for which I am indebted to the philosopher of science Karen Barad. Barad builds on Michel Foucault's approach to the distribution of agency and power. Foucault sees the regulation of behaviour as the result of an ensemble of elements such as discourses, institutions and architectures. The apparatus according to Foucault is the specific set of relations that can be established between them (Deleuze, 1988). Keeping from Foucault the inherent relationality of the concept, she insists on the fact that the apparatus is materially active. For Barad, the material agency of the apparatus has to be researched as much as its diagram of relations. Further, Barad reads Foucault's concept of apparatus through the writing of the physicist Niels Bohr. The experimental apparatus is an instrument that makes measurements legible, but that also binds together the observer, the phenomenon and the knowledge produced (Barad, 1996). In this perspective, the apparatus is performative: entities do not pre-exist the apparatus, they are enacted by it. Barad warns us against taking entities like time, space, the individual etc. for granted as they only emerge in relation with an apparatus that cuts them out and differentiates them. And that they help constitute in return.

My reading will therefore follow two intertwined trajectories in the production of the model of vision: the methods mobilised to make the experiment account-able and the delineation of the apparatus and the entities emerging from the experimental frame. While exploring the apparatus, I will pay a particular attention to the importance given to the stabilisation of photography it enables and its role in enacting seeing subjects.

4.4.1. Micro-time

“It is known that humans can understand a real-world scene quickly and accurately, saccading many times per second while scanning a complex scene” (Fei Fei *et al*., 2007). In the opening line of the *What do we perceive …* article, the authors refer to the ability of human subjects to make sense of a scene at a glance. To study it, they need to construct an experimental frame that can measure and report the behaviour of their subjects. This requires an apparatus that produces a series of micro-temporal stimuli that can be measurably correlated to the reactions of a subject. In the experiment, the subject is seated in front of a screen where images are displayed for an infinitesimal fraction of second. The distance between the subject and the screen, 100 cm, is constant while the subject is exposed to a stimulus at varying intervals. The apparatus isolates a micro-temporal interval in the immobile subject's continuous perception and captures it. Testing the subject's perception in a time frame ranging from 27 to 500 ms is not new. The authors position themselves in an established tradition. The 27 milliseconds presentation time (PT) is considered in cognitive psychology as the moment where consciousness enters the process of visual perception. It represents the delay needed for the visual signal to travel from the retina to the brain. In his article *Is Vision continuous with cognition?*, where he compares the different approaches to visual cognition, the cognitive scientist Zenon Pylyshin notes:

> The early vision system is encapsulated from cognition, or to use the terms we prefer, it is cognitively impenetrable. Since vision as a whole is cognitively penetrable, this leaves open the question of where the cognitive penetration occurs (Pylyshyn, 1999).

To answer experimentally this question, the scientists need an apparatus able to capture a temporal sample that the human sensory system is unable to isolate on its own. Access to this particular time of perception, as Koch (2004) forcefully asserts, can only be made possible through the use of machines: it is only because a machine is able to show an image for a display time of 27 to 500 milliseconds that a world of differences between what is perceived inside this range of time becomes measurable.

Cybernetics and neuropsychology have long considered the eye as an organ whose function is not limited to the mere transmission of the stimulus it receives from the outside world. It already interprets, it already “speaks to the brain in a structured manner” (Lettvin *et al.*, 1959).[57] But the micro-temporal cut operated by the apparatus calls into question the anchoring of the eye in the viewer's body. The eye is made to act as an organ in the experimental configuration, but the entity to which the organ belongs remains uncertain. Does the organ align with the subject or the experimental device? The eye lobe remains an integral part of the physical body of the subject.[58] But wired to the circuitry of the experiment, it becomes unclear if it is to be conceived as an extension of the stimulus device or if it belongs to the subject exposed to the stimulus. The eye is the locus of conflicting dividing lines. Without the ability of machines to produce the stimuli corresponding to infinitesimal intervals, the differences of perception occurring during the intervals would remain ignored. But the existence of such differences, as they depend so intimately on the technical apparatus, raises the question of their construction. Are they artefacts constructed by the machine? Or do they indicate a machinic dimension of human perception embedded in the organism and revealed by a technical apparatus? Whilst the experimental device redraws the borders of its subjects, the authors maintain a strict distinction between the subjects and the apparatus. For them, humans possess a clandestine form of perception that has been detected by machines. The apparatus of observation is for them safely separated from the object of its observation.

The challenge for the experimenters is not only to create a device that isolates perception during a specific time frame. They also need to figure out how to capture the subject's attention. It is not enough to hit the subject's retina with a signal. The experimenters want the subject to describe the

[57] For a discussion of the role of early vision in cybernetics, see Orit Halpern (2014).

[58] In early cybernetics experiments, like Lettvin et al (1959), the animal's optic nerves were physically wired to the controlling apparatus.

stimulus. The subject needs to perceive this ephemeral image as a self-contained unit of visual perception. The device needs to cut through time. It also needs to cut out attention time. It needs to edit perception. To understand how the authors have dealt with this problem, we have to go back to where we left the description of the protocol. The subject is immobile and seated in front of a screen. The control exerted on his body must be extended to his eyes. The administration of the stimulus is controlled by a computer. Just before the image is displayed, a uniformly grey rectangle holding a cross at its centre is drawn on the screen. This image is called the fixation. The subject is invited to focus on the marker to avoid any time-consuming eye movement that would distract her from the task at hand. The fixation disappears and is instantaneously replaced by a fugitive image, the stimulus. The stimulus is in turn replaced by an image made of random noise to cancel residual memory. For each flash, the device generates a sequence of three images, fixation/stimulus/random noise: a montage. And the montage must, in turn, elicit the memory of a fixed image in the brain of the subject.

The authors present us with a subject whose attention lends itself to the sequencing of computer time, and whose eye saccades adjust themselves to the frequency of the stimulus. As the span of attention gradually synchronises with the micro-time segmentation, the eyes of the subject become the visual extension of the machine while the machine's sequencer becomes the subject's clock. And the researchers can record with precision the outputs of such a human-machine assemblage.

4.4.2. The photograph as stimulus

After having analysed the micro-temporal construction of the stimulus, I am turning now to its visual form. Which images are used to make up the stimulus? Where do they come from and more importantly what is the decision mechanism used to select them? To answer these questions, I will concentrate on a second series of cuts, operations of selection and filtering that define a series of agents and their roles in the process of the acquisition of images.

The authors remark that, in experimental psychology, subjects were generally shown images where non-relevant information was discarded (Fei-Fei *et al.*, 2007) such as line drawings, images shot by professional photographers to ensure a “clean”[59] descriptive image. These options are rejected by Li and her colleagues. They are deemed too different from the everyday visual experience of the subjects. They want to show “real-world scenes”. They want to avoid the “sampling bias” of the

[59] All the scare quotes in this paragraph indicate expressions used by the researchers in the report.

professional photographer, or the abstract character of a line drawing. What does a picture of a real-world scene look like? It is a picture like the ones “people take themselves.” Li and colleagues are looking for pictures representing the world as it is “commonly seen”, by “most people.” And the source for such pictures is the internet. This sequence of assumptions requires our attention. It represents a break from the tradition of psychology that may be overlooked today as the internet has become such a popular provider of visual content. In the paper, the authors note the limitations of the image databases used by their predecessors where they find only a narrow choice of scenes or objects shot by a professional photographer. In the previous section, I emphasized the temporal design of the experiment, bearing the mark of the experimental culture of neuropsychology. Here Li and Perona's collaboration on the elaboration of datasets is surfacing. It is not the first time they show a preference for the internet over professional photographers' pictures. And they have justified their decision elsewhere with similar arguments. For instance, Perona and colleagues wrote, concerning the Caltech-256 dataset[60], assembled from photos found on the internet, that its quality lies in the fact that it represents “a diverse set of lighting conditions, poses, backgrounds, image sizes and camera systematics” (Griffin *et al.*, 2007). By choosing images exhibiting such a diversity, the researchers hope to avoid the introduction of bias. But is the professional photographer the only biased provider of images? The researchers assert that every photographer “has a bias”. The role of the search engine is not to eliminate the bias, but to mitigate it. The search engine's role is, as they write, to “average the bias” – which supposes that bias is a dimension of the photograph that lends itself to averaging, through the search engine's algorithmic ranking.

The process of image selection is worth examining in detail as it explains how bias is reformulated through the procedure. The process of selection starts by asking “naïve subjects” to give five scene names. The researchers look for overlapping or converging terms. The researchers describe this operation as an “averaging” of the scene names. These terms are entered as search keywords in the Google Image search engine. The images are then randomly selected from the list of results.
The experimenters do not select the search terms themselves. They delegate this decision to a group of subjects. The terms elicited by the subjects are then submitted to various algorithmic treatments. First, the search terms are submitted to Google Images which internally processes the query using the PageRank algorithm.[61] Next, the query results are randomized by the researchers. And finally, the ten first results of the randomised selection are chosen. This procedure is deeply technologically layered. The PageRank algorithm which orders the results of the search query is by no means

[60] The Caltech-256 Object Category dataset is a collection of 30,607 images culled from Google images and organised in 256 categories by Perona with Gregory Griffin and Alex Holub in 2007.

[61] The researchers only mention the Google Image search as a provider for the photographs.

neutral.[62] It mediates with the many photographers who upload their photos on the internet. It brings other human subjects in while hiding them. It measures the value of their images according to a computation of their connectivity: the more linked an image is, the higher is the image's rank. The output is therefore a grade of sociometrics.

To expect an ordered list of images out of a textual query means to rely of a long chain of translation, metrics and orderings. It also assumes the possibility of a continuity between agents of different natures. The selection process can be understood as a cascade of delegations. First to human subjects (for the selection of search terms) then to techno-mathematical procedures. The subjects propose, the algorithms select, decide and re-order. The researchers design the procedures and do the translations. Viewed globally, the entire protocol can be understood as an algorithm performed by humans and machines. The decision is distributed across subjects and scripts and each of the actors performs an operation where an input is connected to an output. The protocol defines a sequence of discrete steps that produce a finite list of images.

As I have indicated, this intricate procedure is set up in order to mitigate bias, or, in the authors' terms, to average it. Does the procedure guarantee less bias? Or does it introduce more biases? The more agents recruited by the protocol the more distributed the bias. Does PageRank average the image's bias? Is the dynamic of the algorithmic actor only one of mitigation? Couldn't it be one of exacerbation? Or multiple ones that display antinomic behaviour?

My insistence on the question of bias here is not motivated by a desire to make a denunciation of a hidden bias in order to expose a manipulation of the researchers or their carelessness. The indeterminacy of the definition of bias here reveals the seams of the experiment's fabric, the cuts and the stitches. They indicate where the disjoined had to be sutured, patched. They point towards potential zones of variability. They also highlight the assemblage of methods combined in the acquisition process. The researchers outsource the production of the stimulus images to internet users. Whilst they produce an argument against the bias of earlier methods used in the production of stimulus, the positive argument for the acquisition through search engine is barely sketched. Instead they refer to the evidence of a practice of digital photography mediated by the search engine. A certain mediation of photography informs the method whilst remaining outside of the range of decisions the researchers consider themselves accountable for.

[62] For a detailed discussion of the PageRank algorithm, see Pasquinelli (2009).

4.4.3. Perceptual performance

Let's recapitulate what we know about the protocol so far. The subjects see a stimulus for a brief instant and are asked to type a description. For each stimulus this operation is repeated, leading to the collection of hundreds of written statements. Once the descriptions have been collected, the experimenters face the problem of comparing them. Here again, the authors argue with their predecessors. As they did for the choice of photographs, they want to avoid an artifice used in the experimental tradition. The authors refuse to give the subjects a controlled vocabulary as it is customary in this kind of experiment. Faced with a large amount of descriptions using different wordings, they nevertheless need a method to regularize the textual material. To solve this problem, they proceed by taking two steps. The first step is to extract a taxonomy out of the free descriptions. A closed list of terms is obtained by summarizing the mass of descriptions. The taxonomy is conceived as a representation of the corpus, selecting its recurring key concepts and grouping synonymous terms. The second step makes the opposite move. It takes the taxonomy to look back at the free descriptions. The taxonomy is used to evaluate the participants' statements. In the process, the taxonomy becomes an object in itself. It has severed its links to the descriptions and is used to filter them (Bowker and Starr, 2000). The structured vocabulary has become the basis to evaluate the material it is made of. With the taxonomy functioning as a controlling reference, the descriptions are folded upon themselves.

The synthesis operated by the taxonomy on the descriptions has a deep effect on what is kept and discarded. Many terms from the descriptions have not been included in the taxonomy and many others have been replaced by more generic terms. So when the folding is completed, the descriptions are given a different relief. Some parts fade away, others are homogenized, and the specificity of certain terms emerge. And some registers are simply discarded: “cognitive functionalities such as emotions” are simply left out (Fei Fei *et al.*, 2007). The decisions taken here are not explained. The lack of details and justification for this procedure is in itself important. It is crucial to take note of this silence. How to read it? Is it about hiding the motives that have led to the exclusion of categories such as emotion? Or is it more importantly about the fact that such a decision doesn't need an explanation – that discarding an entire category of statements is a decision that is left to the discretion of the researchers? This silence about the decision marks a boundary, another cut. It delineates a border. Again it marks the end of the decisions for which the researchers consider themselves accountable.

To read the descriptions through the taxonomy is more than a convenient recipe to provide terms that can be compared to each other. It is an act of editing. It has also another function, it transforms the nature of the textual entities. And it does so in several steps. To explain this, I will need to pay extra attention to how the descriptions are named by the authors. Reading the Caltech report more closely, one notices that different words are used to mention the subjects' descriptions. My attention goes particularly to two of them: response and perception. After having seen for a brief instant an image on the screen, the subjects are asked to write down what they have observed. This is what the researchers call the response. Later in the paper, they use the term perception to describe the content of the responses. The taxonomy is the key instrument that allows the researchers to move from response to perception. Response in experimental psychology is defined as a direct reaction to a stimulus (Kantor, 1933). In the paper, the term perception only appears to describe the subject's experience when the term response gradually disappears. During the long paragraphs describing the treatment of the responses, the terms of the descriptions are selected and matched with the closed vocabulary. It is only when the responses have been filtered and normalized that they are discussed as perceptions. It is only when all the words that “do not count” (the noise, the style) have been removed, when the specific words used by the subjects have been replaced by those in the list, that perception can be measured and discussed. The taxonomy is instrumental in this conversion from response to perception. A perception for the experimenters is what can be correlated term to term to the taxonomy.

The experiment seemed at first to be privileging the visual and to be adopting a liberal attitude towards semantics. The subjects are invited to look at images and they are free to use their own style of description. But throughout its evolution, from the moment the images are selected to the moment the responses are processed, the experiment follows a trajectory that begins with the search terms processed by the Google search engine to end up with the taxonomic entities. Behind the selected stimuli are the sociometrics and the semantics of the search engine. Behind the perceptions is the closed list of the taxonomy. What begins with a series of scene names queried in a search engine ends with a structured vocabulary through a long chain of translations. The search terms become a list of results. Results bring images. Images become responses. Responses become descriptions. Descriptions are filtered and produce a hierarchical list of attributes. The descriptions are filtered again, this time against the list of attributes, producing what the researchers finally call perceptions. This chain is already long but it is not the end of the semantic journey. There is yet another link to be added to the semantic chain.

In the last section of the report, the trajectory from response to perception continues to develop. As already noted, subjects' responses have become perceptions when all the traces of the subjects have left the text. They have become the perception of an abstract subject using a generic vocabulary. But something else is now happening to the perceptions. The perceptions that originated in the response of a subject become the property of a term. In the last part of the analysis, the researchers do not speak in terms of subjects' perception, but in terms of the perceptual performance of a taxonomy term. Building, chair, sand or rock are endowed with a perceptual performance. What the researchers encapsulate within this expression is a measure of how much, when a chair is depicted in an image, it is perceived sooner and with more accuracy than, say, a tree or a rock. The initial research topic of the paper is the subject's ability to grasp a scene at a glance. But in the course of the paper, the taxonomy becomes a protagonist in its own right and becomes the central entity to which perception is attached. There is a lot happening behind the use of the genitive construction (the perceptual performance *of* a term). This expression seals the black box that encapsulates the hesitant, algorithmically layered process of translation leading from response to perception. A grade of perceptual performance is assigned to a term, a value that synthesizes and obscures all the measurements and conversions of quantities that happened during the experiment.

4.4.4. Supervision and the layers of control.

During the experiment, images have been selected according to certain search terms and shown for a few milliseconds to the subjects who described them. A taxonomy has been built, and descriptions have been processed. Along the way, a concept of perception has been produced and the terms of the taxonomy, distilled from the descriptions, have acquired a “perceptual performance”. During the whole process, a series of devices have been mobilized. Inspired by Barad, I have followed the various moments where cuts are made by the experimental apparatus and how they transformed the entities at play in the experiment. I have said earlier that the subjects and objects were enacted by the design of the experiment. This enactment has more than an epistemic significance. It installs a division of labour and control between different classes of agents.

There are several groups of people involved in the making of the experiment whose roles are clearly separated and who are entitled to very different claims on the knowledge produced in the process. The researchers who conduct the experiments form the first group. The second is made up of the subjects. A third group is constituted of a team of students who assess the descriptions. The researchers are present at all stages of the experiment, the two other groups only occasionally. 22 subjects come to describe the images. Once the subjects have described the images, another team of students is hired and given the task of assessing the descriptions. The subjects are volunteers. The second team is paid. Whereas the subjects in the first group were mentioned by their initials, the second group is anonymous. They perform the tedious and repetitive task of going through all the responses systematically and for each response to check all the categories of the taxonomy. The procedure is demanding and error prone. Additionally, it leaves room for interpretation. A mechanism of control is used to prevent the reviewer's subjective interpretations from biasing the results. The management of the differences between the reviewers is achieved once again through averaging: it is only when a majority of reviewers make the same assessment of a description that the researchers take it into account. And while assessing the subjects' responses, the reviewer is himself submitted to a test. The experimenters have inserted 200 false additional descriptions interspersed with the real descriptions made by the subjects. These descriptions have been made up and are deliberately incorrect. They are used to evaluate the evaluator and check her tendency to “give the benefit of the doubt” (Fei Fei *et al.*, 2007) to a description that may have a vague resemblance to an image or simply evaluate the scorer's level of attention. The work of the evaluator is to assess the subjects' response, but she also contributes to her own evaluation and to the evaluation of her evaluations. The relation between the retina and machine time is taken in a series of control loops. Subjects are tested, and their descriptions assessed, and the scorers who assess them are themselves assessed.

The stated goal of the experiment is to study vision. To accomplish that goal, a system of supervision is created. We move from subjects viewing images at speed to workers evaluating the viewing subjects. Vision becomes an object to be evaluated. Vision in the experiment is the object under study and an object of control. The supervision mechanism is a multi-layered process that delegates vision and monitors it. It creates a layer that isolates the researchers conducting the experiment from the stimulus. In the terms of Peter Gallison and Lorraine Daston (2007), this is an epistemic virtue. The separation of the researcher and the stimulus must be enforced. This separation is materially enacted, it is not just an ideal. It is designed to ensure that the researchers do not work on the raw data, the responses, but on the perceptions, the responses filtered through the taxonomy. To control the scorers, the researchers do not check random samples of their work. They prefer to forge descriptions rather to expose themselves to the raw response.

Supervision is a complex process of delegation and labour management where each layer deals with a different stabilisation of vision and the photograph. Stimuli are administered to subjects, photographs and descriptions are shown side by side to scorers and the statistical summary of the data is displayed on the screen of the researchers. The researchers for their part craft the dataset that serves as the basis for the stimuli shown to the subjects and plant images to test the scorers. The experiment is as much about seeing subjects as evaluating vision and monitoring the evaluation. The scorers are treated as a mediation layer between subjects and researchers. They offer the possibility of separability between the experimenter and the phenomenon. They separate out the observed subjects and observing experimenters.

4.5. Conclusion / transition

Having read the experiment paper carefully, it is time to see how the accumulated observations help articulate the two registers that traverse the report: the emic representation and the performative dimension of the experimental device. As I have discussed in the previous chapter, there is a difference between what the internal definition makes the scientists do and a “performative” understanding of a discipline that sees agency as distributed. These two versions lead to different forms of accountability. One makes the scientists accountable for its objects. The other involves a more complicated approach to the cuts that are being performed and where entities emerge from relations.

Let's begin with the emic representation that is built through the report. By recording and providing a series of precise measurements, the scientists establish the ground on which they intend to respond to their peers. Vision is measured with millisecond precision, coefficients of perceptual performances are calculated. Another form of accountability is related to the position of the researchers. Whenever they can, they extricate themselves from the experimental process. They delegate the choice of images to others, as well as the evaluation. Whenever possible, they choose a distant position that guarantees a separation of the scientist and their objects and subjects. Supervision here is presented as an epistemic virtue. Finally, they insert themselves in a network of references and when they break with the habitual parameters of their experimental tradition, they rationalise their decisions. All these measures and precautions indicate how the researchers anticipate questions and how they define the knowledge that counts in the establishment of accountability. In doing so, they define the terms with respect to which they accept being held accountable. This entails a modulation of the volubilities and silences of the objects they are using and creating.

By contrast, a close reading has identified many other elements for which no accountability is planned. These are habits and forms of conventional knowledge that are assumed in the experiment.

One is the practice of photography mediated by the search engine, which seems to be adopted as evidence. Another is the position of those who create the account. Such a close reading has brought to light how visual perception was understood in the experiment. The reaction of the subject to a brief visual stimulus at the limit of her perceptual threshold is first conceived as a response. It is only through a series of operations that involve the evaluation of the descriptions through a taxonomic filtering that a response becomes a perception. But perception once established by the procedure is detached from the participants and becomes the property of a term in the taxonomy. The subjects have reacted, and the taxonomy's terms have grades of perceptual performance. This raises a double problem. The first is the elision of the subjects in the process as they are not entitled to perceptions but merely to responses. The concept of perceptual performance of a term “black-boxes” the subjects. The second is the distribution of perception as a process spread across all the agents in the experiment. This double elision raises the question of the emergence of the subject and how it can be accounted for. Is there a way to engage with the imbrication of the subject's perception and the apparatus without eliding the participants contribution? The researchers assume many competencies from the subjects. They are able to view images and coherently describe the stimulus. They are able to endure the experiment (it lasts several hours). Yet they don't have a say in the retrospective account. They are not made account-able. They are not part of the voluble commentary.

Reading the experiment closely confronts us with these two intertwined registers. One, the epistemic register, narrates the experiment populated with a series of measurable objects and separable entities. In the other, the material organisational register, we can detect the presence of tacit practices, recycled apparatuses and subjects whose contours are less certain. At this second level, we can begin to sense the materiality of the experiment and how the materiality involves modes of organisation, management and control from which any pretence to knowledge cannot be separated. This “double register” reading calls for a cautious response. If the emic representation can be traced, the question related to what emerges from the experiment, how the objects are cut, what the scientists do and what happens cannot be answered fully. Through the close reading, I have only been able to sense a different world of entities and agencies. To fully explore them will require another form of engagement.

Even if this reading is partial and incomplete, it already gives some elements that help address the problem on which the previous chapter ended: the model of vision that subtends computer vision. It helps address it them on two levels. On one level, (emic epistemic) the experiment can be read as a study of vision with the objective of improving computer models of vision. As the experiment is contextualised in Li's thesis, it sets the goals of what vision can do. It is part of a history of experiments of cognitive science where computational models of vision inform the study of psychological models of vision and vice versa. It sets the standards for what machines need be able to reproduce. And it enshrines these standards in a biologically plausible model.

On the other level, it helps answer the question as to why requesters are confident that seeing at speed can generate appropriate labels. If, as the researchers conclude, “a rich collection of perceptual attributes is represented and rises to conscious memory within a single fixation” (Fei Fei, 2007, p.22), there is no reason the production of annotation cannot meet the levels expected by the AI industry. We can see how an experiment made to understand vision as a model to design algorithms can provide simultaneously a rationale for how vision is mobilised in the annotation environment. The experiment does not merely provide the standards against which the algorithm's behaviour needs to be evaluated, it provides a rationale for modelling the viewer in the annotation environment, or to anticipate the next chapter, a social ontology that renders the viewing subject compatible with the scale of the annotation platform. And as such, it becomes relevant to the photographic elaboration of computer vision.

In both cases, what is crucial is that the experiment assimilates vision to the glimpse and measures perception at the millisecond. In the first case, the experiment adds experimental evidence to buttress the idea that computational vision must strive to emulate human vision as a mechanism of immediate response and not as a long deliberative process. In the second case, it offers useful knowledge in order to construe human viewers as instruments in the elaboration of computer vision. In this second case, it offers more than an intellectual solution. It offers a series of relations materialised in an experimental device. To understand how this material organisation relates the experiment and the annotation environment it is worth considering their architectures. The experiment and the annotation environment's architectures exhibit striking similarities at various levels. Let's begin with their photographic alignments. Following the procedure of selection of images and the cascade of decisions delegated to algorithms helped establish the hybrid nature the visual input of the experiment. The visual input is as much a search result as an image, a mixed construct of representation, socio-metrics and search engine semantics. Similarly, ImageNet is made by correlating search results to WordNet's master list. The photograph is a vehicle travelling between two semantic end points. The photograph is what connects the search to the taxonomy, it makes the taxonomy imageable.

Another point of resemblance is the work of scaling that subtends the process of photographic mediation followed in the experiment and the industrial process of image annotation. The adoption of the search engine to acquire the photos introduces a massive quantitative and qualitative change. The experiment diverges from the tradition of cognitive psychology by expanding the image repertoire using the Google Images search engine. And ImageNet is an attempt to change the scale of the problem of computer vision by extending the visual repertoire to the web. The method of image selection shows how the lab of psychophysics and the platform of annotation are informed by the ways of seeing of the users of networked media. They both integrate the networked image to produce the models of vision that subtend the algorithmic grip on the circulation and consumption of digital photographs. There is a thick layering of photographic practice that impregnates the experiment and the annotation platform. And the motivation behind the choice of a specific source of photographs is in each case expressed as remedies against bias. In the case of ImageNet, the search engine is presented as a device that gives access to a “treasure trove of images” enabling a change of scale that transforms the way computer scientists approach underfitting and overfitting. In the context of the experiment, it brings in a source of imagery that contrasts with the tradition of experimental psychology, an imagery closer to those “people take themselves” breaking away from the sample bias of the professional photographer. In both cases, the choice is justified as a remedy to bias and as a need to reflect the “real world” in terms of representation as well as in terms of variation.

Finally, the experimental device and the annotation platform both implement the same procedures of delegation and averaging of bias, leading in both cases to a succession of control loops. The platform and the experiment are structured around processes of supervision articulating the delegation of vision and the division of labour.

I am refraining for now from jumping to conclusions about the nature of the relation between what has been tested in the Caltech lab and what has been produced with ImageNet. I am limiting myself here to observing that the relation between them can be described using two registers. One is at the level of a model of vision, the other at the level of a mode of organisation and the development of a device. The first register is privileged in the report as it converges with an emic representation. The other provides promising elements with which to conceptualise the material and organisational relations between the experiment and the annotation environment. However, the report only provides some of the elements needed to grasp this relationship fully. To be in a position to explore these elements I will have first to come to terms with what is meant by experiment and how my research can contend with such a question. On what basis should the comparison be made? What do these similarities tell us? How can one engage with the experiment in this perspective? These are the questions that will occupy me in the next chapter.

Chapter 5. The re-experiment

This chapter operates a transition from the theoretical groundwork to the practice. Previously, I have presented computer vision as a process of photographic elaboration. To do this, I have mapped a large terrain and followed several detours. This led me to de-centre computer vision from the notion of code and explore its network of operation. The objects that circulate in these networks of operation and the devices that are crucial to them like the annotation platforms entertain a particular relation with the computer vision lab and psycho-physics experiments. In this chapter, I am taking stock of what has been analysed earlier to achieve two goals: to try to understand better the relation between an experiment and an environment of production, its outside, and to find the appropriate method to engage with an experiment given the complex character of this relation. The method I present in this chapter attempts to mobilise the space left open by the close reading between what the experimenters say about their practice, what they take for granted and the performativity of the experimental device.

What I am presenting here must be understood as a set of methodological precautions and anticipations, the development of a sensibility and attentions rather than a pre-specified set of rules to be subsequently applied. The methodological questioning aims to give weight to the theoretical groundwork whilst not letting it dictate the terms of the practice. The previous chapters have provoked a different understanding of computer vision, its objects, devices and scales. The methodological reflection in the following pages tries to find a starting point to see what new knowledge can be learned from them when such understanding of computer vision's objects gives rise to a practice.

The method relies on a central concept for my research: the re-experiment. For clarity, I will offer here a brief overview of the defining traits of the concept and in the subsequent sections I will develop its different aspects. While doing so, I will revisit the arguments underlying the discussion of mediation initiated in the third chapter and I will make explicit some of the analytical concepts used in the close reading of the experiment.

I use the term re-experiment to refer to a particular form of re-enactment of an existing experiment. A re-experiment is not a replication of an original experiment:[63] it introduces some variations into

[63] The software package Psiturk (Psiturk, n.d.) proposes a “friction free replication” frameworkimplemented on the Amazon Mechanical Turk platform: “Same population, same task code, direct replication, better science.”

the original protocol. In a scientific experiment, a specific set of elements can be changed. By observing the effects of these changes on the properties of given objects, knowledge is produced. In an experiment, change is understood as the variable relation between a set of parameters and a set of properties. The difference between an experiment and a re-experiment is the nature of the changes the latter aims to provoke. A re-experiment explores the effervescence of the original experiment: it tests how much it can amplify tendencies already present in a given experiment or actualise vectors that remained until then only virtually present. It does so by introducing variations, not by starting anew. The design of a re-experiment is informed by the close reading of the original experiment's protocol as detailed in the previous chapter. It requires the development of a sensitivity to the minute details of the practical implementation, a careful observation of the transformation of the different entities at work over the course of the experiment. The design of a re-experiment consists in choosing the elements that produce the variations that reveal its potential for divergence. The re-experiment is a strategy that leverages the inherent instability of the original experiment to find its vectors of self-differentiation.

5.1. Counter-laboratory and re-experiment

To clarify what I mean concretely, a good starting point is to explain what a re-experiment is not. To do this, it is worth explaining in more details how a re-experiment differs from the practice of replication in laboratories.

The Open Science Collaboration project (2015) recently attempted to replicate 100 experiments documented in major psychology journals. As the researchers produced divergent results for 64% of the experiments, Camerer *et al.* observed:

> The deepest trust in scientific knowledge comes from the ability to replicate empirical findings directly and independently. Although direct replication is widely applauded, it is rarely carried out in empirical social science (Camerer *et al.*, 2018).

A pillar of the scientific method, experimental replication is rare. There is little incentive to invest in replications because they are not valued as bringing up new knowledge and their “impact” is difficult to measure. The recent replication crisis (Pashler and Wagenmakers, 2012) brings to mind Thomas Kuhn's notion of “normal science” (Kuhn, 1996). For Kuhn, in the daily course of science, the results of an experiment are accepted by other researchers because they correspond to an established paradigm rather than because the experimental device has been thoroughly tested.

Replication is an exception. It is carried out when researchers consider they have something to gain that exceeds the cost of the replication. According to Bruno Latour (1987), controversies about the validity of an experiment turn laboratories into counter-laboratories. Replication is adversarial in nature.[64] Competing teams are re-doing an experiment to invalidate it. In Latour's view, the competition between rival laboratories is what makes science gain in precision. The more heated the controversy, the more technical the debate becomes. A controversy interrupts the course of normal science. A counter-laboratory is used to launch a polemic against the published results of another lab using the same epistemic framework.

A re-experiment is not a counter-laboratory. It doesn't claim the truth of a model against another. A re-experiment is not seeking the validation or invalidation of a hypothesis but the reformulation of its original questions. It is not a response to the experiment on a quantitative basis. It doesn't fight fire with fire, numbers with numbers. It answers with methodological questioning rather than facts and numbers.

A re-experiment relies on the assumption that an experiment can do more than what is asked from it in the context of evidence-based research that relies on a positivist ontology (Adams St. Pierre, 2013, p.223). It considers that an experiment's potential is not exhausted by the production of quantitative data and its interpretation. It relies on the idea that an experiment always produces an unaccountable excess. The scientists consider this supplement as a problem, a possible source of artefactual data that would interfere with the legibility of the measurements. Rather than repress what is construed as unwanted variables (the bodies, the machine noises, the reflexivity of the participants, etc.) in a typical scientific framework, a re-experiment seeks to address them creatively. The question is not whether an experiment produces verifiable measurements, but whether it realizes its full potential. The re-experimenter is concerned with generativity rather than validity.

The psychologist and philosopher of science Vincianne Despret provides a useful distinction that suggests what a reconsideration of the experimental practice could mean. Despret invites her readers to look at how ethologists have been re-thinking their relation to what is expected from their research, instead of looking at the dominant model of scientific knowledge exemplified by the discipline of physics. According to Despret, to judge the result of an experiment, ethologists

[64] In that they concord with a view of science as an adversarial project as can be found in Bachelard for example: “in fact we know *against* previous knowledge, by destroying badly constructed knowledge, by destroying what in the mind itself is an obstacle to spiritualisation” (Bachelard, 1980)

differentiate between a "foreseeable success"[65] and a "possible favourable outcome" (Jamar and Stengers, 2011). David Jamar and Isabelle Stengers (2011), commenting on the work done by Despret, write:

> Where the success limits itself to note that the animal will have had the expected behaviour when it responds to one or another stimulus, the favourable outcome brings into play the processual construction of the experimental device itself, not oriented towards a pure verification but towards the possible production of an interesting animal. This animal, interesting, active, will have made the researcher do things while he was following the behaviour and was adapting his experimental device.

The counter-laboratory understands the controversy in the perspective of a foreseeable success. It aims to demonstrate that the success claimed by a laboratory was in fact a failure. But it doesn't question the criteria for the evaluation of an experiment. The re-experiment, on the other hand, heads towards a favourable outcome that makes the researcher do things and re-invent an experimental device in a processual manner.

5.2. A device in translation

By differentiating the re-experiment from the counter-laboratory, I have focused on the relations both methods entertain with the original experiment. I will now turn to the relations of translation the re-experiment entertains with an environment of production. To explain this, I need first to construe the laboratory as a device in – as much as of – translation.

In the techno-scientific cycle of production, after a period of experimentation, the experimental object is transformed into a product or a service. The experimental device migrates with the model from the lab to the production environment. The conditions of validity established through the experiment need to be respected in the production environment. As the lab and the workplace are different environments with their respective specificities, what travels from one to the other is not so much a series of objects as a set of relations. And this movement is not uni-directional. For the lab to have any chance of producing knowledge about the world, it also needs to translate the "outside" world into its environment. As Latour (1983) points out, laboratories are devices designed to undo the dichotomy between the inside and the outside. What is relevant for the concept of re-experimentation is the fact that any translation implies a selective interpretation in both ways. The

[65] My translation from the French. In the original language, Despret contrasts "succès" with "réussite". Succès is etymologically "what comes after" and réussite "what comes out".

lab must give a form to a phenomenon to study it. The shaping of the phenomenon is done selectively, making certain relations count more than others. And in turn, the production environment has to invent ways to maintain the relations the lab made meaningful. Therefore, in a re-experiment, the relations that are produced and translated matter more than the historical reconstitution of an original context. A re-experiment is faithful to the relations. The prefix *re* in re-experiment refers as much to relation as to repetition.

To consider the environment as a device in translation means that what is sought through a re-experiment is not to remake the original experiment as it was, but to make a re-enactment that addresses the fact that the experiment was made to be re-invented. It translates a device already in translation. The referent of a re-experiment, to take the example discussed in the previous chapter, is not Caltech in 2007 but Caltech in relation with all the experimental devices in which it reverberates, or, for example, Caltech inasmuch as it reverberates into other devices of annotations.

5.3. The performative dimension of the re-experiment

To understand the nature of change expected to occur in a re-experiment, it is necessary to begin with its performative dimension. I have described the re-experiment as a device in translation and its object as the relations that travel across the different environments. Furthermore, I have contrasted the counter-laboratory opposing one lab to another to the re-experiment, a processual re-invention of the experimental device.

To discuss further the nature of this re-invention, I will make use of the concept of performativity as developed by Karen Barad. Barad's concept of performativity takes its source in an engagement with Judith Butler's thinking. For Butler, feminist and queer theorist, performativity is a conceptual instrument aimed at deconstructing the naturalisation of sexual difference. The body is not naturally gendered. Gender is in Butler's words a “discursive formation” (Butler, 2006/1990, cited in Geerts, 2016). But Butler's account of the construction of bodies goes beyond a mere process of signification. Sexual difference has to be performed. As Barad (1996, p.150) notes, Butler invites feminists to look at bodies not as surfaces but as processes of materialization. Barad at the same time acknowledges and radically extends Butler's concept. Performativity does not relate merely to the materialization of the gendered body but concerns matter itself, the materialization of all bodies, human and non-human. And stresses the importance of the material dynamics of regulatory practices.

Scientific experiments and apparatuses are instrumental to the performative enactment of bodies. In her book *Meeting the universe half-way*, Barad (1996) revisits the experiments of the physicist Niels Bohr and his interest in experimental devices. For Bohr, an experiment is a connected set of several elements: the observer, the equipment and the observed elements. For him, none of these elements can be studied separately. What can be known is the phenomena, or, in other words, the complete experiment. And the only way to know it is to redo it (Barad, 1996). An experiment is not an observation of the behaviour of external entities and the discovery of the rules that govern them. An experiment is ontologically active. An experiment is performative in the sense that indeterminate entities are enacted through it. In Bohr's conception, the observer cannot be detached from what she observes. She is an integral part of what is being studied. Partly, classical physics had already recognized the impact of the observer and the equipment on the result of an experiment. But it was understood that this impact could be quantified and subtracted from the results. Bohr's work ruins such a view. All elements are in a too intimate co-construction for a possible separation to be made. In Barad's words, the observer doesn't inter-act with the experimental environment, instead they intra-act (Barad, 1996, p. 33).[66]

To be more specific to the context of my research, to approach an experiment as performative does not mean to make a reconstitution of the original experiment, and subsequently reflect on it, to separate experiment and reflection. The observer is always too deeply entangled with the apparatus to have the distance to be able to reflect. To engage with an experiment's performative dimension brings about a knowledge coming from a deep entanglement that pays particular attention to the ways an apparatus performs specific cuts and creates boundaries. Performativity doesn't suppose a world already constituted of objects to be studied but an entangled world where apparatuses make cuts that can be undone, enacting subjects and objects whose boundaries are always performed. Performativity here also means that the enactment of subjects and objects is not following a deterministic logic. According to Barad, there is an opening inherent to the experiment, a constitutive effervescence, even if the scientific discourse tends to repress it. A re-experiment elects this performative dimension as its core principle.

5.4. Temporality

[66] Barad takes from Bohr the principle of inseparability of the observed from the agencies of observation. A principle Bohr demonstrated when he showed that an element, light, could adopt incompatible behaviours (wave or particle) depending on the experimental device and the agency of observation.

Acknowledging the performative dimension of the experimental practice means that the contours of its objects cannot be taken for granted, but must be discovered. In a re-experiment, the spatial separation of objects, subjects and apparatus is unsettled. Accordingly, the temporal ordering of the experiment needs scrutiny. I have previously emphasized how difficult it is to isolate the experiment temporally: the experiment continues in the environments of production and what happened in the lab is discovered anew. As I noted in the previous chapter, Li conducted the Caltech experiment in 2007 while creating the first sketch of ImageNet and conducted a variation of the experiment with workers annotating ImageNet's data nine years later.[67] As Barad suggests, we need to complicate the notion of anteriority and replace it by one of iterative becoming (Barad, p.234). What brings me to a re-experimental practice rather than a replication is that anteriority is never given. The precedence of the experiment over its implementation in production is not the only temporal relation that deserves our attention. The experiment itself is also very active in producing its own genealogy. In line with ethnomethodology, I would say that the means by which the experiment inserts itself in a lineage of scientific research has to be treated as a question rather than as a resource.

In the case of the Caltech experiment, the experimenters work hard to find their place in the experimental tradition of cognitive psychology. They are repeating and altering previous experiments. The Caltech experiment is designed to be recognized as an improvement over earlier ones made by vision psychologists. Previous experiments on classification and descriptions of photographs have strongly influenced the experimental protocol. The works of Rosch (1978) on classification, and Tversky and Hemenway (1983) on the categories of environmental scenes are key references for the authors. To align themselves with a research tradition, the researchers do not solely mobilise a series of theoretical references. The work of relating themselves to their predecessors bears on the design of the experimental set-up. The history of experiments in the psychology of vision can be pictured as the formation of a material and discursive experimental template where a generation of researchers modify the same variables and cautiously introduce new ones.

The practice of re-experimenting, as I present it here, is not so much concerned with genealogical relations where the arrow of time is unidirectional and where the past is used to explain the present. It is interested in relations of reciprocal interferences. The predecessors are no more stable than their present counterparts. An experiment acts as a stabilisation of times for both the past and the now. A re-experiment by touching on the malleability of the experimental frame intends to probe the variability of its references.

[67] See *Embracing error to enable rapid crowdsourcing*, (Krishna *et al*., 2016) discussed in the last chapter.

5.5. Non-exclusive relation

If we consider that the experiment is in translation and in iterative becoming, it becomes necessary to question the stability of its relata and its object. The understanding of the experiment as the prototype of a production environment would be limiting in the sense that it would invite an understanding of the experiment primarily through its relation to a specific environment of production. In the characterisation of the scientific work, it is often understood that an experiment is designed with a clearly circumscribed objective. In his work on Pasteur, Latour (1983) portrays Pasteur's lab and the farms as two entities in an exclusive relation. The lab is entirely defined by a series of conditions that can be replicated in the farms to ensure the proper effects of the vaccine. Both entities are entirely consummated in the relation. The two entities must mirror each other. As long as the farm stays parametrised according to the lab, the vaccine works. Whilst Latour's interpretation convincingly matches the model of the chemico-pharmaceutical lab, it doesn't leave much room for a mode of experimentation that finds its object retrospectively and iteratively. Latour puts the emphasis on the stabilisation effort. The experiment is considered as a response to clearly stated objectives. In general, experimental practice uses a goal-centred rhetoric. In practice however, an experiment in cognitive psychology and an environment of production may very well find themselves in contact through a much less exclusive, much looser and less “fixing” relation. To appreciate their relations, one needs to perform a de-centring from the experiment's stated goals. An experiment may very well find its goal only when a practice resonates with it. It is in terms of resonance and affinity between the labs of psychology, the environment of the Amazon Mechanical Turk that the re-experiment will intend to capture their relations.[68] The lab and the environment of production are not in a mechanically causal relation, they are engaged in an iterative becoming.

5.6. Apparatus and photographic alignments

What are the processes of mediation of photography at work in the re-experiment? Which role is photography expected to play and how is it understood? Reading the Glance protocol or following the acquisition pipeline of ImageNet, many details are provided by the researchers or dataset

[68] The Amazon Mechanical Turk workers explain that they are asked to answer so many questionnaires that they become better than average in answering various tests. They are increasingly seen as a particular population that developed specific skills: they become professional experimental subjects (Humphreys *et al.*, 2017).

makers on the apparatuses that are designed to move photographs from one place to another. In these accounts, it seems, the photograph, as it travels, remains consistently the same across the different contexts it traverses. It is presented successively as a search result, a dataset item, a stimulus and a photograph. For a practice of re-experiment where what matters first are the relations, the seeming consistency of an object like the photograph in the experimental apparatus is a question that requires further investigation. To elucidate this question, I will build and expand on the concept of photographic mediation sketched in the third chapter and on Barad's concept of apparatus I outlined above.

An apparatus, for Barad (2010), can be considered as a material arrangement that facilitates the emergence of a series of entities. The apparatus doesn't fully specify nor fully control the entities it enacts. There is always a risk involved and there is always an uncertainty about what exactly a successful operation means for the apparatus.[69] The apparatus is always open-ended, to use Barad's terms. Its mode is performative, not the deterministic application of rules. Likewise, the entities participating in the re-experiment are not fully defined, they are ontologically indeterminate. When the apparatus operates, it enacts a cut that resolves temporarily their indeterminacy in a way that orients their ability to act. The actors and agents in an experiment are not entities with given properties that are turned on or off by the apparatus. These properties emerge through the relation that takes place within the experiment. But as the apparatus is not deterministic and cannot anticipate fully how the entities align, there may be differences in how a resolution can be enacted. The indeterminacy of the entities involved may be resolved in different ways. I use the term alignment to name a resolution that acquires stability and singularity.

The expression “photographic alignment” designates a resolution of a series of indeterminate entities in which the instability[70] of the photograph plays a pivotal role. I do not use the expression photographic alignment merely to note that a series of objects named “photograph” are part of an alignment, but to refer to an alignment in which the ontological resolution of the photograph is crucial in stabilizing all the aligned entities. To study a photographic alignment is to attend to the dynamics that resolve the photograph and through which the photograph iteratively stabilises the other entities at play.

In an alignment, an entity can be resolved as a photograph in many ways. To come back to our initial example, it can be resolved as a search result, as a stimulus or as a dataset item. The variety

[69] “perhaps the real knack is getting to know when the experiment is working” (Hacking in Barad, 1996, p.144).

[70] I use the term instability as an equivalent to ontological indeterminacy in Barad's terms.

of ways an apparatus resolves the entities' indeterminacy is called its effervescence (Barad, 1996, p.235). In each of these resolutions, different entities, techniques and methods are mobilised. A dataset item will not mobilise the same entities and scales as a stimulus. The previous chapter provided another example: an amateur photograph found on Flickr doesn't become a dataset item by magic. It needs to be extracted from an alignment (the Flicker platform in which it is entangled with comments, groups, albums, tags, cameras, api's) to be inserted in another (ImageNet via AMT and WordNet). The photograph is resolved differently in the two apparatuses and properties emerge or disappear.

Let's revisit the analysis of the experimental apparatus with which I opened the previous chapter in the light of these considerations. The Caltech researchers understood very well that vision can be many things and that to be studied it needs to be held in place. The stabilisation of the photograph *as data* is the experiment's method to enact vision in terms it can control. Vision and photography mutually stabilise each other in the experimental alignment. In *What do we perceive ...*, photography is what gives seeing its definition: to see is to see a photograph within a very specific micro-temporal range. The experiment invents vision as susceptible to translation through the stabilisation of the photograph.

The Caltech experiment paper can be read as a long description of a photographic alignment. From this perspective, we can see the authors describe in detail the different strategies and methods through which their apparatus resolves the indeterminacy of photographs and the conditions under which certain resolutions can be converted into others. They start with photographs resolved as visual search queries and subsequently resolve them as visual data, a dataset. The dataset items are then resolved in turn as stimuli where they are displayed for an infinitesimal time span to the subjects. Later on, when the descriptions are assayed, the photographs are resolved as “controls”: there is no time constraint for viewing them and the scorers who assay the descriptions use them to validate the descriptions. The experimental apparatus enacts multiple resolutions of the photograph. Each time a new resolution is enacted, the photograph acquires a different agency and enacts in turn different agents. A stimulus facilitates the emergence of a subject and a control facilitates the emergence of an assayer. Each new resolution is made possible by different technologies (the search engine, the script that flashes the stimulus, the taxonomy) and different techniques of the body (controlling the eye blinking, writing with a keyboard, ticking checkboxes etc.). Furthermore, competences are redefined. The subject's competence is to respond well, the assayer's competence is to evaluate well and for each of these competences, a certain temporal stabilisation is required. Response corresponds to micro-time, evaluation corresponds to a considerably longer time of deliberation.

The apparatus mobilises a set of practices and technologies that attempts to maintain the coherence of these resolutions along an alignment. The work performed by the apparatus is one of composition. The apparatus composes segments of alignments. Segments are compartmentalized and their outputs are regularised and coordinated. The researchers acquire the dataset, the subjects describe the stimuli, the assayers use the “controls”. The alignment is what holds together the multiplicity of the resolutions enacted in the segments.

To re-experiment *What do we perceive ...* as a photographic alignment leads to a reformulation of its original question. *What do we perceive ...* was conceived as an attempt to learn what subjects could see in a glance when a stimulus was shown briefly. The re-experiment is an attempt to produce alignments that resolve subjects, the glance and the photograph iteratively.

What we gain with a Baradian approach is a relational understanding of the processes detailed in the protocol. It gives the possibility of attending to the objects and subjects that populate the experiment without taking them for granted. Subjects and objects in the experiment find their resolution through an apparatus they iteratively enact. Agency is never in the objects or in the apparatus alone. It is a performative relation not a determination. Attention to the performativity of the apparatus guides the design of the re-experiment. The re-experiment is not concerned with the truth claim of the Caltech experiment but with its generativity, the world it is enacting. The variations aim at probing the effervescence of the apparatus, to sense what resolutions it is capable of and to see which ones the participants will take responsibility for.

5.7. Subjects

As it is clear by now, Karen Barad's agential realism provides a key set of concepts that help me articulate the research and the practice. Her views on the agency of the apparatus, the questioning of the centrality of the human and the subject as a given entity are fundamental in both my analysis of the Caltech experiment and the elaboration of my response, the re-experiment. At this juncture, some further remarks regarding the notion of subject are in order. Agential realism doesn't presume that the relevant agency is located in a singularly bounded subject. Intra-action entices us to look for more complex and entangled formations. But if downplaying the centrality of the subject is an interesting strategy to counter a humanist framework, how does it play out in contexts that are actively de-humanizing? This question is particularly sensitive in the context of this study as the Caltech experiment and the environment of annotation share the same trait, the devaluation of the contribution of the human agents. When agential realism decentres the human subject, does it provide enough theoretical resource to account for its exploitation, its mechanisation or its agency? In other words, using agential realism in this context raises a concern: how to reconcile a necessary displacement of the bounded subject while avoiding alignment with the dismissal of workers' and subjects' agency? Or more positively, which role does the subject play in what Despret calls the favourable outcome of the re-experiment? How does it intervene in the alignments?

Understanding Barad (or other strands of posthumanism) too crudely could lead to looking everywhere at the exception of the subject in order to address our problem and simply considering the human contribution as irrelevant. This crude interpretation would double down on the exclusion that the participants suffer and would tell us nothing about the nature of the participants' contribution. This would make us insensitive to the instrumentalisation of workers and subjects. Furthermore it would make us blind to the fact that other forms of agencies are intimately bound to human agency. As we have seen in the previous chapters, understanding the working of computer vision algorithms cannot be achieved without attending to the work of annotation. Understanding the experimental production of computer vision's model of perception cannot be done without understanding how subjects are enacted and black-boxed over the course of the experiment. In *What do we perceive ...*, there is, as I have underlined above, a transformation of response into perception wherein perception emerges only when it leaves a subject who is only entitled to a mere response. The experimental protocol traces a trajectory that elides the subject's contribution. If agential realism does displace the subject, how can it do so without operating the same reduction? A first answer would be that the de-centring of the subject is not meant to dismiss the contribution of humans altogether and cannot be made complicit in the active undermining of the experimental subject's contribution. It should drive instead our attention to the processes through which the subject is enacted. The practical engagement with agential realism in this research is not meant to trade the participants for the apparatus. We find here another version of the problem that relates to the tension expressed in the third chapter, and the difficulty of attending to the agency of the various entities and to the specificities they develop in intra-action.

It would be unfair to criticize agential realism for being blind to exploitation or injustice to human subjects. Agential realism is committed to "a new materialist understanding of power and its effects on the production of bodies, identities and subjectivities" (Barad, 1996, p.35). The problem lies in the fact that, in contrast to a critical tradition in which the subject is taken either too much for granted or too much as a subject of language, Barad often construes the subject as a negative notion. It is (rightly) criticised as an entity that is given too much stability, and tied to a relation to the world that is one of distance. With a focus on subjects, Barad suspects, scholars too readily assume the primacy of language and symbolic meaning to which too much power has been granted (Barad, 1996, p.132). It remains that, in a context (i.e. psychology, cognitive science, education) where subject formation remains a key question, agential realism offers less substance about the subject in emergence than about the apparatus from which subjects emerge. Therefore, to account for the existence of a subject beyond a representationalist framework, it feels logical to turn to theoretical frameworks that converge with agential realism to supplement it. One pre-requisite for this kind of articulation is to elect a framework that does not rely primarily on consciousness and the consumption of symbolic meaning. There is an affinity and a potential to articulate the subject of affect with agential realism, even if these frameworks are not reducible to each other. This articulation is productive in that it offers a rich set of concepts and sensibilities to attend to the emergence of the subject. In what concerns me here, the work of Brian Massumi (2002) and Erin Manning (2016) constitutes a reservoir of resources to apprehend an embodied subject whose involvement in the resolution of the re-experiment mobilises the sensible vectors and intensities circulating through the apparatus. Affect does not simply account for subjects of sensation within the apparatus, affect can be used to innervate the apparatus of agential realism. Furthermore, the subject of affect being an open entity proposes a mode of collaboration that accounts for multiple modalities and sensibilities in the collaboration. The subject of affect is from the onset a participant, never just an “experimental subject”. For these reasons, affect theory accompanies me in attending to subjects’ emergence in the re-experiment.

However, such a move brings its own set of limitations, as Lisa Blackman (2014) points out. In affect theory, a place is given to the subject only when considered in a relation of a-signifying semiotics, in relation with signs that tune directly to the body and trigger affective processes, that capture pre-individual elements. On the basis of the assumption that the phenomenologically experiencing subject should be treated in the same manner as other entities, there is a tendency to replace the subject by “a variety of neurophysiological concepts (including distinctions made between will and automaticity)” (Blackman, 2014, p.373). If cognition is seen as distributed, there is a strong incentive to look into the modes of cognition that do not privilege consciousness and reflexivity and rather focus on automaticity and affect:

> Subjectivity is replaced with the somatic, which easily reduces to the neuro-physiological aligned to concepts such as the automatic, the subpersonal, the pre-programmed, the non-conscious and non-cognitive (Blackman, 2014, p.373).

Blackman's words resonate with my discussion of the re-experiment for several reasons. The first is the primacy given to automaticity in both cognitive psychology and affect theory. Even if the subject of affect is not reducible to the black boxed entity providing a mechanical response in the experiment, there is only a fine line between the two. To understand the nuance, let's sketch provisionally the different understandings regarding the experimental subject and automaticity. For the Caltech researchers, automaticity is a given. The subject is treated like a response machine automatically providing responses to stimuli. From a humanist perspective, this is simply a denial of the richness of the subject's interiority and an example of instrumentalisation. Affect theory provides an alternative to both. There is no subject fully equipped with interiority that pre-exists the experiment – it emerges through it. It is the superject, the subject of the event, whose behaviour does not require an explanation in terms of conscious processes. Yet it does not equate to an uncomplicated response mechanism either because its emergence is the site of a rich experiential moment that does not correlate neatly an input to an output. There is newness in the emergence. Affect theory offers a subject that avoids both the mechanical depiction we find in the Caltech experiment and a rational self-bounded subject who extricates himself through meaning making. But why are affect theorists so confident of finding enough in this emerging subpersonal subject?

This question brings us to the second reason why Blackman's cautionary words need our attention. The reason why there is an affinity between the subject as it is enacted in the Caltech experiment and the subject of affect is that the latter is built on proto-theories of the self inherited from neuro-psychology. The convergence is not accidental. Reading the Caltech experiment through the lens of affect theory brings in the complicating factor that affect theory in many ways finds its concepts in one of the disciplinary lineages of the experiment. We need to be aware of a potential blind spot as affect theory tends to use the same sources to produce its notions of subjectivity. Moreover, we should be doubly wary of this genealogical link: as we have seen in the Caltech experiment, to treat subjects as automata is by no means innocent and corresponds to a division of labour in the experimental frame. Both Caltech researchers and affect theory's notion of subjectivity relies on a dividing line between the mind and the site of affect, the body. The body senses and enacts a form of thinking before the mind intervenes. I understand the fertility of desolidarising subject and individual, giving prominence to the non-discursive in order to “avoid capture and closure on the plane of signification” (Massumi cited in Leys, 2011). Yet as Ruth Leys points out, if this stance is taken literally it leads us to another form of mind and body dualism. Representation, signifying semantics, intention, are relegated to the province of the mind and dispensed with. And the division of labour that treats the experimental subject as a mere responsive body may become difficult to criticize from a perspective that traces too hard a line between affect and consciousness.

In our case, this separation creates an additional difficulty. Semantics, the production of labels, perceptions and classifications are key objects of the experiment. Symbolic processing, from making descriptions to producing classifications, is central to it. Agential realism engages us to look into the material production of these symbolic operations, to see them as not merely language games but inscribed within a more distributed machinery. But we need to be careful not to make the opposite reduction and ignore the engagement of the subjects with symbolic meaning making. The experiment follows a course that begins with flashing stimuli on the subject's retina, and continues with various layers of semantic intervention ending in a taxonomy.

There is a lot to lose in taking for granted the separation of the material-semiotic, the affects and the representations. The re-experiment should not to play them against each other and neither should it assume that they are isolated from each other. Therefore, I am not ending this theorization of the subject of the re-experiment on an unconditional allegiance to the theoretical frameworks that resonate through and inspire the design of the re-experiment but on a cautious note, an awareness of the tensions that the research creates. In this sense, the re-experiment does not inhabit a smooth theoretical terrain. It represents an attempt to resist a too absolute demarcation between affect and other dimensions, an effort to integrate them rather than oppose them, with the tensions that entails. It leaves room for the subject of affect to eventuate in consciousness and articulate meaning even if those do not represent its primary concerns.

5.8. Supervision, division of labour, social ontology of the experiment

The experiment as seen in the previous chapter was conceived as a contribution to the fields of cognitive psychology and computer vision, and uses the conceptual structure of psychophysics. The subjects in the experiment are essentially perceptual entities. The experiment however carries with it a social ontology (Danziger, 1992). The population of subjects is treated as a collection of interchangeable individuals understood as individual atoms. Social phenomena are reduced to individual behaviour and individual behaviour to a written response. In such a conceptual framework, dialogue between the subjects is deemed irrelevant. Furthermore, the structure of the experimental device prevents it. The experiment's social ontology includes more than experimental subjects. It distributes the different people involved into separate classes and prevents their interaction. The scorers have no contact with the subjects. Therefore, the decision mechanism is never established through a collective discussion among the subjects or in dialogue with the scorers. It is reached through an averaging of the responses processed by the scorers. The only ones to be able to function as a team are the researchers. Those under study or under control, those supervised, are considered as discrete units whose outputs can be levelled and averaged. The individual subject as the result of an atomistic reduction is one element in the chain of discretisation that makes the operations of counting and averaging possible through the whole experiment. The population of subjects is constituted of unrelated individuals producing unambiguous output that can be measured. The statistical convergence of these outputs counts as consensus. The experiment naturalises this social ontology and provides tools and references to measure it.

If the experiment's social layout excludes any form of interaction within a group, the idea of the individual subject as the fundamental unit provokes another form of exclusion. With agential realism, we cannot take for granted the delineation of an entity as the individual. There is a process of individuation that is intimately related to a cut performed by an apparatus. In this sense, an individual is always a collective because its resolution as an individual is intimately related to the resolution of other entities (devices, gazes, measurements, intervals etc.). To question the social ontology of the experiment is therefore to operate on the constructions of human-nonhuman collectives that give rise to this experimental subject. The subject of vision exposed to a stimulus of a few milliseconds is in extremely intimate co-constitution with the device that flashes light on her retina. To consider the subject as an individual unit that can be imported into the experiment and whose properties are affected by the experimental device blinds the experimenter from the magnitude of what the emergence of the subject involves.

This however is of great importance for the experimenter as the collective that emerges in the co-constitution of the subject and the alignment in which it is resolved also involves the researcher herself. The experimental device binds together the experimenter and the subject, it carries instructions, defines positions, enacts movements and enables propensities. As Despret remarks, subjects never simply respond to stimuli, they always try to figure out what being a subject of an experiment means. They are interpellated and this interpellation goes far beyond the decoding of instructions or simple forms of cueing. Devices are conducive and subjects and devices are in affective atunement. Even if subjects cannot decode the stated intentions of the experimentor, there are no naïve subjects, Despret (2009) concludes after in-depth studies of laboratory experiments.

The naïve subject pertains to a social ontology of the experiment that postulates the separation of entities and their isolation from the device and the experimenter. Contesting this social ontology requires attending to the problematisation of separability and an engagement with the subject's involvement in the experiment that goes beyond a production of stimulus. Therefore from the perspective of the re-experiment, attention given to the performativity of the apparatus does not lead to a dismissal of the subject, but a renewed curiosity for its enactment.

This brings us back to the conclusion of the previous chapter where parallels were being drawn between the experiment and the annotation environment. The social ontology offers a level of abstraction at which they can be related beyond a mere similarity of forms and beyond their apparent differences.

Ontology	**AMT**	**Caltech experiment**
The social is a collection of bare individuals or atoms	Worker identified by ID	Anonymous naïve subject isolated from her peers
Readiness of perception and availability of the world	Ability to identify concepts *imaged* in photographs	Ability to respond to stimuli *representing* transparently things in the world
Perception is accomplished through self-contained tasks that can be apprehended in isolation	Mirco-task, fixation as a measure of attention	Discrete stimuli, fixation as measure of attention
Consensus through averaging	An entry is kept if a minimum of 3 annotators take the same decision	An entry is validated if a minimum of 3 scorers take the same decision
Static knowledge	Workers do not learn from previous task, mechanical repetition.	Subjects do not learn from exposure to stimuli, mechanical repetition.
Basic knowledge	The worker is a specialist at being generic	The only requirement is to be fluent in English
Recognition as aperception	Workers apply already learnt categories	Subjects apply already learnt categories
Levelling	Levels of perception are defined by the depth of	Levels of perception imply a correlation between micro-times and

	WordNet categories.	taxonomy levels
Separability through supervision	Workers only concerned by a segment of the task. No involvement in the complete process. Turker vs. requester	Students only concerned by the response to the stimuli. No involvement in the complete process. Subject vs. experimentalist.

Table 1. The social ontology.

Let's review the defining features of the social ontology I have analysed above and see how the environment of annotation and the experiment converge through them. The first axiom postulates the social as composed of individuals isolated from their peers. These individuals can be treated as atoms. In the experiment, the social atom is the subject and in the annotation platform, it is the crowdworker. These bare individuals are interchangeable units. They are characterised by their readiness to perceive a world available to them. These individuals have value because they are able contribute to processes of semiosis and perception. The object of their perception is a concept imaged in photographs. In both cases, these photographs are amateur photographs exchanged on photo-sharing platforms such as Flickr. To a large extent, the recognition of the concept happens through a transparent process. The mediation inherent to the process can be ignored as photographs are said to “image” concepts and concepts represent things in the world. Their perception is dissected in self-contained units. It can be decomposed and quantified in terms of fixations, brief moments where the gaze remains focused on a single point in space. These fixations can be actioned on demand through the injunction of an apparatus. They respond to stimuli or to HITs but they don't negotiate the terms of this response with others. The consensus is obtained through averaging**.** Sociality is constructed through the mediation of an apparatus that collects the responses and processes them. The mode of consensus that can be obtained from these individual units is a process of averaging that dispenses with direct interaction between them. Subjects must not influence each other and workers cannot collaborate and even less unionize. Neither subjects nor workers are given any explicit feedback. The subjects produce descriptions but they have no occasion to review their work or to engage in a discussion with the experimentors. The annotators do not receive explicit feedback from the requesters either. From their perspective, their work is sent in a black hole. Nobody expects them to improve, to learn from their past interactions, to get better at what they do. Both experimental subjects and workers are expected to mechanically respond to stimuli. There is no expectation of an evolution from their part. In fact only a basic knowledge is expected from them. Recognition presupposes that the work of cognition has already happened prior to perception.

Recognition is understood as a way of matching what is given to the senses with categories already acquired. The expectations regarding the kinds of categories already acquired by the individuals are low. There is no expectation of expert knowledge of any sort that would distinguish them. The Turkers are even required by the ImageNet team to tell when they have specialist knowledge in a given area to avoid a penalty if their response does not coincide with other workers. As stated above, recognition can be dissected temporally through discrete fixations. To each of these fixations correspond a level of description. Levels of descriptions and presentation times can be correlated. Different levels of description require a different effort from the individual. To perceive at a basic category-level is easier than for sub or super-ordinate levels. Therefore effort can be evaluated in terms of attention time and productivity: categories requiring more effort are more costly to complete. Finally the experiment and the annotation environment share a division of labour. Subjects and experimentalists are separated as are workers and requesters. Supervision is the mode through which this division is enforced.

This social ontology is the site where objects are made readily available and relations already established for both the experiment and the annotation environment. In both cases, the ontology is implicit and its materialisation is distributed over myriads of entities, rhythms, scales and devices. To make it explicit as I do in this chapter, is to prepare the perspective through which what can be probed in the experiment may be used to resonate in the annotation environment. The hypothesis behind the variation that will inaugurate the next chapter is that this social ontology is a site where an horizon of change for computer vision and its photographic elaboration can be probed.

5.9. Conclusion / transition

In this chapter, I have moved from questions concerning the analysis of computer vision's elaboration to the practical question of how to engage with it. The experimental work of computer vision scientists gives me a starting point to develop a practice. Engaging with an experiment, however, raises difficult methodological questions. The experimental paradigm underlying the scientific method carries with it a host of readily available objects, an understanding of relations and a social ontology that need to be questioned. What makes this questioning difficult is the fact that these objects, relational principles and ontologies are not simply exposed in the form of instructions or explicit methodological guidelines, they are also embedded in devices and practices. Building on Karen Barad's work, I am trying to move my understanding of the experiment from a notion wherein entities pre-exist their relations and already acquired properties are measured, to a notion wherein indeterminate entities acquire a resolution and find a temporary stability. This move towards a relational ontology leads me to interrogate the conditions of emergence and the resolution of these objects, relations and devices and stresses the importance of imagining ways to attend to the performativity of the devices and practices articulating the Caltech experiment. Along with this questioning, the chapter lays out the fundamental principles of the practice that will occupy me in the next chapters.

The practice, the re-experiment, is a form of re-enactment conceived as a means to engage with a series of devices, relations and scales that are crucial to computer vision. To re-experiment means to engage with an experiment to probe what this experiment can do, its potential, rather than a form of replication that seeks to validate or invalidate it. Re-experimenting is to explore a mode of variation of an experiment that differs from the canonical variation of parameters an experiment expects. A re-experiment tests the effervescence of a given experiment, its potential for divergence when variations are applied to its protocol. A re-experiment expects another kind of outcome, a possibly favourable outcome where the phenomenon makes the researcher do things and forces her to adapt her experimental device.

The re-experiment doesn't consider the relation between the experiment it re-enacts and a site of production as a simple relation of exteriority, where what is invented on one side is applied on the other. The where and the when of the experiment are questions to be researched in the framework of the re-experiment, not givens. Experiment and site of production are engaged in a form of iterative becoming, not a simple genealogy. The experiment and the site find each other retrospectively and iteratively. To re-experiment is to inhabit a space where vectors of relations are never fully actualised, where relations are better captured through resonances and dissonances rather than as linearly connected entities.

As the entanglement of the different entities is the rule, and as the where and the when are in question, there is no outside from which to observe. This conditions how a re-experiment relates to the difference it creates with the experiment it re-enacts. Agencies of observation must attune to the difference within. To re-experiment doesn't mean to make an alternative experiment, neither “next to” nor against, but to engage with the manner an experiment self-differentiates and diverges from within. To engage in a re-experiment means to consider an experiment as already more than itself.

Entities must be discovered rather than assumed as readily available and self-standing. This is true for photography as much as for other entities: the photograph is not a given object, it is an indeterminate entity finding a resolution within the apparatus and retrospectively stabilising the apparatus. The re-experiment treats the Caltech experiment as photographic in that it attends to the work of stabilisation of photography as one of the structuring forces of the experiment. To re-experiment means to engage with the different resolutions of the photograph and how they relate to each other, as well as how they enact other entities. Binding the indeterminacy of the photograph to the micro-temporal dimension of vision is at the core of the experiment. Re-experimenting here means to address practically the relation between the multiple resolutions of photography and vision such binding entails.

At this stage, these considerations may seem far removed still from a practice. They are however suggestive and function more as an opening rather than normative methodological guidelines. The relation with theory in the thesis is not one where practice applies the concepts theory provides. Theory works more as a tool to expand the map of inquiry, to de-naturalise concepts rather than provide strict precepts. The concept of re-experiment draws heavily on Barad's agential realism. This framework is extremely helpful in attending to the performative dimension of the apparatus. It helps dissolve the sedimented layers of assumptions, positions, representations and hierarchies making up the positivist substrate of the experimental method. In that sense, agential realism is a formidable tool to start engaging with a world whose coherence is not given.

The relation with theory also involves a certain distance, a certain caution. Whilst agential realism helps undo many notions underlying the experimental method, it also comes with a series of assumptions that could limit the inquiry prematurely. The context in which agential realism is developed means it is used to downplay the importance of the subject rather than elevate it. Essentially, it provides the elements to attend to the subject's emergence but the corpus of research it generated tends to treat the subject primarily as a negative notion. To attend to the stabilisation of the subject in the re-experiment, I read agential realism with a compatible framework, affect theory, in the vein of radical empiricism. Affect theory provides material to nurture a sensibility to the stabilisation as an event not just a cut. The subject in affect theory is the subject of the event, which it doesn't pre-exist. To read the experiment using agential realism therefore requires taking in hand the difficulty inherent in such a framework and finding room for manoeuvre, to open it to dialogue with others.

Finding the right distance to theory is also motivated by the relations that theory entertains with scientific knowledge itself. For instance, affect theory aligns with proto-theories of the self coming from cognitive science that insist on the separation between consciousness and affect. At that level, as affect and the disciplinary distribution of the Caltech experiment align, it may too readily confirm if not the findings but at least the objects or methodological principles that structure the experiment. If radical empiricism is a wonderful tool to attend to the event of the subject's emergence, it needs also to be treated with caution as it may impose too readily a series of distinctions it inherits from the discipline that informs the design of the experiment. Therefore if theory informs the methodological considerations, re-experiment also forces me to read back theory from within the practice. Theory is used to cultivate sensibility, orient attention, dissolve what is sedimented in the scientific method. But the practice will also require careful attention to its own methodological framing not to let theory foreclose too early an avenue of investigation or pre-empt the research process.

Finally, the difficulty of approaching the emergence of the subject also impacts how the social ontology of the experiment is taken in charge. The re-experiment considers the experimental device as more than a collection of instruments producing the measurement necessary for the production of knowledge. The experimental device provides a managerial template, a set of relations in material form that distributes and organises labour. A social ontology informs the experiment: groups are an aggregate of individuals whose responses can be averaged. A collective is a sum of discrete individuals. Re-experimenting is finding ways to question this distribution and finding ways to think together with the subjects. The importance given to emergence, to affect, to the technological articulation of the experiment and how they intervene in the formation of consensus inform crucial decisions in that regard.

Chapter 6. Re-experimenting

The previous chapter operated a transition between the analytical groundwork and the practice. This chapter and the next provide an account of the practice and what I learned from conducting the re-experiments. The concept of re-experiment is the pivotal element of this transition. With the notion of re-experiment, I am attempting to find a way to engage with the Caltech experiment while problematizing its methods. My objective, in this account of practice, is to reflect on how a re-experiment works concretely. I proceed to the analysis of the concrete circumstances of re-experimentation in practice. My intention is to share as vividly as possible the process of discovery of what the re-experiment can do, to explore its potential and its repertoire.

Re-experimentation is eventful. It creates an oscillation between the experimenter and his re-experiment. I am making the re-experiment do something but, in return, it also puts me to work. This chapter and the next document the unfolding of the practice through which I discover what I have set in motion. The core of these chapters aim to give a first sense of what is experienced, the atmosphere of a re-experiment, but also to convey a sense of the immanent relationship of the methodology to the practice (Stansfield, 2009, p.19). The following pages reflect on what happens to the research when a re-experiment is conducted rather than thought, when it happens in the room rather than on paper.

As I don't want to lose the reader in the meanders of the process, these pages offer a balance between a textured description that shows the re-experiment's seams and a selective presentation of the evolution of the practice. This is why this text needs to be read together with another set of documents (the four appendices) and the companion website located at *http://functionariesofthecamera.net/algorithms-of-vision* (for access credentials, see the section Website, p.234). In several instances, I refer to specific moments of interaction that are described in more detail in appendices. As the voice and the aural dimension plays an important role in the process, the website offers access to many annotated recordings. It also allows the reader to interact with a web-based version of the apparatus used in the re-experiment and experience various stimuli at different presentation times.

The account of practice develops over two chapters. In this chapter, I concentrate on the presentation of a variation of the Caltech experiment's protocol bearing on its social ontology. The focus is on the relation between the alteration of the social ontology of the experiment and the participants' dynamics. I reflect on what it means to change the relational diagram and my intervention in the protocol invites new techniques, skills, and forms of embodiment in the re-experiment. The next chapter concentrates on the role of the apparatus and the photographic elaboration at work in the re-experiment. And it ends with a discussion of a specific device, the taxonomy and its role in the stabilisation of the re-experiment.

6.1. Main variation

After the last chapter's theoretical considerations, the time has come to consider how re-experimentation has been carried out. The theoretical considerations were an attempt to question what an experiment is and its ontology, to start to understand it as a relational and performative process. And to problematise its relation to its outside. A re-experiment is based on a series of changes in an experiment's protocol. What to keep and what to change are equally difficult decisions. Over the course of the re-experimentation, I have introduced several variations. However, the one I am presenting here is a variation tested in the first pilot session that has been consistently enacted on many occasions and is representative of the work I have carried out more generally.

There is a huge difficulty at the design stage to choose the kind of alteration in the protocol that will follow a line of divergence inherent to the experiment. In the context of this study, it means to introduce a variation relevant to the common problems identified in the experiment and in the environment of annotation. What seemed most important for me was to understand the implications of the social ontology of the experiment, what it prevents and how a change in the protocol would create an opening, different enactment of the subjects. The motivation is to explore a different configuration of agency among the actors with the hope that it will lead to what Despret called a favourable outcome: the iterative transformation of the subjects, apparatus and experimentalist. And concomitantly, by contrast, to understand which exclusions are performed: what is left aside when subjects are enacted in a specific way. An intervention in the social ontology of the experiment is an attempt to understand the relations between the various entities making up the experiment. Beyond the observation of the evolution of the individual subjects, it is their mode of relation that motivates the intervention. Which division of labour ensues and what changes in the resolution of the experiment: how the granularity of what is perceived relates to the mode of consensus that takes place. And as the social ontology has been presented as what relates the annotation environment and the experiment, the work of interpretation in these pages will be to identify how altering the social ontology helps rethink the terms of the photographic elaboration of computer vision more generally.

The main variation introduced in the re-experiments bears on the social ontology of the experiment. The participants are not considered mere subjects responding to a stimulus. The re-experiment proposes a different configuration in which they are expected to contribute actively. The participants are not recruited as subjects, they are invited to perform as subjects. Moments of shared accountability are integrated within the experiment. The new configuration stresses the importance of the participants' own understanding of the situation they are in. The participants adopt different roles. They are describing the stimuli, and they are invited to assay the descriptions in a second phase. In the Caltech experiment, the distinction between subjects and scorers is enforced. Those who produce the responses and those who map them on to a taxonomy were distinct groups. The subjects had no overview of the whole process and ignored how their responses were processed. In the re-experiment, they contribute to the entire process. And they are invited to discuss how the assessment of their responses affects the outcome of the re-experiment.

Another variation affecting the mode of participation concerns the co-presence of the participants during the re-experiment. In contrast with the Caltech experiment where subjects were isolated, the participants are always together physically in the same room. The principle of segmentation that guides the Caltech experiment is abandoned. The participants annotate orally, first, one after the other, listening to each other's description. And in a second phase, collectively as groups of increasingly larger sizes. They experience together the “reaction machine”, the device that delivers the stimulus. They are invited to discuss this experience at various moments and are encouraged to enter into dialogue with each other. The key intention is to cultivate the ability of the participants to account and respond to the experimental device. As Despret writes: “To think with is at the same time what is at stake in - and the condition of - the experiment” (Despret, 2009). The intention behind these moments of dialogue is to elaborate with the participants a shared account of the process.

The collective dimension of the annotation and the spatial de-compartmentalization introduce two more changes. The participants do not describe the stimuli in written form, but orally. And their oral descriptions are transcribed by the experimenter which brings two modes of archiving into the process: the written transcription by the experimenter and the recording of the participant's voice. Another modification bears on the screen's dimension. If the participants are seated together, listening to each other and describing together, it makes sense that they see the same stimulus at the same time and therefore to see the stimulus on a large screen rather than a monitor. It changes the display conditions from the viewing of an image that the eye can grasp at a distance to an image in which the viewer is immersed. The participants not only listen and talk to each other, but they also all experience the same stimuli.

The decision to introduce this set of changes was clear from the outset and has been at the core of the re-experimentation for the whole duration of the project. The motivation behind the re-experiment however was also to understand better - and engage with – the agency of the visual apparatus. A point of departure to think about the photographic apparatus would be the selection of photos for the re-experiment. Yet, to rehearse an argument developed in the first chapter, photography is not reduced statically to a medium defined by a series of pre-given objects, but is understood as an alignment: a process of mediation that temporarily resolves the relation between objects and practices whose contours were not stabilised before entering the relation. This meant that attending to the selection of photographs would have to go beyond a mere concern for the representations introduced in the re-experiment. In this sense, I looked for ways to make a variation on the photographic alignment of the experiment.

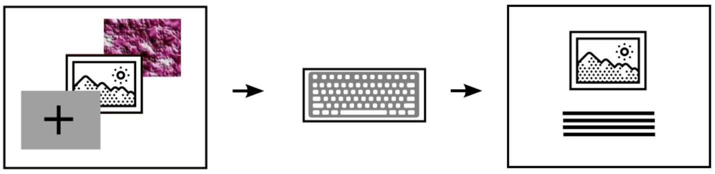


Illustration 4: In the Caltech experiment, the subject writes a description

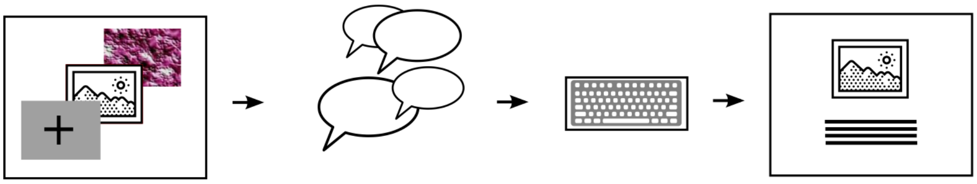


Illustration 5: In the re-experiment, the description is made collectively, orally and transcribed simultaneously

With this in mind, I turned to the ImageNet database. This decision was not based on certainties; I was following an intuition that was nurtured by the early stage of practical re-experimentation. ImageNet was in many ways informed by the Caltech experiment and as I explained already in the previous chapter, the relations between ImageNet and the experiment were multiple and indirect. Introducing ImageNet into the re-experimental device was an attempt to find ways to learn more about the relations between the dataset and the experiment. An attempt to extend the re-experiment to reach into something that is more than a convenient set of photographs, but the specific device introduced in the first chapter, the dataset. The significance of the variation bearing on the participants dynamics appeared early in the process and the account of practice in the next chapter reflects this. The relevance of the variation on the source of stimuli has been more indirect and diffuse. Over the course of the analysis, I will come back to their relations and how they mutually reinforced each other's effects.

The following table summarises the differences between the Caltech experiment and the pilot variation.

Experiment Caltech	Re-experiment
Stimuli	Stimuli
Dataset of 44 indoor images and 46 outdoor images collected from Google Image	The images need are selected from ImageNet
Stage 1	Stage 1 (variation)
Subjects	Subjects
22 students from California Institute of Technology proficient in English (from 18 to 35 years old). Subjects are naive about the purpose of the experiment.	Participants are invited through the communication channels of the institution hosting the session. No particular expertise is required. The subjects receive information about the process and purpose of the re-experiment.
Apparatus	Apparatus
Dark room, Mac OS 9 computer, screen with refresh rate of 75 Hz. Custom made software and Matlab.	The light is dimmed but the room is not dark. The set-up must permit an optimal control on the micro-time. The code has been re-written.
Procedure	Procedure
An image from the dataset is presented for one of seven different possible presentation times. Subjects are given unlimited time to write the description. Subjects are isolated, one subject per room.	An image from the dataset is presented for one of seven different possible presentation times. Participants are given unlimited time to describe the stimulus. The participants describe a first series of photographs alone, then in groups of different sizes.
Stage 2	Stage 2 (variation)
Five paid volunteer students from different schools in the Los Angeles area (from 18 to 35	The same participants review the descriptions.

years old) serve as scorers.	
Apparatus	Apparatus
The scorers evaluate and classify image descriptions using custom made software using MATLAB. Same computer specs.	The participants receive a copy of the taxonomy on paper.
Procedure	Procedure
The research team constructs a standardized and hierarchical list of attributes based on the descriptions (105 terms). The scorers evaluate the descriptions using the list of attributes. For each response, they check which attribute is referred to and if the description of this attribute is accurate.	The participants are shown the stimuli one by one. They review the results collectively using the taxonomy.
Stage 3	Stage 3 (variation)
No debriefing	Debriefing discussion with the whole group
Session	Session
22 subjects participate in 5 sessions. 1980 descriptions are produced.	The pilot experiment is organized with 12 LSBU students, BA Photography. 12 descriptions are produced.

Table 2. Summary of the differences between the two experimental protocols

6.1. Overview of the sessions

The changes described in the previous section are not spectacular and they are not negligible either. In themselves, they are only tentative. It is only when the re-experiment is conducted that one can realise what they meant, where they lead, what kind of risks they force me to take and what the re-experiment becomes in return. This chapter explores how I learned from them in practice.

Over the course of the research, I have conducted 11 re-experiments. Five sessions took place in the UK while the others were organized in various European countries. The re-experiments have been conducted in several universities and an art institution, The Photographers' Gallery (TPG). Additionally, they have also been hosted by Algolit, an art and research group[71], for two sessions.

[71] Algolit is a collective working on e-literature. The group, influenced by the principles of the Oulipo-meetings, organises regular work-sessions that concentrate on the practises of computational ways of reading and writing. See

The participants were invited through the communication channels of each hosting institution or organisation. Mostly the participants had an interest in photography or media more generally. As the greater part of the venues were institutions of education, the majority of the participants were students, teachers and researchers. They were students of photography at London South Bank University and Universidad Politècnica de València, students of Computational Aesthetics in Aarhus University and Mediawissenschaft in Potsdam. As three sessions took place at TPG, the visitors (artists, photographers, writers, amateurs of photography, educators) and the Gallery team participated actively in the sessions. The following table offers an overview of the iterations and their timeline.

	Date	Location	
0	02/2016	ERG, Brussels	Pre-pilot. 20 students, Fine Arts, BA level. 2 sessions of 4 hours. I presented my initial ideas for a re-experiment and a first version of the Stage 1 was tested and discussed.
1	16/05/2016	London South Bank University	Pilot session. 12 students, Photography, BA level. 3 hours session. Stages 1 and 2.
2	14/10/2016	The Photographers' Gallery	First public session. 15 visitors to TPG. 3 hours session. Stages 1 and 2.
3	14/11/2016	Aarhus University, Denmark	Pilot session. 14 students, course of Computational Aesthetics, BA level. 2 hours session. Stages 1 and 2.
4	09/05/2017	London South Bank University	Session for my colleagues of the Centre for the Study of the Networked Image reading group. 8 researchers and teachers among which 4 active participants and 4 observers. 30 minutes session. Stage 1.
5	23/05/2017	The Photographers' Gallery	Session for the Gallery's team. 15 members of TPG's team. 2 hours session. Stage 1.
6	27/06/2017	Ways of machine seeing, Cambridge	In an event organised by the Cambridge Digital Humanities Network, Cultures of the Digital Economy Research Institute (CoDE) and Cambridge Big Data. 4 symposium participants volunteered. 20 minutes session. Stage 1.

https://www.algolit.net/index.php/Main_Page

7	11/11/2017	Maison du livre, Brussels, Belgium	A session with attendees at Algoliterary Encounters, an event dedicated to machine literacy. 12 participants. 6 hours session. Stage 1.
8	12/11/2017	The Photographer's Gallery, London, United Kingdom	The re-experiment as interfering device. 15 visitors of TPG. 3 hours session. Stage 1 and an incursion into the exhibition room.
9	22/01/2018	Potsdam University, Germany	8 students, Mediawissenschaft, MA level. 1-hour session. Stage 1.
10	09/02/2018	WTC, Brussels, Belgium	The re-experiment as echo chamber. 8 members of the Algolit group. 6 hours session. Echo Chamber and Stage 2.
11	08-10/05/2018	Master fotografia UPV, Valencia, Spain	A session with the UPV 15 students, MA level. 3 hours session. Stage 1 and an incursion into the exhibition room of Bomba Gens art centre.

Table 3. Overview of the sessions.

During the first phase, I have explored the consequences of transforming the protocol of the original experiment as described at the beginning of this chapter. Sessions 1 to 6 and 9 belong to this phase. The variations on the social ontology of the re-experiment have generated rich and complex dynamics. The details of this phase are the main concern of the next section. What is important for now is the fact that the introduction of a collective description transformed the re-experiment in different ways and challenged my position regarding the re-experiment. During the first year, the general configuration of the re-experiment was relatively stable. There are two core components that are common to all as established during the design of the pilot. The *stimulus component* where the participants describe the stimuli individually, then in groups, then all together in several rounds. And the *feedback component* where they discuss their experience. Depending on the circumstances, the participants were presented with more or less stimuli and they had more or less time to reflect together afterwards. With the observations accumulated through the project's progress, I felt the need to share the process of analysis with the participants. This led to the design of an *echo chamber* for the re-experiment, a session dedicated to listening to the recordings of earlier re-experiments. The analysis is presented in the section “The game of making sense” later in this chapter. But more generally, many insights into the experience of the participants originate from this particular session. In three longer sessions, an additional component was integrated, the *taxonomy* *component* in which the participants were invited to filter and classify their descriptions. The last section of the next chapter builds on the interactions of the participants with the taxonomic device.

The majority of the sessions were attended only by participants, with the exception of sessions 4 and 6, where a few volunteers described the stimuli in front of an audience. The duration of the sessions varied considerably. The shortest one lasted for 15 minutes while the longest lasted for 6 hours. The practice could never have developed without the generous involvement of the participants and their intense contribution.

If the variations bearing on the social ontology have been introduced in all the re-experiments, other variations have been introduced and abandoned over the course of the research. Variations are of different natures and each iteration of the project probes different configurations of the experiment. Over time, I have learned to make an essential distinction. The variations bearing on parameters, and the variations bearing on the structure correspond to critically different forms of engagement. Variations on parameters are variations that provoke differences without altering the experimental paradigm. Extending or restricting the range of presentation times is typically a change of parameters. Certain parameters can stay "dormant" for a few sessions, keeping the same values, and then "activated": i.e., a new set of presentation times is added. When the variation bears on the structure, the object of the re-experiment is also the experimental paradigm itself. It means engaging with the whole relation between experiment and observer, environment and context. When the re-experiment is brought to bear on the paradigm, one runs the risk – and takes the chance – of an alteration. In this case, we touch upon the effervescence of the apparatus, its ability to resolve differently the entities at play. The difference between these two levels of intervention is not as clear cut as it may appear at first. A modification of a parameter can trigger an unexpected interference with the paradigm. The account of practice that unfolds in this chapter documents the gradual realization of the concrete effects of change and a reflection on their nature. Re-experimenting proved to be a difficult exercise. In what follows, I attempt to convey a sense of the complexity of the process and the many deadlocks I have reached. I am trying to avoid the pitfall of giving a retrospective coherence to a process that was devised through trial and error. And I want to convey a sense of the time it took for the re-experiment to acquire its own consistency.

6.2. The temporal unfolding of a session

Until now, I have described the re-experiment through theoretical considerations and through the design of its protocol. It is time to turn to how it concretely unfolded and its experiential and eventful character. This section offers an account of the temporal unfolding of a session. In complement, the reader is invited to consult appendix 1 which offers a fine-grained account of the interaction between a participant and the experimental device. Additionally, in the recordings section of the website, the reader may browse a large selection of recordings of various sessions.

Thus, what happens in a session?

After a brief introduction, the light is dimmed. At that moment, the concentration increases, only a faint murmur can be heard among the participants sitting in the room. Over the course of a few minutes, 5 stimuli are displayed on the screen with various presentation times, so the participants get used to the protocol, the size of the photograph on the screen and the display times. The participants make different breathing noises and sometimes muffled laughter is heard.

Then, each participant describes two images individually while a volunteer or myself (I will use the letter *S*, for scribe, to refer to this person from now on) transcribes the description. The transcription appears on the screen so the participant can check if it is complete and correct. The participant describing the stimulus holds an audio recorder and passes it on to the next when he has finished. The descriptions are usually short, and the description mainly concentrates on sensations of colour and generic objects or situations. The participants' tone is circumspect and defensive. Their words are separated by long silent intervals. The words come slowly, and every statement is carefully pondered. The statement does more than describe the stimulus. The participants test what is expected from them, waiting for a clue from S or the other participants. While the descriptions accumulate, the participants' collective takes shape little by little. The collective is built indirectly. It is in the background noise, the signs of approval or puzzlement, laughter, or minor interventions that one can find the signs of its construction. The correction of the transcript plays an important part in binding the participants together. They watch the screen and intervene in the writing. They guide S' interaction with the keyboard. They sometimes spell the words for S, signal a mistake or remind S of an element of the description he failed to include.

After a round of individual descriptions, the participants describe the stimuli collectively. They form small groups, three people or more depending on the number of contributors to the session. The member of the group describing the photo holds the recorder. The other members tend to come closer to her; there is a small re-organisation of the space and increasing physical proximity. The dynamics change. The descriptions become more comprehensive. Multiple voices are being heard.

The participants tend to echo each other, repeat what the others say, although they are not necessarily in agreement. They deal with multiple versions of what they saw with sometimes a participant trying to speak for the group. They are increasingly talking to each other. This leads to a conversation about the content of the description, but other avenues towards finding common ground are explored. For instance, after having agreed on the transcript, they attempt to guess the presentation time. Guessing the presentation time shows how familiarity with the experimental frame is growing and how they are looking for ways to coordinate themselves with the experimental device's clock. Now, the length of the descriptions contrasts dramatically with the brevity of the first individual depictions. Details abound and contradictory elements co-exist. The participants use the conditional to speak of their impression ("it could have been") or hedges ("I think") and the word "or", which extends potential alternatives rather than negates other versions. They confirm and reinforce previous statements. They repeat statements from other participants when they are missing in the transcript. The description resulting from their interaction gains an increasing presence, to the extent that some of them remark that one risks losing one's own recollection of the stimulus while the others are talking. And they keep the memory of previous descriptions alive, making reference to other photographs already described.

When all the groups have finished, all the participants contribute to the descriptions at the same time. The dialogue is still increasing as well as the complicity between the participants. The participants cultivate an intimacy with the devices (i.e. guessing the presentation time has become a routine) and their familiarity with the set-up (constant back and forth between their dialogue and the correction or the expansion of the transcription). They continue to refine a form of dialogue that allows them to build a common description including various versions. And especially at the end, as the group consolidates, the description process takes a long time and its progression is eventful. The participants are taking more risks and try different forms of depiction to explore their recall. They improvise micro-narratives, they test the affective resonance of the stimulus and they speculate on the use value of the image, its context and its purpose.

6.3. Seeing on the bias

Now that the general dynamics of a session have been laid out, I will approach the re-experiment from another angle. I will examine the intimate relation between the participants and the stimuli before they even describe what they have seen, what happens just before the words are uttered.[72] I

[72] In the stimuli section of the website, the reader may experience different stimuli at various display times.

am now turning to the details of what it means to see in the context of the re-experiment. To do so, I will first go back to the notion of bias as interpreted by the Caltech researchers and redefine it in the light of my observations.

As I mentioned in the analysis of the Caltech experiment, when the researchers are looking for images, they want a visual material that represents the world as it is “commonly seen”[73], by “most people”. In order to avoid the “sampling bias” introduced by the photographer, they downloaded images from a search engine to “average the bias”. It is a strategy of mitigation. Essentially, bias is seen as a factor that may jeopardise the objectivity of their experiment, introduce unwanted and uncontrolled variables. I have already expressed scepticism regarding the real effect of the researchers' strategy. But my scepticism concerned the efficacy of the method rather than the notion of bias itself. Now, I would like to come back to this question from within the re-experiment with fresh insights. The re-experiment puts me in a position where I can attempt to disentangle the question of bias from the researchers' anxiety and explore the relation between image and bias anew. What makes this question resonate differently now that I am in the midst of the re-experimental practice?

To answer this question, let's get closer to what a participant experiences at the onset of the stimulus. She is concentrated, trying to catch the image, but it always takes her by surprise regardless of her degree of concentration. After a few attempts, she understands that her attention needs to be diffuse rather than focused. The more she wants to see something, the more the image's content eludes her. She needs to let the image come to her. To catch, she needs first to be caught. She gradually acquires a sense of the speed of the image display. The sense of a dark animal lying down on a couch suddenly strikes her as a dog, in a kind of after-thought. She doesn't see a dog as much as gets a feeling of canine agility. She is enthralled by a movement of wheels, an intuition of speed, a certain reflection of light, a brightness, a weight. She doesn't know if it was a car, a bus or an airplane. Over the course of the session, she is learning the discipline of opening herself to the flow of images, to accept taking in rather than grasping. And she learns the discipline of converting the apparition of an animal, a car or a landscape into a muscular contraction: do not blink. Feeling surprised at the apparition of what she is on the lookout for, transforming this jolt of energy into pressure and finally converting pressure into a slightly more persistent impression. Negotiating the delays, the afterthoughts. At first only sometimes, and then, progressively more often, succeeding in the coordination of the various stimuli, impulses and decisions. Managing with increasing agility

[73] All the scare quotes in this paragraph indicate expressions used by the Caltech researchers in their report (Fei Fei *et al.*, 2007)

the conversion of sensation into impression.

This fragile process of taking in the fleeting stimulus presents two risks. The first is to keep the most salient elements of the impression, and discard what doesn't emanate with enough clarity. The other risk is to elaborate a description from the faint indices without feeling their weight, to take the looseness of the sensation, the obliqueness of the experience, as a warrant for improvising: “anything goes”. In both cases, what is missing is a sensitivity to the insistence of the lightest sensation. In both cases, what prevails is what psychologists characterize as a cognitive bias (Wilke and Mata, 2012). Pre-conceived ideas and expectations dominate an incipient perception, neutralise the sensitivity to the tangential experience that took place.

I am suggesting here that this subtle process of taking in the fleeting stimulus, when it avoids these two risks and succeeds in giving way to a rich description, also relies on a biasing. But to elaborate this notion further, I need first to reformulate the question of bias that I have uncritically taken as given until now. Bias as understood in the Caltech paper, is posited as an object that pre-exists the encounter between the image and the onlooker. The photographer “has” a bias or the bias “is” encoded in the image. This classical approach to bias posits an object, the photograph, and a subject in front of it. Either the subject is aware of the image's bias or not. Similarly the subject herself can “have” a bias. In both cases, it is a property that predefined entities carry with them prior to the relation. This separation of viewer and image however is challenged through the re-experiment. There is something radically indirect in the visual experience described above. The image seems to turn its back half-way to the viewer. Here, the relation of subject and object cannot be easily taken for granted. The expression “the subject looks at the image” describes poorly what is happening. In such a statement, all the agency is located on the side of the subject. But the re-experiment shows that the condition for the subject to see anything is a deep entanglement with the device. It is to shape one's own body according to a distributed form of agency. Once we accept that the border between subject and stimulus is redrawn, we can say that the image and the viewer are in a biasing relation. The viewer here is more correctly defined as an opening to something that appears whilst cancelling itself. The interstitial presence of the image induces an oblique viewer. A viewer who is letting herself be traversed by what is traversing the image.

What is experienced at the firing of the stimulus is a biasing and the viewer obliquely relates to its interstitial presence. This ephemeral traversal leaves the viewer with a faint, impermanent impression. To avoid cancelling the sensation, the participant must learn to make it more permanent without freezing it, to “not lose sight of the horizon of duration” to borrow Kember and Zylinska's

(2012) felicitous formula. The participant needs to learn how to stretch it without breaking it. The participant needs to learn how to be *on the bias*.

The meaning given to the term bias in the textile industry may help us conjugate the notions of biasing as an oblique movement and the imperative to care about the stretch of a fragile fabric. I am invoking here the meaning of the term as it is used by seamstresses to describe a particular grain of a piece of woven fabric. A piece of garment is composed of perpendicular threads. The grain refers to the orientation of the threads. The straight grain is oriented parallel with the warp threads and the selvedge while the cross grain runs perpendicular to them. Straight and cross grains describe the directions of actual threads. The bias grain or bias is a virtual line that runs at 45 degrees to its warp and weft threads (Wikipedia, no date). To cut a pattern on the bias means to cut a piece whose main seams align with this line of 45 degrees. In an article encouraging its readers to become “grain rebels”, a journalist from Thread Magazine (2008) lists the qualities of a bias cut:

> It has the most stretch and gives fabric a flowing drape over the body. Because of the inherent elasticity of bias, it requires special care in cutting and sewing to utilize the stretch without distorting the fabric.

We are here provided with a vocabulary to complicate and nuance the notion of bias. Bias is what gives fluidity to the cloth and the ability to flow over the body. It provides elasticity, but this elasticity is not infinite, it requires care, it always risks distorting the fabric. Cutting on the bias is a responsible approach to cutting. To perceive on the bias would then mean to embrace the elasticity of bias as constitutive of an oblique perception and take responsibility for the cuts and the distortions that are at risk.

Now let's come back from this detour from the seamstress' workshop to the re-experiment and test what it means concretely. How is the oblique dimension of the flashing image handled? How is the sensation's stretch tested? The answer lies in the details of the interactions.

6.4. Composite echo

Just after the stimulus has been administered, the participants start speaking. During the collective description that ensues, every time a participant explains what she recalls, a mental image is generated. Confused or precise, every contribution elicits mental representations among the participants. The composite echo is the overlay of the multiple images that circulate during the description of a stimulus. It is made of all the mental images present. At first, the participants know very well that this image is as fragile as their own recall. As noted above, the participants tend to present their depiction as versions, prefixing or affixing their statements with “I think” or “my guess is”, extending the speculative image with “or,” suggesting co-existing alternatives. By doing so, they avoid cancelling the fragile echo that takes shape and respect the tangential nature of their encounter with the photograph.

As the participants' own recalls fade over time, the composite echo gains consistency. The composite echo is nevertheless compared regularly by each participant to her own. The composite is the site of high social intensity in the re-experiment. It is the result of a collective work. No participant claims it as her own. It cannot be reduced to an average of the images. Participants paint with small brushes, underlining a limit, focusing on a detail, or shading off, blurring a contour. While creating the composite, they also alter, deform, distort and metamorphose existing characters, spaces or objects. The composite echo can incorporate different interpretations of the stimulus, sometimes even include incompatible elements. It can simultaneously integrate a person that is declared present by one participant and absent by another. Places like airport and stock market are not understood as mutually exclusive. An airport-stock market is a valid instance in the ontology of the composite echo. It has a dreamlike quality. A house with moving parts, a basement morphed into a courtyard, a fire escape ending in a window, a wall made of nineteenth century bricks with a disproportionate amount of cement, all these elements blended together give birth to a building that emanates from a collaborative effort to produce a multi-layered architectural depiction.

What is at stake in the composite echo is more than an overlay of points of view. Stemming from a biasing, the participants' impressions don't have the authority of a stabilised perception. Each attempt at a description seeks an echo in others. And each echo is a trial that tests the elasticity of the sensation that traversed the viewer. How much does another description add to mine? Deform it? The composite echo is the cultivation among the participants of a sensitivity to the stretch. It gradually takes its own identity, acquires its own centre of gravity and, as we will see later, when it consolidates, not everything can be added to it any longer. After a while, some participants will either contest the direction it takes or discard their own recall because the composite doesn't give space for it, or because it is stronger. With time comes the decision to cut, to agree on the fact that the new layers of depiction have lost their sensitivity to the insistent lightness of the sensation.
In a form of closure, the composite echo is marked by the uttering of the precise micro-temporal interval of the stimulus it echoes. “Was it 100 milliseconds? - I'd say 65.” The participants conclude the creation of the composite by guessing the display time, as if they were adding a micro-temporal fingerprint to the fleeting object of their common effort.

6.5. Rhythms, queues and cues

At this point, I have presented an overview of the re-experiments and their contexts and zoomed inside a session to give a first outline of its temporal unfolding. From the bird's eye overview, I have moved to the detailed observation of the act of seeing in the re-experiment. I have analysed what is happening in a glance and the crucial entanglement of the viewer and the apparatus. I need now to show how these two levels of description are related and how seeing is distributed over the temporal unfolding of the re-experiment and progresses through the punctuations that rhythmically structure it.

I have noted how the gradual evolution from an individual annotation to an annotation performed by groups of increasingly larger size influences the dynamics of the re-experiment and how the various actors are enacted and interact. There is a second factor that modulates the unfolding of the re-experiment: the variation of the display times. Every time a stimulus is administered, it is displayed on the screen for a duration varying between 27 to 500 milliseconds. These cuts in a micro-temporal scale constantly reconfigure vision and its subjects. At 27 milliseconds, the stimulus is barely resolved as an image with some participants doubting either the apparatus (“was there an image?”) or their senses (“I blinked”, “I don't compute shape at that speed”). At 27 ms, we are at the limit of the equipment's speed. The refresh rate of the computer that performs the stimulus, the image cache, the system's scheduler, the bandwidth of the cable connected to the projector and the projector's framerate, all need to align and synchronise strictly. If the participant blinks or doesn’t focus on the centre of the screen, the stimulus may very well not even affect her retina. To resolve the stimulus as an image at this lower limit of the apparatus, the minutiae of the jpeg decoder, the conversion from ROM to RAM, the constriction and dilation of the pupils, the secretion of the lacrimal glands are all critical. This represents a moment of high tension and results in the description of intensities. A reddish thing is “going on”, “continuous colours” blend into each other. Every beginning of shape or colour intensity is about to take consistency and ultimately vanishes into formlessness. When the display time reaches the range between 50 to 100 milliseconds, fog dissipates, the descriptions tend to stabilise in the enumeration of generic objects and a sense of the scene (indoors, outdoors) where they find context (e.g. a dining room with chairs, an aircraft departure gate). And around the half second, the participants are able to count the objects, elaborate on the details of the scene and comment on the stimulus as a photograph whose point of view and composition are taken into account.

These observations however must be carefully nuanced as many exceptions arise from the fact that the effort to stabilise the stimuli is intense and must be sustained all along the session. Even if some regularities can be observed, there is no such thing as a neat evolution from 27 to 500 milliseconds. A participant can still at 100 milliseconds blank on a stimulus and in some case already at 27, a group of participants is able to enumerate the chairs in a coffee house, or even at 100 milliseconds be overwhelmed by the intensity of a colour ("a stone on acid"). The faint regularities that correspond to the display times are not a given property of the human visual apparatus that simply respond predictably to specific stimuli. The intense mobilisation of the participants and their engagement with the apparatus are necessary and always at risk of waning or of losing their synchronisation. To be able to absorb the photograph appearing so briskly on the screen, the participants need to be on the look-out, on the watch. A participant needs to focus on the screen at the stimulus' onset. The general rhythm of the re-experiment is given by S who addresses the participant "Okay?" and triggers the stimulus. The practices of synchronisation require a particular form of attention. One prepares oneself, opens oneself to the stimulus, and makes oneself ready to resonate with what is coming. A particular tension is palpable as one needs to be outside oneself, on the look-out, as well as looking inward. The act of seeing, in the re-experiment, is as much an inscape as an outlook. The participant needs to be poised for what is about to come and, immediately after, turning inside, preserving the immediate trace of what has already left.

To make sense of what they saw, the participants use more than what the stimulus provides. They are reading the clues given by the apparatus. As the experimenter, the participants made me realise the subtle signs that were sent to them during the sessions. Which kind of priming was in place? How did it condition the participants' emergent awareness? In the Caltech paper, the authors use the word priming to refer to the act of giving a sample description to the participants before starting the experiment. In this context, priming is related to a textual content, a mental suggestion, an intellectual framing: this is what we expect you to see. In the re-experiment, priming functions by inducing a rhythm, guiding the behaviour and the corporeal disposition, the attention. Each visual shock conditions the body and prepares it for the next. "Ready?" asks S. There is a clock in the re-experiment, the stimulus is administered at regular intervals although not measured by a chronometer. But the priming does not originate solely from the administration of the stimulus. It is distributed over the whole apparatus including the participants. The participants take turns. The session is divided in rounds of description. It is orderly. One can feel one's turn nearing and anticipate it. Taking cues from what the others say, from the repetition of certain categories of images. Even when it is not their turn, the participants keep listening to what the others are saying.

Will I have the time to see? Will it be too fast? Will I see anything? Would I have done better than him or her?

The more the re-experiment unfolds, the more it gives cues about its periodicity and the composition of the collection of photos used for the production of the stimuli. And the more it takes the viewer off-guard as, interspersed in the identifiable series, photos resisting an easy apprehension sporadically make irruption. It is here that I can begin to understand the consequences of my choice of using a subset of categories of ImageNet as a source of stimuli. After having been exposed to a series of photographs from the dataset, the participants "detect" regularities ("again a supermarket", "another pine wood interior"). They understand that the photos belong to a series of limited registers and can make inferences on that basis. But if the categories exhibit regularities and photographs within the same category are somehow homogeneous, this homogeneity is not strict. Sometimes a photo with a very different composition or colour scheme appears in the same category. ImageNet categories contain a certain amount of "outliers" contrasting with other images in the same synset. The dataset exhibits nested variations. It varies from category to category. And within each category, the distribution of photographs is not homogeneous. Therefore, the experience of the participants is already influenced by the organisation of the dataset as it contains strong regularities and intrinsic "contradictions" or surprises. In addition, the step function (the part of the software that selects which photo will be displayed and for how long) gives also an ambiguous signal of repetition and surprise. There are two random functions in the software that affect particularly the priming. One function selects randomly a photograph among those that have not yet been displayed. Another randomly selects a presentation time that has not yet been selected. Once all the display times have been selected, the process starts again. The participants understand after a few descriptions that the amount of categories is limited. They understand that the random choice is made between already familiar categories and this helps them compensate for the lack of visual clues when the stimulus is flashed at a low display time. At that point, the question becomes: from which familiar category will the next image be? This narrows considerably the frame of expectation for the participants and conditions heavily the inferences they make. But this also gives a strong power of disturbance to those images that are not typical of a category as they confuse the representation of the dataset the participants have built over time.

At this point we have already several elements that help us understand that the participants are involved in something more than a visual response. They are involved in the resolution of different scales. They are not merely viewers punctually responding to a stimulus, they are actively part of multiple processes of synchronisation and anticipation. They resolve the difference between the dataset's scale and the micro-temporal dimension of the stimulus, the scalar difference between the categories' regularities and the various clocks of the re-experiment.

6.6. Embodying the scales

As we see now, to resolve a scale is a very embodied form of engagement. In the paper that documents the Caltech experiment, the subjects' bodies are hardly mentioned except for their eyes. In the protocol, one learns that they were seated and were typing on a keyboard. From the Caltech report, one can only form a vague idea of a person's body. As the organiser of the re-experiment, part of my role was to make sure the material conditions for the proper conduct of the re-experiment were met. In collaboration with the institutions and groups hosting the sessions, I had to take care of the participants' bodies. The participants needed to be comfortably seated. I brought some food and drinks for the longer sessions. I have been evaluating how much they could concentrate, planned a break and, as customary at TPG, informed them about the location of the fire exit and toilets. But, even if I had to make sure the logistics were properly taken care of, once the re-experiment had started, I primarily focused on their words and the textual content of their descriptions.

However, as I was transcribing their descriptions, the headphones over my ears, my attention shifted from the verbal content to the aural quality of the sessions. I was becoming increasingly intimate with the group's voices, intonations, with the participants' rhythm, their hesitations and impatiences. By scrolling back and forth through the timeline of my audioplayer, I could locate the zones in the audio channel corresponding to a moment of silence or the peaks provoked by the coughing or a deep exhalation. As I was typing up their words, I came to the realization that the participants were breathing. Observing the participants in situ, I hadn't noticed how much the breathing manifests itself in the act of describing the recall. Participants cough, inhale and exhale loudly, keep their mouths open during moments of silence, laugh and even scream while they are trying to remember a fugitive sensation.

The recall comes from a ventral practice, and from the lungs. The bodies do not move much, as the participants are seated. But a lot of air goes in and out of the body. Listening to the recording of a re-enactment, the soundtrack is punctuated by the air movements, the rhythms of the laughter and the bursts of anxious sighs. The participants hold their breath when looking at the stimulus and exhale multiple times when recalling. While the time of the computer clock measures milliseconds of display time on the screen, the body incorporates rhythmically the flashing image into a series of episodes of inhaling and exhaling. The eye saccades not only affect the retina, they also contract and expand the lungs.

How could I have missed it? And what made me notice it? The answer lies partly in the process of transcribing and the technology used for it. The slowness of my typing, the intimacy with the sound provided by the headphones, the visualisation of the audio timeline coincided to make me feel with more than one sense, with my ears, fingers and eyes, that there was an incorporation of the micro-temporality. Far from being an experience of disembodiment, an extraction of a temporality that doesn't belong to the body, the re-experiment was producing a strong embodiment of the fugitive sensation. Looking back at the report of the Caltech experiment, it became clear that my understanding of the temporality of the experiment had begun to diverge. Asking the participants to describe aloud had made it impossible to take into account exclusively the machine time of the millisecond. The inhalations and exhalations, the laughing and the coughing are articulated with the infinitesimal interstices of display time. The perception time was indeed sequenced by the clock of the computer but was also anticipated by the inhalation. And the recall was accompanied by the exhalation. The session was punctuated by the display of images as it was by the coughing. Perception time was sequenced by the lungs as much as by the eye saccades.

The attention to the breathing took me further. The voice had emphasised the inter-relation of rhythms in the experiment. An attention to the sound had made me sensitive to ways in which the measurement of the millisecond was itself taken in the larger sequencing of the session. It made me attentive to how the flashing was anticipated, how the cumulative views of the images created expectations, how the repetition of the breathing, viewing, describing, coughing, laughing, viewing and describing again created their own flows and their own cuts. The body was redrawn: a breathing eye, attached to the lungs, not only to the brain, negotiating the discrete time of the machine with the longer loops of an expanded rhythm. Calling for a reformulation of the experiment's temporality in terms of rhythm rather than as a separate collection of instants defined by punctual display times. This reformulation was bringing up very practical questions. What is, in the existing experimental device, an instrument for compressing and expanding perception time and recall? In an interview with Rick Dolphijn and Iris van der Tuin (2012), Barad proposes thinking of time as an entity that is not given, that needs to be continuously re-articulated and synchronised through a variety of practices. In this perspective, the breathing, the laughter, the inhalations and exhalations are practices of synchronisation that re-articulate constantly the temporality of the re-experiment as much as the micro-temporal visual jolts to which the participants are exposed. The re-experiment engages with different practices of time, their articulation and synchronisation as well as their disjunction, interference, conflict. Breathing in, seeing out, seeing in, breathing out. I mentioned earlier that the viewer was defined as an opening in a relation that traverses him. In a nearly literal sense, to view is to inhale the stimulus, to breath it in.

To attend to the breathing, the waiting, and the priming requires a different form of listening than the decoding of the contents of the conversation. It requires paying attention to the nuances of silence and faint noises. A continuous susurration of “hms” and “yeahs” encourages the speaker. These fragments of sound repeatedly renew an implicit encouragement: “I am listening” or sometimes “I agree”. But besides these indirect indications, the participants' attention is hard to “locate” with certainty. There is an inherent difficulty in hearing the act of listening, its nuances and its depth. Perhaps the listening can be heard in the voice that speaks. If the speaker doesn't sound anxious to finish, if she doesn't permanently check if she is allowed to continue, if she sounds authorised, part of her confidence comes from the fact that others are paying attention. The listening of others is an amplifier, it is present in all the voices that speak. Listening is also an exercise in self-backgrounding to offer the foreground, the floor, to somebody else, another voice. To hear the self-effacement, the self-backgrounding, the attention given to somebody is a nearly impossible task. Only indirectly can one hear attention. Yet this dimension is crucial. If it is not present, the course of the re-experiment is dramatically altered.

6.7. The game of making sense

As stated earlier, the more I was transcribing and taking notes, the more I felt the need to reconnect the analysis with the participants. The more I felt the need to bring the collective dimension of the research into the listening itself. I didn't want to reproduce the schema of the original experiment where the participants are treated as mere description providers and are subsequently disconnected from the analysis. This led me to propose to the participants to join me in a session dedicated to listening to past sessions' recordings. The session was called an echo chamber to stress the multiple layers of repetition that such an exercise would involve and the potential for distortion that it entails.

Another element that motivated the creation of an echo chamber is that I was hitting the limits of my individual listening. The motivation to find a way to listen with others was not only to propose to others to do what I was doing at home in front of my sound editor, but to produce other sounds to produce another listening. I was looking for an active listening that assumes that listening is transformative and productive. And to produce these echoes, I needed another room. A room where the affects produced by the re-hearing could be embodied, re-experienced as much as re-experimented with.

In some ways, the echo chamber evokes what sociology calls member checking, a process by which a researcher goes back to her informants to validate her findings (Turner and Coen, 2008). Yet it diverges from member checking in that member checking considers that validation can happen in a neutral space where both members and the researcher can reflect at a distance. The echo chamber is on the contrary an extension of the re-experiment, a different iteration of a process of intra-action. Contrarily to member checking, in the echo chamber the object of the study has not ended but continues unfolding through the listening.

Practically the echo chamber sessions consisted in shared moments of listening to the recordings of previous re-experiments. During these sessions, the participants and myself engaged in different modes of aural attunement, reading transcripts aloud, replaying fragments of recordings or browsing through the archived descriptions. The detail of these sessions is documented in Appendix 2 and a section of the website gives access to the material used during these sessions. The principal objective of these moments of shared listening was to find ways to be affected by what was recorded, to identify the moments where affect surges. The voice became an instrument to listen and to read the dynamics of the sessions. Some of the participants had contributed to earlier sessions, others did not. A key technique we used was to let another participant's voice inhabit our own, to let the voice travel among the group. The exercises in the echo chamber helped to concentrate on the dynamics and the tensions at play in the work of responding to the stimuli.

One important take away distilled from these sessions is a better understanding of some implicit rules emerging from the implication of the participants in the re-experimental apparatus. A complex set of rules regulating the hierarchy of statements is encapsulated in what the participants describe as the “game of making sense”, a competition between “those who perceive well”. In the competition, none of the rules are stated explicitly. They are nevertheless experienced by the participants. The competition is not scripted in advance. It emerges out of the use of the experimental device. In this game, all statements do not have the same value. A statement about a blob of colour or a general shape has less value than the expression “counter with two people”. The closer one gets to a certain level of description, the more value it is granted. If a hybrid like an airport-restaurant-stock market is indeed accepted, it is not the case for all potential associations and transformations. As a participant reports, once somebody has described the scene at a certain level of precision, it is hard to contribute at a more general level. The contribution at a different level is considered redundant or irrelevant. This process of levelling that combines levels of description and grades of value is at work in the regular course of the re-experiment. But on certain occasions, the re-experiment heats up and the dynamics change. An example of this moment of inflection can be found in a recording where one participant breaks this implicit rule and states he was not able to go further than "shapes and colours", the most general category of description. After the participant made this declaration, the conversation collapses. This statement is more than a mere remark. After it is made, the participants stop adding to the description. The statement has an ambiguous consequence. At the same time, it downplays the participant's own contribution, and it simultaneously discredits the description that was collectively elaborated before the participant spoke. The composite echo is fragile and must be collaboratively supported either by adding to it or by staying silent. Negativity when expressed has an impact on the dynamics. Therefore, certain statements have more value than others but their value is volatile and they may lose their value when somebody withdraws his trust and questions, even implicitly, the level of description the others implicitly had agreed upon. The compatibility of levels is what defines exchangeability between descriptions. The composite echo doesn't accommodate different contributions if they are not at the same level or growing in precision. An indoor space can transform into a cafeteria, and a cafeteria can transform into a baggage reclaim area. But the cafeteria cannot be brought back to the level of an indoor space without endangering the composite echo and triggering complex negotiations.

The collective production of the composite echo is a site of high investment for the participants who have to negotiate tensions and figure out the right resolution (understood as the ratio between the level of description and grade of values) of the apparatus they are part of. The game of making sense has stakes for the participants. As a participant who considers herself as someone for whom "it's super difficult to see anything" explains, the re-experiment is stressful because one is supposed to see "at least something". There are different kinds of participants: those already mentioned for whom it's difficult and the others "who see a lot" and who are competing with each other. Those who do not see, then, are facing a three-pronged alternative: shutting up, making up ("am I going to invent something?"), or going with the flow ("when people then start confirming things that you just hook on to something that's something vague corresponding I can imagine myself doing it"). Whatever the option chosen, to participate feels risky. This dimension of risk is not apparent in the descriptions transcribed by S. It is through listening and repeating, remembering and being affected that the tensions inherent to the game have surfaced.

6.8. Conclusion / transition

In the previous chapter, the variations were introduced as a means to provoke more than a local change: to probe the effervescence of the experiment, to take advantage of its inherent indeterminacy. Now we have more elements to understand the impact of a variation in practice and the nature of the changes it enacts. To bring all the participants into the same room and to ask them to respond to the stimulus in groups increasingly larger provokes a change that reverberates across the different registers of the re-experiment. The dynamics that lead to their production brings about a change in the social ontology of the re-experiment. The table below shows side by side the defining features of the social ontology of the experiment and their transformation when the variation discussed in this chapter is introduced.

Caltech experiment / AMT	**Re-experiment**
The social as a collection of bare individuals	Composite echo
Readiness of perception and availability of the world	Thickness of mediation
Perception as collection of fixations	Perceiving on the bias
Consensus through averaging	Game of making sense, resolution
Static knowledge	Evolving knowledge
Basic knowledge	Large array of competences and skills pertaining to the apparatus and the re-experiment itself.
Separability through supervision	Dialogue, listening in the echo chamber

Table 4. Social ontology, comparison.

The participants, when they are not treated as social atoms, are responding to each other and are seeking consensus in a composite echo. The descriptions produced together through the composite echo are considerably longer and exhibit a different texture than those produced alone in front of a screen. But the length of the descriptions and their texture are only a few of the elements that change. The social porosity introduced by the variation facilitates the circulation of speech. The dynamics of the composite echo contrast with the dynamics of the experiment where no description has the power to affect another. The composite echo reveals the thickness of the mediation process. The “readiness to perceive” expected from the subjects cannot be taken for granted. The world represented through photographs cannot be simply understood as available for experience without coming to terms with the participants' entanglement with the apparatus. In the report of the Caltech experiment, perception is conceived as a collection of independent fixations. The subject is defined as applying previously learned categories onto the visual input automatically. In the re-experiment, the participants perceive on the bias. Bias is understood here as a complex relation to the cut that attempts to adjust to an horizon of duration. In opposition to a mechanistic understanding, bias is here defined as temporal, embodied and rhythmical. In the social ontology of the experiment, consensus was obtained through the averaging of outputs. In the re-experiment, the composite echo provides the ground for a process of negotiation that the participants described as a game of making sense. The game of making sense makes room for disagreement and adjustments. It generates discussions and learning rather than a tally of responses where the parties do not learn anything from their differences.

The variation opens the possibility for a dynamics where the knowledge of the participants evolves and encourages them to share it. Their knowledge is not limited to the response they produce, it extends to the apparatus and the re-experiment itself. The large array of competences and skills displayed by the participants reveal the complexity of a task that in the re-experiment is presented as purely mechanical. This knowledge is not limited to their performance during the re-experiment itself. The role of moments of reflexivity as well as the participants' involvement in the echo chamber sessions demonstrate their interest in making sense of the process they contributed to. And without their reflexive contribution, many aspects of the game of making sense would not have been noticed.

After this outline of the effects of the variation on the social ontology of the experiment, it is worth examining in further detail some of them in particular. The re-experiment introduced a different definition of bias. Bias is not understood here so much in terms of position and prejudice, of aperception, of applying previous knowledge. It is construed as the negotiation of a horizon of duration. In the chapter, the emphasis is not placed on a subject observing representations of the world in a detached manner. Instead, the subject is deeply intertwined with the micro-temporal apparatus at work in the re-experiment. My analysis emphasizes the dynamic nature of bias. The participant is not a subject that is easily triggered towards a predictable response nor an independent viewer unaffected by the situation. The re-experiment complicates the definition of bias as it shows that it cannot be treated as a mere mechanism of encoding meaning but needs to be understood as a process of priming the embodied emergence of the objects that will be available for perception. This is an understanding of bias that is rhythmical and embodied. Bias here operates at the inferior threshold of meaning-making, when semiosis is barely in formation. To understand bias as a deliberative process where an individual exerts her judgement in the abstract leads to a truncated view on the problem. The re-experiment forces to understand bias in a scalar and rhythmical perspective. Judgement does not happen in a who but in a when. Bias needs to be approached not merely in objects, subjects and their properties, but in the apparatuses, the cues, the speeds and the scales they enact. I propose to think about bias in relation to specific embodiments of scale and speed. And in the case of the Caltech experiment, I propose to relate bias to an embodiment compatible with the scale and speed of the annotation environment.

The more the participants are able to compose with the apparatus and as a group, the more they complexify their relation to the stimuli. A process of learning takes place in the re-experiment that goes against the reductive understanding of perception as a response mechanically bound to a stimulus. As the division of labour changes in the re-experiment, as the participants are treated differently than simple instruments, their relation to bias changes. They explore bias and try to figure out the consequences of what they discover. At a perceptual level, the problem of the participants is to accompany the sensation, let the stimulus come to them in order to describe it appropriately. But their problem is ultimately to take responsibility for the cuts and the distortions that are at risk. This has importance for the participants who acquire competences through the process. And the very fact that they acquire competences is of uttermost importance for the larger questions of this study. The participants engage in a process where they learn and develop a common responsibility towards the cuts they make. Their engagement in a learning process and their growing sense of responsibility show that selected variations in the apparatus have an effect and that the social ontology of the experiment changes in response to these variations. This gives us a whole new vocabulary to approach these changes and the relational texture that gives rise to them. It gives a series of coordinates to apprehend the readily available objects of cognitive psychology, the relational principles underlying the Caltech experiment and the social ontology. As none of these are simply exposed in the form of instructions or explicit methodological guidelines, the practical engagement and the affective and reflexive collaboration of the participants are crucial to identify them in devices, practices, techniques and rhythms.

Now that the active role of the participants in the production of knowledge in (and about) the re-experiment has been examined, it is worth highlighting the aural dimension of their interaction in the re-experiment. As the participants describe orally, a recording device is introduced. The audio recorder brings the vocal dimension of the experiment to the fore as well as the affect circulating through it. As their voices leave their grain on the recordings, the embodied engagement of the participant makes itself felt. They are on the look out. They are breathing heavily, they are bursting into laughter, their speech accelerates or decelerates. Affect in the re-experiment is more than traces in recordings, something that emanates from the enaction of the subjects in the apparatus. It is also a method mobilised by the listeners to make sense of what has been recorded. As I wanted the participants to be able to intervene in the different stages of the process, I invited them to listen with me to the recordings. In these listening sessions, the listeners are affectively attuning to the recordings.

The affective listening produces an account of temporality that exceeds by far what could be grasped from the Caltech report. Indeed the micro-temporal interval is a key temporal structure, but it doesn't operate in isolation from other rhythms. The eye saccade is not isolated but taken in other durations and accelerations. For instance, repetition brings a temporal course that takes the subject into a larger timeframe. The micro temporal flash enacts a specific subject of vision and so does the temporality induced by the dataset and the repetition of certain categories of photos that produce a priming of the subject. The subject is on the look out for the next hit on her retina, she is at the same time expecting gradually the return of certain categories and patterns. Waiting, breathing, saccading, finding repeating patterns are all embodied techniques that add to the temporal depth of the re-experiment and leave traces on the recordings.

Paying attention to the textured temporality of the re-experiment leads to a consideration of the introduction of ImageNet. Beyond a resource for the stimuli, ImageNet in the re-experiment introduces a scale and a sense of repetition. The stimuli are not self-contained units to be perceived in isolation. As the re-experiment progresses, there seems to be an infinite number of stimuli whose patterns give a sense of coherence and variation to the participants' visual experience. When the stimulus hits their retina, they explore at the same time the density of an infinitesimal scale of visual perception and the depth of the ImageNet collection. They perform a resolution of scale that articulates the large number of stimuli, the perceived depth of the image's pool and the infinitesimal interval of perception to capture the discrete stimuli. The viewing subjects are enacted through the micro-temporal cut of the stimuli and through moving across a potential choice of photographs between ImageNet's billions of photos.

The listening as it has been performed with the participants traverses a large affective spectrum which doesn't preclude analytical attention. If the affective listening revealed the embodiment and the affective engagement of the participants in resolving a scale, it also revealed implicit rules in what a participant called the game of making sense. These rules did not concern primarily the content of the descriptions. They concerned the level of description and the degree of details it contains. And in return, they established a hierarchy between those who, according to these criteria, see well and those who don't. If the composite echo is collaborative and additive, there is also a normative dimension to it. This dimension would have been missed if I had not spent a session listening to the recordings with the participants. It is worth stressing again that to exclude the participants from a given stage of the process is to preclude a certain knowledge from being produced. Changing the social ontology of the experiment changes what can be learned from it. It is by attuning to the surge of affect, the pressures and the silences in the recordings that this grey protocol of the re-experiment could be addressed. But it also could be addressed because we, the listeners of the echo chamber, didn't turn a blind eye to the details of the semantics. In the listening practice, affect was not seen as a radical rupture from other forms of cognition. In the practice, the affective and the analytical were experienced more as different moments in a spectrum than incompatible registers. The scale of value inherent to the composite echo was at the same time affectively active and rationalised as an element of reference according to which the participants evaluated their contribution.

The presence of this scale of value indicates that even if the re-experiment follows a divergent path to the experiment, it doesn't necessarily lead to a perfect utopia or an ideal world where the social ontology can dispense with all relations of power. There are emergent exclusions and descriptions implicitly evaluated against a scale. But at least the re-experiment, by including the participants at later stages of the process, has the potential to address and listen to its own limitation. It takes one step towards making the process more account-able.

Chapter 7. Alignments

The previous chapter has provided an account of practice gravitating around the complex enactment of the subjects of vision and its relation to the social ontology of the experiment. In this chapter, I am making a complementary move and concentrate on the agency of the different devices and alignments at work. Here the photographic elaboration of the re-experiment is discussed through an analysis of the performative dimension of the apparatus. The chapter is organised in two parts. The first revolves around the stabilisation and destabilisation of photographic alignments. It concentrates on the effervescence of the apparatus, its ability to perform different forms of inclusion and exclusion. As the effervescence redraws the boundaries of the re-experiment, the chapter reflects on its potential of change and the constant work required for keeping it within spatial and temporal limits. The second part of the chapter is dedicated to the analysis of a device especially active in the stabilisation of the re-experiment, the taxonomy. The taxonomy is posited as a device that cuts into more than words. Its role in enforcing the organisation of labour is revealed through the examination of the tense interactions with the participants. I conclude the chapter reflecting on the tensions arising from the resistance of the participants to the taxonomic device. These tensions are a sign that a *thinking with* the participants is occurring together with the alteration of the social ontology and the effervescence of the apparatus.

7.1. Alignments in the re-experiment

In the previous chapter, I observed that the value of a description corresponds to some extent to a scale where descriptions gain in granularity. But this observation had to be amended. I ended considering that correlating the value of a statement to a certain level of precision is not an absolute rule. Its value could be contested and trust could be withdrawn. In what follows, I will go further and see how the whole apparatus, where it begins and ends, is at stake in the attribution of value to certain descriptions.

To give some ground to launch the discussion of the intricate role of the devices and participants, I will start with an exchange that happened during a session hosted by TPG. Four people are responding to a stimulus that has been displayed for 53 ms. They suppose that the scene they identify with relative certainty is a shop. They cannot confirm whether there is more than one person in the scene. They mention the potential presence of two elements in particular: a brown paper bag and an object identified as Captain America's shield.

Until one of them mentioned the presence of Captain America in the stimulus, the participants were debating over a brown bag and a human presence or absence, evaluating their performance with criteria such as accuracy or exhaustiveness. The participants were behaving as if they were in a lab, contributing to an experiment where the key competence is to provide accurate descriptions. But, as soon as an element from mainstream commercial entertainment enters the conversation, a spike of affect is noticeable. When the participant hesitates and then claims to have seen Captain America in the stimulus, he behaves as a player making a risky move in a game. In the debriefing after the session, a participant states that having seen Captain America defines the subject as someone who watches “this kind of movie”.[74] Another participant listening to the exchange suggests that the person who saw the comic book character is “ashamed in this gallery context”. This remark interprets the “shame”, the hesitation, the “confessional tone” of the subject as a surge of affect that indicates the presence of another element in the room: the gallery. It foregrounds the presence of the institution that, in the minds of the listeners, stands for a set of cultural values. Recognizing an element in a picture may therefore become a disadvantage, or at least has become less self-evident. Another set of criteria to evaluate the descriptions comes to the fore. Somehow, the complexity of the exercise increases. It is not only the participant's ability to detect precisely an element that makes her a good player in the “game of making sense”. She becomes more than a perceiving retina, more than a generic subject perceiving generic objects in standard settings with a unique scale of values. Something is changing in the configuration of the re-experiment. The course of recognition of the re-experiment has evolved: to recognise means to be recognised.

To understand the nature of the switch happening here, it is necessary to come back to the notion of alignment presented in the last chapter. The eruption of affect and the change of rules that ensues are intimately related to the photographic alignments at play in the re-experiment. To understand what this means, it is worth differentiating formally the alignments present in this fragment. I will name the first alignment, the “composite echo” alignment. I have described above the effort of stabilisation of the stimulus into an aural image retaining some of its vibration but gradually limiting the proliferation of statements by levelling them out. Through this process, a series of objects and presences are taking consistency. In this case, the brown bag as an object provides an anchor for a scene that involves a commercial transaction in a small shop. The brown bag is on a counter, the human presence spotted nearby is “filling a bag”. This description is not just an

[74] The scare quotes in this section refer to comments made by the participants listening of the recording of the exchange during the echo chamber session. See **Appendix 2**, for more details.

immediate response stemming from a stimulus. It is the adjustment of a series of parameters. The viewer is enacted as a participant and shown a scene with objects featuring basic categories[75] (bag, counter, person). The description is calibrated to respond to the general lab-like circumstances as the white walls and the general silence in the room attest. Everything is made, from the systematic way the stimulus is administered to the removal of any visual or auditive distraction, to enact the participant as a focused entity concentrated on the control of her eye saccades. All these parameters and entities mutually enable each other and are necessary for the stimulus to be resolved as "visual data" as a participant puts it. The term *alignment* emphasises the solidarity of all these elements and the fact that they gain a specific consistency. It names the emergence of a specific distribution of agency between subjects and objects. This alignment with different nuances remains relatively stable during the course of the re-experiment. It is important to note that stability doesn't mean uniformity. On the contrary, as I have already mentioned earlier, the re-experiment evolves during a session. With the changes from an individual response to the stimulus to collective responses of larger groups, the dynamics transform. Another parameter that provokes noticeable change is the increase or decrease of presentation times. When the stimulus is shown for 27 or 43 ms, there are obvious differences in how the participants react to the stimulus compared to longer display times like 500 ms. In one case the stimulus barely stabilises into an evocation of shapes and colours whose contours are evoked with great difficulty. In another, the stimulus consolidates into a reference where objects can be counted and enacted as a photograph with a point of view, a frame and even a context of circulation ("an E-bay photograph"). At one end of the scale, the participants try to adjust to the energy of the stimulus acting as a vibrating source, whilst at the other end, they are constituted as deliberating subjects equipped with opinions and even taste. If there are contrasts and differences across the micro-temporal scale, there is also an enduring sense of continuity. The re-experiment navigates a scale that helps it maintain its coherence even if it crosses various thresholds. These differences concern the variability of the apparatus. An alignment is what holds these differences together, providing a sense of integrity across differences.

Nonetheless, the emergence of an alignment in an apparatus is never fully pre-determined, and some ranges of display times have a particular influence over the re-experiment. Display times between 53 and 100 milliseconds in particular open the apparatus to changes of another nature. If at both ends of the scale (27 and 500 ms), the alignment stabilises the stimulus more easily either as a raw vibration or as a well-formed image, in the 53-100 range, the resolution oscillates between an intensity of colours about to give way to distinct shapes and a representation that does not

[75] Basic categories, a term coined by cognitive psychologist Eleanor Rosch (1970), refer to entry-level categories in a classification. They are the concepts used in daily life when a description is made in a general context. A speaker will say, have you seen the *cat*? rather than *tabby cat* or *carnivorous animal*.

consolidate fully. At that moment, the stability of the alignment is weakened and the effervescence of the apparatus increases. To understand this, let's return to our example and consider the emergence of a second alignment. The group was describing a brown bag and a counter in a shop when suddenly a participant mentions a logo, the Captain America logo. I will name this alignment the “logo” alignment. Let's take a moment to measure how much this alignment differs from the previous one. This time, the description revolving around an object that grounds a scene is disrupted by the mention of a logo. Perhaps the most striking difference can be located in the architecture of sight of the re-experiment. When the participants are describing the nuances of the brown bag and the counter, their eyes are concentrated on the screen. In the exchange about the presence of Captain America, the participants turn their eyes to each other. The word shame, as used by a participant to describe the hesitation of the person who makes the Captain America statement, doesn't denote so much embarrassment than what Martin Jay defines as an inversion of perception and a reification of the viewer (Jay, 1993, p.289).[76] To perceive is to be perceived with all the risks it entails. Before making his statement, the participant is measuring that risk and hesitates. Yet this surge of affect, this moment of arrest is quickly superseded by a sense of triumph when the participant decides to utter the statement. And, later, when the participants, at the end of the session, look at the photographs without time constraints and compare them with the descriptions, his interpretation is confirmed. Everybody can see Captain America's logo on the tee-shirt of a person on the photograph. More than the fact that he saw the photo correctly, there is a shared pleasure among the participants to have brought the comic character into “this Gallery context”. They celebrate the fact that one of them dared to make the statement. And the participant who dared increases his visibility among his fellow participants. The course of recognition of the re-experiment has changed: to be recognised as the participant who saw this kind of movie has another value than being a provider of accurate descriptions.

It is worth insisting here on the fact that the shift of alignment that occurs does not depend on the sole mention of the name Captain America on the premises of TPG. This contrast alone would be anecdotal. To notice for instance the logo of Captain America on the tee-shirt of a person in a photograph on the media wall of the Gallery wouldn't have produced the same effect. When the stimulus is flashed on the screen, the participants try to resolve it as a coherent image with a spatial organisation. The participant has seen the logo, but he doesn't know how to relate it to the other elements. It is neither in the foreground nor in the background. He cannot resolve the stimulus as a coherent scene. He juggles with different visual patches that do not fit together. The logo shatters

[76] “Shame is the recognition of the fact that I am indeed that Object and that the Other is looking at and judging.” (Jay, 1993, p.289)

the scene, it is not perceived as integrated but alien to it. It refuses its coherence to the nascent image. To mention the comic book hero is an operation that works at multiple levels. It does speak to the context of the cultural institution, but as importantly, it works against the process of stabilisation of the stimulus that is already under way.

7.2. Diverging alignments

What happens here goes beyond a daring move made by a participant. It shows a difficulty for the apparatus to fully resolve the photograph as stimulus or as icon. This denotes something different than the variability of the apparatus. The surge of affect, the abrupt change of scale of value, the inversion of the gaze are the signs of something more profound than a variation. They are the signs of the effervescence of the apparatus, its ability to resolve differently the entities at play. Indeed, the Captain America event does not affect the stimulus alone. All the aligned entities are affected and resolved differently. The architecture of sight in the re-experiment is remade, the participants move their chairs to look at each other, they turn their heads and their shoulders. The silence is broken, and the white walls of the lab become the walls of the white cube. The boundaries of the re-experiment change. In one case the re-experiment is circumscribed to the limits of a staged lab situation, in the other, the external limit of the re-experiment, the gallery that hosts it, reverberates through the apparatus. The logo alignment branches out to reach the Marvel Comics character and pulls in the photographic institution.

As shown in this contrasted example, alignments vary in reach as the apparatus doesn't end at the same place. In one case, it reaches out to the Gallery. In the other, it closes itself upon the staged lab environment. Although both alignments are deployed within a similar presentation time, their horizons of duration vastly diverge. The brown bag, although stemming from a fugitive sensation, gradually unfolds into a description that gains in depth and has the time to mature. It is aligned with the same kind of photographs that have been shown repetitively. Captain America corresponds to a shock and an arrest, a punctuation that locks the viewer in a discrete instant and breaks from the monotonous set of photographs expected by the viewer. This shift of alignment points to the complex situation which the re-experiment has to deal with. The re-experiment is at the same time engaging with a computer vision experiment (Caltech in 2007) while being an event taking place within an institution of photography (TPG in 2018). The documents handed out to the participants testify to the complex spatial enfolding of the re-experiment that the Captain America's irruption exposes. When they enter the room, they are given a folio containing the guidelines of the

University's ethic committee. And when they leave it, they are handed the Gallery's standard feedback form to evaluate the activity. They enter the room as subjects and leave it as visitors.[77]

Such an alignment shift in the re-experiment shows how the central question of the original experiment is reformulated. In the case of the Caltech experiment, the problem is framed in terms of subjects, stimuli and referents: a subject can describe an object at a given presentation time to a certain degree of precision. In the re-experiment what becomes clear at this stage is that the question is not to study what the subject sees but how, by being enacted as seeing subjects, the participants contribute to the stabilisation of an alignment or to its bifurcation. The re-experiment makes the alignment shift available to experience. There is no place for it in the Caltech experiment. In the original experiment, all the differences are interpreted as different levels of descriptions, as variability of the apparatus. In contrast, the re-experiment makes room to account for the effervescence of the apparatus, not only for its variability.

7.3. The resolution of the apparatus

Where the apparatus ends is a question that never ceases to be raised. As we have seen with the participant who risked the Captain America “move”, the definition of the limit of the re-experiment greatly impacts the course of recognition of the session. How will a participant be recognised when she recognises an item in the stimulus? To recognise as we have seen is to be recognised as well. And the terms of this recognition change according to where the participants are: in a computer vision lab or in a cultural institution? If this affects the participant's dynamics, it has consequences for the resolution of the apparatus as well: what it is able to produce in terms of granularity and level of description. Two more examples may help to grasp how the contours of the re-experiment are moving and affects what can be recognised and how.

In a session conducted with a group of students, it becomes very quickly apparent that the participants do not display the same intense attention that others have shown in previous iterations of the project. The participants resist the re-experiment. They do not do it overtly. The difference with other groups lies in their lack of listening to others. As I have written previously, in almost all

[77] Even in the documentation of the session, the two alignments seem to have been recorded in separate registers. The participants who follow the session through the recordings react to the surge of affect following the utterance of the words Captain America. Whilst another group that tracks the session through the transcript treats the brown bag as the central thread. The alignment shift continues to find an echo in different recording devices. And the colours in the stimulus continue to flicker while the listeners follow the session unfolding three months after it took place at TPG. The tension remains unresolved. The bifurcation is re-activated and amplified through the listening.

iterations of the re-experiment, the participants, when they are not asked to respond to the stimulus, are observing others. When S makes a mistake while transcribing a participant's response, it is not rare for a participant who is waiting for her turn, to correct typos or gaps in the transcript. The participants actively listen, they nod, they offer support to those whose turn has come to respond. In this session, the participants respond to the stimuli, but those who are waiting are not paying attention. They are only present to the re-experiment when it is their turn to respond. There is a continuous murmur, a diffuse distraction. The stimulus is resolved as an assignment. The lack of attention is a constant reminder that the alignment doesn't end at the staged lab. By modulating intensities, establishing distances, changing the aural texture of the background, the participants are redrawing the limits of the re-experiment. They are changing the background's texture, rendering it unconducive to concentration. We are at school. The subjects are not resolved as participants but as students. And as such, they must be careful not to display too much zeal to comply with the assignment. The lack of attention, whilst not being hostile to the re-experiment, provokes a form of restraint. And this restraint affects in turn the content of the descriptions which turn out to be much shorter, less granular, more generic. However, even if this tendency is noticeable, it doesn't mean all the participants play the game the same way. Some of them, sometimes unpredictably, engage in the game of making sense, inflecting the course of recognition in another direction. And the participants are not the only ones to have agency in the process. The various rhythms and surprises produced by the administration of the stimuli may pull the participants back in the premises of the lab environment, changing the borders once again. These dynamics should not be interpreted as a pre-given structure, the Class in a School System, taking over the re-experiment, but as a class discovering itself iteratively through the re-experiment, the students exploring their relations by taking distances, lowering their involvement, monitoring one another. Exploring also their differences, as some engage as experimental subjects and others as students. And testing the limits of their agency when they are pulled in, surprised, or interpellated by the various devices at work in the session.

The distribution of speech and silence, the protocols that authorise speech are subject to interferences which produce a zone of indeterminacy where the site of the re-experiment is in flux. Through the re-experiment, the lab and the photography institution (or, in this case, the institution of higher education collaborating to the project) are constantly making and un-making each other. They may converge and reinforce each other. But it is important here to realise that they are not two entities meeting each other but multiplicities. There isn't a speech of the lab and a speech from the school. There are modulations of speech coming from both sides which may align and the enfolding of the two spaces may consolidate or they can interfere and the alignments diverge.

Silence here should be considered as more than non-speech. I have mentioned earlier that the listening was difficult to hear. The silence of the listeners was heard in the confidence of the voice that speaks. With this example, we see again how silence is active. But in this case, the silence is not the silence of subjects refraining from speaking but the silences that emanate from certain spaces. In this sense, silence can be understood here as connector between spaces. There are few privileged spaces where silence can be heard. It requires material conditions of isolation, of distancing. The lab, the photographic institution, the class room have all their ranges of silences, their distributions of speech. And the silences carry with them a certain load. It is only in spaces where silence can be made that mumbling is heard and quickly gets noticed and acquires meaning. In this sense, the silence that opens up the lab to the classroom invites the classroom's soft chatter. The silence invites practices of resistance from one place to enter another.

Moving from one place to another, or better said changing the spatial enfolding of the re-experiment, silence also inscribes the re-experiment in a different temporal sequence. When the re-experiment's space is fielded as a classroom what comes before and what comes after are other educational activities. The chatter and its interference in the silence re-aligns the re-experiment with the school programme and schedule.

The limits of the re-experiment are moving, spatially as different sites are making and unmaking each other and temporally, as the re-experiment is given a before and an after. The distribution of speech and silence are loading the background and reconfiguring the course of the re-experimentation through seemingly minor cuts. This brings me to a second example where such a form of reconfiguration again affects the resolution of the re-experiment.

In most of the re-experiments, the tempo is set by S who administers the stimulus. S makes sure that the participant is ready to receive the stimulus and the participants answer with a positive sign. After a participant has finished, S checks with her if the transcript is complete then turns to the next participant. The questions uttered by S (“ok?”, “ready?”) are markers that isolate a contribution from another. These expressions mark the limits of a segment of interaction with the next. There are a whole range of nuances involved in the skill of marking the end of a sequence and the beginning of another. They give a sense of closure to a participant and help the next to concentrate on the task. Marking the end of a description implies checking if the participant is comfortable with the transcript, giving extra time for an eventual development of the description. The silence after a participant's description, the attention given to the participant, even after she seems to have finished, sometimes encourages the participant to say more. It is especially important when the participant is making the description alone and tests how long a good answer should be, or when the participant tries to figure out S' expectations. Having played the role of S many times, I have come to realise the decisive importance of letting a moment of silence grow after the description, letting the silence weigh on the participant to sense if there was something more to be said. This moment of extra silence either leads to a clear sign from the participant that her contribution is over, or to additional observations that have not yet coalesced into clearly defined contours. When the descriptions are collective, S intervenes regularly to check if she understood well the different contributions (“Are you happy with that?”), especially if many participants are speaking together. All these interventions distribute the descriptions in sequence; they reinforce the relations between the descriptions and the participants. It is a subtle work of decomposition and composition, of marking sequences and units, addressing agents.

During a session organised with The Photographers’ Gallery's team, this mode of sequencing is altered by the participants. The work of articulation performed by S is re-doubled by two participants (I will name them for convenience's sake A and B). When a participant is engaged in a description, A (or B) punctuates the sentences by approving laughter, signs of validation or surprise. When a sequence of description is about to end, A intervenes (“you two finished?”) before S has the time to ask for confirmation. Or A repeats the instructions given by S. Where the students of the previous example were retrieving in the background behind a murmur, here two participants intervene actively in the marking of the sequences and the turn-taking. A and B introduce a secondary temporal articulation in the re-experiment. They duplicate S' position and emphasize S' role. By doing so, they add legibility to S' active presence. S' position of coordination becomes more apparent as two entities occupy the same position.

A and B mediate between S and the participants. The participants talk to S directly, but his interventions are repeated by A and B. This redoubling of the ending of the sequences, and at times their anticipation, have the effect of limiting the expansion of the descriptions. It is a work of containment. It doesn't allow time for the silences to expand. The “waiting a bit too long for comfort” that helps undecided participants to give a chance to an incipient sensation, this moment that makes the difference for daring to say something that is on the tip of the tongue, is minimized if not suppressed. The intervention the redoubling occupies, fills the interstice, preventing the still fragile parts of the description from being uttered. At times, A or B even attempt to accelerate the process (“What's next?”) forcing the participants to keep moving on.

Again here, these interventions are met with partial resistance. A participant may insist and continue to elaborate a response and engage with the uncertainty of the sensation the stimulus has provoked. Or the dynamics of the collective description may cancel these interventions, the majority of the participants ignoring them or not even hearing them anymore. To which A and B may also respond with more direct interventions ("now you are projecting").

A and B's work is a soft calibration of the participants' involvement. And as in the previous example, it contains the flow of description, it limits its granularity. To a member of the team who shows enthusiasm and a great ability to catch the gist of the stimuli in fine-grained descriptions, A says: "perfect you've got a new job". The remark made with a tone of gentle irony is nevertheless telling. The competences the participant displays in the re-experiment are not the same as those expected from him in his current job. The difference is highlighted, and the border re-drawn with a smile. The limits of the alignment, involvement, competences and the grade of resolution come together. And highlight the power that resides in the role of coordination performed by S.

Again here, this analysis should not be interpreted as a pre-given structure, the Workplace, taking over the re-experiment. But the course of recognition of the re-experiment gives signs of challenging a border. Competences are exposed and exposing. Recognising too much makes you the candidate for a new job. Depending on where you are, what qualifies can disqualify. By increasing the levels of description or resisting the pressure to shorten the time given to less articulated responses, the participants are moving between the lab and the workplace. This movement is constrained by the calibration of silences and the emphasis on the sequences' limits. Hardening temporal limits within the re-experiment, some participants enforce its border. In this last example, the workplace has stakes in what counts as photography and photography-related competences. The re-experiment is used not only to explore computer vision but to differentially figure out what count as photographic competences. And to figure out where one is, in the lab or at work.

7.4. Taxonomic devices

Over the course of the chapter, I have outlined different levels of agency. The participants are more than mere "reaction machines", they create composite echoes, negotiate consensus, figure out scales, make risky moves in a game of making sense, they intervene in the coordination, measure their involvement, they reflect on what happened and contribute to the account of the re-experiment.

They experiment with a new body that is enacted through the various rhythms and intervals of the apparatus. They see on the bias. Their lungs, voices or ears become instruments to attune to the micro-time of the re-experiment, or to resonate with its recorded archive. They are much more than black boxes producing generic descriptions. However, their agency is not absolute. They are traversed by different speeds, they are taken into accelerations, surprises. They are cued, forced to anticipate and take risks. Their competences arise from the intra-actions with the apparatus. Alignments hold them in place. Their relation with re-experimental apparatus is one of mutual enactment. Central to this mutual enactment is the resolution of the photograph. The photograph is not resolved in isolation. The whole apparatus is mobilised for its resolution. Rhythms, retinas, lungs, flashing light, screen and keyboard need constant re-synchronisation, repair, attunement and coordination to form alignments. Technologies and techniques of the body are engaged in the stabilisation effort. The stakes are high for the participants. Competences, consensus, trust arise from it. It affects the course of recognition, what can be recognised in the stimulus and how the participants are recognised in return. And the resolution opens a space for the participants' strategies.

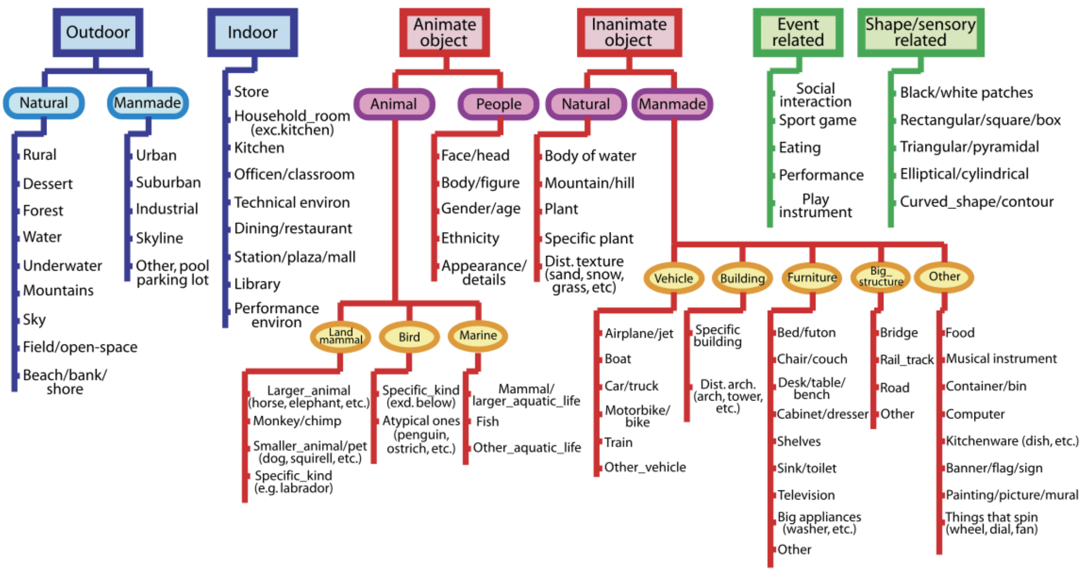


Illustration 6: The controlled vocabulary used by the Caltech researchers

As importantly, the resolution of the photograph moves the limits of the re-experiment. Where its alignments begin and where they end is an open game. There are many stakes in holding in place an alignment. When the contours of the apparatus are changing, so are competences, positions, forms of involvement and scales of values. The work of stabilisation performed within the re-experiment concerns stimuli, subjects, echoes and descriptions.

The work of stabilisation can also be performed post-hoc and from without. This is the role given to the taxonomy in the Caltech experiment. Yet as the place where the apparatus ends is not a given and can be contested, the exteriority of the taxonomic device is in itself a question. To study how the taxonomic device operates in the experiment and its relation of exteriority to it, I have proposed to different groups of participants to apply the taxonomy of the Caltech experiment on the descriptions produced during the research. These sessions correspond to what I have called above the filtering stage. In this last section of the chapter, I set out to analyse the filtering stage. In addition further details of the protocol of these sessions as well as a description of the sessions' dynamics are available in Appendix 3.

In the Caltech experiment, as we have seen, the classification is a way to transform the responses, the descriptions as they were written by the subjects, into perceptions, where they are reduced to the terms of a taxonomy. This operation had an important consequence: the subjects were unable to participate in the process of interpretation of the results. To challenge the social ontology of the experiment and engage the participants in the whole process, during these sessions, I have asked them to classify the content of the descriptions made previously by other participants. In particular, I have asked them to mark the words of the descriptions corresponding to an entry in the taxonomy differently than those that were not included in the taxonomy. The participants produced lists of terms that were validated by the taxonomy and lists of rejected terms. Examples of the results of the process can be found in Appendix 4. Facing these two lists, the participants for the first time are confronted with the grid of intelligibility the Caltech researchers have used to filter the process. This exercise provokes a strong reaction of shock among the participants. This feeling of shock resonates with the shifting limits of the re-experiment. At this new stage, the relations of the participants and entities involved are being dramatically altered.

The distinction the Caltech researchers made between response and perception is being felt as profoundly misleading. It cancels the possibility of keeping track of the process of classification and filtering which was at work earlier in the process. And by doing so, it denies this competence to the participants involved in the description stage. Here we touch upon a critical aspect of the compartmentalisation of the alignment. The separation between the process of description of the stimuli and the taxonomic filtering seem to imply that classification happens a posteriori and is a distinct process. Whilst the filtering stage contrasts with the previous stage, if we follow the descriptions from their inception to the filtering process, much of the work of classification could be interpreted as a continuity rather than a rupture in the re-experiment. The filtering stage constitutes a second wave of filtering, an intensification of a process that had already started. To paraphrase a participant, in the filtering stage, the hierarchies of the classification tree are applied onto the “hierarchies in the participants' heads”. The taxonomy is not a device applying a structure over a supposedly raw input. In the example discussed above, Captain America or brown bag implicitly brought in the re-experiment different taxonomic trees. On another register, the question “is this an AirBnB picture” that came repeatedly in the longer sessions, also refers to categories that inform the way the picture is taken and understood. Finally, as I explained earlier, the process of levelling prevents certain descriptions from being made, already filtering out the responses.

However, if the process of sieving and levelling has started earlier in the re-experiment, it is worth noting that the process of elimination increases significantly when it reaches the filtering stage. The attributes list is limited and radically excludes entire segments of the descriptions. For instance, there is no category to hold temporal data, in contrast to the interest showed by the participants when they give sophisticated indications about the historical condition of the objects they identify (e.g. “a contemporary object that is designed to look like something made in the seventies”). Neither is there any place for irony or double meaning. All relational elements are severely sieved out: pronouns, articles, most of the verbs, but also indications of point of view or affect. There is no place for the words that describe the stimulus as a photograph: terms like frame, perspective or composition cannot find their place in the tree. Not to mention the photograph as an element that circulates on a platform (“an E-bay picture”). Moreover, the filtering makes the re-experiment tone-deaf: the nuances and precautions used by the participants to soften the affirmative character of their statements (“maybe a cup”, “a tree, I think”, “the car would be next to the van”), are ignored and the statements interpreted as straightforward assertions. Finally, the expressions that convey the intensity of a sensation are flattened out (“a reddish thing going on” is mapped to “Colours and shapes”).

And yet, there is more to the classification stage than elimination. Remarkably, the participants do not complain much about the limited size of the vocabulary. Most of the resistance to the device focus on the architecture of the taxonomy. The engagement with the taxonomic device reveals a much more complex operation than just a sieving of the statements. It is the confrontation, capture, folding, or even deflection of a system of levelling against/into/to another. As I just explained, the previous stage does not consist of raw responses to the stimuli. There is a game going on. And the descriptions are already evaluated by the participants when they produce them according to a level of precision. They are already levelling. The lively exchanges have already been regularized by the conversation in the previous stage, the discussions have already been condensed in brief descriptions. There was already in the description stage a taxonomic process at work. Reaching a consensus on the scales and levels of descriptions affected the granularity of the apparatus. But this process was implicit, there was no object representing an organised structure according to which a statement would be accepted or rejected. The difference between this stage and earlier ones is that the participants are facing an imperative to classify whereas, before, classification and filtering were ongoing processes that were organically handled.[78]

What happens at this stage also goes along with the stabilisation of the photograph. It is here again necessary to insist on the fact that the compartmentalisation of the Caltech experiment in separate stages implies that the problem pertaining to the stabilisation of the stimulus was a problem of the description phase. At the filtering stage, the researchers do not mention any display time for the photograph used as a control.[79] Yet while re-experimenting, the stabilisation of the photograph remains an issue. We can hear the changing agency of the photograph in the reactions of some participants engaged in the classification exercise. A participant says “I want to do it through the text”. By this, the participant means that, when the participants are classifying the descriptions, they have a printed copy of the stimulus photograph next to the written description. And this, for him, implies an equivalence between the printed photo and the stimulus. And on this basis, the printed photo serves as a reference for the evaluation. To “do it through the text”, without looking at the reference photo, is to resist this equivalence. The stimulus and the printed copy are two distinct forms of stabilisation of the photograph pertaining to two different alignments. Those who produced the description were part of an apparatus where the stimulus was enacted in a very different way and where the display of the photograph elicited a different horizon of duration. Another participant adds “we see more than the people who made the description”. His remark does not stem from a purely formal concern. He points to a change in the conditions of reciprocity and to an asymmetry.

Changing the stabilisation of the photograph re-orients the epistemic compass of the experiment. To understand better the problem, it is worth examining the difficulty faced by a participant trying to code the expression “man with white shirt”. The participant wonders if he can place “white shirt” in the *man* category[80] (see illustration 7). This requires asserting that the shirt is “on that person”

[78] Moments when participants “suspend” their criticism. At times they even embrace it literally. The process becomes a game where the artificiality of the classification is pushed to its limits. There is a form of humour and an irony in taking the system at its word, respecting its principles at the letter. To classify becomes the discovery of a simplified version of the universe. Exploring a process of dumbing down as a participant puts it. There is a seduction in letting go, lowering one's vigilance against the reduction of the descriptions' meaning. Yet these moments of embrace are short.

[79] Control here refers to the use of a printed version of the photograph that was used as a stimulus to check at a later stage if a description had correctly described the scene.

[80] This would locate shirt in the “appearance of a person” category rather than in the “manmade object” category. See **Appendix 3** for the details of this discussion.

which is not verifiable from the description alone. As another participant rightfully observes, the description mentions a man with white shirt without further detail. Assuming that the description meant that the person was wearing the shirt is an appeal to common sense. If the person is said to be with a shirt, in the absence of any other information, one can assume the person is wearing it. But to assume that the description obeys the rules of common sense is to take a leap of faith. In the previous sessions, the participants have experienced first-hand how the delusive stimulus produced dissociated images. The relations between objects were shaken by a patchwork of impressions corresponding to a stimulus. Nothing is more fragile and elusive than a “with” in such a description. The shirt could be worn by the person but “with” could allude to the proximity of the person and the shirt just as well, or simply indicate that the viewer perceived them together without any ground to establish their relation. The inference that “man + with + shirt” corresponds to the activity of wearing the shirt relates to a competence of the participant. To perform a classification, the participant is assumed to be able to make this kind of reflection in the absence of detailed instructions. But the re-experiment has left a scar on the confidence in commonsensical assumptions. The participants face a double problem: to match “their” common sense with the classification AND doubt their common sense because of their experience of previous sessions. This leads us to the agency of the photograph when it is enacted as a control. The shirt cannot be made the “appearance/detail” of a person because the only way to infer that the person wears the shirt is to refer to the current position of the person doing the assessment and who sees the printed photograph without time constraint. The complex and uncertain relations exhibited by the objects and the shapes detected at the stimulus stage are hardly reducible to the hierarchical relations of the taxonomy.

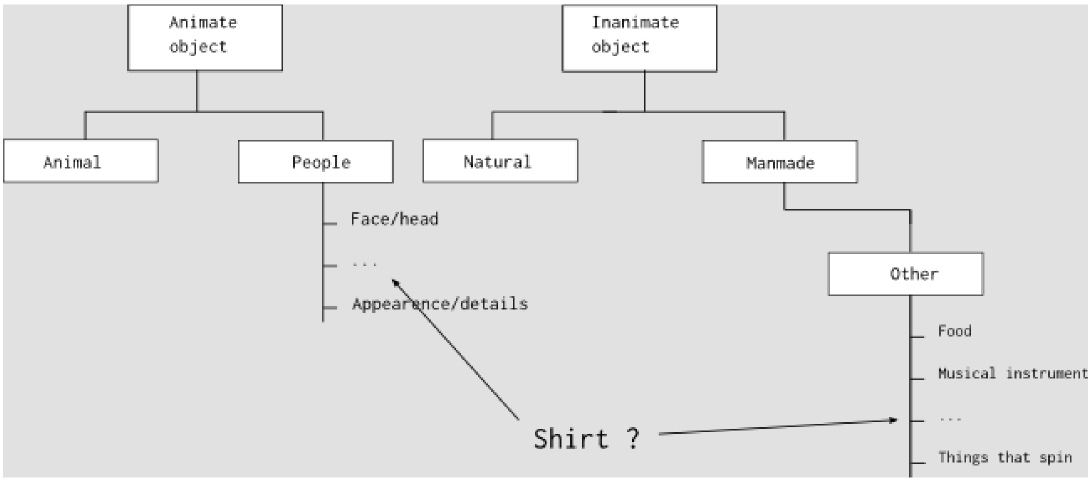


Illustration 7: Is shirt related to the appearance of a person or a manmade object?

Furthermore, the agency of the photograph has dramatically changed. In the previous stage, enacted as a stimulus, briefly present on a screen, the photo was enabling a seeing on the bias, provoking a composite echo where statements were proliferating. In the current stage, enacted as a printed reference, always available to the taxonomist, the photograph becomes a "control", a document used to resolve the indeterminacy of the descriptions. The participant contests the effect of the separation between a description stage and a filtering stage because it implies that the photograph travels from one stage to the other as an unchanged object pointing towards an unchanged referent. The control, the participants say, cannot be used to disambiguate the description. The description does not refer to the photograph, it refers to the photograph-enacted-as-stimulus. And in this case, "wearing" is not a relation that makes more sense than any other. When the photograph is enacted as a stimulus, the relation between man and shirt is very often simply indeterminate. When the participant says "I want to do it through the text", he resists the conflation of a resolution of the photograph with another. And by doing so, he resists what the separation of stages seems to imply: classification is separated from description, perception is not response and the photograph travels unchanged.

All the above reflections lead to the understanding of the taxonomy as an apparatus that performs cuts in more than words. It redefines the contours of the actors in the re-experiment and enables or disables agencies. The first cut lies obviously, in its structure, in the partitioning of the world it operates. Outdoor, Indoor, Animate object, Inanimate object, Event related, Shape/sensory related are disjunctive entries at the top of the tree. Beings and objects are distributed among hierarchies of kinds. This makes certain descriptions available for coding and others not. It defines what is filtered out and what is kept.[81]

The second cut enacted by the taxonomy enables certain actors to become judges over the content. They are in a position where they are expected to decide. A taxonomy implies decision and therefore power over the future of a description. A power of selection is given to the participants. Yet this power is conditional to the use of the taxonomy and a certain resolution of the photograph. This is where the participants show the most resistance. What is at stake for them is to maintain and defend the common ground of experience, to keep the experience they shared with other participants and their common set of references, the common protocol over which they implicitly agreed at the description stage. Removing all the relational elements from the description is not a simple question of filtering noise, it is removing the traces of the patient effort of finding an echo in each other's descriptions. With the classification process, the tacit contract that underpinned the game of making sense is being re-written. The participants do not oppose the imperative of deciding

[81] See **Appendix 3** for a sample of the descriptions filtered by the participants according to the taxonomy.

the fate of the descriptions out of a purism that seeks to preserve the entirety of the subjects' words. “I think”, “maybe” and “perhaps” are not mere decoration but the signs of an intense collaboration. Additionally, the participants refuse to position themselves according to a set of corporeal coordinates prescribed by the classification device. Their refusal to use the photograph as a referent to interpret the description is not so much a refusal to be “influenced” by the picture, but to adopt a position that alters the nature of the relation between the participant and the device. To do it “without seeing the picture” means to refuse to use a frontal seeing to evaluate an oblique glance. Their resistance is their response to the re-alignment that the classification operates, changing, more than words, the set of relations that hold together, enact and confer agency to subjects, viewers, photographs, levels and hierarchies. As we have seen in the previous chapter, the re-experiment's apparatus was never fully stabilised, and its indeterminacy was resolved in divergent alignments. At this stage, the taxonomy vigorously holds in place the apparatus and re-aligns firmly the entities at play. As the changes of alignments cannot be taken in charge by the taxonomic device, what is filtered out is the apparatus' effervescence.

The participants do not oppose the classification exercise per se, they are struggling against the taxonomy's lack of sensitivity to the alignment and its lack of sensitivity to the performative character of the apparatus. The roughness of the classification scheme, its absence of sophistication, are not just lack or absence. They are also actively defining the competence they expect from the subject. The participants complain about the work of filtering. The object of their complaint is not the fact that the exercise is too complicated, too abstract or too demanding. On the contrary, it is the fact that it only requires a bare minimum of attention. A sentence like “It is a middle-aged guy, bald in a good mood.” is reduced to “head, appearance”. “Woman in a heavily institutionalised context” becomes “woman, interior”. A participant calls this a process of “dumbing down”. Entries in the classification are there because they are the lowest common denominator among interchangeable subjects rather than the results of a sophisticated organisation of knowledge. At the description stage, the common denominator was the highest denominator obtained through a highly invested process of levelling. At the filtering stage, the common denominator is the lowest, replaced by a generic grid that requires dis-involvement. The lowest common denominator responds to an imperative of management: the management of the subjects that are averaged and can always be replaced. The taxonomy is a storage device adapted to a knowledge that de-motivates the subject. It enacts a dumb subject, not an inherently stupid subject but a subject whose interest is deemed irrelevant to the process, the antithesis of a participant to the game of making sense. Whereas, at the description stage, the participant's intelligence was mobilised, at the filtering stage it is dumbed. Dumbness holds the participant at a distance. But distance here does not correspond so much to a condition for an objective classification. Distance here means simply that the participant should not care too much. Distance means the opposite of involvement. The filtering stage is not the only place where we encounter the calibration of the participant's involvement. I have commented upon other mechanisms that imposed restraint like the change of background texture with the group of students or the redoubling of the coordination work with TPG's team. But these attempts at restraining involvement emerged as resistance from within the process of description and they were originating from the participants' strategies. Here they are built into the apparatus. Filtering is therefore not solely a semantic affair. It is part and parcel of the management strategies that hold the participants in place and at a distance.

7.5. Conclusion / transition

The previous chapter concentrated on the re-experiment at the level of the participants, how the subjects emerge and what this means for the larger apparatus in return. Here we made the complementary move and focused on another level of understanding: the alignments. The alignments are performing the work of stabilisation of the entities at play. They are the level at which the dance of stability and effervescence can be studied. Effervescence is the moment at which the re-experiment switches between alignments in a way that redefines its borders. Typically its “where and when” are at stake when alignments are redrawn. Through several concrete examples I have shown how the changes occurring in alignments are more than mere experimental variations, how they bear on the very definition of what is happening in the re-experiment, its social ontology and the scales of values and competences.

Interestingly by paying attention to the details of the interactions of the sessions, the elements that provoke change in alignments are not necessarily dramatic or obvious. They can correspond to very vocal surges of affect as in the Captain America's example. But crucial changes that pertain to the paradigm can also be provoked by discrete modulations of rhythms. It is by modulating the murmurs in the background or by intervening in the silences that follow a description that the contours of a re-experiment are being challenged. By affecting the tempo, the general coordinates of the session are under variation. The alignment is at the same time a site of brittleness and solidity. This allows for the continuity (dozens of re-experiments have taken place in a rather coherent fashion) and effervescence (each session in its own way extends the re-experiment's repertoire). As the fielding of the re-experiment shifts, so does the resolution of the photograph. As the re-experiment extends to reach Captain America, it brings the Gallery into the premises of the re-experiment. The resolution of the photograph changes from stimulus to commercial logo and back, flipping the subjects' positions and redirecting their gaze towards each other. In another example, the photograph is resolved alternatively as stimulus and as assignment, creating a friction between laboratory and the classroom. In this sense the resolution of the photograph involves far more than vision and the deciphering of content, it branches into tempos, saccades and spatial continuities. Every time the stimuli flashes onto a retina, the question asked is not only “what do you see?”, it is simultaneously “where are we?” and “when are we?”. And every time the subject enacted at this occasion may be attached differently to the other entities in the alignment, it can be taken in a different space and time, it is enacted in a field that opens a different potential.

Taking part in the re-experiments gave me first-hand experience on the performative dimension of the apparatus and the strength and effervescence of the alignments. The stabilisation and the continuity is the result of much work distributed over a large chain of agents active inside the re-experimental device. However, other devices are tasked to effectuate this work of stabilisation from without. This is the role given to the taxonomy.

If the classification of the descriptions can be seen at first as a stage of semantic filtering, re-experimenting this part of the Caltech experiment revealed that the taxonomy is performing cuts in more than words. When the participants engage with the taxonomy, there is a realisation that the process through which they were already performing a classification and a levelling, the composite echo, is being taken over. The reason this matters to them is that the composite echo is a site of exchange through which the participants define the rules of their collaboration. The taxonomy by being tone-deaf simply removes the texture of the descriptions in which the participants' positions were engrained. The careful phrasing elaborated through finding an echo in each other's words is replaced by a radically limited set of concepts. Furthermore, as seen in the example of the white shirt, problems arise when a description produced as a response to a stimulus containing an exploded array of shapes and objects must be mapped onto a representation of the world that expresses relations as fixed properties. A description where “white shirt” and “person” may coexist in an indeterminate relation can hardly be mapped onto a taxonomy where shirts are properties of a person's appearance without raising difficult questions. Is the ontology of the exploded stimulus at all reducible to the taxonomy's worldview? Is it possible to assume that the stimulus is related by a common referent to the “control”, the printed photograph used by the person who performs the classification? And that this referent is the photograph's content remaining untouched when moving from the lab to the room where the filtering is made? To do it “through the text” means to refuse an interpretation of the words that would imply that the description can be assessed without taking into account the alignment from which the description emerged and the irreducibility of one alignment to another. If the alignment shift between the photograph as stimulus and the photograph as control cannot be taken in charge in the classification, the descriptions are resolved as mere labels to be passively matched to other labels in a grid, a process of dumbing down.

I would like to close this account of practice by pondering the question of involvement and intensity that was raised once more with the classification. If to think with the participants[82] is what is at stake in - and is the condition of - the re-experiment, the taxonomic device can be considered as the most difficult challenge I have faced. What impeded the “thinking with” was not so much the content of the classification per se as the feeling of dumbness and de-motivation it generates. Yet, the fact that the participants expressed their frustration with the taxonomy is in itself an accomplishment. There is all through the research a sense of engagement. To think with is not something one can take for granted. It is an invitation that can be accepted or refused. There have been different grades and modalities of acceptance. Moreover, the two terms of this invitation have themselves been reformulated. To think has not been a purely mental, rational, subject-centred affair. Thinking has happened through priming, seeing on the bias, composing echoes, attunement, waiting, backgrounding. And the *with* has been simultaneously a *with whom* and a *with what*. With other participants, various devices, alignments, the enactments have been varied and the forms of relations have evolved across scales.

When the participants resist a “dumbing process”, they make clear that the question at stake at the filtering stage is not limited to the fate of the descriptions and which words are selected in the end. It is the larger question of what will be made with the knowledge accumulated along the re-experiments. By showing resistance to the classification process, the participants take responsibility for what mattered during the previous stages and affirm that the taxonomy cannot take charge of what constitutes the richness of what they experienced when they responded to the invitation to *think with*.

The fact that the participants take responsibility for what matters to them in the process does not concern the participants alone. It has a high importance for identifying an horizon of change in computer vision. To open up an horizon of change doesn't mean to find an alternative, a ready to use replacement, that would solve machine vision's problems. To open up an horizon means to do the work of clearing up the channels along with changes can flow. It shows that an experiment studying early vision is not doomed to produce subjects that merely respond to stimuli. It shows that the fate

[82] As I proposed in the introduction of this chapter quoting Vincianne Despret (2009).

of such experiment is not to fatally reproduce a subject compatible with – or even optimised for – the scale of the machine vision industry. It shows how it can enact a subject in divergence with it. This doesn't mean outside of it. In the re-experiment, the participants are still engaged with early vision, and they are still operating within a micro-time scale.

The account of practice running across this chapter and the previous one offer a different set of conceptual tools to rethink the photographic elaboration of computer vision. More than in the contents of datasets, this study has looked into rhythms, scales and how they are embodied. It has located the enforcement of the division of labour in unexpected places like the taxonomic device or the repetition of the instructions of the experimentalist. Thinking with alignments rather than the readily available objects of experimental psychology allowed to account for the expansion or the shrinking of the borders of the re-experiment. Listening to the surges of affect, the silences or the frustration expressed by the participants regarding the controlled vocabulary has been crucial to sense that the experiment can do more than what was originally stated by the Caltech experimentalists. The account of practice offers a new vocabulary and a sensibility to the performative dimension of the apparatus. This is of uttermost importance as the social ontology permeating the experiment (and the environment of annotation) is never explicitly stated. On the contrary, it is deeply entrenched in the material configuration of the experiment, its various technical devices, the limits of its alignment, the rhythms that traverse it, the various temporal registers it synchronises and the many resolutions of the photograph it enacts.

By changing parameters, introducing variations, the re-experiment allowed to understand better the nature of the changes that may occur in such environment. To make alignment switches available to experience makes it possible to understand how certain differences make a difference and how other only add a grade of variability. All differences are not equal and some are worth resisting the course of the re-experiment. In a sense, the reaction of the participants anticipates the double challenge that awaits me in the final chapter: to duly recognize the nature of the knowledge generated through the re-experiments in collaboration with the participants, and to identify what can be done with this knowledge and where it can be taken. This means how it can be used to read back the annotation environment analysed in the second chapter with another set of concepts and an enhanced sensibility. And how it can be used to discern more lucidly the changes and conditions that may open up an horizon of change for computer vision.

Chapter 8. Conclusion

This study began with two vignettes: the hollowed code presented by Jeff Bezos and a psychophysics experiment. The research journey led from one to the other. It has traced a pathway connecting the environments of annotation of machine learning, the psychology lab and the photographic institution. The research mainly functioned in three modalities: first, a mapping of the terrain whose objective was to understand how computer vision operates through a series of detours and displacements, secondly the development of a practice, the re-experiment engaging with the role of the experimental practice of computer scientists in resolving computer vision's scale and thirdly a theorization of the entities, alignments and apparatuses involved. If the practice has been the focus of the research, it wouldn't have been possible without an exploration of the network of operation of computer vision, a rethinking of its objects and the method to engage with them. Together, the mapping, the theorization and the re-experimentation have furthered the understanding of the photographic elaboration of computer vision. Re-experimenting has been challenging, difficult and rich. It simultaneously produced a form of knowledge and disseminated it. It has been informed by theory and asked questions to theory. It found its place through following the displacements of computer vision. In this last chapter, I will make the journey back to the environment of annotation with new insights. In these pages, I am reflecting on the kind of knowledge that has been produced and how it can help rethink and renew the photographic elaboration of computer vision. However, to ground these reflections, it may be useful to start with a review of the itinerary I have followed in the previous chapters.

8.1. Itinerary

Bridging the semantic gap, computer scientists are opening up the digital photograph, once considered a black box, to algorithmic scrutiny. In computer science papers and in public platforms where tech companies announce their products, the photograph is portrayed as a passive object mined by an active algorithm. Photography is said to be transformed by algorithmic agents. “No more relying on tags” (Garun, 2017), “Google Photo can tag your pictures for you” (Rosenberg, 2013). Many authors in media studies or cultural theory seize on these recent developments to investigate anew the technological context of photography (Lister, 2007; Hoelzl and Marie, 2013; Lehmuskallio, 2016). Whilst the thesis acknowledges that photography as a discipline needs to take the developments of computer vision seriously, it stresses that the influence is mutual. In response to many accounts of the technological “framing” of photography, I am proposing here to study the photographic elaboration of computer vision: how the mediation of photography transforms computer vision from the inside. To understand the importance of photographic practices and methods at work in computer vision, one needs to make first a double move: computer vision must be considered as a discipline that is outsourcing the production of the domain knowledge it relies upon and photography must be understood as a mediation process rather than a limited cast of readily legible objects (the photo and the camera) and actors (the photographer and the spectator)

In the second chapter, I proceeded with the analysis of one of the largest datasets to date, ImageNet. Through the discussion of ImageNet, I have adopted an expanded approach to computer vision that included a whole network of practices and methods. Algorithms do not act alone (Cox, 2017). De-centring the understanding of computer science from code is a pre-requisite to understanding where photographic processes inform the elaboration of visual algorithms. As Wendy Chun puts it, code is not source, it is a re-source; programs and algorithms pertain to larger horizons of programmability (Chun, 2011). Code needs to be made relative (McKenzie, 2006), “down to earth” (Bogost, in Jaton, 2017). If computer vision is the invention of human vision as susceptible to translation, this invention doesn't happen in front of a text editor only. It is made through the production of objects like the dataset, an obligatory passage point; it is made in multiple sites as the environment of annotation and the lab of cognitive psychology. As importantly it requires labour, not just pure mathematical speculation. And the workers who annotate the datasets contribute to the epistemic production of the models machines are learning from.

To understand the active contribution of photography to computer vision, a same movement of expansion must be made. Photography must be de-centred from an understanding that too closely limits itself to the photograph, the photographer and the camera. It is again a network of practices and methods that needs to be explored. We have to avoid a replacement trap where computer vision provides a new camera and a new kind of photograph, where the same old objects are “upgraded”. It is necessary to concentrate instead on the different alignments, cuts and methods that are mobilised in the processes of photographic mediation as they actively take part to the invention of vision as susceptible to translation. Photography can therefore be understood as more than a mere provider of visual material (photographs), but as a set of active processes and methods constantly instantiating computer vision's objects and subjects in different ways. The question then is: how do photographic practices intervene, orient, the invention of vision in terms that make it amenable to computer vision?

Having made this double move, we have on one hand photography enacted within a process of mediation, and on the other the annotators actively involved in those processes in order to produce computer vision models. I have called this articulation the photographic elaboration of computer vision. To study this articulation the thesis starts from the hypothesis that the photographic elaboration of computer vision is experimentally tested in the labs of cognitive psychology, an environment to which too little critical attention is given. The experimental practice of computer vision and its alliance with cognitive psychology are enacting the objects and the relations needed to conceptualise the model of vision of artificial vision and to configure the settings of the annotation work. This leads me to inquire into the experimental construction of vision in the work of Fei Fei Li. The Caltech experiment is studied precisely for its focus on the articulations of the scales necessary to coordinate vision, annotation, semantics and large image collections. Using as a theoretical frame science and technology studies and agential realism, my analysis of the experiment concludes that what is probed in the lab is not only a model of vision. What allows me to connect the experiment to the photographic elaboration of computer vision is not just that concepts and ideas move from the lab to the environment of annotation. It is that the apparatus, or what I call here photographic alignment, is translated. Furthermore, what is produced in the lab is the managerial template for the annotation environment that regulates the subjects and their epistemic contribution.

At this juncture, one must resist however approaching the experimental work of computer vision as the original site of knowledge production explaining mechanically the model of vision used by engineers. To see the lab as a site where a blueprint is created and subsequently applied in the environments of production is reductive. The relation between the experiment, the crowdsourcing platform and the model of vision is one of a fluid genealogy, at the same time more intimate and less constraining than a linearly causal one. The relation can be best understood as one of assonance and resonance with the model of vision underpinning the apparatus of industrial annotation. The cadence of the Amazon Mechanical Turk may be understood as a match to the speed enacted in the model of early vision. There is convergence between the imperative of productivity and the glancing annotator. If Li has never asserted a direct causal relation between the experiment and the methods of annotation implemented for ImageNet, the experiment intimately relates to the annotation platform in a more subtle manner. The nature of this fluid, ambiguous and iterative relation of the experiment with the other entities informs the practical engagement at the core of this study, the re-experiment.

My intervention is conceived using insights from the branches of science and technology studies and feminist studies of science that pay attention to the performative character of scientific work. The intervention takes the form of a re-experiment, a form of re-enactment of an experiment, questioning the definition of entities, emphasizing co-construction of subjects and apparatus, opening up the experiment to the contribution of the participants (as opposed to mere subjects), questioning the relation of genealogy between the experiment and an environment of production.

Additionally, in the design of the re-experimental device, particular attention is paid to the photographic apparatus at work in the Caltech experiment. This doesn't mean scrutinizing the photographic documents selected for the experiment. It means emphasizing how different methods stabilise the photographs into distinct entities endowed with specific properties and agency: the dataset, the stimulus, the composite echo, or the imaged concept. Altering the experiment, submitting its protocol to variations, crucially alters the photographic alignments. Different photographic alignments enable different relations between participants and entities involved, different ways of cutting, seeing, speaking and classifying. They facilitate the coming into being of different forms of relations, the development of different proximities. In the photographic alignment, the indeterminacy of the photograph finds a resolution, and so do the subjects and the scales.

Re-experimenting is not counter-experimenting. It doesn't aim to expose the shortcomings of the Caltech experiment. Grounding vision in a temporal micro-scale is a very provocative act. There is at the core of the Caltech experiment something begging to be explored further. The original experiment draws from the ability of the subjects to resolve vision at speed. They explore the material conditions in order to dissolve the “seeing subject” / “viewed object” division, they study their entanglement through vision. Participating in the re-experiment offers the opportunity to see on the bias, to question one's own assumptions about vision, to experience a de-centring of one's own relation to sight. It is also the occasion to contribute to a work of synchronisation and de-synchronisation, a discovery of multiple micro-temporalities.

At the same time, re-experimenting means being confronted by the multiple ways the original protocol attempts to restore the positions the re-experimentation undoes. The re-experiment attempts to resist a project of optimisation embedded in the Caltech protocol that pays little attention to the parties involved. This resistance is not frontal. Re-experimenting is to explore the difference within, not the difference against. The re-experiment's postulate is that an experiment's stability is not given. Working from within the experiment itself, it doesn't aim to offer a radically different alternative. The term variation is chosen to emphasize that the path chosen is not to position oneself outside, but to look for the differences that matter, to cultivate the divergences that every experimental apparatus at the same time produces and contains: its effervescence.

The space where the re-experiments took place is rather elusive. It wasn't in the lab, it wasn't in the crowdsourcing platform, it wasn't at the gallery or the university. Yet it was in all these places at once, in alternation, in oscillation, incompletely and uncertainly. Now that it is time to conclude the research journey, the uncertain and fluctuating nature of this space of re-experimentation needs to be acknowledged. The very limits of the apparatus have moved considerably, sometimes in the same session. Generated through these shifting settings, the knowledge produced through the re-experiment is a knowledge that stems from a performative practice. It is therefore not a knowledge that makes guidelines and norms. It is not a knowledge that isolates causes and effects. It is a knowledge made of assonances and dissonances, insistences. The research doesn't feed into a template of instrumental reason but offers a renewed attention to potential and active divergence. Its conclusions do not lend themselves to prescriptive considerations. However, the research did not happen in isolation with disregard for the realities of machine vision and the production of its models. The space of the re-experiment has not been hermetically closed to the world. It established a singular mode of relation with computer vision and photography, not a parallel speculation. At this conclusion stage, I am attempting to capture the specificity of this indirect relation.

The practice of re-experimentation has brought to light the articulation of two methodological levels. At one level, there are my intentions, the design of the apparatus, based on an analysis of the existing experiment, to create an environment seeking the collaboration of the participants, the instructions I give them, the archival strategies, etc. The second methodological level concerns the accidental architecture. There is a constant oscillation between what is planned, pre-conceived, anticipated and what happens. The re-experimental frame is by nature open-ended and relies on the participants to bring their own methods into the frame. To speak, to describe, to adapt one's body to a situation where a stimulus is displayed briefly, to attune to the various rhythms of a session, to navigate a taxonomy, to listen, to perform a transcript etc., these are a few examples in the thesis of the participants' techniques. They structure the re-experiment, transform the method, change the frame and they are central to a textured understanding of what seeing means in this environment.

Seeing in the re-experiment is the product of a biasing relation. Bias should not be understood as a property of the viewer or the image. A property inherent to subjects or objects. The re-experiment and the retrospective conversations with the participants forced me to requalify the viewer/image relationship in terms of biasing, a relation by nature incomplete that enables an oblique viewing. As perception is interrupted in its inception, the viewer may be vulnerable to a desire to fill the gaps in her perception. Filling these holes with what a subject already knows confirms the traditional notion of bias and terminates the state of suspension that the participants experience when they are briefly exposed to a stimulus. Here, the re-experiment provides the conditions for the suspension to be prolonged. As the participants hear the others, they refrain from sealing their perception. They engage in the production of what I have called a composite echo as it is made from the oral descriptions made by the group and generates a conjunctive aural image that gradually takes a life of its own. During the production of the composite echo, the participants test the elasticity of their perception. For a while, the sensations of the participants can resonate and be discordant, and the fleeting aspects of what has not yet coalesced into a stabilised percept can be attended to.

The production of a composite echo builds on the participants' intuition. It is not for that matter a process free of constraints. It is, as one participant terms it, “the game of making sense”. As a game, it follows an implicit rule, the rule of levelling. This rule, made explicit by the participants through the exercise of listening, specifies the condition for a partaker to contribute to the common description. The contribution must be at the same level of precision or more precise. This rule excludes participants who, as one of them says, “do not see well”. The production of the composite echo also remains subject to the agreement of the participants to validate an oblique vision, to accept the biasing as a constituting part of perception. When a participant refuses to adopt this position, the game ends or stalls. Importantly, the notion of making sense does not merely refer to the activity of guessing the semantic contents of the stimulus. This “game” also serves to establish the relational map of the re-experiment, to organise together the production of the composite echo. It is at the core of the collaborative dynamics. The sense of the re-experiment is not given to the participants, it is through the intra-actions that they make sense of what they are doing.

All through the Caltech paper, disparate units of measure accumulate. The ranking, the averaging, the randomization, the perceptual performance coefficient, all represent different forms of counting across the experiment. Counting the rank in the search engine and the numbers of occurrences of a term in the descriptions are the results of the application of different rules of calculation. Some of these rules are algorithmic procedures and others are rules that allow for a margin of interpretation. All these techniques of partitioning, measurement and ordering produce a perception prepared for the scaling operation necessary to generate en masse annotations. The re-experiments give a vivid account of the performative nature of these techniques. To perceive below the millisecond, to respond with a series of words, to pin these words onto a taxonomy require from the participants a total engagement of the body and attention. There is a form of athleticism particular to the re-experiment. The participants produce an intense effort that lasts on several occasion two or three hours. The exercise mobilises the retina and the eyelid. It also mobilises the lungs as the participants are breathing the stimulus in and out. To a stimulus that cuts through attention time, the participants respond with an echo that prolongs the sensation and integrates fragmentary oral images. The rapid stimulus is absorbed and slowed down. The complex act of listening adds density to the tenuous shards of perception. There is a rhythm proper to the Caltech protocol where the display screen flashes one image after the other. There is a complementary rhythm in the re-experiment that suspends the stimulus in a lapse of shared attention that gives way to a dialogue made of waiting, taking turns and repeating. This rhythm is not merely a response to a speed, it is the resolution of a scale. No stimulus is ever taken in isolation, and the response is an anticipation of images yet to come based on a sense of the repetition of similar images and an intuition of their series. "A supermarket again." "Another bedroom." Further, it is an attempt to figure out a level of description, with the effect of excluding descriptions which do not match the required level. The participants negotiate speed, redundancy and the granularity of their responses. At their level, they resolve the scale, they internalise it and adapt their methods of figuring out the stimulus rhythmically. The resulting descriptions are not raw responses, they are performed resolutions of the multiple dimensions the participants are engaging with. As such they carry the temporal signature of the stimulus, its seriality (if the image belongs to a category previously seen before) as much as its semantic content. The resulting descriptions also offer a resolution of the apparent incompatibility of the responses, they "resolve" the sometimes contradictory versions by juxtaposing them and integrating their assonances and dissonances.

What the study narrates is the co-construction of another photographic alignment departing from the original experiment. There is a patient elaboration of other alignments of practices, methods and devices, synchronisation and accelerations between the participants and me. At the core of the practice there is the realisation that a photograph is not a docile container of meaning. It has an agency that gives too much and never enough, it is always traversed and enacted by various alignments and series. The transformation of the experimental apparatus over the iterations of the project follows the need to make room for the complexity of the photographic object. Departing from the Caltech experimental device where the photograph is seen as a biased object that imperfectly images concepts, the re-experiment welcomes its fundamental indeterminacy and takes it as a generative principle. Through the different phases of the research, I made various attempts to address the fluctuating nature of the photograph and its related practices. The question of the resolution of the instability of the photograph is central to the dynamics of the re-experiment. The term alignment emphasises the solidarity of all the entities at play and the fact that they gain a specific consistency. It names the emergence of a singular distribution of agency between subjects and objects. Taking various examples, I explained how alignments can shift, redistributing all the entities at play in a different set of relations. A photograph resolved as a stimulus becomes a logo or an assignment, a subject becomes a consumer of entertainment products or a student. And the borders of the re-experiment are remodelled, pulling in the gallery of photography or the classroom. The shifts of alignments are deeply invested by the participants as they enable competences and scale of values. Every shift sets in motion a different course of recognition, a trajectory affecting what is recognised and how the recogniser gains recognition. All these elements affect the resolution of the apparatus. Resolution is here understood with the cumulative meaning of resolution as precision, as a manner to deal with differences and as a strategic attempt to stabilise the indeterminate entities taking part to the re-experiment. The different shifts are attempts to explore what the re-experiment can do, where it ends, how it extends, how it retracts, what it rubs against and which competences matter. There are therefore many stakes in holding in place an alignment. Participants' methods, the various cuts performed by the stimuli, the rhythms and scales traversing the re-experiment all concur to the stabilisation of the alignment internally.

The work of stabilisation can also be performed post-hoc and from without. This is the role given to the taxonomy in the Caltech experiment. The participants live the transition from the description stage to the filtering stage as one of rupture. There is a re-alignment taking place. They resist the re-alignment as it cancels what mattered to them in the process. Yet their resistance is not absolute as they hint at a potential remapping of the taxonomy and attempt to lend it their flexibility. It is at this stage that the micro-politics and micro-strategies of the re-experiment are expressed the most clearly. When the participants engage with the taxonomy, there is growing awareness of the rules and emerging protocols at work in the re-experiment. The taxonomy is a site of tension and of clarification for these rules as much as it is a site of triage of the description. Nowhere more than in their struggle with the filtering and classification work do the participants realise this and are ready to defend their understanding, their way of making sense of the work. Through the filtering, the participants make clear the kind of agreement they reached when they were making the descriptions and resist the taxonomy when it threatens the common ground upon which their collaboration relies.

What the participants aim to preserve from the filtering is the position that is engrained in the descriptions. This is where their engagement towards the resolution of the description (the degree to which they are described, precision and levels) coincides with their resolution of the social dynamics (how they managed to work out their differences and reach a consensus). When a participant says “I want to do it through the text”, he means that he doesn't want to evaluate a description using a different mode of perception. The descriptions are the products of a biasing relation, an oblique vision where the subject, the apparatus and the stimulus are entangled. In the filtering stage, the model of vision used as a reference is the orthodox subject-image division. To do it “through the text” means to refuse an interpretation of the words that would imply that the description can be assessed without taking into account the alignment from which the description emerged. The description is produced in a biasing relation. The participants insist that the criteria and methods for its assessment should reflect this. The biasing relation has been the base on which the resolution of the experiment has been built among the participants. Resisting its erasure at the filtering stage, the participants aim to preserve what has temporarily bound them together.

There is another aspect that emphasizes the dimension of rupture at the filtering stage, a certain distance from the material is enforced. Compared to the sophistication of the apparatus of the first stage where stimuli are controlled to the millisecond, the taxonomic device is coarse and arbitrary. The tool the Caltech researchers give us to classify not only implies a reduction of the diversity of the vocabulary or a reduction of precision of the information contained in the responses. It also, as a participant puts it, “dumbs” them down. Dumbness implies a distant subject. But distance, here, does not correspond so much to a condition for an objective classification. Distance means the opposite of involvement for the participants. Their ability to make sense is hampered by the dumbness of the process. The managerial template of the experiment takes over. The filtering process does not simply average the descriptions. It detaches the subjects from the experiment and manages their affective involvement.

8.2. Theorizing photographic alignments

To complete the overview of the research journey, this section rehearses the global theoretical argument. In permanent questioning with the practice, a series of theoretical notions have consolidated under the rubric of photographic alignment. Theoretical elements have been exposed in different places in the thesis and at various stages of development. This section presents their articulation.

Representationalism runs through computer vision from end to end. Taxonomies like WordNet are constructed around the idea that language represents things in the world. And, in the dataset pipeline, photographs are said to “image” concepts. In turn, vision is defined experimentally as the ability to view and describe photographs. Datasets are collections of “normal” photographs by which the Caltech authors mean they transmit a homologous picture of objects out there in the world. These are only a few examples of representationalism accumulated in the previous pages. Representationalism treats photography as a transparent window to the world. If we take computer vision scientists at their word, we find ourselves limited to a discussion about the merits of given representations and photography's role in computer vision is tied to discrete objects called photographs carrying the said representations.

The theorizing of photography in this thesis is not based on what computer vision engineers say but on their practices, their apparatuses and the scales they relate to. To engage with those in terms of representation is too narrow. Computer vision scientists and engineers spend considerable time designing systems to stabilise photographs, recognizing implicitly that entities named photographs prove recalcitrant in fulfilling their roles as docile conveyors of representations. And in turn photographs are enacted in myriads of ways: as data, as image, as stimulus. Photographs are part of multi-scalar alignments that resolve their indeterminacy temporarily.

To account for this proliferation of apparatuses, I turned to Barad (1996) and her concept of apparatus as a material arrangement facilitating the emergence of entities. For Barad, entities do not pre-exist an agency of observation, but emerge through it. The indeterminacy of the entities involved may be resolved in different ways. I use the term alignment to name a resolution that acquires stability and singularity. The expression photographic alignment designates a resolution of a series of indeterminate entities in which the instability[83] of the photograph plays a pivotal role. I do not use the expression photographic alignment merely to note that a series of objects named photographs are part of an alignment but to refer to an alignment in which the ontological resolution of the photograph is crucial in stabilizing the alignment.

The apparatus is a material arrangement that stabilises an alignment. It performs cuts and composes with disparate segments. Where an alignment begins and where it ends is a question, not a matter of fact. Engaging with the apparatus means exploring its boundaries both temporal and spatial. Its spatio-temporal enfolding is never given once and for all. Entities emerging in the apparatus are not enacted mechanically. If they are not ontologically separated, they nevertheless acquire agency. This agency leads to alliances, strategies, consolidations or conflicts.

For these reasons, an apparatus is not deterministic, it is performative. Effervescence, divergence,

[83] I use the term instability as an equivalent of ontological indeterminacy in Barad's terminology.

difference from within are different terms pointing to the fact that an apparatus is not a static entity imposing its rule from the outside. An apparatus and the alignments it stabilizes cannot be known in advance. An apparatus is capable of multiple resolutions. The Captain America episode in the sixth chapter gives an example of an alignment switch that re-aligns entities in different configurations. An apparatus cannot be known either by a distant observation. As agential realism questions radically the separability of observer and phenomenon, it suggests that, to study an apparatus, one must begin to acknowledge and fruitfully engage with one's own intra-action with it. To engage with an apparatus means to seek the differences within, the vectors of divergence in an alignment of entangled parties.

The theorizing in this research finds its root in the necessities of practical engagement. Practical engagement with an apparatus made me realise the importance of relying on other modalities than the visual to study photographic alignments. Temporal cuts like series, sequences, repetitions, saccades, air intake are crucial to the stabilisation of the photograph as experienced in the sessions. Practical engagement brings forward the need to account for the experiential and affective dimension of the emergence in terms that do justice to the large repertoire of intensities and propensities that are generated. Emergence and stabilisation relate to embodied and affective labour. Along with retinas, lungs and mouths open in anticipation, on the look-out. Bodies are absorbing shock not just taking in information. Engaged agents discover a new body for themselves in the apparatus. Affect theory with its emphasis on the pre-individual offers a rich palette of concepts and sensibilities to attend to the intense work required from bodies to resolve macro and micro scales, attune to rhythms and priming.

Crucially, affect theory helps to open up the notion of automatism and contest it from the inside. Its insistence on the adaptive quality of non-reflexive mechanisms provides the basis for a powerful alternative to the reductive notion of automatism that subtends the Caltech protocol. Adaptive automatism consists in a potential opening that is not exhausted by a logic of stimulus and response. Whilst it accounts for the possible mechanisation of bodies, the mechanisation does not exhaust what the pre-individual body is capable of. Adaptive automatism exceeds mechanical repetition, it is equally sensitive to what is about to change, to incipiency and to tendency. This difference between a repetitive automatism and an adaptive understanding of automatism, their overlap and incommensurability is another factor that increases the potential that lies at the heart of an apparatus. The apparatus cannot work without the affective labour of the involved agents. The difference within the apparatus (or its active divergence) depends on the degree to which – and the manner with which – the affective automatism relates to a repetitive template. In this sense, microperception is as much a site of incipient knowledge as an engine of repetitive recognition of sameness. An understanding of bias as aperception (the repetitive recognition of sameness according to pre-existing categories) is based on a notion of automatism that denies its incipiency. Automatism means more than repetitive response to input. What is commonly called bias is automatism reduced to repetition. To accomplish this reduction requires the mobilisation of a whole array of aligned entities. To study a photographic elaboration is to explore and learn to distinguish the different forms of incipiency they enact rather than to take supposedly inherently biased entities for granted.

With all these elements in mind, we can come back to the notion that is central to the theorizing, resolution. The resolution performed by an apparatus occurs at different levels. As I said, the expression photographic alignment designates a resolution of a series of indeterminate entities in which the instability of the photograph plays a pivotal role. At stake here is the ontological resolution of an indeterminate entity such as a Flickr amateur photograph, a JPEG, a dataset item or a stimulus. This resolution involves in turn the resolution of divergences and conflicts. As I said, alignments are not stable by nature and are not unified. They are only temporarily resolved. This entails that some form of convergence or consensus must be found. For example, the composite echo or the confrontation with the taxonomic device do not only enact differently the photograph. They also entail different forms of decision making through which statements are accepted, transformed or discarded. The consensus is inherently tied to a third form of resolution: resolution as ability to show things to a given level of granularity. In this last sense, the resolution of an apparatus relates to a perceptual potential and to levels of definition. To continue with the composite echo and the taxonomy, both are examples of resolutions where a different levelling, a different grade of definition, becomes available. A different form of consensus means a different vision. In short, in the apparatus that stabilises a photographic alignment, these three levels of resolution (ontological resolution, resolution as consensus or convergence and resolution as granular definition) are always enacted together. In the composite echo or the confrontation with the taxonomic device, resolution as consensus is always also ontological (how participants and devices are enacted) and granular (the levelling of perception is at stake).

With the theorizing of the photographic elaboration, my goal is not to provide a fully-fledged theoretical system but a constellation of notions to prepare a practice, open up objects, follow trajectories and become sensitized to different natures of difference. The theorizing in the thesis allowed me to create a relation of interest and resistance with the emic representations of scientific work and the experimental method. It has helped me to find the pulse of the living practice of engineers in computer textbooks and was crucial to understanding the marks of the division of labour in experimental reports that presented themselves as pure epistemic objects. It has been nurtured by the practice of re-experimentation which forced its development. It is also now that we reach the end of the study one of the means by which the knowledge produced in practice can be, if not transferred, at least translated.

8.3. Social ontology

Crucial to the understanding of the photographic elaboration of computer vision is the social ontology that permeates the experiment and the environment of annotation. In the thesis, the social ontology surfaces in three places with different grades of legibility. It is implicit in the analysis of the AMT and the role of workers in the annotation environment. In the chapter dedicated to the experiment, it specifies the role of the subjects in the apparatus. And finally in the re-experiment, it is redefined by variations brought into the experimental protocol. In this section, I summarise the different versions of the social ontology and examine how they are addressed by the variations in the re-experiment. Following this and building on the concepts emerging from the theorising (apparatus, photographic elaboration, resolution, bias), I revisit the initial questions I raised about the critique of computer vision and its datasets.

The research identifies the fundamental axioms of this ontology as follows. The social is composed of discrete individuals that can be treated as isolated atoms. The bare individuals taking part either in the annotation process or the experiment are interchangeable. These individuals are ready to perceive and the world is available to them. The world in this case is made of concepts “imaged” through vernacular photography. To a large extent, the recognition of the concept happens as a transparent process hence mediation can be ignored. Their perception is defined in atomistic terms too. Perception can be broken down into discrete units, fixations, that can be quantified. The subject offers a response to the administration of a stimulus, the worker responds to a HIT on the annotation platform. Both perform their tasks in isolation. Sociality is obtained through the mediation of an apparatus that collects the responses and processes them. The mode of consensus that can be obtained from these individuals is a process of averaging that dispenses with direct interaction. Neither the experiment nor the platform offer any explicit feedback. Subjects produce descriptions but have no occasion to review their work or to engage in a discussion with the experimentors. The annotators do not receive explicit feedback from the requesters. Their knowledge is understood as basic and static. Nobody expects them to improve, to learn from their past interactions, to get better at what they do. Both experimental subjects and workers are expected to mechanically respond to stimuli. The course of the experiment and the annotation work is reduced to a concatenation of discrete sequences. Recognition in both cases presupposes that the work of cognition has already happened prior to perception. Recognition is understood as a way of matching what is given to the senses with categories already acquired. The privileged categories are what Rosch termed basic categories: categories that correlate with the lowest perceptual effort.

In the social ontology of the experiment, individuals engaged in perception perform fixations that can be correlated to levels of descriptions and presentation times. As perception can be decomposed in units and effort can be quantified, the subject of the experiment provides the criteria to measure economically the labour of the annotator. If labour can be measured by the same criteria in the experiment and the annotation environment, they also share a division of labour. Subjects and experimentalists are given different roles as are workers and requesters. And labour management is achieved through supervision.

These axioms inform the experimental apparatus as well as the environment of annotation. The thesis' response to the presence of this social ontology is two-fold. First, it states that even if these axioms permeate both environments, it means that the relation between the abstract dimension of the ontology and the performative dimension of the apparatus is a question to investigate. To re-experiment has been a necessary step to uncover this ontology as it is never transparently stated and it never takes the form of an explicit intention from the part of the experimentalist. And the requesters do not have an explicit knowledge about the genealogical relation of their environment to the experiment. The ontology permeates devices, scales, rhythms, readily available objects and subjects rather than explicit guidelines. This is why it needs to be probed and cannot simply be read in documents such as journal articles or reports. This ontology is inseparable from the apparatus that enacts it. There is no subject of perception exterior to an agency of observation. A perceiving subject whose fixation neatly correlates with a stable category does not exist in a state of nature but can only be enacted through a specific material assemblage that cuts through time, image series, perception and scales. The thesis emphasizes the relation between the social ontology and the performative nature of the apparatus through which it is enacted.

Secondly and more importantly, it mobilises the performative dimension of the apparatus to actively challenge the social ontology. This is achieved by introducing variations in the experimental protocol that directly interfere with the different dimensions of the ontology. In response to an understanding of the social as a collection of bare individuals, the re-experiment treats the participants as subjects emerging from relations. The composite echo of their interaction contrasts with the isolation of the annotation environment or the Caltech experimental device. The composite echo reveals the thickness of the mediation process and exposes the extensive work required to enact the “readiness to perceive” taken for granted in the ontology. The world cannot be simply understood as available to experience through discrete intakes. A perceiving on the bias responds to the understanding of perception in the experiment where seeing is reduced to a collection of fixations. Perceiving on the bias implies a complex relation to the cut and the entanglement with the apparatus. In the re-experiment, the subject of perception is not an automaton triggered at will but an embodied participant involved in the resolution of an horizon of duration. Rhythmical attunement replaces mere mechanical response. Consensus in the re-experiment is achieved through negotiation and a game of making sense rather than an averaging of discrete outputs. Consensus is obtained iteratively in a process wherein the participants affectively attune to the stimulus and to each other. The re-experiment follows a course, it is not a concatenation of mechanical responses. The knowledge of the participants constantly evolves and they are encouraged to share it. The knowledge produced through their interaction with the device extends to the apparatus and the re-experiment itself. Recognition is more than the identification of already known categories, it is a course, it follows different steps in which the viewer gets recognised too.

The knowledge at stake in the re-experiment is far from basic, it requires a whole range of competences and skills. And these competences are differently valued depending on where the re-experiment begins and ends. In the social ontology of the experiment, the dimension of levelling implies that there is a coherent view in which the disparity of the world can be resolved. In the re-experiment, the participants oppose clusters of notions to a unified taxonomy and resist the simplification of the descriptions that would be required to map them onto a seemingly coherent hierarchy. This resistance is possible because participants and scorers are not separated entities. The division of labour in the re-experiment is remapped and the model of supervision is altered. The role of moments of reflexivity and the echo chambers sessions contribute to undo this separation as well. The reconfiguration of the division of labour is the pre-condition for a different involvement of the participants. It is more crucially, the pre-condition for the production of a different knowledge about what the experiment does, how its apparatus performs and what the re-experiment can do. To change the relational layout of the experiment is to enable subjects, objects and forms of knowledge that are otherwise pre-emptively foreclosed by the social ontology.

What this discussion of the social ontology establishes is that there is a factory in the experiment and there is an epistemic device at work in the annotation environment To identify the social ontology means that there is more than the observation of the perceptual properties of human subjects at stake in the experiment. As the experiment needs to meet an imperative of production and prepare subjects to a scale, it provides a management template for the annotation environment. This ontology is an abstract device that allows to apprehend the texture of relation in which both the experiment and the platform are woven. Taking together the reflection on the social ontology and the examination of its performative resolution, I will now come back to the environment of annotation from which the research journey started.

8.4. The photographic elaboration of computer vision

Drawing from the accumulated experience of conducting the re-experiments and from the theorizing that accompanies it, I will now come back to the questions I raised at the onset of the thesis and the context in which they were formulated. In contrast to a bias critique based on representations, I have asked: what could be learned from the annotation process by a consideration of its middling, its stabilisations and alignments? How can problems such as bias be re-appraised by taking into account the radical entwining of viewer, apparatus and visual data? In contrast to a critique of labour, I have affirmed the importance of reconsidering what is going on under the name of automatism and repetition. I have postulated that these questions required undoing the boundaries of the annotation environment and studying its spatial enfolding with the computer vision lab. After having conducted the re-experiments and clarified the theorizing they gave rise to, I am returning to these questions and propose a questioning of the current framing of the controversies about computer vision datasets. Acknowledging the problems raised by both the bias and labour critiques and the importance of challenging the current state of computer vision, I will present the experience and theorizing developed through the research as relevant resources to rethink the definition of the horizon of change of computer vision rather than a means to provide ready-made alternatives to its problems.

For decades datasets have been assembled without much consideration for their cultural and political implications. Fuller and Goffey (2012, p.93) wrote “In the 'bureaucracy of statistics', few objects are grayer than the dataset.” Liminal objects, outsourced to a workforce whose epistemic contribution is not acknowledged, datasets were considered low priority for a critical thinking of technology and have proliferated below the radar of public scrutiny. As I explained in the second chapter, the fact that computer scientists are beginning to realise the various prejudices stemming from the classification operated by the datasets is the result of a combined effort of different strands of activism and a critical understanding of the discriminatory practices made possible by AI and computer vision, on the part of tech workers, journalists and the general public. As the debate reached a larger audience, the problem has often been framed as one of bias. The datasets are criticised for encoding multiple forms of discrimination and have been blamed for producing toxic representations of the world. In this section, I will use the problem of bias as a means to discuss more largely the photographic elaboration as well as the horizon of change of computer vision.

The accumulated experience of re-experimentation and the analysis that prepared it give me a ground from which I can question the limits of an approach to computer vision in terms of bias. Bias, as I proposed, is a term that comes loaded with a series of assumptions. It is worth recapitulating here how I presented in this study the discussion between computer scientists and the communities attempting to address the issue of bias in terms that are computationally tractable.

Algorithmic bias is defined in computer science as the degree of autonomy of an algorithm with reference to its training data. Bias is treated as a measure of overfitting or underfitting. It limits how much an algorithm is able to capture the significant signals in the data. Algorithmic bias is presented as a problem of statistical interpretation of the data. The ability of an algorithm to generalise is not due to its mathematical structure alone. The dataset represents the variability of the domain from which it needs to extract enough regularity and enough differences. At this level, computer scientists speak of dataset bias where bias is framed as a problem of statistical representation within (instead of statistical *interpretation* of) the data. Both algorithmic bias and dataset bias affect the ability of an algorithm to generalize on the basis of its training. In this study, the focus is on dataset bias as I approach the problem from the perspective of the elaboration of the dataset.

For Fei-Fei Li, Internet is a source of diverse poses and lighting, offering considerably more variations than in controlled environments such as datasets made of scarce professional photoshoots. Changing the size of the training set is seen as a response to the problem of dataset bias. Diversity is obtained by a laissez-faire, liberal approach to curation. In opposition to the “toy” dataset, Li focuses on the “real”, the “wild”. To solve bias requires the acquisition of millions of photographs, the introduction of a new scale in computer vision.

In the Caltech experiment, Li understands bias on the same basis as in her work with ImageNet. The traditional provenance of images of cognitive psychology experiments, the professional photographer distributing samples on CDs, offers only a narrow world-view. The Internet is a source of more diverse imagery. It is also the source of a more “normal” photography (ie. amateur rather than professional). There is a convergence between size, a factor supposed to provide more diversity, and the prevalence of a certain practice of photography.

In contrast to these definitions, Buolamwini and Gebru introduce bias as a measure of fairness, a definition that will be picked up by computer scientists willing to address bias in their systems. Buolamwini and Gebru argue that systems are discriminatory because the composition of data reflects an asymmetric distribution of power in society. For instance, face recognition systems are discriminatory because they perform badly on images of people with darker complexion. This kind of error can be correlated to the lack of phenotypical diversity in the data sets. In their view, the remedy is the creation of inclusive datasets where all parties are given equal treatment[84]. Diversity is obtained through voluntarism, curation. In opposition to the wild, Buolamwini and Gebru defend a “civilised”, “curated” approach.

With these different views on the problem of bias in mind, let's see how the critique of bias raised by the Fairness community has been understood by dataset makers. Indeed important dataset projects and commercial companies have not ignored the *bias as discrimination* critique. They have responded on the same terms. Photographs are taken as representations with fixed attributes and according to the values of these attributes, datasets are re-balanced and harmonized. Further, photos with offensive attributes are removed. The response takes the form of a process of dataset mainstreaming. The researchers maintaining ImageNet, one of the datasets most exposed to criticism both on public forums and within the computer vision community, recently published a progress report on their attempts to increase the “fairness” of their creation. Their “curatorial” approach that tracks discrimination in the photographs responds to what Anne-Marie Mol (2002) has named a “politics of who.” Such a politics considers that discriminatory choices are made by rational subjects that can distance themselves from the problem at hand. Furthermore, a politics of who isolates moments where choices are being made and concentrate on these isolated moments to propose remedies.

The problem of removing dataset bias is understood by dataset makers and their critics as a means of restoring the correct perspective over its data. Shocking as they are, misrepresentations in renowned datasets like ImageNet, I contend, are symptoms of larger problems that will not be solved by patching up representations or taking away entire categories. These symptoms stem from

[84] This also entails more inclusive teams of developers who would have different sensibilities and prevent the dominance of a view pertaining to a certain group.

the photographic elaboration of computer vision not from the selection of individual photographs with wrong attributes. As an elaboration, they pertain to the conditions in which the work of producing datasets is performed. To a mindset that asks who is missing, one should add a mindset that asks: which environment is the annotator entangled with and which rhythm is she able to get attuned with? What are the scales resolved by the apparatus? Which course of recognition does it follow? To answer these questions with a politics of who reduces the problem to moments of choice where the involved agents exert punctually their faculty of judgement. Instead, the experience accumulated during the research leads me to respond with a “politics of when” where moments of choice are not isolated but distributed all over the alignments that have taken consistency. The selection of photographs is made within a specific apparatus that enforces certain forms of consensus and a way of seeing that affects its resolution, what it is able to distinguish and at which scale. Changing post-hoc the most controversial categories and removing those not deemed imageable do not address the apparatus and labour conditions in which choices are made.

Furthermore, such a criticism and its corresponding response do not address the photographic dimension of this elaboration. By photographic here, I mean the processes of mediation that stabilise the photographic object into a dataset item. Bias, as it is understood by prominent critics of machine vision like Buolamwini and Gerbu (2018) who had an impact on computer vision, is the property of a photographic object. It is not thought of in terms of alignments and processes. The use of the search engine to extract the photograph from its context, the use of the AMT to resolve the indeterminacy of the photograph and hold it in place and pinning the photos to a taxonomy are seen as discrete problems in such a critique and remain treated in isolation in various attempts to restore fairness. When they respond to a critique of bias, engineers treat each problem separately and design a corresponding solution. The search engine bias must be fixed by inserting queries in more languages, AMT workers should not be asked to classify non-imageable categories of photographs and offensive labels should be removed from WordNet.

Perhaps the most striking example of the limits of resolving the dataset's challenge using the lens of the critique of representation is to be found in the remedies tested by ImageNet researchers to reduce the negative consequences of using the WordNet ontology. As I noted in the second chapter, the researchers call WordNet a *stagnant* vocabulary (Yang *et al.*, 2020), the age of WordNet supposedly explaining its conservatism and some of its blind spots. The relationship the researchers maintain with WordNet is deeply ambivalent. They claim that the ontology structures the semantic organisation of the dataset and that the acquisition pipeline is driven by the thesaurus: a candidate image is included at the condition of matching a given synset.[85] At the same time, they always treat it as provisional. This ambivalence can be perceived in the Caltech experiment, too. The experimenters involved in *What do we perceive ...* in 2007, and Li's team two years later for ImageNet, are imposing a specific set of categories upon the visual data. Yet, they want the classification to be interchangeable and treat it as a movable part. Li declares that she wants structure (GoogleTechTalks, 2011) and is clear about using WordNet as such, a structure that can be borrowed opportunistically as it comes for free. She is not interested in WordNet as a classification reflecting her personal worldview. The criticism against ImageNet's data reduces the classification to a vocabulary where the choices of classes introduce discrimination. But WordNet does more than provide a set of classes. To understand this, one must take heed of the spatial partitioning operated by the classification. WordNet presents itself as a grid, it is partitioned in rows and columns. When a cultural critique treats the classification as a logical set of classes with acceptable or offensive labels, it emphasizes the importance of the vertical partitioning (inanimate versus animate, human versus animal, male versus female, etc.). However, for Li what matters is the dimension of the hierarchy that divides the world into a stack of rows. The “vertical” partitioning is secondary. The wide range of columns of the classification is convenient because they cover a large range of classes, but their meaning is treated with relative indifference. Another set of classes would do as well. The horizontal partitioning, the number of rows of the structure is what is deemed of critical importance. The insistence of the Caltech researchers on a hierarchical organisation of the vocabulary and the investment of the participants on the levelling of descriptions in the re-experiment tells us why. There are high stakes in establishing a relation between level of description and perception time. The experiment's claim that there is a match between the display times and the levels of the taxonomy is relevant to the calibration of the apparatus of annotation. The assertion that 200 milliseconds of visual attention correspond for the subject to the amount of time she needs to associate a basic category label to a stimulus provides a reference for the organisation and coordination of the annotation work. To correlate a level of classification to a display time allows one to estimate the classification's cost and speed.

At this point it is worth insisting on the influence of the psychologist Eleanor Rosch (1978), famous for her work in classification theory, on Li and her work. Rosch affirmed that a taxonomy is a device that calibrates the viewer's attention and regulates a cognitive economy (Rosch, 1978). The differences between levels, the rows, is not approached as one of semantics so much as one of duration and attention. The interest in the truth claim of the classification, in its ability to capture the essence of a knowledge domain, is relative. What counts is to optimise the speed of the process, to

[85] A category in WordNet's parlance

correlate the time given to the subject to see a photograph and a level in the taxonomy. A critique narrowly based on WordNet as a vocabulary that includes or excludes misses the point that WordNet is used to do much more than filter words. WordNet pertains to an articulation of scales where the speed of vision and the platform's productivity relate to the amount of photographs that need to be classified. The place of a term in the taxonomy's hierarchy helps the dataset makers to evaluate how much time and attention should be dedicated to the evaluation of its corresponding image candidates.

To understand this mechanism with a concrete example, let's go back to the way consensus is reached among annotators. The dataset makers anticipated human errors of judgement and the difficulty of deciding whether a candidate image contains a particular object. For a candidate image to be included in the dataset, it is not enough that one annotator selects it. A certain amount of annotators must make the same choice. So, the same images are shown to different annotators. This factor increases the cost of the operation. Therefore, it must be calculated precisely. To determine the amount of annotators who need to make the same choice for a given candidate, the ImageNet team established a rule that defines the inter-annotator agreement. The inter-annotator agreement is modelled according to a scale that varies as a function of the “semantic difficulty” (Deng *et al.*, 2009) of the terms. The semantic difficulty[86] is a grade of imageability, the “ease with which a term elicits a mental representation” (Paivio, 1986), that varies according to the location of the term in the hierarchy. The ImageNet authors give the example of the different level of difficulty of reaching a consensus for a synset like “Burmese cat” in comparison with “cat” (Deng *et al.*, 2009). As cat is deemed more imageable than a Burmese cat, the number of annotators who need to agree on the label “Burmese cat” for the same photo has to be higher (5 annotators need to agree) than the number of agreeing annotators for “cat” (3 is enough). Therefore, to identify and agree on the label Burmese cat costs more to the system than the label cat as the candidate images need to be shown to more annotators and the consensus takes more time to be reached. As we noted already in the previous chapter, the taxonomy is a device that cuts into more than words. It is a key part of an alignment where vision, display time and taxonomic hierarchies are correlated to provide the managerial template of the annotation's environment.

Labour cost, attention time are key aspects of the annotation process that exceed an understanding of classification in terms of a selection of labels happening on a purely intellectual plane. As I have said, computer vision follows a process of elaboration eliding the workers' contribution. The bias

[86] Semantic difficulty is a notion somehow similar to the perceptual performance of a term. A term with a high semantic difficulty is a term with a low perceptual performance.

critique also ignores this dimension. In the introduction chapter I explained that, parallel to the criticism of bias, another strand of activism concentrates on the conditions of exploitation dominating the platforms of crowdsourcing. An effort is made by workers, activists and researchers to develop fairer models of micro-tasking. A constellation of projects and organised communities such as TurkerNation, MturkGrind or Reddit groups like /r/HITsWorthTurkingFor share the objective of rewriting the platform's social contract. Tactical media projects like the Turkopticon made interventions in the platform to "interrupt workers' invisibility" (Irani and Silberman, 2013). Under the umbrella of platform cooperativism, the worker-owned Daemo (Gaikwad *et al.*, 2015), a self-governed crowdsourcing marketplace has been launched with the support of the Stanford's HCI Research Group in 2018. These efforts are concentrated on the crowdsourcing platforms, they are addressing a larger population and not strictly limited to computer vision. Their focus is on the labour conditions. Fairness in this context doesn't refer to fairness of representation but to fairer wages. Workers and activists reached various degrees of success in their struggle against the owners of the micro-labour platforms. Contrarily to a critique based on representation, these projects make room for the labour involved in the elaboration of computer vision. They concentrate their efforts on reducing exploitation and the side effects of repetitive work.

As it was the case for the bias critique, it is worth noting that the dataset makers are not ignoring it. They are responding, although moderately, to the calls for fairer wages. For instance, when they hired workers to annotate Visual Genome, a dataset of 108, 077 images, the researchers agreed to new academic guidelines for a fairer compensation of the annotators, the result of a long struggle (Krishna, Zhu, *et al.*, 2016). Even if the raise was rather symbolic, a Turker working continuously made 6-8 dollars an hour. Whilst the accomplishments of these various projects need to be acknowledged, their framing of the nature of the annotation work doesn't differ much from the employers'. Little is done to gain recognition for the workers' epistemic contributions, let alone to figure out what they consist of. The thesis provides insights into how the issues of decision, scale, attention and speed can link the concerns about classification and conditions of work. To understand how the research can contribute to the epistemic role of the workers, I will propose to read the annotation process with the experience accumulated through the re-experiment. What I will do is to find how – by asking questions pertaining to the key elements that came out of the research like scale, the flow of kinetic energy or the work of alignments – I can complicate and enrich the critiques of bias and labour.

To annotate at speed, as we have explored with the re-experiments, is not a mechanical response issuing from a passive subject. To expand the understanding of the annotator's contribution, I will ask what changes when we understand the process as one of photographic elaboration and we pay heed to the annotator's involvement in the apparatus. I am asking how the reading of the annotation pipeline is affected when embodying a scale, figuring out rhythms and levels, understanding and refraining from involvement are considered as crucial to the epistemic relevance of the process. The experience accumulated through the practice sensitizes me to other areas and dimensions to which the workers greatly contribute. It helps me not to focus solely on the semantic decision, the pivotal moment where the annotators assert the meaning of a photograph but the complex methods through which they synchronise within an alignment and embody a scale. What does it mean for the process of annotation and its critique if synchronisation, scale embodiment, attunement, listening, seeing on the bias are placed at the core of the photographic mediation of computer vision? If they intervene crucially in the resolution of the photograph in a given alignment, to understand the annotator's contribution is therefore to be attentive to how she intensely relates with the apparatus, how she probes where the apparatus begins and ends. How she walks the path of recognition it provides from Turker to perceiving subject and back.

To give some flesh to such discussion, I will turn to a series of examples of annotation in ImageNet to see how these insights gathered from the practice resonate with the daily work performed in the annotation platforms. Through the example's discussion, I will not try to replace a critique of labour or a critique of bias by another critique that makes them irrelevant. I will not claim either that I have an alternative version of the process that competes with a critique of labour in the description of the actual circumstances of platform's work. The reading interrogates both critiques by concentrating not on the extensive but on the intensive dimension of platform's work. What I will do is to bring the intensities, the movement of kinetic energy, the tensions of stabilisation and effervescence from the re-experiment and look for resonance in the annotation pipeline. I will take the dynamics of the re-experiment to question where incipiency and propensities are at work in the platforms.

This exercise focuses on several examples of classification where the singularity of a reading based on alignment and intensity increases progressively. To begin, instead of choosing an example among the synsets already marked by controversy, I will concentrate on a mundane one. Indeed, most of the controversial examples used to denounce dataset's malpractice focus on historically fraught categories and have triggered a response that presented these categories as exceptional and consequently consolidated the view that they could be “fixed” in an ad hoc manner[87] or removed from the dataset (Yang *et al.*, 2020). In selecting a mundane case, my aim is to show that the

[87] For many AI apologists, racist discrimination is considered an edge case. Yonatan Zunger, AI architect at Google, apologizing to a Google Photo user whose photo portraits had been tagged “gorillas”, explained that new updates would include fixes for words “to be careful about in the photos of people” (Kasperkevic, 2015)

examples of the shocking controversies are more intense occurrences of a deeply entrenched problem: that they constitute a difference in degree of a same problem, not a difference in kind.

The category I have chosen, “ratatouille”, contains 1024 items and is filed under Misc → food, nutrient → nutriment, nourishment → dish → stew. The description given to the workers for the ratatouille concept is “a vegetable stew; usually made with tomatoes, eggplant, zucchini, peppers, onion, and seasonings” (WordNet, 2010b). The items in the synset however exceed the limits of the given definition. Various forms of stew are included (e.g. meat stew) or tomato mozzarella salads. It even contains a Pixar character from the eponymous movie. The cause of these classification mistakes are attributed to the workers. Engineers develop techniques to track “satisficers” (Hata *et al.*, 2016, p.1) who accomplish their tasks with too little care. Other problems related to the annotators are evoked. They are considered culturally incompatible with the request, or requesters suspect they don't read the definition with enough attention. The insights from the practice lead me to propose a different perspective on the question of choice: exchanging the decision moment from the “isolated who” (Mol, 2008) for a more distributed consensus of actors and devices. The question becomes: how do the Turkers achieve consensus without explicit coordination? And the answer suggested by the practice is: because they figure out the rhythm they have to follow.

How can such a proposition change our reading of the annotation pipeline? The first level at which it makes a difference is the importance given to rhythm and speed. These are not circumstantial factors remaining external to the process. They become integral to the epistemic contribution of the workers. As I showed in chapter two, the cadence of the platform comes from the remuneration. To secure a minimal income, the Turkers have to perform at high speed and they need to keep up this pace for hours. They have to calibrate their involvement. They need to figure out a balance between speed and precision. If I am looking for an assonance with the re-experiment, I will say that the Turker is performing the work of finding her place in the alignment. From this perspective, she is contributing to the resolution of a scale, to the correlation between semantic hierarchies and speed. In the re-experiment, we have seen how the subjects were not blind to the repetitions of the stimuli, to the various modes of suggestion and priming. The whole spatio-temporal environment speaks to them. This is where the re-experiment can help complicate the understanding of how a consensus may emerge without direct deliberation. In the annotation environment, the Turkers do not debate together to decide whether a candidate image should be considered “tomato mozzarella” and should consequently be included or excluded from the set. If we consider that speed is integral to the annotation work, we need to ask how momentum, acceleration and deceleration take part in the decision. From this perspective, we would need to think that the Turkers have to judge whether it is worth slowing down to give attention to this candidate or if, at first glance, it can be assimilated to the concept ratatouille without threatening their remuneration. They would have to intuit if the difference they notice is worth changing pace or if they can remain indifferent to this difference. From this perspective, it becomes possible to envisage how Turkers develop a common sense of the apparatus' resolution without having to explicitly negotiate. They are lock stepping, not deliberating. If we accept allowing the re-experiment and the annotation environment to resonate, we can start considering that the pace of the AMT defines its resolution, its grade of precision in the descriptions. And concomitantly, it creates a zone of indiscernibility for the apparatus.

As we have seen in the re-experiments, a consensus is not always a matter of explicit argumentative deliberation. It is a matter of levelling, finding a common grade of involvement. The composite echo is not produced through argumentation. Based on the re-experiment practice, I am not asking what the arguments that lead to the selection of some items are rather than others, I am asking how the echo is propagated through the AMT. If I let the intensities registered through the re-experiment bear on my reading of the annotation process, I will propose that, to build the consensus, the workers have to develop a sense of the speed required by the apparatus. Speed as I have shown previously is not a mere measure of acceleration or deceleration. It is an intensity expressed by the resolution of different scales. As such speed affects the whole stabilisation of the entities taking part in an alignment, following in my time-critical approach, I will look more closely at the photographic alignment of the AMT and ask how it induces speed. For instance, AMT stabilises the candidate images as thumbnails detached from their original context. They are search engine results and as such the search engine already induces a sense of consensus. If many candidate images look similar, it suggests that a consensus exists over them. The interface shown to the Turkers mirrors the search results page. The Turkers do not mechanically validate the dominant representations they are given by the interface. But the regularities that spring out from the grid of thumbnails function as a cue. From a time-critical perspective, the grid is not merely a visual arrangement of photographs passively awaiting the Turker's gaze, it acts as an accelerator. To choose against the coherence emerging from the thumbnails' grid requires more work and time. It requires looking at the candidates that do not stand out. An alignment does not juxtapose discrete elements, it enables trajectories. As the viewer is enacted in a trajectory, to parse a visual input means to discern occasions of acceleration or deceleration as much as correlating a label to a pattern. From what we have seen in the re-experiment, cues do not propagate only from one moment to the next, one stimulus to the other. There are different durations involved, a complex coming together of times.

AMT provides its own texture of duration. Workers jump from one screen to the next. But another important cue comes at a different periodicity. This signal comes from the rate at which a worker's jobs are being accepted or refused by the requester. The reasons why a job is accepted (understand as remunerated) or refused (the worker has no recourse against the requester) are extremely rarely given by the requester. It is therefore not always possible to correlate a decision made by the worker and a rejection from the requester. But as they have significant economic consequences for the annotator: she evaluates her work based on an interpretation of the requester's decision. Therefore, there is never a fully explicit causal relation that can be established but an echo where different forms of feedback and cues resonate with each other. If we agree to construe AMT as an alignment, we need to consider the workers' contribution as intimately related to a listening of the apparatus. They have to attune to different cycles and temporal trajectories as much as they have to read the labels and the definitions of the synsets.

To come back to my first concrete example, to confuse “ratatouille” and “stew” may be considered as a mere distraction or a problem of catching the nuances of visually similar meals. Seen in its globality, the synset still offers visual samples that in majority represent a ratatouille. This problem however can grow as two further examples will attest. In the synset Parisian, “a native or resident of Paris” (WordNet, 2010a), a large portion of the selected items are photos of the socialite Paris Hilton (in various forms such as candid or press photos, 3D models or selfies). Hilton becomes a Paris resident and Parisian takes the form of the American idol. With the synset “reformer”, the problem is even more acute. The synset's gloss reads “An apparatus that reforms the molecular structure of hydrocarbons to produce richer fuel; a catalytic reformer” (WordNet, 2010c). However, only one synset photo actually depicts the apparatus. All the others refer to another kind of reformer, the apparatus used in Pilates classes and exercise rooms. The discrepancy between the gloss and the image selection is even more astounding if we consider the precautions that are in place to make sure the definition is read. Before seeing the thumbnails, the annotator is presented with the definition of the synset. Then subsequently with another screen where she has to choose the description she just read among several others. The annotator, to access the HIT, has to confirm explicitly that reformer was a chemistry-related apparatus and not a gym class prop. Even with such precaution, the synset barely features one correct photograph.

The Parisian or the reformer cases make clear that the composition of the synsets cannot be explained by a process of workers carefully reading the gloss and selecting the candidate images accordingly, by focusing on a pivotal moment of an “isolated who” making a decision. If we see the work as enacted temporally within an alignment, vision, attention and decision are related differently. In such a view, for the annotator, glancing is not only a mode of perception related to the rapid scanning of thumbnails, text too is read in a glimpse. The Turkers hover over a visual configuration that dominates the interface. Rushing through the pages, workers see an overwhelming presence of the shining blondness of a familiar icon whose name matches the synset's label. If we consider the Turkers as enacted in a trajectory and listening to the cues propagated through the apparatus, they do not necessarily have to believe that Paris Hilton is a resident or born in Paris to select a photograph that represents her. They only need to decide if there are enough responding echoes in the apparatus to validate the selection without having to spend time verifying. My contention here is that if we consider the annotation pipeline from an alignment perspective, what changes is not only how the consensus emerges but the nature of its object. The object of the consensus is not the fact that Hilton is a Parisian, but that such an approximation will not isolate the worker who makes it. As we have seen in the practice chapters, a course of recognition traverses the apparatus. To consider the Turker as enacted in a trajectory requires asking how this trajectory enables her not only to recognise but also to be recognised. To be recognised as a Turker, in such a view, one doesn't have to exhibit the ability to recognise things that are factually right. It is to recognise the grade of approximation that is expected from her. The Turker's competence is multi-scalar, not just a semantic affair. As a Turker wrote on a forum, it is to have a sense of what is "enough to get away with". To get away is indeed the right term as a Turker must always have an eye on the previous Hit and another on the next. To have an acute sense of trajectory. In an alignment, speed is conductive to consensus.

A late experiment made by Li in 2016, building on the Caltech experiment, indicates how the pressure to produce more data for the machine learning industry leads to increasingly aggressive attempts to instrumentalise the annotator's ability to resolve scales and propagate consensus through the apparatus of annotation[88]. Using a technique of rapid serial visual presentation, Li and colleagues have tested an approach that produces extremely fast judgements (Krishna *et al.*, 2016). Their idea was to immerse the annotator in a flow of images that she has to classify. She is given a binary task (e.g. flag all images depicting a horse) and once the start button is pressed, photographs are displayed successively on the screen at the rate of 100 ms. When the annotator perceives the target (e.g. the horse), she presses the space bar. The rhythm is so fast that in the time she takes to react and press the key, other images have already been displayed on the screen. As a form of feedback, at the bottom of the screen, the last four images that have been displayed since she pressed the key are briefly shown. Pressing the bar does not interrupt the flow, the moment of

[88] It is worth noting how the AMT platform actualizes the constitutive circularity of the relation between the environment of production and the lab of cognitive science. Sarah Kember in her study of face recognition, using Foucault, interprets this circularity as a feature of biopower , a milieu in which "a circular link is produced between effects and causes" (Foucault in Kember, 2014). Using the AMT as both a work platform and an experimental device, Li increases dramatically the circularity between the model and the modelled, effect and cause.

selection doesn't give her a pause. While her selection appears among the four images, new candidate images keep coming dividing her attention towards the two areas of the screen. Literally, the pivotal moment of choice is dissolved in the visual flow.

The experimental apparatus accumulates selections of four candidate images. To assert which one was the target the annotator meant to select, the same images are sent separately to other annotators in a different order. Cumulatively the workers confirm each other's decisions making a candidate image increasingly more likely to correspond to a given label. Their decisions cannot be interpreted in isolation, they need the confirmation of others to become legible. In such a system, a larger error rate is accepted from the participants. They are informed from the start that a certain (unspecified) amount of error is accepted. Yet they are told at the same time that some candidate images are known to the requesters and failing to select them may lead to rejection. Here again, finding a consensus and figuring out a degree of approximation come together with an attunement with the rhythms and scales of the apparatus. This extreme experiment, led by Li when she accepted the position of chief scientist for AI and machine learning at Google Cloud, exemplifies what it means for an annotator to resolve a scale.

From the perspective of the re-experiment, I propose that workers are not so much biased than on the bias, they do not lack precision, but their precision is elsewhere. Or to borrow Mark B. N. Hansen's (2014) formula, it is *elsewhen.*[89] Workers are calibrating attention, adapting to rhythms, perceiving with the horizon of duration of the apparatus they are entangled with. They seek an attunement with the apparatus, they are sliding into alignment. To say that workers are not paying attention is to believe that the objects of their attention are discrete entities such as a written definition or a singular picture. If we think of attention in the context of an alignment, we can see they are attentive, but not to the same objects. They are attentive to a less legible element that they are able to grasp by listening to the cues and echoes reverberating through the platform: the platform's currency of attention. The platform is literally paying attention in terms of remuneration. But it is a weak currency. It rewards a low grade of attention to the individual objects of annotation work. What the platform attempts to extract are not so much consensual decisions than interchangeable ones. As we have seen in the last chapter, interchangeability is obtained by lowering involvement.

The nature of my reading is a reading through. In this case the object of the reading is not a text

[89] In his book *Feed-Forward*, Hansen (2014) produces a detailed analysis of the current attempts (commercial and military) to work around deliberative consciousness through sensory micro-temporal solicitation. Manipulation doesn't happen so much elsewhere (through a hidden meaning) then elsewhen (at another time scale).

through another, but two devices through one another. During the account of practice, the annotation environment was always at work within the device of the re-experiment. Here the circulation of affect, incipiency, propensities that were active in the re-experiment are used as vehicles to probe the annotation environment. I am not trying to impose a competing claim about what the process of annotation is in its actuality. It is not an attempt to challenge studies based on fieldwork or direct observation on their grounds. The thesis doesn't provide a competing factual assessment, but it brings a series of intensities and trajectories to bear on factual assessments of annotation work. It contributes to form a sensibility to what may not register in interviews or statistics in current surveys. It is an attempt to keep these facts from closing upon themselves. With this exercise, I try to provide a reading expressive enough to make room for the kinetic energy the environment of annotation feeds off and the alignment that stabilises it temporally. It should not be substituted to other kinds of interventions, it doesn't intend to dismiss, make redundant or cancel the important studies tracking discrimination or trying to improve the labour conditions. It intends however to enrich and complicate their objects, to extend and multiply their trajectories.

Workers do not have the luxury of a deliberative judgement. They glance, they rush to the next task and they receive no explicit feedback. Decision is distributed throughout the apparatus. Bias functions as an accelerator that helps keeping up with the global pace of the apparatus. In the environment of annotation, bias is inherent to a viewer who is exploited and whose contribution is unacknowledged. Therefore the relation between bias and discrimination should be seen as intimately bound to the division of labour formalised in the ontology. Bias is intrinsically related to the discrimination between those considered as knowledge producers and those who are merely performing micro-tasks. Those who have a name and those who do not. The latter are put in a position where they are induced to relay and amplify bias by the various cues, rhythms and signals administered to them by the former.

This understanding of bias emphasizes the necessity of gaining knowledge about the embodiment, the apparatus, the entanglement, the distributed character of the process. By intervening in the social ontology of the experiment, it also shows the potential of learning and the ability to address the complexity of vision, description and classification. It suggests that the lab of cognitive psychology as well as the environment of annotation could be sites of learning and developing sensibilities, attunements with high relevance to machine vision. As the social ontology changes, participants demonstrate and acquire what counts as competences. The end of bias as discrimination requires ending micro-worker's exploitation.

At this point, it is useful to come back to the discussion of chapter three and Buolamwini and Gebru's argument which defined bias as an asymmetric distribution of properties reflecting an imbalance of power, as well as racial and sexual discrimination. This argument has lead to a critique of the selection of objects in the system (ie. photographs of face) and a corrective. The corrective was a benchmark dataset containing better and more equal distribution of phenotypes. The benchmark had a huge influence on how computer scientists attempt to remedy to bias in datasets. By formulating the remedy in the terms computer scientists evaluate the technical performance of their products, it provided a means to operationalise their critique. With the beginning of the institutionalisation of AI ethics, fairness metrics offered a solution compatible with the current paradigm of AI. Fairness could be translated as a measure of under or overfitting.

One can sense now the huge difference between a critique of bias as a critique that pertains to objects and attributes and a critique of bias as scalar and temporal. There is a sense in which the former frames bias as a problem in need of a fix, whilst the latter implies a holistic critique and a questioning of the general research effort. The recent evolution of Timnit Gebru as well as other actors of the Fairness community illustrates this point. Whilst computer scientists respond to their critics by an object-replacement strategy, influential critiques of bias have recently moved towards issues of scale, in convergence with what is being argued in this study. The limits of remedies based on fairness metrics are increasingly perceived in the Fairness community. A recent controversy sparked by a publication authored by Gebru shows what is at stake when one moves from a critique of representation to a critique of scale. As stated above Gebru, in collaboration with Buolamwini, began by providing benchmark to assess bias in computer vision systems. Her work on algorithmic discrimination as well as her collaborations with Li[90] qualified her for a position in the newly formed Ethical Artificial Intelligence Team at Google. Her recent work followed a different inflection and shifted towards issues such as the polysemy of data, the provenance and curation of samples as well as infrastructural and environmental problems. The article co-authored with Emily Bender and colleagues at Google, which triggered the controversy leading to the termination of Gebru's employment, *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* (Bender *et al.*, 2021) argued that the size of current datasets used to train models concentrates power in the hands of a few major players, the only ones able to afford financially to train them. The sheer size of these datasets comes with huge problems and prevents a proper screening of their contents even for well-financed companies. Worse, attempts to remedy the problem by documenting

[90] Li was the doctoral supervisor of Gebru. The two are co-authors of several articles including *Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States* (Gebru, Krause, Wang, Chen, Deng, Aiden, *et al.*, 2017) and *Fine-Grained Car Detection for Visual Census Estimation* (Gebru, Krause, Wang, Chen, Deng, and Fei-Fei, 2017).

the data, Gebru observes, leads to further asymmetry as it adds to the already vertiginous costs of data management and reduces even more the number of players able to pay for the work. And finally as the size of datasets make them unfathomable, there is little chance that, under this paradigm, harm and discrimination can be avoided once systems trained at such scale are operating in real conditions.

The evolution of Gebru's position shows a path of convergence with has been discussed in these pages. Even if, in Gebru's, scale is understood in terms of cost, harm and risk assessment, it converges with the importance given to scale in the thesis. The termination of Gebru's employment in the wake of the writing of the *On the Dangers of Stochastic Parrots* article also shows the strong resistance to critique when it is directed at the scale of the problem and the difficult work that lies ahead for those who want to effectuate concrete change. If a critique of bias has found a certain echo in machine vision, a critique of scale is met with a blanket refusal by major players such as Google.[91]

The shift from representational bias to issues of scaling and the tension that ensues indicates that the definition of bias is intrinsically linked to the question of the horizon of change in computer vision. Each definition of bias comes with a corresponding perspective of transformation.

When bias is defined as as over or underfitting, the condition for change is the selection of more data harmoniously balanced. It is a technological fix that merely concerns the composition of data. ImageNet is an attempt to solve this problem by integrating data “in the wild”, and changing the scale and provenance of data.

In the experiment, Li justifies the preference for photographs culled from the Internet as a source of stimulus that corresponds to normal imagery as opposed to the studio photographer. Bias is defined in an opposition between artificial and natural where natural images correspond to the codes of amateur photography and the representation of basic categories. Here the horizon of change is again a question of changing the provenance of photographs with the expected consequence of moving cognitive psychology closer to how people see the world and how they spontaneously categorize it.

When bias is defined as discrimination, as in Buolamwini and Gebru, the horizon of change is

[91] This is not to say that interesting counter-arguments do not exist. See, for instance *The Slodderwetenschap (Sloppy Science) of Stochastic Parrots -- A Plea for Science to NOT take the Route Advocated by Gebru and Bender* (Lissack, 2021) and *A criticism of On the Dangers of Stochastic Parrots: Can Language Models be Too Big* (Goldberg, 2021). What matters here however is the celerity and the brutality of the reaction from the company to a critique of scale.

conditioned to the curation of data that respects the diversity of the world and doesn't reproduce hegemonic representations. Bias is understood in terms of risk and harm. The horizon of change is a fairer and more just selection of data where the regularities are expressed unambiguously in benchmark datasets containing representations in harmony with the diversity of populations.

Finally when bias is defined as *on the bias*, it is understood in its embodiment and more precisely the rhythmical embodiment of a scale. To be on the bias is to understand that to perceive is to labour at scale and at speed. Apparatuses and their related social ontologies may enforce bias as aperception or help exploring a perception on the bias. To embrace the complexity of interpretation and engagement with perception are possible even in a glance. To be on the bias defines automatism more generously than a mechanical application of filters. This definition of bias implies another understanding of labour in the photographic elaboration of computer vision. The value of labour cannot be bracketed, understood abstractly or only in terms of effort. Fairer labour conditions must come with an acknowledgement and celebration of the competences of - and the knowledge generated by - subjects and annotators. To perceive on the bias means that perception is labour entangled with a temporal and rhythmical apparatus in the midst of a thick process of mediation. The horizon of change here concerns the potential reconfiguration of this entanglement.

The thesis does not offer anything near to a solution to the industrial annotation bias problem. And should it? In a Medium article, Helen Nissenbaum, author of the landmark essay *Bias in Computer Systems* (Friedman and Nissenbaum, 1996) provocatively affirmed with Julia Powles, “Bias is real, but it’s also a captivating diversion” (Powles and Nissenbaum, 2018). For Nissenbaum and Powles (2018), recognizing and acknowledging bias is a strategic concession that “subdues the scale of the challenge”. The thesis contributes to identifying what needs to be changed in the elaboration of computer vision for it to continue being a project worth pursuing at all. It provides important elements to figure out the “scale of the challenge”. What are the conditions to make a difference that makes a difference? The thesis contributes to discerning a horizon of change for computer vision. Change is not synonymous with object replacement. Change requires engagement in mediation processes. The research contributes to the reformulation of the questions that must be addressed in order to avoid what Kember and Zylinska (2012) called the replacement trap: an update of the dataset's objects (a cleaner taxonomy, more suitable photos). Fixing bias means “updating to remain the same” as Wendy Chun (2016) puts it, going from one bias problem to the next whilst keeping the underlying system untouched. If the dataset is a scale, as Li claimed, and as I have explored in my analysis and in my practice, its problems won't be solved when those who try to solve them consider it as a mere collection of images and labels. ImageNet, says Li, is an attempt to resolve the scale of computer vision. This scale is increasingly showing its limits. At the time of writing, all the photographs have been removed from ImageNet's website. The researchers are currently reducing by half the size of its People category. But as Olga Russakovsky, the researcher leading the revision of ImageNet, conceded in a recent interview: “this represents only a fraction of the whole dataset” (Yang *et al.*, 2019).

Computer vision needs to go beyond the mainstreaming of the datasets and the damage control of the stagnant taxonomies, because datasets are more than a collection of photographs, and taxonomies are more than lists of names. Taxonomies are managerial devices and datasets are modelling devices. And they are traversed by scales produced through the processes of mediation of photography. The problem faced by computer vision is not just one of improvement of contents nor of an adjustment of the sample quality. What the practice of re-experimenting suggests is that computer vision needs a new resolution. Li is right that the key question to ask is how to resolve the scale of the problem. With ImageNet, Li has given a lesson to her field. ImageNet has been instrumental to re-inventing the computer vision paradigm. Yet to learn ImageNet's full lesson is also to realise the limits of the scale she has constructed. It requires understanding afresh how a scale cuts computer vision's objects and subjects. Local patches and fixes are ways to preserve the existing scale. What we need instead, the thesis suggests, is a re-invention of the scale. A scale that is built on the acknowledged contribution of the workers and subjects. On the acknowledged active role of the mediation of photography. A scale produced in the full awareness of the photographic elaboration of computer vision.

Reading the annotation platform through the re-experiment gives arguments to contend that computer vision's failures and problems are symptoms of how it represses and conceals its elaboration. However, this reading through is not just a critique. It also makes a case for considering these flaws as the signs of an unstable condition, an effervescence. As Anne-Marie Mol (2002) reminds us, nothing is solid enough to be a self-contained unity. What the research did, within its limits, was to probe this effervescence. To stubbornly insist that if more was possible in the Caltech experiment, more should be demanded from the dataset's elaboration. And just as importantly, it makes the argument that demanding more doesn't mean to restore a slow vision or to scale down. It means to explore with a renewed attention what is diverging within the model. Indeed, nothing is inherently wrong with a large collection of photographs, nor with a vocabulary per se, nor with seeing at speed. It is a specific stabilisation of aligned entities and a singular articulation of scales that I am contesting here. A scale is not reducible to a question of quantity. It is an articulation, it is an intelligence of relations, it is a “perspectival optics” (Fuller, 2008) begging to be further challenged and researched.

The practice of re-experimenting takes a step in this direction. With the participants, we have probed different articulations of scales and engaged in the discovery of the effervescence of the apparatus. The multiplicity of its resolution has led to various strategies. We have changed the division of labour that was implicit in the experiment and the annotation platform. And gradually the apparatus became itself an object of experimentation. The work I have done has been to open up the photographic elaboration of computer vision in the experimental apparatus. To do this has required the development of a sensitivity to a whole repertoire of bodily gestures, attunements, ways of apprehending, waiting and even breathing. It drew attention to rhythms, gaps, repetitions, redoublings, and echoes. If the thesis doesn't offer a ready-made solution to the many problems of computer vision, it enlarges considerably the register through which its elaboration can be probed. The practice at the core of the research has not been a mere critique of computer vision but the enactment of relations that makes the latent potential for change in the social ontology and experimental apparatus available to experience.

Furthermore, to trace a horizon of change for computer vision has required to unsettle many boundaries and straddle dividing lines. The demarcations between viewers, subjects, participants, annotators, experimenters and engineers have been fluctuating, and the differences between photographs, images, echoes, stimuli have shifted accordingly. In what follows, I will concentrate on one dividing line in particular, the ones that separates the institution of photography from computer vision, I will reflect upon what is at stake in the demarcation between the two and the potential for change that opens up for the photographic institution when the borders erode.

8.5. The photographic institution in the vision community

My research is situated in institutions of photography, at TPG primarily and within a loose network of institutions of education of photography, and media art. In this section, I reflect on what it meant to have articulated, answered and reformulated my research questions in close relation with the photographic institution. And I am pondering the effects and potential consequences of this collaboration for TPG, the principal collaborating institution.

As computer vision outsourced the modelling work and opened itself to a photographic elaboration, it created indirectly an opportunity for an institution like TPG to inflect its development. The thesis has gradually clarified the terms through which this opportunity can be understood. This required performing a series of shifts: a shift from an understanding of computer vision as code-centric to a process that requires important detours through, for example, the computer vision lab or the annotation environment. And a shift from an understanding of photography as the domain defined by an established tri-partite (photographer, photograph and subject) to an understanding of the flows of mediation and the alignments that resolve the indeterminacy of the photograph. We are now at the limit of what the research has accomplished and we are entering the horizon of change it draws for the photographic institution. The research cannot decide for the institution but makes the case that an opportunity exists, clarifies the terms and connects together a series of elements that allow understanding of the timeliness of such an opportunity, how the opportunity presents itself in a specific conjuncture.

The opportunity for the institution is to claim a place in the epistemic space of computer vision. To claim the relevance of the institution for computer vision on its own terms. What the thesis proposes is a position in a detouring circuit, a position that problematizes the inside/outside dichotomy of the field. It doesn't mean to abide by the rules of a discipline or to adopt emic representations. The research offers an example, a prototype, of the kind of activities that an institution willing to intervene in machine learning could adopt.

Along with the practice comes an invitation to the institution to adopt a certain role. The institution's long interaction with photography makes it potentially a candidate of choice to force the discussion about the photographic elaboration of computer vision. It has a cultural standing that can in part counterbalance the authority of the engineering community over its products. Its status as a long-standing institution could be used as leverage to provoke an opening in the vision community.

There is, in the practice I have developed, an intention to offer a platform for an encounter to happen concretely and actively. The research shows TPG is a space where the “experimental” apparatuses of computer vision can be questioned, engaged with, in ways that are not possible within the narrow margins of the computer vision lab. The research outlines the nature of the relevant knowledge such an institution could produce. The research contributed to making this knowledge available to experience for the team and the audience. Not as a lecture, or as an artwork, but through an embodied involvement with a particular device, the re-experiment. Its mode has not taken the form of a policy document but the exploration of the active potential of textures and its rhythms.

During the research, the institution has been hosting a practice where computer vision was engaged in divergence. The question that comes up now is: will it claim its membership to the vision community and support practices such as this one in their intervention into the photographic elaboration of computer vision? Will it position itself, not as a provider of an expertise that fits what is expected from an arbiter of the profession (give expert opinion on what a good photo is), but as a site where methods, apparatuses, the middling of computer vision are questioned and engaged with? Not through a circulation of critique or aesthetic appropriation of computer vision's objects, but by inserting itself into the detours of computer vision?

This can be achieved on condition that the institution takes on an expansive understanding of its role. It must conceive of its role not exclusively in terms of being an art institution, as an authority on the artfulness of the medium, as an arbiter of photographic professionalism, as a protector of such standards as good photography[92], but as a platform where the various alignments of practices, devices and ways of seeing coming and going under the name photography can be explored in their active role in the modelling of the technologies of vision.

Whether an institution like TPG wants to seize this opportunity to play an active role in the shaping of machine learning dynamics is a question that will only be answered by the institution itself. But the research indicates one way the education department and the digital programme as formations can intervene and challenge the ways of seeing in computer vision. It addresses the institution, its team and its public, as legitimate actors not so much to claim a responsibility as to cultivate response-abilities, a role in working out account-abilities more than the traditional forms of accountability for the processes of teaching machines how to see. This practice has done so, not by importing other forms of accountability (not by inviting lawyers or experts to lecture), but by drawing on an interest in photographic elaboration and by engaging with the institution's own audience and staff, its own apparatus and material configuration. The practice proposed in this research doesn't function as an import, but as a translation, and gives importance to the direction of the translation in a current context where computer science is too often understood as defining photography without acknowledging how the latter affects the former. Photography and computer vision are in a relation of mutual instantiation, and the research offers a means for the photographic institution to contribute to the elaboration of this relation.

It is worth insisting on the current conjuncture and the crisis of trust undergone by computer vision.

[92] Occasionally, this role is reinforced by the Computer Vision community. Photographic experts are invited to give the criteria defining the standards of good photography for photo apps that aim to improve the aesthetics of users' photos. For an analysis, see Katrina Sluis's *Still Searching* posts (Sluis, 2020).

The limits of the model of peer review of computer vision have been exposed crudely to the public these last years. As noted above, datasets have been at the centre of the controversies that gave rise to public scrutiny and outcry. The various critiques aimed at ImageNet have been amply debated in the press, but ImageNet is only one example of the many datasets wherein bias has been exposed, and many of these datasets have been simply moved offline as a consequence. But what is particularly worrying with ImageNet is how it gained such an honorific status within the computer science community while being plagued with glaring problems. How can the recipient of the prestigious PAMI Longuet-Higgins Prize[93] classify *Transexual* below *Anomaly*, *Unusual Person* and next to *Aberrant*, *Zombie* or *Ugly Duckling*[94]*?* How can one of the highest cited papers of the discipline appropriate images without referring to their authors, include people portraits without permission, and recycle racist, homophobic and mysoginistic imagery?

The discrepancy between the understanding of the dataset by computer scientists and the public has never been as painful than these last years. It is important to acknowledge that computer vision has failed and to discuss why it has failed. Computer vision has failed to produce an understanding of the complexity of the very object it claims to be able to interpret: the photograph. To classify a person as a monkey because of the colour of her skin is a problem that cannot be reduced to a mere question of over or underfitting. To diagnose such issue requires considerable cultural skills and to take responsibility for the world such technology enacts. The computer vision community has shown a surprising lack of understanding of the scope and ramification of the problems inherent to representation, mediation and the roles of apparatuses in the production and circulation of images. One reason for this is that these questions exceed what computer scientists consider the knowledge relevant to their domain. ImageNet has been the object of a considerable amount of computer vision papers. Nevertheless, it took 10 years for obvious problems of misclassification and discrimination to be noticed, and, even when noticed, to be addressed.

It is crucial here to acknowledge that the vision community has only started to address these issues because it has been held accountable from the outside. This means that the mechanisms of mutual supervision of colleagues have failed. As noticed, it is only when activists and artists - and scientists evolving at the margins of the field in what came to be named the *Fairness Accountability and Transparency in computing* community - started to point out the outrageous racist slurs and sexist representations that the vision community began to question its methods. This is important because it teaches us something about the disciplinary composition of computer vision and how little its peer

[93] The Longuet-Higgins Prize recognizes computer vision papers from ten years ago for their significant impact on the field (The computer vision foundation, 2020).

[94] For a more extensive discussion of ImageNet's classification problem, see Crawford and Paglen (2019).

review mechanism has done to open the eyes of the community. Countless authors have discussed improvements in accuracy without even examining the composition of datasets. During ten years, peers have reinforced the epistemic contours of the discipline, what counts as a serious problem and what can be ignored.

What I have been contesting during my research are not the contents of the datasets but the mode of knowledge production of computer vision and how it defines what counts as knowledge. It is important here to insist on the issue at stake when discussing the role of The Photographers' Gallery: to provide a context where computer vision's mode of knowledge production, the effects of epistemic boundaries and exclusions can be probed. It supported a research whose premise is that computer vision is a techno-cultural project all along. It offered an environment where the epistemic fallacy of computer vision was not the rule unlike in the most celebrated research labs where the line of research that critiques its objects and scale has been vigorously discouraged[95].

The fact that the institution dedicated time and financial support to develop a reflection and activities around the relation between technology and photography needs to be highlighted and appreciated. And in turn it is equally important to acknowledge that the practice at the heart of this study could exist because there is an alliance between an institution of higher learning (ie. CSNI at South Bank University) and the Gallery. The relation between the two opens the possibility of a research carried out over a time period that differs from the quick pace of an art institution submitted to the imperatives of cultural programming. While the research develops, issues have the time to mature and prepare the ground for an item of programming such as an exhibition programme or a conference. It introduces a dialogue between different modalities and temporalities. This particular articulation between academic research, intervention in situ and public events has been a condition for the research process to develop in a way that draws from the resources of the academy and unfolds in public at the same time. By providing an anchor, the partnership also made it possible to collaborate with other institutions and groups and make the research circulate and benefit from the knowledge produced in other environments of higher learning and independent research groups and collectives.

In conclusion, the existence of such an inter-institutional construction is instrumental to have a chance to interrogate and imagine an horizon of change for computer vision in a context that does not pre-emptively foreclose what contradicts the emic representations of computer science. At a time where massive influx of money coming from the part of the government and the industry

[95] For example, Gebru's termination of employment discussed above.

supports and re-inforces an AI curriculum designed according to the disciplinary principles of computer scientists[96], it is important to underline the relevance of the contributions coming from networks of knowledge production such the collaboration between TPG and the CSNI during these last years. When enquiring into the photographic elaboration of computer vision is deemed either irrelevant or actively discouraged in the disciplinary precincts of machine vision, these questions cannot be asked and researched without an alternative network of actors and institutions that support such research. The very existence of the present study shows how the consolidation of such partnerships is the pre-condition for critical questions to be asked about the experimental formation of machine vision, its scales and its photographic elaboration.

In short, as computer vision increasingly relies on a process of photographic elaboration, the photographic institution could become a relevant actor that would support a renewed understanding of the thick mediation that informs computer vision's objects and subjects. The photographic institution would consider itself as a platform where the various alignments of practices, devices and ways of seeing coming and going under the name photography can be explored in their active role in the modelling of the technologies of vision rather than a site for the celebration of the artistic character of the medium. As importantly, in the current circumstances where computer vision as a discipline has shown its inability and even at times, its refusal to engage with a critique that goes beyond representational bias, it is of crucial importance that the institution continues to operate as a node in a network of knowledge production that contributes to rethink the horizon of change of machine vision.

[96] A good example in the UK is the AI Sector Deal based on a report written by Regius Professor of Computer Science, Dame Wendy Hall and Facebook's Vice President of AI, Jérôme Pesenti. As an outcome of the deal, the government committed to invest £406 million in “maths, digital and technical education, helping to address the shortage of science, technology, engineering and maths (STEM) skills” (Gov.uk, 2017)

Appendices

Appendix 1. The tank top

The writing in this section contains excerpts from the notes I have taken while listening to the recording of the third session conducted at The Photographers' Gallery. This section concentrates on the effort of a participant named J, during the first round where she responds to a stimulus. I have chosen these notes to emphasize the importance of listening as a means to attend to the embodiment of the participant and the micro-dynamics of the situation. The fragment begins after a participant inadvertently turned off the microphone.

Notes from the recording

00:00:00. The microphone is turned back on.

00:00:00 → 00:00:18: During the first 18 seconds the microphone changes hands and the presence of soft noises indicates a contact with the object. The participant receives the microphone (“thanks” 00:00:03.297). Chairs are moved, footsteps in the background indicate the incident has used as a short break to fetch tea or biscuits offered by the gallery. These movements resonate in the room. The room doesn't absorb sounds and produces cold reverberations.

00:00:18 → 00:00:21: Three seconds of silence. Around 00:00:20 the feint noise of S clicking with the mouse to trigger the stimulus. The room is silent. No external noise interferes with the participant's concentration.

00:00:21 → 00:00:35: J laughs. She apologizes, she has been distracted by the turban of the person in front of her partially masking the screen. She asks to be shown another stimulus. She addresses S and the person in front of her in the same sentence: “Can I have another … XX you have to stay still because your … your ...”. The request for a new stimulus and the remark addressed at the other participant are flowing into each other interspersed with laughter and the reactions of the two addressees. While S accepts to launch another stimulus (“no problem”), the other participant finishes her sentence … “the turban?” J apologizes again while laughing. Neither S or the other participant join her laughing. J asks to the other participant to stay still and concludes with “euh ok” to mark the end of the incident. To which S responds (background: “ok no problem”)

00:00:35 → 00:00:43: Eight seconds of silence. Feint noise of visitors passing by in the corridor. (not sure). No external noise interferes significantly with the participant's concentration.

00:00:43 → 00:01:57: The main description of the image. J breaks the silence: “OK”. J speaks slowly and articulates very punctiliously. She is aware she is dictating her text and S is typing for her. She is also aware that English is not S's first language and that she delegates the transcription to another body. She makes sure that S transcribes the words in extenso. And repeats willingly, making various pauses to ensure S has enough time to write everything down. The transcription is a constraint that slows her down. She has internalized the constraint. S nevertheless has to ask her to come back to some details and create a little confusion as she has to “rewind”, to come back to an earlier stage of the description. There are long pauses between the sentences. To let the transcription work happen but also to a certain extent as a way to let her impressions crystallise. She checks if the description meets the experimentalist's expectation “It's too much information isn't it?” as well as if she doesn't exceed S's capacity (too slow to type what is said integrally) The relation of delegation is double. S writes for her and she describes for S:
- as S is typing, she delegates the writing, she “dictates”.
- as she describes the image, she checks if she behaves correctly, meets the expectation.
The slowness of the typing forces J to maintain her memory of the fugitive instant alive. Even if there are long pauses between the sentences, the tone she uses to end the sentences suggests that the description is not finished. She uses different markers to punctuate her description. Three sentences out of four are prefixed with a sequence of “hum” and a click made with the lips. They mark intervals, they are giving the tempo, the beat. The feint presence of the keyboard is barely audible in the background during the pauses between the sentences. Although the keyboard is not audibly present in the foreground, it is addressed many ways indirectly. The speed of the transcription, the delegation, the slowness of the voice, all these elements relate to the keyboard as a rhythmic device. Additionally, the keyboard as a tool for data entry, enables the transformation of what is said in what is written. The presence of the keyboard emphasizes the fact that the descriptions are archived. The apparatus of memory is shown in action. Even performed. By S and by the participants. But this performance, this emphasis on the writing, on the conversion from speech to text also conceals the fact that another recording, the audio recording is taking place. Even if the participants have the microphone in their hands, their attention is continuously directed on the screen, towards the keyboard.

00:01:57 → 00:02:13: Some noise is produced by contact with the microphone. It had not been touched during the description. A sign that J has finished and is ready to hand it over to the next participant. But before she does it, another participant reminds her that she mentioned the woman in the picture was wearing a white top. The participant notices that S had not written it and wants J to be aware of it. Instead of talking to S directly, he speaks to J. J immediately remembers and repeats the information. She specifies the top's style. And repeats for S who didn't get the nuance. The description is added on the description. Maybe J considers this information secondary: she continues to play mechanically with the microphone. Another participant coughs while she repeats. Different signs that the description is about to conclude.

00:02:13 → 00:02:31: J hands over the microphone while one can hear the sound of the keyboard in the background finishing the note-taking. Twenty seconds of silence while the participant waits for S to resume writing and deliver the next stimulus.

Additional remarks

J does a lot of work in this fragment. She repairs the experimental set-up at different levels. The experiment has been interrupted by the previous participant who turned off the microphone. J's role is to respond to a stimulus. Additionally, she takes on herself to put the experiment back on tracks, and to help everybody refocus on the work after the previous participant had interrupted the course of the re-experiment. When she receives the microphone, she sets the tone for the rest of the session. When she is about to do that, her vision of the screen is partially masked by another participant. Again she has to repair the experiment. She asks the participant to stay still. To do so, she has to use different skills. She needs to be firm while being tactful with the other participant. At the same time she has to convince S to give her another "chance". To convince S that what happened was something in need of a fix. Which she skilfully does by expressing the two requests at the same time. While laughing to nuance, soften the importance of the request.

J ensures the continuity of the experiment in more than one way. She has to deal with the slowness of the typing as mentioned earlier. To keep a pace that is sufficiently slow to be correctly written down whilst not loosing the thread she is following. Speaking slowly is the solution she finds to cope with the writing speed and avoid interrupting herself. J works on the continuity of the experiment and the integrity of her own visual memory at the same time. She negotiates the interruption provoked by the turban in her visual field, she ensures the transition with the previous participant. S's request to repeat something she already said destabilizes her: "the woman and the camera?". Going back to a previous thought confuses her for a moment. As if to come back to what she just said would jeopardize the rest of her description. Conscious that the recall cannot be summoned at will[97].

The other participant's intervention shows how they contribute to the maintenance of the integrity of the experiment. They are not just “subjected” to it. They are also its “janitors”. They perform checks and balances, they maintain and repair. There is no formal instruction regarding the content of the transcription. The level of exhaustivity of the transcripts is variable and the participants notice it. The intervention of the participant who notices that a part of the description is missing from the transcript shows how engaged they are even when it is not their turn to respond to the stimulus. The participant “does the exercise” silently. The intervention shows that the participants observe the experiment and its conditions and react to it. And they feel authorised to take the initiative to comment on somebody else's description.

[97] As if juggling with several balls. As long as they are in the air, she can play with more balls than hands. But when the movement stops, the balls are falling.

Appendix 2. The echo chamber

The re-experiments as they have been conducted in the first phase of the research were constructed in different stages. Two stages were closely related: a first phase during which the participants were asked to recall what they had seen on the screen (I referred to this stage as *Stimulus-stage,* or *Stimulus-component,* in the Re-experimenting chapter, especially sections 6.6 and 6.7) and, a later stage, during which they were shown the stimulus-photograph for a longer time, asked to comment on their recall and more largely on the experiment (I referred to this stage as the *Feedback-stage,* or *Feedback-component,* see section 6.1.). These activities have been recorded by various means. The oral description of the photograph as well as the discussion of the feedback stage have been audio recorded. A transcript of the oral description of the photograph has been kept in a database.

The echo chamber proposes a protocol to listen to the recordings of previous re-experiments. As I explained in the Account of practice, I felt the need to share the listening with the participants and to open the analysis to a collaborative questioning. The echo chamber takes the close listening described in appendix 1 into an environment where it can be shared and where the modes of engagement with the recorded material are extended.

In the Account of practice, the argument feeds on the observations, analyses and intuitions that emerged from the echo chamber session. In this appendix, I am entering in the details of the configuration of this apparatus of listening, in the details of the interactions that took place as well as in the reflexive process that lead to formulation of the observations and analyses used in the thesis.

Echo chamber set-up

The echo chamber is a room in which the following procedure is implemented:

- Five participants, who volunteered for the task, are given a sheet of paper with a transcript of a discussion from an earlier session. They have not experienced the stimulus the subjects describe in the transcript. Each volunteer impersonates a participant of the earlier session and lends his voice to this participant. The five participants are sitting in a row in front of the others who listen to them. This is the *Replay-stage.*

- The five participants join the others at a table and they experience the stimulus together. They are discovering it visually for the first time. They discuss briefly what they imagined from the transcript and their own recall. This is the *Discovery-stage*.
- The participants listen together to the recording of the conversation of the earlier session. They only had access until now to the transcript. They hear the voices, the accents and the laughter of the participants to the earlier session. This is the *Listening-stage*.
- The participants discuss what they have heard and seen based on their experience of re-enacting, experiencing the stimulus, reading the transcript and listening to the recording. The stimulus remains visible on the wall during this phase, the notes they used for the replay are available and when needed, the recording of the earlier session is replayed. This is the *Comment-stage*.

The echo chamber in action

To show the echo chamber in action, I have chosen in this appendix to develop in more details a fragment of a session already introduced in chapter 7. My angle is different as I am focusing here on the processes, the dynamics and the organisation of the device rather than on the content of the analysis made by the participants. The echo chamber session is based on the recordings of an early re-experiment that took place at TPG. To briefly recapitulate the content of the recordings, four participants are trying to recall a stimulus they have experienced for 53 ms. They suppose that the environment depicted in the stimulus is a shop. They cannot confirm whether there is more than one person in the scene. They mention the possible presence of two elements in particular: a brown paper bag and an object identified as Captain America's shield. After these hypotheses have been formulated, a participant explains that he could not make sense of what he saw although he recognizes to have perceived “colours and shapes”.

This appendix concentrates on an instantiation of the echo chamber in Brussels wherein other participants engage with the recorded material. I have chosen to follow how the three elements (the brown bag, Captain America and the participant who “couldn't picture”) are echoed, discussed, translated by the second group of participants. And how these elements are reconfigured through – and interfere with – the different devices that are made to listen to them.

The choice to listen this particular fragment with other participants was motivated by various elements. I wanted to find a way to hear the full affective spectrum of a session. This fragment contained more than high or low intensities. It also provided an example of how negativity was at work in a session: what was retracting, avoiding, interrupting. In complement to the very explicit ways in which Captain America and the brown bag were discussed, I was interested to hear what the participants had to say about the intervention of the viewer who “couldn't picture” what he saw. I wanted to address the difficulty expressed by this person. I was interested to hear their thoughts and see how they would respond to the expression of the negative statement, as negative statements had rarely been uttered. Furthermore, I was interested to listen with them to what surrounds the statement, to collectively intensify the listening of the non-verbal, the way such an elusive statement is given importance. More than for the brown bag and Captain America, I was curious to experience together another form of listening, to test what collective attention, collective hearing could bring to the fore when the semantic content was so minimal and “retractive”.

In what follows, I examine how a description is read and listened at in different ways in another location, how each re-reading or re-hearing helps different elements to gain presence and intensity. Finally it follows how the participants involved mobilise the listening and the various devices to reflect upon the dynamics of their collective descriptions: the game of making sense.

Replay stage

Several weeks after the London session, five participants volunteer to re-perform a fragment of the session in Brussels. They are given a sheet of paper with a transcript of the Stimulus-phase. Each Brussels volunteer impersonates a London participant. The five participants are sitting in a row in front of the others who listen to them. The volunteers are reading from the transcript.

When the participants read these lines, they do it neutrally, mechanically. They are deeply concentrated on their character's lines. They give the impression of reading against the grain, carefully trying to avoid stumbling over their words. Even for a simple line of four words as “a brown paper bag”, they carefully insert pauses between each word. Each sentence is clearly delineated. This gives a great importance to silence during this moment of re-enactment. The turn-taking is emphasized, each phrase resonating distinctly. The transcript has minimized the overlaps between the speakers. It gives the impression of listening to a slow mechanical reproduction. This is the first effect of the echo chamber: a speed reduction. A particular rhythm and a careful listening. But this slowing down has a disruptive effect. The rhythm is too slow to convey meaning seamlessly. The sentences are disarticulated. Read too evenly, they become harder to grasp. The pace of the reading reaches a threshold below which the meaning of the phrases dissolves.

Even if the transcript contains indications of laughter, the participants keep using a neutral tone and continue with their careful rhythm. But an element disrupts the reading. In the transcript, an indication ("question from the back") provokes a moment of wavering as the participants cannot make sense of this information. A sequence of laughter ensues. When they start reading again, something has changed. The laughter has breathed life into the reading. This dialogue fragment is on the second page of the transcript. One can hear the participants turning the page and this introduces a small delay before they continue reading. In what follows, the tone remains rather mechanical but much less than when they were reading the first page. The accident, in some sense, reinvigorates their reading, re-injects emotion and expression in the performance. They recover motility. After the last line "Captain America it's very good", the participants burst into laughter again.

Discovery stage

When the reading of the transcript ends, the performers join the other participants at the table and they all experience the stimulus together. The photograph is displayed on the screen next to its description. First, the participants see in the photograph a confirmation of what they heard: "definitely a brown paper bag". Even if they are looking at the photograph without time constraints, the performance has affected them. It is only after a few minutes that one of them realizes they have hallucinated the brown colour and corrects the description: "white, no?".

For the group, Captain America is not an obvious reference. Two participants are more familiar with the Super Hero character and its distinguishing feature: "he has a shield I think and on the shield there [...] the star". Even with this information and the photograph in front of them on the screen, the various elements making up Captain America's identity do not really coalesce into a coherent image: "is it a star?" Whilst, for the TPG participant, 53 milliseconds were enough to spot the comic book character. Neither the bag or the Super Hero are given a conclusive definition. The composite echo continues to propagate.

Finally, at that stage, the participants do not refer to the "I couldn't picture" sequence. They mention the bag and Captain America, but the difficulties expressed by the participant who had merely perceived colours and shapes are overlooked.

Listening stage

After having re-enacted the description and seen the photograph, the participants listen together to the recordings of the London's session. The participants are silent. The recording of the stimulus-phase is played. The amplified sound fills up the room. The comfort of listening in this condition helps to grasp the meaning of the sentences and to concentrate on the grain of the voice, the hissing sound of the recorder. Some participants close their eyes to listen more carefully. The voices of the London participants can be heard saying the same words again “a person in a dark shirt I think filling a bag” … The words “Captain America” resonate several times in the room as the recording goes on. The participants do not give any sign of amusement or surprise when they hear them even when they hear bursts of laughter. Neither do they express any reaction when they hear the participant's words describing his inability to disentangle his perception.

This time, the words they know from the transcript are experienced through the technical reproduction of the voices. The words come with the tones and rhythms of the speakers' voices. The performance had de-synchronised and slowed down the words at the risk of loosing their coherence. The audio player re-synchronizes them.

Comment stage

The participants comment on the reading of the transcript, on the recordings and the photograph, after having engaged in different forms of reading, hearing and speaking in the previous stages.
At this stage the participants have access to recordings and transcripts, and they are in charge of the course of the conversation.

Listening to tone

Perhaps at this stage, the biggest challenge for the participants is to find a manner to listen with the same concentration to the content of what is being said and to the more general dynamics of the conversations in the recordings. They spend great effort finding a way to balance attention to the contents of the descriptions with sensitivity to the tone, the intensity of what is being recorded. As they have experienced, the transcript is tone-deaf and the recording can be overwhelming. There is an inherent tension in the listening.

During the comment stage, the participants listen to the fragment where the participant is about to say that he saw the Captain America logo. The recording is played and introduces another form of presence through the amplifier. AM, a listener of the echo chamber, suggests that the participant is “announcing a childhood memory”, “something intimate”, “a confession”. I have shown in chapter 7, how important was this moment, how alignments shifted, how the logo named more than an odd item in the stimulus but was the symptom of a larger shift in the stabilisation of the re-experiment. What is important here is to consider that what stops the listeners and makes them notice that something is happening does not pertain only to the meaning of the words but also to the tone. It is the hesitation, the intimate inflection, then the embrace, and then the joy and the explosive sound that comes along with the mention of the logo that catch the listener's attention.

At this point the participants encounter a double difficulty: to attend to the tone, the intensity conveyed through the recording and to resist to “psychologize” what they are hearing, to flatten what is occurring, the aural event, to the personal feeling of a subject. For that reason, hearing the confession is a difficult exercise as it seems to limit the event to the protagonist. The mention of Captain America troubles the whereness, the location of the event. As I said earlier, it brings the gallery in the lab, it complicates the re-experiment's border. But there is, at the same time, a temptation to ignore this movement in the spatial enfolding of the re-experiment and locate everything in the speaking subject. This difficulty remains present throughout the session. It remains as a tendency. But this tendency is compensated by another. The participants concurrently develop an awareness of the fact that the recordings are not giving them a direct access to the subjective states of the people speaking in the recordings. They develop techniques to relate to the various recording devices.

Techniques of listening, Insistent repetitions

The participants are seeking to relate themselves to what they heard in ways that involve intensive forms of repetitions that may appear “compulsive”. There is a sense in which the repetition never ends. For instance, in this fragment, the participants go back to the transcript and try to make the words come to life again through their own voice in a sort of repetition of the Replay-stage:

C: colours and shapes
MB: hm?
C: colours and shapes
MB: yeah
S (nodding): hm
MB: does he say colours and shapes?
C: no it was his first
S: yeah
MB: ah yeah colours
AM: colours
S: colours
C: colours and shapes I think there was
MB: ah yeah
S: colours and colours I think there was
S: then he says shapes
MB (overlapping): shapes

The participants are taking a first pass at the words of the subject who “couldn't picture”. There is no particular attempt to try to make sense of the words. The participants are hovering over the words. At the surface nothing happens besides what sounds like an absurd form of repetition. Reading the words, here, on this page, it is difficult to feel what is happening in the room. And even in the room, it is easy to overlook the importance of such a moment. Yet there is a process of collective synchronisation, attunement that is perceptible. A form of chorus, a mutual attempt to let the same words resonate across the different bodies. It seems like the meaning of the words matter less than their rhythm. “Colours and shapes” sounds like empty words. But it is exactly the charge of their emptiness that is at stake at that moment. How their emptiness is bearing on the conversation but also increasingly bearing on the composite echo that the participants had created together. The composite echo continues to propagate, it doesn't stand still.

Quoting and ventriloquing

Contrarily to the words of the participant who “only saw colours and shapes”, the mention of Captain America lends itself almost immediately to analysis. Yet even if the participants are eager to interpret the content of the Captain America exchange, they do so by trying to relate affectively to the fragment. The group mobilises a large range of techniques to manifest Captain America's presence: ventriloquism, deep quoting and paraphrasing. These techniques are not designed in the echo chamber, they are brought in by the participants. There is a co-elaboration. And each of these techniques has a specific impact on what the repetition produces. Let's observe these techniques in action.

In an exchange that begins as a sort of echo from the earlier phases, the participants are repeating the words that have been read from transcripts and then heard on the recording.

> MB (overlapping): No I think she was saying that to person two
> MB: “Ah Captain ... Captain America that's very good that you spotted that ”
> H: ok
> H: yes because there was also timing and a comment on that
> S: No but you are pointing to thing that I ... but ... got my attention also ... because hm he says line 18 “I think I saw a logo like a company logo” and then you hear the smile in his voice and he is
> AM: (nodding)
> S: preparing to announce “Captain America” and hm and then hm the next participant says “I saw it too” but the woman who comments on Captain America at the end just continues on her thread and hm “the counter is in the foreground” she doesn't ... catch up with the ...

When they are commenting, the participants make a heavy use of repetition. Sometimes literally, or from memory, at times with intonation or neutrally. Before the above fragment begins, two participants, H and MB are trying to understand the context of the recording. Had the participants of the London session seen the image? Did they talk to each other? To answer these questions, MB slightly paraphrases a character in the transcript: “very good that you spotted that” and wonders to whom she is speaking (“I think she was saying that to person two”). S goes back to the transcript and quotes it directly, mentioning even the line number in the transcript (“he says line 18”).

It is important to notice the nuances in the way the two different participants, MB and S are “repeating” the words from the recording and the transcript. In essence, MB and S deploy two different techniques: MB paraphrases and S quotes. When she paraphrases, MB attempts to do several things at once. She tries to breath life into the words while saying them aloud, to make them count, matter, vibrate. She also reformulates the written sentence to clarify its meaning. Or more simply to recall what she heard and the sensation she got from listening to the recording. Or to revisit her own sensation when she was reading from the transcript at the replay stage. With the “ventriloquy”, she performs a work of translation and interpretation grounded in affect. She aims to make H, S and herself understand what the original words meant, rather than refer to the transcript.

S's quoting is different. It is a reading of the transcript. In this case, the authoritative source is the text, not his memory. It is also an attempt to reconnect what he heard from the recording to the transcript (“you hear the smile in his voice”). The quotes are articulated along a storyline that insists on a linear account with a succession of events: “and then”, “continues” etc. MB and S are both “sounding” the recorded discussion, but they engage in a different bodily relation with the words. MB makes them hers through the voice and ventriloquism, gauging their intensity while S quotes them as excerpts in his own narration, his retrospective reconstruction. They make the words matter differently.

The recording is not a transparent reportage. MB and S are part of different alignments. One where words, apparatus, and transcript mutually stabilise each other provoking a linear reconstruction of the London session. And another where a surge of affect, the intensity of a sound, the vocalizing are aligned with the audio recording device. This represents more than two interpretations but two engagements enrolling devices, rhythms and affects. Their conversation does not resolve their differences. The voice that uttered Captain America moves from one alignment to another. It is not fixed and requires a negotiation, a passage from an engagement to another.

Affective attunement, assonance sensing

Engagement, negotiation imply something different than a distant listening of the recorded objects. Affect is not just a variation of intensity captured in the recording. It is also a mode of attunement to what is being heard from the part of the listener. Intensity cannot be contained in the recording, it overflows. It does something to the listeners, it is also among them, it moves them. There is a difficulty to hold affect in place. The echo chamber seeks to give room for that displacement. The attunement experienced in the echo chamber is not just a relation to peaks in volume. The attunement experienced in the echo chamber also relates to more subtle forms of intensity, tensions and pressures felt by the participants.

To be more concrete, I will turn to how the listeners approach a fragment where a participant talks about his inability to picture or resolve the stimulus. In this example, affect is not related to the contents of the stimuli but to a general pressure exerted on the re-experiment's participants: the pressure to see something.

Right after the fragment had been played, a comment comes from AM who participated in an earlier session. From her experience of being one of the participants for whom “it's super difficult to see anything”, she considers the re-experiment stressful because one is supposed to see “at least something”. There is an implicit imperative. According to her, there are different kinds of participants: those already mentioned for whom it's difficult and the others “who see a lot” and who are competing with each other. Those who do not see, then, are facing a three-pronged alternative: shutting up, making up (“am I going to invent something?”), or going with the flow (“when people then start confirming things that you just hook on to something that's something vague corresponding I can imagine myself doing it”). Whatever the option chosen, to participate feels risky.

This observation by AM is far from benign. AM speaks for a long time uninterrupted and the only other voices heard are mine and two other participants nodding and encouraging her. There is a change in mood, of atmosphere. AM expresses something difficult and the general tone of the conversation is shifting. The listening has intensified. Attunement is not disconnected from analysis. Instead analysis matures with the repetitions and the intensive attunement. Getting sensitive to the aural movements nurtures the process of sense-making. Sensing movements and pressures are pre-requisites to understand how the participants are taking decisions, how they are making moves in the re-experiment. Assonance sensing as a method to engage with the recordings is crucial for the participants to discuss how levelling operates in the re-experiment.

As I explained in chapter 6, there is an implicit game where the description of an object has more value than an impression of colours and shapes. When a description reaches a given level of description, the participants who cannot give more precise details refrain from speaking. If “you have missed that level”, then “you feel outside of the discussion”. For instance, when two participants discuss how to interpret the presence of people at a counter in a shop, they may add details or objects. But a statement like “this is an interior” is already implicitly contained in the idea of a retail store, therefore adding it is considered irrelevant. And if it is the only thing a participant could add, he has nothing to add: “nothing nothing that you're gonna say then is going to contribute to describe the image”. Interior is already contained in shop. MB calls the dynamic of collective description a “game of making sense”. Without AM's intervention, these dynamics would have simply stayed implicit.

Uncovering this rule is important to follow how the composite echo is elaborated. It is also crucial to attend to other levels of interactions in the recordings. C observes that she applied this implicit rule even when she was reading the transcript or hearing the recording. She was paying more attention to the “Captain America and paper bag thing” that were giving themselves as already legible or audible objects. It is only after a long and insistent series of repetitions and desynchronisations that she directs her attention to the “I couldn't picture” intervention. She concludes: “I realize myself that I am biased towards what is clear”. This understanding is inherently connected to the ability to hear other things, to be affected by other moments in the recordings.

Here it is important to nuance that what is at stake in assonance sensing is not only hearing what was said but not heard. It is also to be able to attend to the active force of negation. The understanding of the game of making sense progresses by understanding better how the refusal to enter the game affects it. What needs to be listening at, is not just that a participant could not picture and therefore does not level. But that the very act of saying it transforms the dynamics of the session. According to AM, by uttering these words, the participant discredits the others. He is not only commenting on his inability to picture, he is questioning the validity of any other description. The negativity of “I couldn't picture” has the power to discredit the others:

JL: distinguish anything then it completely kills that process
S (nodding): hm
AM: Yeah
JL: of collaboratively
MB: yeah to build ...
JL: picturing the image

Summary

This appendix is a window into the Echo chamber of the re-experiment. The echo chamber implements a protocols that runs through different stages to explore the recordings of past sessions. Each stage proposes different forms of engagement with the material and different modes of listening and attuning with the transcripts and audio-recordings. I have shown how each stage addressed the re-experiment specifically and how the process culminated in a stage where the initiative was left to the participants.

The echo chamber is a device to share the listening and the process of analysis with the participants who marked their interest in being involved deeper in the process. The echo chamber proposes a mode of feedback or response to the re-experiment. It seeks to elaborate a mode of response-ability that is congruent with the re-experiment’s method.

As a listening device, the echo chamber faces the difficulty of facilitating a listening where the objects are not taken for granted, where the recordings are not treated as transparent reportages of what has been said and where the relation of voices and subjects is not one of evidence. Through its various stages, it proposes a mode of listening that gives importance to affect, that forces resonance, that explores de-synchronisation. It also makes room and time for what apparently lacks meaning. The listeners are encouraged to go through countless phases of meaningless repetitions, through moments where they cannot reconcile meaning and transcript, where they cannot stabilise in a linear narrative the explosive nature of what they hear or its nearly inaudible presence.

The analysis does not come from a transparent reading from the transcript or straightforward hearing from the recordings. It is the fact that these forms (the transcript and the recording) cannot be reconciled that provokes a different listening and a different reading of the transcript. This invites the participants look elsewhere and listen elsewhere, displace the ear and the eye. Contrarily to a classical procedure of member checking, the participants are asked to move away from readily legible and audible objects. There is a provocation at the core of the echo chamber, an intention to set the listening in motion. But the listening is on the move only at the condition that the participants engage with it and bring their own methods to sense the recordings and transcripts.

I have identified three participant's methods at work in the echo chamber. Through a process of insistent repetitions, the participants gradually familiarise themselves with the intensities traversing the recordings. Through quoting and ventriloquing they inhabit the gap between the recordings and the transcript and attend to the difficult task of following the voices. And finally by sensing assonances, they affectively attune to the different forms of pressures and movements that leads them to objects that are not readily audible or to the work of negativity (the refusal to participate) to the re-experiment. These techniques, mobilised in response to the provocation of the echo chamber, inform the analysis they produce and extend the scope of what contributes to the dynamics of the re-experiment.

Appendix 3. Taxonomic devices

In the main body of the thesis, the taxonomy was introduced mainly as a means to discuss the stabilisation of the experiment. I analysed it as a device that holds the experiment in place and limits the effervescence of the apparatus. Essentially, I have presented the filtering stage as a moment where the epistemic compass changes and where the ground on which the participants had established a consensus is altered. I have emphasized the importance of this process for two reasons: for its role in the transformation of the photographic alignment and because it dis-involves the participants. The participants are not confronted with highly sophisticated device that requires them to "think with" but with a device that presents arbitrary and limited categories and expects the participants to "dumb down" the descriptions.

In this addendum, I introduce several elements that help to understand the circumstances, the process through which this analysis has been carried out and how the participants were involved. I do this first by looking more closely at the visual layout of the taxonomy, its formal and logical design and how it distributes differences in depth and precision. Having done that, I introduce an exercise through which the participants use the taxonomy to filter the descriptions and experience the varying degrees of granularity of the classification. Through this exercise, I am giving concrete examples of how the participants experience the classification's shortcomings. As already observed in other stages, the participants bring their own methods and mobilise the device in unexpected ways. However, the taxonomy resists the efforts from the participants to make it more flexible, extend it or repair it. These methods were only alluded to in the 7th chapter, this addendum offers concrete examples to support the chapter's argument.

The attributes list

We know already, from the reading of *What do we perceive ...*, that the classification is used as an instrument to turn response into perception. In the thesis, I have shown the classification tree in action and pondered its effect on the course of the re-experiment. It may be useful here, as a complement, to have a closer look at its semantic organisation and observe how it manifests itself in the form of a graph, as an image of structured knowledge, separating areas, connecting others, and decomposing categories in discrete units.

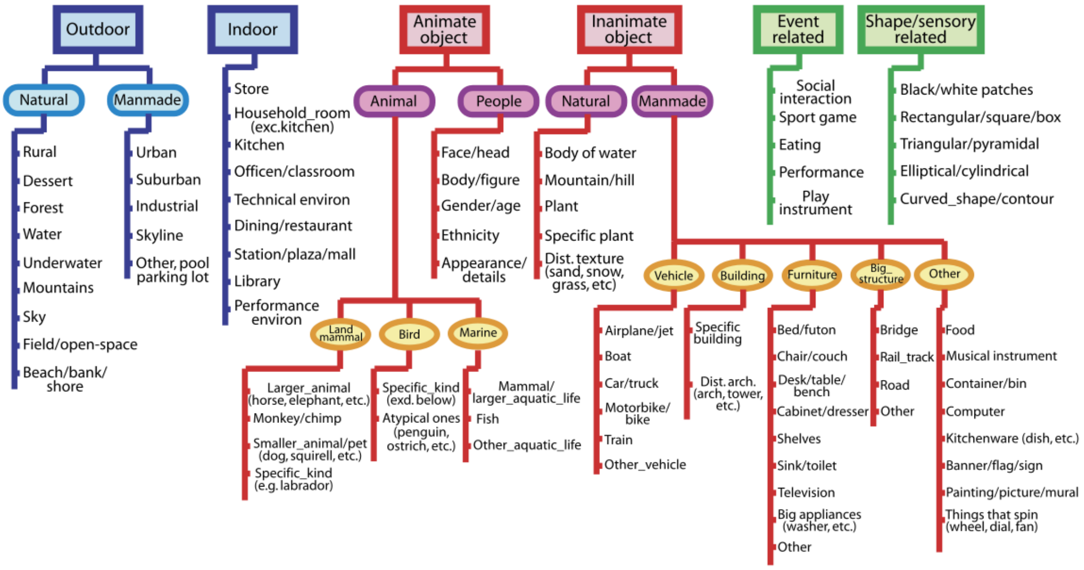


Illustration 8: The vocabulary tree

The classification (illustration 8) is composed of three pairs of categories: *Outdoor* and *Indoor*, *Animate object* and *Inanimate object*, *Event related* and *Shape/sensory related*. These categories are displayed in a tree-like fashion where sub-categories are connected along vertical axes. Each pair is given a different colour. There are striking differences between the categories: two branches of the same level may contain different amount of attributes, and two categories may have considerably different depths. They differ in content (quantity of items) and extent (ramification). The researchers conceived the taxonomy as an abstract set of relations. But once the taxonomy needs to be displayed in a visual form, material considerations come into play. The tree needs to be drawn so as to efficiently occupy a limited surface. The difference between the various branches guides their arrangement on the page.

Located in the top right quadrant, in green, *Event related* and *Shape/sensory related* are the smallest superordinate categories. They do not contain sub-categories and fit in a limited space. The green pair exhibits a perfect symmetry both structurally and in quantity (5 attributes each). On the other side of the image, the blue pair, *Outdoor* and *Indoor*, diverges structurally. *Outdoor* is subdivided in two axes, *Natural* and *Manmade*, whilst *Indoor* contains a single axis of 9 attributes. *Outdoor* seems to require a more complex structure than *Indoor*. A central pair holds *Animate* and *Inanimate* objects. Here, each category contains a branch with two levels of subdivision and a branch only one level deep. These latter categories, *People* and *Natural*, are drawn side by side at the centre of the image. Their structural and quantitative resemblance (5 items each) and the fact they are coded with the same colour suggests an analogy whilst their lineage separates them out. The deeper branches of the central categories are sprawling in opposite directions. *Manmade* with its 5 sub-categories occupies the space below the green categories while *Animal*, with three sub-classes, steps into the *Indoor*'s territory to colonise the empty space below it.

The imbalance of the categories and their opportunistic distribution over the page surface are the first signs that convey the impression of a provisional structure. The closer one looks the stronger this impression becomes. The categories are deployed vertically. On the diagram, the attributes of a same category are stacked on top of each other. They also develop horizontally by multiplying siblings. *Inanimate object* with its nested structure of 41 elements (nearly half of the total number of items) reaches from the top to the bottom of the diagram and radiates from its centre to its right extremity. The attributes are labelled in part with single words (*Forest*, *Water*), multiple words (*Musical instrument*, *Specific plant*), and more rarely adjectives (*Rural*) or verbs (*Play instrument*). The forward slash character expresses alternative labels with varying degrees of synonymity (*Body/figure*, *Dining/restaurant*, *Desk/table/bench*, *Sink/toilet*). And finally using parenthesis, the researchers attempt to define some labels or to give a sense of their extent: *Atypical ones (penguin, ostrich, etc.)*, *Things that spin (wheel, dial, fan)*.

Now that we become more familiar with the taxonomy, we can begin to sense how the eclectic nature of the labels relates to a patchwork of classification criteria. Animals are defined in turn according to their biological taxa (*mammal*) or their size (*larger animal)*, or their typicality (*typical/atypical*), three criteria corresponding to distinct classification traditions: biological taxonomy based on species, a classification based on measurements (big versus small) and the archetype theory of Eleanor Rosch (1978) where typicality plays a key role. The tree does not attempt to articulate its heterogeneity. Categories issued from different classification schemes are juxtaposed. Not only are these criteria juxtaposed but there seems to be no overarching principle that would define the extent of the categories. This question becomes crucial when terms such as “others” or “etc” appear in the columns. Does it mean that categories explicitly denoting the logical continuation of a series (the categories that include terms such as *others* or *etc*) must be understood as non-exhaustive? And therefore offer some leeway to the participant to extend them? And as importantly, for the categories that don't, should the list of items be therefore understood as strictly limited to what is specified? This question is made even more pressing when considering the discrepancy between two branches of a same level. How can *natural inanimate* objects with its 5 entries compete with the 34 entries of *manmade inanimate* objects? Finally, the eclectic nature of the classification scheme raises another regarding the relation between the items within a category:

are they disjunctive (urban or suburban, larger animal or smaller animal)? Or can they be combined (gender/age and ethnicity)?

The visual diagram contains the whole taxonomy and provides an overview by administrating drastically the visual space. It displays structural differences and resemblances between the categories. Paying closer attention to its distribution over the page shows their imbalance. In the limited space allocated to the labels, it condenses names, information about synonymity and disambiguation, and indications of extent. The tree offers a map to locate levels of discrimination. Its dense and nested structure, its minimalist labelling compose a visual device introduced in the re-experimental apparatus. The whole drawing gives an impression of solidity and structure, yet under close scrutiny, many loose ends begin to appear. Lists are incomplete, and at times ambiguous. The graph presents itself as a transitional sketch that has become a reference: evocative and yet rigid. If the first part of the experiment presented itself through the sealed black box of the micro-time management device, the second part conveys a different feeling, less formal, ambiguous and imperfectly "regularized". The same remark could be made here as in the opening description of the Faces1999 dataset. We are in the experiment's back-end where the sense of accountability and the disciplinary standards are loosened.

Engaging with the taxonomy

As in the previous stages, the different ways by which the participants mobilised and engaged with the experimental device have been crucial to understand the experiment as a process that was both actively shaping the action of the subjects and in return very much transformed by it, a process of entanglement rather than one of mechanical response. The taxonomy however has a particular way of soliciting the participants' contribution and minimising it. More than in the other stages, the taxonomy operates as a script, a device whose purpose and strength is to reproduce behaviour and re-affirm boundaries (Akrich, 1992). By giving a close look at the taxonomic tree, I have seen how many categories were left incomplete or were opening space for interpretation. This close look at the graph gave a sense of the arbitrary nature of the size of certain categories or the kinds of items they include. To understand better the active nature of the device however it is when the device is mobilised and how it responds to use that we can understand best its agency and its role in the stabilisation of the re-experiment

Here, I will use the example of a variation introduced in the protocol of classification to look more closely at the participants' involvement with the device and the methods they bring in at this stage of the re-experiment. To re-experiment with the process of classification and submit it to variation, I have tested different configurations. Sometimes the same group of participants produced a description and assayed it, sometimes the filtering was done months later in another context. Different spatial organisations, different forms of social interactions have been tried: round tables, small groups, collective discussions. And different tools and forms of visualisation of the description have been actively solicited. Much more than for the first part of the re-experiment, I had to re-invent and learn from my mistakes. If most of the groups actively had engaged with the description of flashed images with enthusiasm, the resistance to participate to the filtering of the results was palpable. The structure of the tree with its hierarchy exerting a heavy constraint over the participants even when they tried to bend it or modify its logic.

I am presenting here a re-experiment whose design is the result of these many trials and errors. It is an attempt to give space for the filtering process to develop and the dynamics to unfold including its difficult aspects. In this version, I chose to make room for the participants' frustration and to confront the obduracy (Akrich, 1992, p. 206) of the apparatus.

After the participants have annotated images and discussed the descriptions, the set-up changes and they are invited to assay a series of descriptions using the taxonomy. Instead of asking the participants to filter their own descriptions, I am giving them a series of descriptions produced by other participants in earlier sessions. They have acquired a familiarity with the process of description (they have done it for half a session or more) but the descriptions they have to assay are new to them.

The spatial organisation is as follows: the participants are seated around a table facing a screen where different stimuli are projected during the session. The microphone is at the centre of the table suggesting a collective discussion. Each participant receives a bundle of documents[98] that contains:
- a copy of the taxonomy presented above
- a sheet of paper with a photograph, the time constraint and the transcript of its description, followed by three empty sections named respectively *Selected*, *Replaced* and *Discarded.*
They are instructed to filter the description using the taxonomy and to fill in the three fields:

- In *Selected*, they write the words or expressions from the transcript that found a match in the taxonomy.
- In *Replaced*, they write the labels from the classification corresponding to the words or

[98] See **appendix 4**.

expressions listed in *Selected*.

- In *Discarded*, they write the words or expressions that did not find a match in the taxonomy.

The participants repeat this procedure for three descriptions. When they have finished, they comment the descriptions and the decisions they have taken. As they have worked on the same descriptions, they compare their findings and questions.

To understand better the nature of the exercise and the dramatic impact of the filtering, it is worth looking at what happens to the descriptions. Using two samples, let's look at which words are selected, which ones are discarded and what remains of a description when the discarded words have been removed from a description and the selected ones replaced by the terms of the taxonomy.

Example one. Subject RW PT 500 ms

A room full of musical instruments. A piano in the foreground, a harp behind that, a guitar hanging on the wall (to the right). It looked like there was also a window behind the harp, and perhaps a bookcase on the left.

Discarded.

A ... full of ... A ... in the foreground, a ... behind that, a ... hanging on the ... (to the right). It looked like there was also a ... behind the ..., and perhaps a ... on the left.

Replaced

A INDOOR HOUSEHOLD full of MUSICAL INSTRUMENT. A MUSICAL INSTRUMENT in the foreground, a MUSICAL INSTRUMENT behind that, a MUSICAL INSTRUMENT hanging on the INDOOR (to the right). It looked like there was also a INDOOR behind the MUSICAL INSTRUMENT, and perhaps a SHELVES on the left.

This first example expresses strikingly the role of the taxonomy in filtering out entire dimensions of a description. A look at the discarded words reveals a heavy cut. All the spatial organisation of the scene is discarded. The discarded version of the description offers a precise description of a space haunted by invisible objects. Unnamed objects are “hanging”, “behind”, “on the left” or “in the foreground”. This spatial articulation comes along with a series of precautions (“looked like”, “perhaps”). Neither the spatial organisation nor the precautions find a match in the taxonomy, they are simply filtered out. The hesitations of the subject is ignored (and therefore the grade of confidence in what she asserts) and given a full positive value. What is *perhaps* a book case becomes a book case and what *looks like* a window becomes one. There is no space left for doubt[99].

Example two. Subject EC PT 107 ms

This is outdoors. A black furry dog is running/walking towards the right of the picture. His tail is in the air and his mouth is open. Either he had a ball in his mouth or he was chasing after a ball.

Discarded.

This is ... A ... furry ... is running/walking towards the right of the picture. His tail is in the ... and his mouth is open. Either he had … or he was …

Selected.
outdoors, black, dog, air, a ball in his mouth, chasing after a ball

Replaced

OUTDOOR BLACK PATCH SMALLER ANIMAL SKY SOCIAL INTERACTION SOCIAL INTERACTION

This example reinforces an observation that was introduced in the analysis of the first example. Very generally, the taxonomy's representationalism has for effect to discard a a whole class of semantic elements. Here items like “this, the, a” are considered purely syntactical and, as they don't represent an entity in the world, can be simply dispensed with. Also, any reference to the stimulus as a photograph has to be abandoned. The “picture” in the description is considered as a transparent window. The only way to include a photograph in the taxonomy is as an object depicted in the

[99] Or negativity for that matter. How would a sentence like “definitely not sea” would be treated in such a context?

stimulus. The photograph has a place in *inanimate object / manmade / other / painting /picture / mural*. The photographic mediation hasn't.

The example also shows how the taxonomy translates what it recognises in its own terms and the difference of treatment between elements that belong to different branches of the tree. The semantic treatment of the dog reveals drastic differences of classification between human and animal. Dog has a place under *Animal → Land mammal → Smaller animal*, but further details of the dog description as tail and mouth cannot be matched. *Face, head, appearance details* are only properties of the class *People*. Animals are not endowed with body parts in the classification. Additionally, as the dog is in movement, its description falls further in the classification's grey zone. *Running/walking* has a place in the taxonomy only when interpreted as sport or social interaction. And the direction of the dog's movement also falls outside the grid, the expression "towards the right" is being doubly excluded as no spatial information exists in the tree and its position relative to the frame is impossible to register as the stimulus cannot be defined as a picture.

Finally this second example also demonstrates the effect of separating out sensory information from the rest of the categories. "Furry" doesn't find a place in the grid. As an animal's property, it is excluded as only humans are granted appearance details. And as a sensory related information, it is not abstract enough. The category *Sensory related* only includes geometric information. This level of detail is too fine-grained as the category *Sensory related* is not created to supplement a property or an object with sensuous details but to code what cannot be named as an object. *Sensory-related* offers term to match an expression like "something blue" to "colour patch", but not to match the softness of a tissue or the dryness of a skin.

When I looked at the graph, I could form a first sense of imbalance and inconsistency about the classification. Now I can begin to apprehend how they affect what can find its way in the final result, what is discarded or how different branches are levelled unevenly. More than an intuition, there is now a deeper sense of the depth of the filtering, its wide-ranging effect and the discrimination it operates between items it recognises or discards.

Moving through the taxonomy

At this point, we have a better picture of what happens to the words when they travel from the description to the classification tree. To classify however is not a purely semantic affair. It is also a process of elaboration. How does that work? What kinds of decisions, interactions with the taxonomy are required? What kind of mobilisation and trajectories are enacted? To answer these questions, I am turning now to the close engagement of the participants with the device and follow their movements.

Responding to the hierarchical nature of the tree, the participant who assays a description identifies first the terms which are the closest to those used in the description and then moves up to the highest level of abstraction in the tree. This is what a participant, MB, calls "looking for the highest level word". If MB finds the word person, she will select the entry *People* and climb her way up to *Animate object*, the most abstract entry of the branch. A problem arises when, once at the top of the taxonomy, the participant realises that the top category conflicts with the attributes she needs to attach to her entry. An example of this problem can be found in a description such as "a rock sandy in colour". The participant AM, attempting to classify the expression "sandy in colour", selects *Appearance/details*, but is blocked on her way up because only *Animate objects* are given the attribute *Appearance/details*. And, rock, located in the taxonomy under *Natural Inanimate Object,* cannot have animated objects attributes[100].

As we have seen in chapter 7, the same question arises when another participant attempts to code the expression "white shirt". For this, the participant GDG needs to know whether a shirt can fit in the category *Appearance/details*. This brings a series of questions about what conditions the access to this attribute. GDG suggests that the shirt should be considered as part of somebody's appearance. To acquire the attribute *Appearance* the shirt needs to be under *Animate object → People*. But shirt, as rock in the previous example, is in the *Inanimate object*'s branch. GDG managed to get to *Appearance/details*: not by moving down from the most abstract level, but moving up from the lowest elements ("if I go from below"). Entering from the bottom of the tree, GDG may reach the category he wants and stop there, ignoring the upper categories. This operation implies to circumvent the hierarchical descending order that constitutes the taxonomy. It is not about finding *the* path, but finding *a* path to unlock the desired attributes.

If we concentrate on how the participants move, we see that they are not merely following the lines of inheritance, they are also scanning the tree. If a large part of their comments address the logical coherence of the taxonomy, their movements on the tree image are global and detached from the

[100] As the discussion continues, AM finds an additional problem:
AM: It needs to be animated
S: breathing with difficulty Ahh
AM: a rock is animated. There is no reason a rock should be considered inanimate in the first place.

paths structuring it. Classifying is as much about finding pathways and trajectories than applying a logical schema.

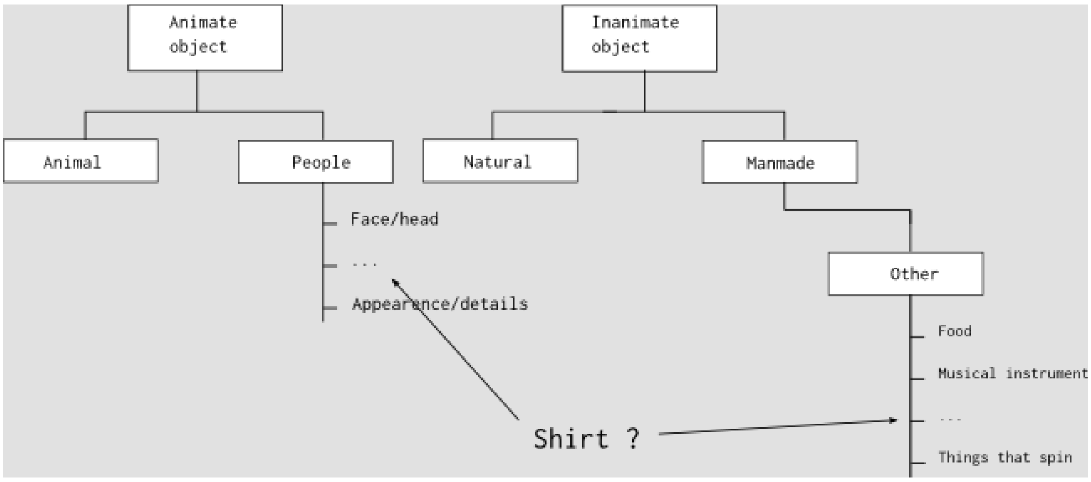


Illustration 9: How to move through the taxonomy?

To convince ourselves further, let's follow GDG while he moves visually through the tree image while trying to pin the expression “sandy in colour” that puzzled AM earlier:

GDG: sandy in colour I made it mountains I made it. It's the inanimate object natural mountain. This is one versus the other
AM: I made it [...] mountain.
GDG: It's only one I was not able to find it.
MB: Mountain is there twice.
GDG: But there is mountains from Outdoor that's the enormous scale and there is mountain/hill
MB: This is quite [...]
GDG: Actually there is also water twice [...] what it was
AM: Several mountains
MB: Yeah
AM: Mountains that's why I chose mountains

“I *made* it mountains”. The use of the verb make expresses again the weight of the operation of coding. There is a dimension of re-creation, of construction in the act of mapping a statement onto a taxonomy. GDG and AM both use the verb “make” and their interpretations of the same textual material “sandy in colour” converge. They chose *mountain* because it is the closest term available. Where is it located? GDG has a problem to find back the term on the taxonomy. MB looking at the tree remarks she has seen the entry *mountain* twice. Indeed the classification proposes two kinds of mountains. One, in its plural form under *outdoor* and another, coupled with *hill*. A difference of scale, interprets GDG. Their eyes are moving across the tree. GDG sees that *water* is also in these two separate branches of the classification. *Mountain* and *water* are available through distinct paths. GDG has chosen *Inanimate object → Natural → Mountain/hill* against the "enormous scale" of *Outdoors → Mountains*. It is the plural form that helped him to choose, not the parent categories. AM, on the contrary, motivates her choice by the parent category, a mountain cannot be inanimate.

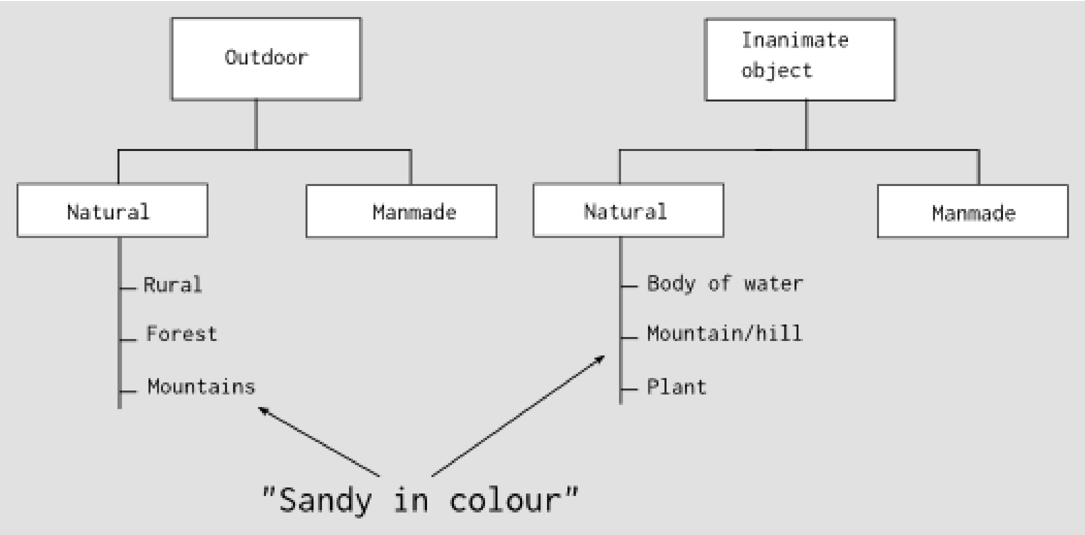


Illustration 10: A dilemma: Mountains or Mountain/hill?

Again here, the substance of the discussion cannot be separated by how the participants are moving across the classification. The participants retrace other paths through the tree. They connect different branches. They jump from one side to the other. They look for patterns and redundancies in the tree. They locate a potential candidate in two places, a typical feature of the taxonomy (*water* as well as *mountain* are repeated). There is, here, as in so many other occasions, a form of repairing at work. They are struggling with the fact that the taxonomy is never complete, always needs to be interpreted. Gaps must be filled, approximations need to be made. The entries are indicative and must be expanded. The taxonomy imposes a structure but when used, the structure is under negotiation. It is a map without legend and as such it leaves room for interpretation. The participants bring their knowledge and ability to extend categories that they find too narrow. They test whether there is an implicit "other" or if a list should be considered closed.

Observing the movements, the participants are also doing more than expanding and filling the blanks. They remodel the tree. They invent new pathways. They even suggest the shift of a whole branch to another part of the tree (plant should move from *inanimate* to *animate*). As the descriptions are mapped onto the taxonomy, the taxonomy itself is in turn, re-mapped bit by bit. This happens through a complex choreography, an ample set of movements of the hand and fingers, with the ball pen. They point to certain zones of the screen, the eyes crossing over the taxonomy, gazing over the taxonomy's surface, moving the words verbally from the paper to the projected map and back again. They look at each other or they listen while looking at the screen.

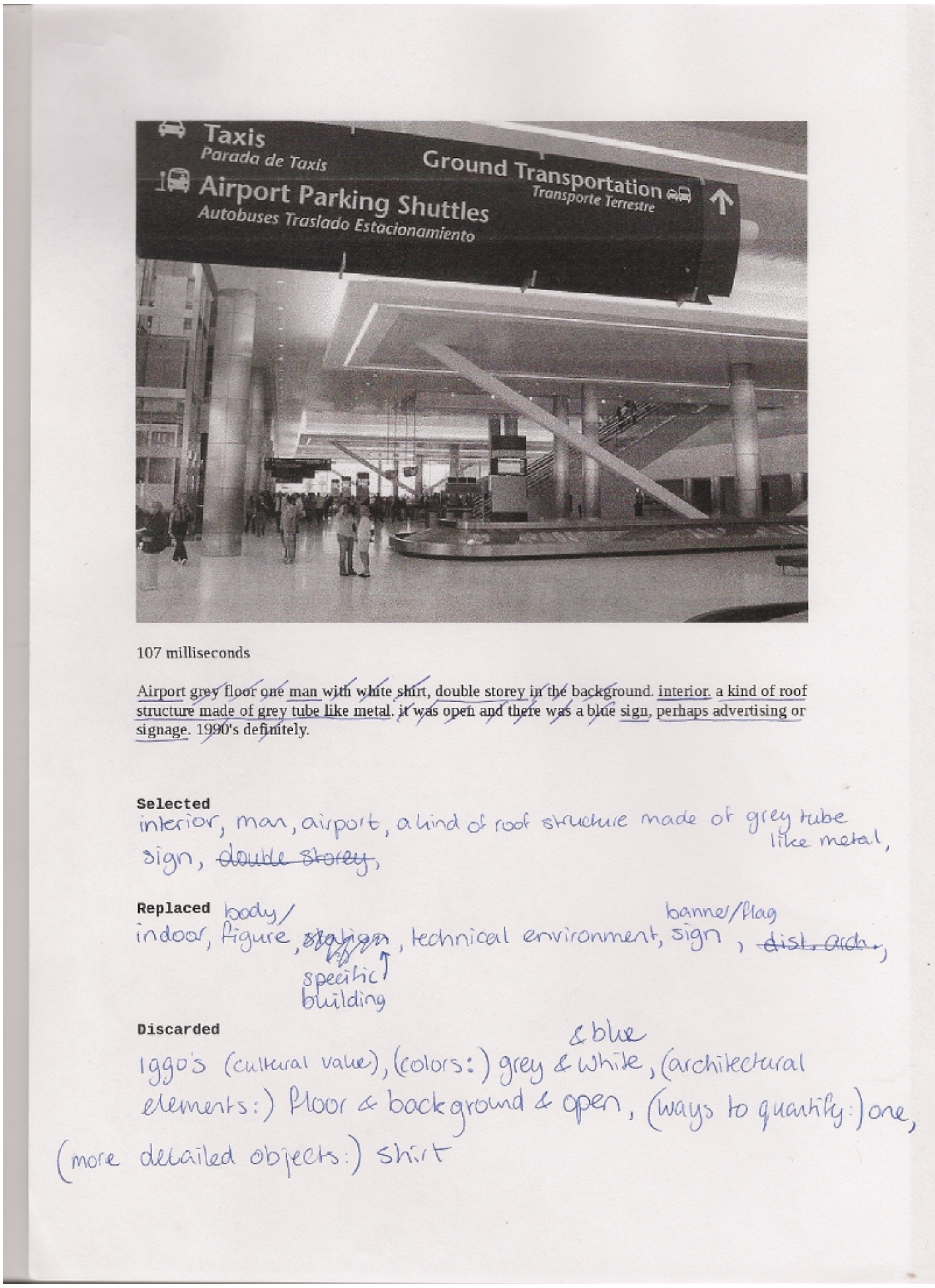
However, even if the participants go to great length to amend the taxonomy and bypass its limitations and shortcomings, they cannot ignore that its contradictions are beyond repair. There is simply no place for spatial information or the mediation of the photograph. Certain kinds of properties are arbitrarily denied to entire segments of the tree: an animal has no appearance or body part. And assumptions about the nature of buildings are hard-coded in the classification. To be a building with a commercial function, a building has to be qualified as indoor. An outdoor market can only be coded as *manmade → urban*, losing its commercial function. This combined sense of incoherence and rigidity affect the participants. And lead them to gradually abandon the search for alternative trajectories within the taxonomy. Regarding the general mood of the participants, there is a stark contrast with the previous stages. Participants' enthusiasm is fading and they manifest suspicion. Even with a set-up that gives space for the discussion, the atmosphere remains different from the earlier sessions. The participants only with time allow themselves to explore different strategies to obliquely interact with the device, but remain overall unable to adjust it to their needs. More than their ability to redefine the taxonomy through use, what they learn is the obduracy of classification device.

Summary

In the 7th chapter, I introduced the taxonomy as a device that makes explicit a process that had started earlier: the decision to include or exclude certain elements of the descriptions, and the levelling of the statements. Whereas in the previous stages of the re-experiment, this had been accomplished through consensus, here a specific device and a closed list of terms become the obligatory passage point for the descriptions. In the chapter, I have insisted on the importance of the shift that happens at that stage for the epistemic course of the re-experiment, for its underlying consensus and for its photographic elaboration. In this appendix, through a reading of the graph and the discussion of a variation on the protocol, I have given more details about the semantic structure of the taxonomy, and commented on the huge differences in size and depth of its categories as well as its logical inconsistencies. By observing what happens to two samples descriptions I have shown how drastic was the process of elimination and how little nuanced was its handling of a wide range of semantic categories.

In the chapter, I have discussed the classification process as one lived by the participants as dumbing down that demotivates the subjects, changes the parameters of the resolution of the re-experiment. As a complement, in this appendix, I have shown that the participants attempt to resist this effect of “dumbing down”. They deploy various methods to try to repair, expand or simply make sense of the arbitrariness or limitations of the taxonomy. They are “making” mountains, not just mechanically applying the grid over the text. They are constantly filling the gaps left in the tree. By looking for different trajectories, trying to find crossways in the taxonomy, starting from the bottom or the middle rather than following the top-down logic of the device, they try to make room for interpretation and to find a means of categorization that does justice to the nature of the process that lead to the descriptions: the composite echo. However these attempts fall short of circumventing the limited logic and the shortcomings of the classification. At that level, the appendix shows the extent to which the participants actively attempt to change the device without succeeding in doing so. In this sense, the appendix offers more elements to consider the operation of stabilisation performed by the taxonomy as one that conjugates the elimination of meanings in the description with the demotivation of the participants. It remains deaf to tone and nuance, as well as to the various attempts to engage with it beyond a mechanistic application of the grid. It holds the participants in place even when they attempt to find new trajectories.

Appendix 4. Controls


Taxis
Parada de Taxis
Ground Transportation
Transporte Terrestre
Airport Parking Shuttles
Autobuses Traslado Estacionamiento

107 milliseconds

Airport grey floor one man with white shirt, double storey in the background. interior. a kind of roof structure made of grey tube like metal. it was open and there was a blue sign, perhaps advertising or signage. 1990's definitely.

Selected

interior, man, airport, a kind of roof structure made of grey tube like metal, sign, ~~double storey~~,

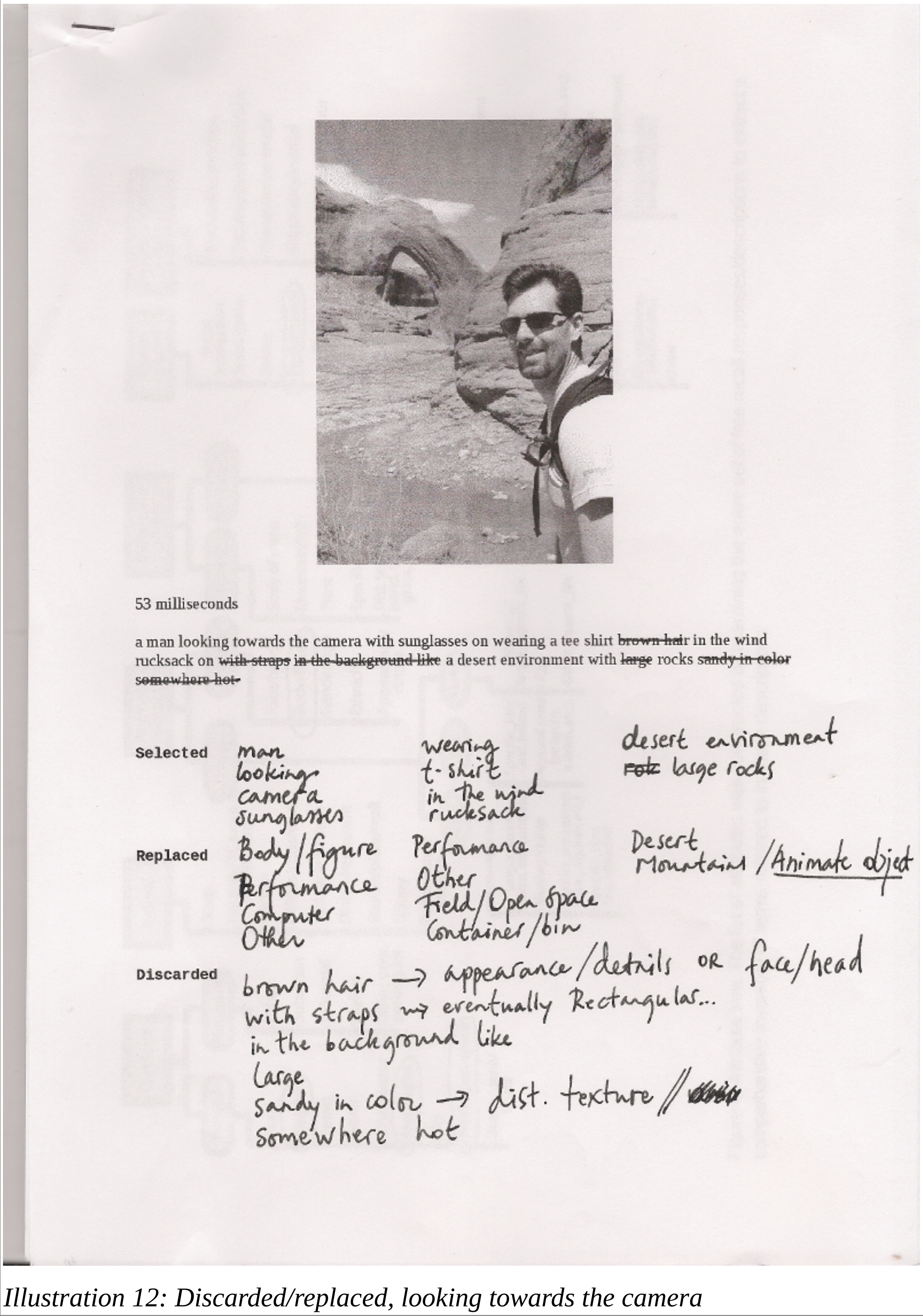
Replaced

indoor, body/ figure, ~~station~~ specific building, technical environment, banner/flag sign, ~~dist. arch.~~,

Discarded

1990's (cultural value), (colors:) grey & blue & white, (architectural elements:) floor & background & open, (ways to quantify:) one, (more detailed objects:) shirt

Illustration 11: Discarded/replaced, airport



53 milliseconds

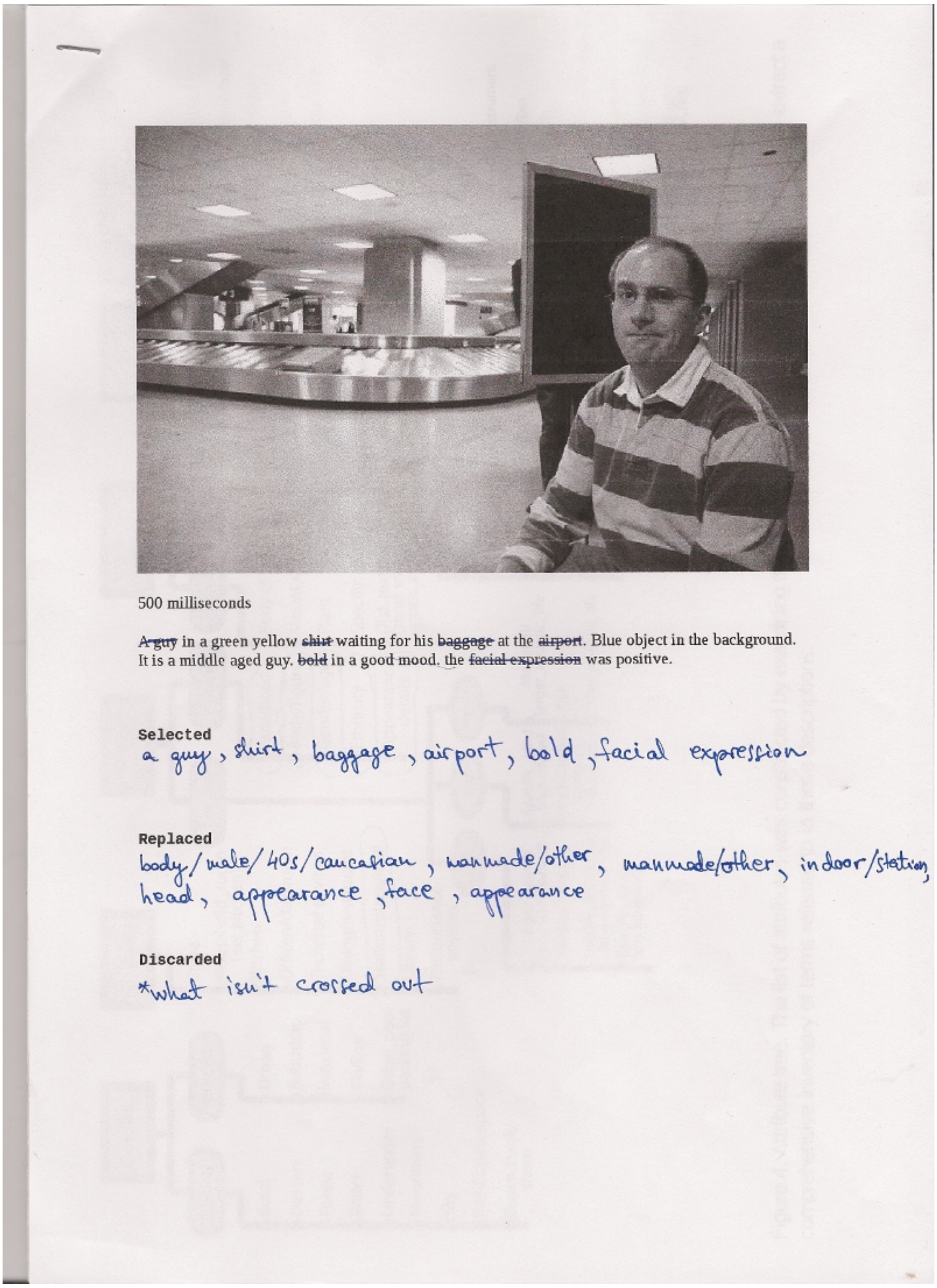
a man looking towards the camera with sunglasses on wearing a tee shirt ~~brown hai~~r in the wind rucksack on ~~with straps~~ ~~in the background like~~ a desert environment with ~~large~~ rocks ~~sandy in color~~ ~~somewhere hot~~

Selected	man looking camera sunglasses	wearing t-shirt in the wind rucksack	desert environment ~~rok~~ large rocks
Replaced	Body/figure Performance Computer Other	Performance Other Field/Open Space Container/bin	Desert Mountains / Animate object

Discarded

brown hair → appearance/details OR face/head
with straps ⇝ eventually Rectangular...
in the background like
large
sandy in color → dist. texture //
somewhere hot

Illustration 12: Discarded/replaced, looking towards the camera



500 milliseconds

~~A guy~~ in a green yellow ~~shirt~~ waiting for his ~~baggage~~ at the ~~airport~~. Blue object in the background.
It is a middle aged guy. ~~bold~~ in a good ~~mood~~. the ~~facial expression~~ was positive.

Selected

a guy, shirt, baggage, airport, bold, facial expression

Replaced

body/male/40s/caucasian, manmade/other, manmade/other, indoor/station, head, appearance, face, appearance

Discarded

*what isn't crossed out

Illustration 13: Discarded/replaced, a guy in a green yellow shirt

Website

A selection of the descriptions made by the participants during the sessions can be accessed by the examiners at the following URL:

http://functionariesofthecamera.net/algorithms-of-vision/

To view the page, enter the following information:

User Name: preview

Password: behindthescenes

References

Adams St. Pierre, E. (2013) The Appearance of Data, *Cultural Studies ↔ Critical Methodologies*, 13 (4), pp. 223–227. DOI:10.1177/1532708613487862.

Akrich, M. (1992) The De-Scription of Technical Objects, in: Bijker, W. E. and Law, J. (eds.) *Shaping technology/building society : studies in sociotechnical change*. Cambridge, MA, pp. 205–224.

Apprich, C., Hui Kyong Chun, W., Cramer, F. and Steyerl, H. (2018) *Pattern Discrimination*. Lüneburg.

Arar, S. (2017) An Introduction to the Fast Fourier Transform. Available from: https://www.allaboutcircuits.com/technical-articles/an-introduction-to-the-fast-fourier-transform/ [Accessed 22 January 2021].

Augustin, M. D., Leder, H., Hutzler, F. and Carbon, C.-C. (2008) Style follows content: On the microgenesis of art perception, *Acta Psychologica*, 128 (1), pp. 127–138. DOI:https://doi.org/10.1016/j.actpsy.2007.11.006.

Bachelard, G. (1980) *Épistémologie*, Lecourt, D. (ed.) . 3rd ed. Paris: Les Presses universitaires de France.

Barad, K. (1996) Meeting the Universe Halfway: Realism and Social Constructivism without Contradiction, in: Nelson, L. H. and Nelson, J. (eds.) *Feminism, Science, and the Philosophy of Science*. Dordrecht: Springer Netherlands, pp. 161–194.

Barad, K. (2010) Quantum Entanglements and Hauntological Relations of Inheritance: Dis/continuities, SpaceTime Enfoldings, and Justice-to-Come, *Derrida Today*, 3 (2), pp. 240–268. DOI:10.3366/drt.2010.0206.

Barbican (2019) Trevor Paglen: From ‘Apple’ to ‘Anomaly’. Available from: https://www.barbican.org.uk/our-story/press-room/trevor-paglen-from-apple-to-anomaly [Accessed 22 January 2021].

Beheshti, S.-M.-R., Tabebordbar, A., Benatallah, B. and Nouri, R. (2016) Data Curation APIs, *CoRR*, abs/1612.03277. Available from: http://arxiv.org/abs/1612.03277 [Accessed 22 January 2021].

Bender, E. M., Gebru, T., McMillan-Major, A. and Shmitchell, S. (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? □, in: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: Association for Computing Machinery, pp. 610–623.

Bengio, Y., Lee, D.-H., Bornschein, J. and Lin, Z. (2015) Towards Biologically Plausible Deep Learning, *CoRR*, abs/1502.04156. Available from: http://arxiv.org/abs/1502.04156 [Accessed 22 January 2021].

Bergen, M. and Wagner, K. (2015) Welcome to the AI Conspiracy: The ‘Canadian Mafia’ Behind

Tech’s Latest Craze. Available from: http://www.recode.net/2015/7/15/11614684/ai-conspiracy-the-scientists-behind-deep-learning [Accessed 22 January 2021].

Berry, D. (2011) *The philosophy of software*. London: Palgrave Macmillan UK.

Blackman, L. (2014) Affect and automaticy: Towards an analytics of experimentation, *Subjectivity*, 7 (4), pp. 362–384. DOI:10.1057/sub.2014.19.

Bogost, I. (2015) The cathedral of computation, *The Atlantic*.

Bolton, R. (1992) The contest of meaning: critical histories of photography, in: Bolton, R. (ed.) *The contest of meaning*. Cambridge, Massachusetts: MIT Press, pp. ix–xix.

Borji, A. (2017) Negative Results in Computer Vision: A Perspective, *CoRR*, abs/1705.0. Available from: http://arxiv.org/abs/1705.04402 [Accessed 22 January 2021]

Bowker, G. C. and Star, S. L. (2000) *Sorting Things out: Classification and Its Consequences*. Cambridge, MA, USA: MIT Press.

Bradski, G. and Kaehler, A. (2008) *Learning OpenCV*. Sebastopol: O’Reilly Media.

Buolamwini, J. (2016) InCoding — In The Beginning. Available from: https://medium.com/mit-media-lab/incoding-in-the-beginning-4e2a5c51a45d [Accessed 22 January 2021].

Buolamwini, J. and Gebru, T. (2018) Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, in: Friedler, S. A. and Wilson, C. (eds.) *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. New York, NY, USA: PMLR,81, pp. 77–91.

California Institute of Technology [no date] Computational Vision: [Data Sets]. Available from: http://www.vision.caltech.edu/archive.html [Accessed 22 January 2021].

Calyskan, A., Bryson, J. J. and Narayanan, A. (2017) Semantics derived automatically from language corpora contain human-like biases. Available from: http://science.sciencemag.org/content/356/6334/183 [Accessed 22 January 2021].

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., *et al.* (2018) Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015, *Nature Human Behaviour*, 2 (9), pp. 637–644. DOI:10.1038/s41562-018-0399-z.

Casilli, A. (2019) *En attendant les robots*. Paris: Le Seuil.

Chun, W. H. K. (2011) *Programmed Visions, Software and Memory*. Cambridge, Massachusetts: MIT Press.

Chun, W. H. K. (2016) *Updating to Remain the Same: Habitual New Media*. The MIT Press.

Cox, G. (2017). Ways of Machine Seeing: An Introduction. *A Peer-Reviewed Journal About*, 6(1), 1-9.

Copeland, M. (2016) What’s the Difference Between Artificial Intelligence, Machine Learning, and Deep Learning? Available from: https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/ [Accessed 22 January 2021].

Crawford, K. and Paglen, T. (2019) Excavating AI , The Politics of Images in Machine Learning Training Sets. Available from: https://www.excavating.ai/ [Accessed 22 January 2021].

Cruz, E. G. and Meyer, E. T. (2012) Creation and Control in the Photographic Process: iPhones and the emerging fifth moment of photography, *Photographies*, 5 (2), pp. 203–221. DOI:10.1080/17540763.2012.702123.

Danziger, K. (1992) The Project of an Experimental Social Psychology: Historical Perspectives, *Science in Context*, 5, pp. 309–328. DOI:10.1017/S0269889700001204.

Daston, L. (1994) Enlightenment Calculations, *Critical Inquiry*, 21 (1), pp. 182–202. Available from: http://www.jstor.org/stable/1343891 [Accessed 22 January 2021].

Daston, L. and Galison, P. (2007) *Objectivity*. New York, NY, USA: Zone Books.

Davis, H. (2019) A Dataset is a Worldview. Available from: https://towardsdatascience.com/a-dataset-is-a-worldview-5328216dd44d [Accessed 22 January 2021].

Deleuze, G. (1988) *Foucault / Gilles Deleuze ; translated and edited by Sean Hand*. Athlone London.

Deng, J., Dong, W., Socher, R., Li, L., Li, K. and Fei-fei, L. (2009) Imagenet: A large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition

DeSantis, K. and Housen, A. (2007) *Highlights of Findings -San Diego: Aesthetic Development and Creativeand Critical Thinking Skills Study*. Available from: https://vtshome.org/wp-content/uploads/2016/08/4HighlightsSanAntonio.pdf [Accessed 22 January 2021].

Deselaers, T. and Ferrari, V. (2011) Visual and semantic similarity in ImageNet, *CVPR 2011*, pp. 1777–1784.

Despret, V. (2009) *Penser comme un rat*. Cemagref,. Paris: Editions Quae.

Dewdney, A. (1995) Curating the photographic image in networked culture, in: Lister, M. (ed.) *The photographic image in digital culture*. London: Routledge, pp. 95–112.

Dolphijn, R. and van der Tuin, I. (2012) *New Materialism. Interviews and Cartographies*.

Dourish, P. (2016) Algorithms and their others: Algorithmic culture in context, *Big Data & Society*, 3 (2), pp. 2053951716665128. DOI:10.1177/2053951716665128.

Dulhanty, C. and Wong, A. (2019) Auditing ImageNet: Towards a Model-driven Framework for Annotating Demographic Attributes of Large-Scale Image Datasets, *CoRR*, abs/1905.01347. Available from: http://arxiv.org/abs/1905.01347 [Accessed 22 January 2021].

Fei-Fei, L. (2005) *Visual Recognition: Computational Models and Human Psychophysics*. California Institute of Technology.

Fei-Fei, L. (2010) ImageNet, crowdsourcing, benchmarking & other cool things, in: *CMU VASC Seminar*.

Fei-Fei, L (2012) Computers that see. Available from: https://www.youtube.com/watch?

v=viwpTTvSQKM [Accessed 22 January 2021].

Fei-Fei, L. (2020) Where Did ImageNet Come From? Available from: https://unthinking.photography/articles/where-did-imagenet-come-from [Accessed 22 January 2021].

Fei-Fei, L., Iyer, A., Koch, C. and Perona, P. (2007) What do we perceive in a glance of a real-world scene?, *Journal of Vision*, 7 (1), pp. 10. DOI:10.1167/7.1.10.

Fei-Fei, L., VanRullen, R., Koch, C. and Perona, P. (2002) Rapid natural scene categorization in the near absence of attention, *Proceedings of the National Academy of Sciences*, 99 (14), pp. 9596–9601. DOI:10.1073/pnas.092277599.

Fellbaum, C. (ed.) (1998) *WordNet: an electronic lexical database*. MIT Press.

Friedman, B. and Nissenbaum, H. (1996) Bias in Computer Systems, *ACM Trans. Inf. Syst.*, 14 (3), pp. 330–347. DOI:10.1145/230538.230561.

Fuller, M. (2005) Media Ecologies: Materialist Energies in Art and Technoculture. The MIT Press.

Fuller, M. (2008) *Software Studies: A Lexicon (Leonardo Books)*. The MIT Press.

Fuller, M. and Goffey, A. (2012) *Evil Media*. The MIT Press.

Gaikwad, S., Morina, D., Nistala, R., Agarwal, M., Cossette, A., Bhanu, R., *et al.* (2015) Daemo: A Self-Governed Crowdsourcing Marketplace.

Garun, N. (2017) Facebook’s AI now lets you search for photos by their content. Available from: https://www.theverge.com/2017/2/2/14486034/facebook-ai-update-photo-search-by-keyword [Accessed 22 January 2021].

Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L. and Fei-Fei, L. (2017) Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States, *Proceedings of the National Academy of Sciences*, 114 (50), pp. 13108–13113. DOI:10.1073/pnas.1700035114.

Geerts, E. (2016) Performativity. Available from: https://newmaterialism.eu/almanac/p/performativity.html [Accessed 22 June 2021].

Gibson, J. J. (1979) *The Ecological Approach to Visual Perception*. Boston, Massachusetts: Houghton-Mifflin.

Gillespie, T. (2014) The Relevance of Algorithms, in: Gillespie, T., Boczkowski, P., and Foot, K. (eds.) *Media Technologies*. Cambridge, MA: MIT Press, pp. 167–193.

Goffey, A. (2008) Algorithm, in: Fuller, M. (ed.) *Software Studies: A Lexicon*. Cambridge, MA: MIT Press, pp. 15–20.

Goldberg, Y. (2021) A criticism of ‘On the Dangers of Stochastic Parrots: Can Language Models be Too Big’. Available from: https://gist.github.com/yoavg/9fc9be2f98b47c189a513573d902fb27 [Accessed 17 June 2021].

Goodman, B. and Flaxman, S. (2016) EU regulations on algorithmic decision-making and a “right to explanation”, in: *ICML workshop on human interpretability in machine learning (WHI 2016), New York, NY. http://arxiv. org/abs/1606.08813 v1*.

GoogleTechTalks (2011) Large-scale Image Classification: ImageNet and ObjectBank. Available from: https://www.youtube.com/watch?v=qdDHp29QVdw [Accessed 22 January 2021].

Gov.uk (2017) AI Sector Deal. Available from: https://www.gov.uk/government/publications/artificial-intelligence-sector-deal/ai-sector-deal [Accessed 24 June 2021].

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y. and He, K. (2017) Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour.

Griffin, G., Holub, A. and Perona, P. (2007) Caltech-256 Object Category Dataset, *CalTech Report*.

Halpern, O. (2014) *Beautiful Data, A History of Vision and Reason since 1945*. Durham: Duke University Press.

Hand, M. (2012) *Ubiquitous Photography*. Cambridge: Polity Press.

Hansen, M. B. N. (2014) *Feed-Forward: On the Future of Twenty-First-Century Media*. University of Chicago Press.

Hansen, M. B. N. (2015) The Operational Present of Sensibility, *Nordic Journal of Aesthetics*, 24 (47).

Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C. and Bigham, J. P. (2018) A Data-Driven Analysis of Workers’ Earnings on Amazon Mechanical Turk, in: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, pp. 449:1—449:14.

Hardesty, L. (2018) Study finds gender and skin-type bias in commercial artificial-intelligence systems. Available from: https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212 [Accessed 22 January 2021].

Harel, J., Koch, C. and Perona, P. (2006) Graph-Based Visual Saliency, in: *Adv. Neural Inform. Process. Syst*.19, pp. 545–552.

Harris, C. and Stephens, M. (1988) A combined corner and edge detector, in: *In Proc. of Fourth Alvey Vision Conference*. pp. 147–151.

Harvey, A. (2019) Microsoft Celeb Dataset. Available from: https://megapixels.cc/msceleb/ [Accessed 22 January 2021].

Hasnain M., Rishabh S. and G., S. (2019) Smart Home Automation using Computer Vision and Segmented Image Processing, in: *2019 International Conference on Communication and Signal Processing (ICCSP)*. pp. 429–433.

Hata, K., Krishna, R., Fei-Fei, L. and Bernstein, M. S. (2017) A Glimpse Far into the Future: Understanding Long-term Crowd Worker Quality, in: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. New York, NY, USA: ACM, pp. 889–901.

Hentschel, C., Wiradarma, T. P. and Sack, H. (2016) Fine tuning CNNS with scarce training data — Adapting imagenet to art epoch classification, in: *2016 IEEE International Conference on Image Processing (ICIP)*. pp. 3693–3697.

Hietaniemi, J., Macdonald, J. and Orwant, J. (1999) *Mastering Algorithms with Perl*. Sebastopol: O'Reilly Media, Inc.

Hoelzl, I. and Marie, R. (2017) From Softimage to Postimage, *Leonardo*, 50 (1), pp. 72–73. DOI:10.1162/LEON_a_01349.

Housen, A. [no date] Research and Theory. Available from: https://vtshome.org/research/ [Accessed 22 January 2021].

Hu, G., Peng, X., Yang, Y., Hospedales, T. M. and Verbeek, J. (2016) Frankenstein: Learning Deep Face Representations using Small Data, *CoRR*, abs/1603.0. Available from: http://arxiv.org/abs/1603.06470 [Accessed 22 January 2021].

Huang, X., Shen, C., Boix, X. and Zhao, Q. (2015) SALICON: Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks, in: *The IEEE International Conference on Computer Vision (ICCV)*.

Humphreys, M., Greenaway, K. H. and Bentley, S. (2017) Psychology turns to online crowdsourcing to study the mind, but it's not without its pitfalls. Available from: http://theconversation.com/psychology-turns-to-online-crowdsourcing-to-study-the-mind-but-its-not-without-its-pitfalls-74070 [Accessed 22 January 2021].

Irani, L. C. (2015) The cultural work of microwork, *New Media & Society*, 17 (5), pp. 720–739. DOI:http://dx.doi.org/10.1177/1461444813511926.

Irani, L. C. and Silberman, M. S. (2013) Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk, in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, pp. 611–620.

Iyer, A. (2008) *Planning Goal-Directed Actions: fMRI Correlates in Humans and Monkeys*. California Institute of Technology. Available from : https://thesis.library.caltech.edu/5232/

Jahanian, A., Keshvari, S. and Rosenholtz, R. (2018) Web pages: What can you see in a single fixation?, *Cognitive Research: Principles and Implications*, 3 (1), pp. 14. DOI:10.1186/s41235-018-0099-2.

Jamar, D. and Stengers, I. (2011) Les médiateurs sont dans une position semblable à celle qu'ils imposent ou construisent avec leurs publics. Available from: http://www.iteco.be/antipodes/De-l-individuel-au-collectif/Respect-etc [Accessed 22 January 2021].

Jaton, F. (2017) We get the algorithms of our ground truths: Designing referential databases in digital image processing, *Social Studies of Science*, 47 (6), pp. 811–840. DOI:10.1177/0306312717730428.

Jay, M. (1993) *Downcast Eyes: The Denigration of Vision in Twentieth-Century French Thought*. Berkeley, Los Angeles, London: University of California Press. DOI:10.2307/2168472.

Kantor, J. R. (1933) In defense of stimulus-response psychology., *Psychological Review*, 40 (4), pp. 324–336. DOI:10.1037/h0073851.

Karpathy, A. (2014) What I learned from competing against a ConvNet on ImageNet. Available from: http://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/ [Accessed 22 January 2021].

Kasperkevic, J. (2015) Google says sorry for racist auto-tag in photo app. Available from: https://www.theguardian.com/technology/2015/jul/01/google-sorry-racist-auto-tag-photo-app [Accessed 22 January 2021].

Kastner, M. A., Ide, I., Kawanishi, Y., Deguchi, T. H. D. and Murase, H. (2019) A preliminary study on estimating word imageability labels using Web image data mining.

Kember, S. (2014) Face Recognition and the Emergence of Smart Photography, *Journal of Visual Culture*, 13 (2), pp. 182–199. DOI:10.1177/1470412914541767.

Kember, S. and Zylinska, J. (2012) *Life after New Media: Mediation as a Vital Process*. Cambridge, Massachusetts: The MIT Press.

Kirchdoerfer, T. (2018) *Data Driven Computing*. California Institute of Technology, Pasadena, California. Available from : https://thesis.library.caltech.edu/10431/8/Kirchdoerfer_Trenton_2017_Thesis.pdf [Accessed 22 January 2021]

Kitchin, R. (2017) Thinking critically about and researching algorithms, *Information, Communication & Society*, 20 (1), pp. 14–29. DOI:10.1080/1369118X.2016.1154087.

Knorr-Cetina, K. D. and Malkay, M. (1983) *Science observed : perspectives on the social study of science / editors, Karin D. Knorr-Cetina and Michael Mulkay*. Sage London.

Knuth, D. E. (1997) *The Art of Computer Programming, Volume 1 (3rd Ed.): Fundamental Algorithms*. Redwood City, CA, USA: Addison Wesley Longman Publishing Co., Inc.

Koch, C. (2004) *The Quest for Consciousness*. Roberts & Company.

Kowalski, R. (1979) Algorithm = Logic + Control, *Commun. ACM*, 22 (7), pp. 424–436. DOI:10.1145/359131.359136.

Kremer, J., Stensbo-Smidt, K., Gieseke, F., Pedersen, K. S. and Igel, C. (2017) Big Universe, Big Data: Machine Learning and Image Analysis for Astronomy, *IEEE Intelligent Systems*, 32 (2), pp. 16–22. DOI:10.1109/mis.2017.40.

Krishna, R., Hata, K., Chen, S., Kravitz, J., Shamma, D. A., Li, F.-F. and Bernstein, M. S. (2016) Embracing Error to Enable Rapid Crowdsourcing, *CoRR*, abs/1602.0. Available from: http://arxiv.org/abs/1602.04506 [Accessed 22 January 2021].

Krishna, Poddar, M., Giridhar M K, Prabhu, A. S. and Umadevi V (2016) Automated traffic monitoring system using computer vision, in: *2016 International Conference on ICT in Business Industry Government (ICTBIG)*. pp. 1–5.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., *et al.* (2016) Visual Genome:

Connecting Language and Vision Using Crowdsourced Dense Image Annotations.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012) Imagenet classification with deep convolutional neural networks, in: *Advances in Neural Information Processing Systems*.

Kuhn, T. S. (1996) *The structure of scientific revolutions / Thomas S. Kuhn*. 3rd ed. University of Chicago Press Chicago, IL.

Kurenkov, A. (2015) A 'Brief' History of Neural Nets and Deep Learning, Part 1. Available from: http://www.andreykurenkov.com/writing/a-brief-history-of-neural-nets-and-deep-learning/ [Accessed 22 January 2021].

Lapedriza, À., Pirsiavash, H., Bylinskii, Z. and Torralba, A. (2013) Are all training examples equally valuable?, *CoRR*, abs/1311.6. Available from: http://arxiv.org/abs/1311.6510 [Accessed 22 January 2021]

Latour, B (1987) *Science in Action. How to follow scientists and engineers through society*. Harvard University Press.

Latour, Bruno (1983) Give me a laboratory and I will move the world, in: Knorr, K. and Mulkay, M. (eds.) *Science Observed*. Science Ob. Sage, pp. 141–170.

Lehmuskallio, A. (2016) The camera as a sensor: The visualization of everyday digital photography as simulative, heuristic and layered pictures., in: Gómez Cruz and Lehmuskallio (eds.) *Digital Photography and Everyday Life. Empirical studies on material visual practices.* London/New York: Routledge, pp. 243–266.

Lettvin, J. Y., Maturana, H. R., McCulloch, W. S. and Pitts, W. H. (1959) What the Frog's Eye Tells the Frog's Brain, *Proceedings of the IRE*, 47 (11), pp. 1940–1951. DOI:10.1109/JRPROC.1959.287207.

Leys, R. (2011) The Turn to Affect: A Critique, *Critical Inquiry*, 37 (3), pp. 434–472. DOI:10.1086/659353.

Lissack, M. (2021) The Slodderwetenschap (Sloppy Science) of Stochastic Parrots - {A} Plea for Science to {NOT} take the Route Advocated by Gebru and Bender, *CoRR*, abs/2101.10098. Available from: https://arxiv.org/abs/2101.10098 [Accessed 24 June 2021]

Lister, M. (2007) A Sack in the Sand: Photography in the Age of Information, *Convergence*, 13 (3), pp. 251–274. DOI:10.1177/1354856507079176.

Lister, M. (2009) Photography in the age of electronic imaging, in: Wells, ed. L. (ed.) *Photography: a critical introduction*. 4th ed. Abingdon: Routledge, pp. 311–344.

Lister, M. (1995) *The photographic image in digital culture*. London; New York: Routledge. Available from: http://lib.leeds.ac.uk/record=b2112956 [Accessed 22 January 2021]

Lynch, M. (1994) *Scientific Practice and Ordinary Action: Ethnomethodology and Social Studies of Science*. Cambridge University Press. DOI:10.1017/CBO9780511625473.

MacKenzie, A. (2006) *Cutting Code , Software and Sociality*. Berlin: Peter Lang.

Mackenzie, A. (2015) The production of prediction: What does machine learning want?, *European Journal of Cultural Studies*, 18 (4–5), pp. 429–445. DOI:10.1177/1367549415577384.

Mackenzie, A. (2017) *Machine Learners: Archaeology of a Data Practice*. MIT Press.

Madge, C. and Harrisson, T. (1937) *Mass observation*. London: Frederick Muller.

Malabou, C. (2009) *What Should We Do with Our Brain?* Fordham University Press.

Manning, E. (2016) *The Minor Gesture*. Durham: Duke University Press.

Marr, D. (1982) *Vision. A Computational Investigation into the Human Representation and Processing of Visual Information*. Cambridge, Massachusetts: MIT Press.

Masi, I., Tr▯n, A. T., Hassner, T., Leksut, J. T. and Medioni, G. (2016) Do We Really Need to Collect Millions of Faces for Effective Face Recognition?, in: Leibe, B., Matas, J., Sebe, N., and Welling, M. (eds.) *Computer Vision -- ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*. Cham: Springer International Publishing, pp. 579–596.

Massumi, B. (2002) *Parables for the Virtual: Movement, Affect, Sensation*. Duke University Press.

Massumi, B. (2019) Immediation Unlimited, in: Manning, E., Munster, A., and Stavning Thomsen, B. M. (eds.) *Immediation II*. London: Open Humanities Press, pp. 501–543.

McKim, J. (2008) Of Microperception and Micropolitics An Interview with Brian Massumi. Available from: http://www.inflexions.org/n3_massumihtml.html [Accessed 16 June 2021].

Miceli, M., Schuessler, M. and Yang, T. (2020) Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision.

Miller, G. A. (1995) WordNet: A Lexical Database for English, *Commun. ACM*, 38 (11), pp. 39–41. DOI:10.1145/219717.219748.

Mol, A. (2002) *The Body Multiple: Ontology in Medical Practice*. Duke University Press.

Murphy, M. L. (2003) *Semantic Relations and the Lexicon: Antonymy, Synonymy, and Other Paradigms*. Cambridge University Press.

Nerakae, P., Uangpairoj, P. and Chamniprasart, K. (2016) Using Machine Vision for Flexible Automatic Assembly System, *Procedia Computer Science*, 96, pp. 428–435. DOI:10.1016/j.procs.2016.08.090.

Open Science Collaboration (2015) Estimating the reproducibility of psychological science, *Science*, 349 (6251). DOI:10.1126/science.aac4716.

OpenCV [no date] Harris Corner Detection. Available from: https://docs.opencv.org/master/dc/d0d/tutorial_py_features_harris.html [Accessed 22 January 2021].

Paivio, A. (1986) *Mental representations : a dual coding approach.* New York (N.Y.)▯: Oxford university press.

Papadopoulos, D. P., Uijlings, J. R. R., Keller, F. and Ferrari, V. (2016) We don’t need no bounding-boxes: Training object class detectors using only human verification, *CoRR*, abs/1602.0. Available from: http://arxiv.org/abs/1602.08405 [Accessed 22 January 2021].

Pashler, H. and Wagenmakers, E. (2012) Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence?, *Perspectives on Psychological Science*, 7 (6), pp. 528–530. DOI:10.1177/1745691612465253.

Pasquinelli, M. (2009) Google's PageRank Algorithm: a diagram of cognitive capitalism and the rentier of the common intellect, in: Becker, K. and Stalder, F. (eds.) *Deep Search*. London: Transaction Publishers.

Perona, P. (2010) Vision of a Visipedia, *Proceedings of the IEEE*, 98 (8), pp. 1526–1534. DOI:10.1109/JPROC.2010.2049621.

Powles, J. and Nissenbaum, H. (2018) The Seductive Diversion of 'Solving' Bias in Artificial Intelligence. Available from: https://onezero.medium.com/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53 [Accessed 22 January 2021].

Psiturk: crowdsource your research (no date). Available from: http://psiturk.org/ [Accessed 22 January 2021].

Pylyshyn, Z. W. (1999) Is Vision Continuous with Cognition? The Case for Cognitive Impenetrability of Visual Perception, *Behavioral and Brain Sciences*, 22 (3), pp. 341–365.

Read, J. C. A. (2015) The place of human psychophysics in modern neuroscience, *Neuroscience*, 296, pp. 116–129. DOI:https://doi.org/10.1016/j.neuroscience.2014.05.036.

Rogers, B. (2018) A Director's View. Available from: https://thephotographersgallery.org.uk/viewpoints/outside-arriving-photographers-gallery [Accessed 22 January 2021].

Roman, V. (2019) Supervised Learning: Basics of Classification and Main Algorithms. Available from: https://towardsdatascience.com/supervised-learning-basics-of-classification-and-main-algorithms-c16b06806cd3 [Accessed 24 June 2021]

Rosch, E. (1978) Principles of Categorization, in: Rosch, E. and Lloyd, B. B. (eds.) *Cognition and Categorization*. Hillsdale, NJ: Erlbaum, pp. 27–48.

Rosenberg, C. (2013) Improving Photo Search: A Step Across the Semantic Gap. Available from: https://ai.googleblog.com/2013/06/improving-photo-search-step-across.html [Accessed 22 January 2021].

Rubinstein, D. (2009) Towards Photographic Education, *Photographies*, 2 (2), pp. 135–142. DOI:10.1080/17540760903116598.

Rubinstein, D. (2018) What is 21st Century Photography? Available from: https://thephotographersgallery.org.uk/content/what-21st-century-photography [Accessed 22 January 2021].

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., *et al.* (2014) ImageNet Large Scale Visual Recognition Challenge, *CoRR*, abs/1409.0. Available from:

http://arxiv.org/abs/1409.0575 [Accessed 22 January 2021].

Shah, M. (2012) Lecture 17: Bag-of-Features (Bag-of-Words). Available from: https://www.youtube.com/watch?v=iGZpJZhqEME&index=17&list=PLmyoWnoyCKo8epWKGHAm4m_SyzoYhslk5 [Accessed 22 January 2021].

Saito, S., Chiang, C.-W., Savage, S., Nakano, T., Kobayashi, T. and Bigham, J. P. (2019) TurkScanner: Predicting the Hourly Wage of Microtasks, *The World Wide Web Conference on - WWW '19*. DOI:10.1145/3308558.3313716.

Salehi, N., Irani, L. C., Bernstein, M. S., Alkhatib, A., Ogbe, E., Milland, K. and Clickhappier (2015) We Are Dynamo: Overcoming Stalling and Friction in Collective Action for Crowd Workers, in: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, pp. 1621–1630.

Sanchez del Rio, J., Moctezuma, D., Conde, C., Martin de Diego, I. and Cabello, E. (2016) Automated border control e-gates and facial recognition systems, *Computers & Security*, 62, pp. 49–72. DOI:https://doi.org/10.1016/j.cose.2016.07.001.

Schmieg, S. (2016) Decision Space. Available from: http://decision-space.com/ [Accessed 22 June 2021].

Scholz, T. (2017) *Uberworked and underpaid : how workers are disrupting the digital economy*. Cambridge: Polity.

Screenwalks (2020) Screenwalks.com. Available from: https://screenwalks.com/ [Accessed 22 June 2021].

Seaver, N. (2013) Knowing algorithms. Cambridge, MA. Available from: http://nickseaver.net/papers/seaverMiT8.pdf [Accessed 22 January 2021]

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J. and Sculley, D. (2017) No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World.

Shore, S. (1999) *American Surfaces, 1972*. Munich: Schirmer/Mosel.

Shore, S. (2004) *Uncommon Places: The Complete Works.* London: Thames & Hudson.

Silverio, M. (2020) Google AI for breast cancer detection beats Doctors. Available from: https://towardsdatascience.com/google-ai-for-breast-cancer-detection-beats-doctors-65b8983352e0 [Accessed 22 January 2021].

Simonite, T. (2019) Fei-Fei Li Wants AI to Care More About Humans. Available from: https://www.wired.com/story/fei-fei-li-ai-care-more-about-humans/ [Accessed 22 January 2021].

Simpson, E. (2012) *War From the Ground Up : Twenty-First Century Combat As Politics.* Oxford University Press.

Slodzian, M. (2000) WordNet : what about its linguistic relevancy ?

Slow Art (2017). Available from: https://www.ucpress.edu/book.php?isbn=9780520285507

[Accessed 22 January 2021].

Sluis, K. (2013) 'The Canon After the Internet’: (Discussion with Katrina Sluis, Christiane Paul and Julian Stallabrass), *Aperture*, 213, pp. 37–41.

Sluis, K. (2018) 'Katrina Sluis'. Interviewed by Lewis Bush for *1000 Words*, 4 February. Available from: http://www.1000wordsmag.com/katrina-sluis/ [Accessed 22 January 2021].

Sluis, K. (2020) Survival of the Fittest Image. Available from: https://www.fotomuseum.ch/en/explore/still-searching/articles/157058_survival_of_the_fittest_image [Accessed 22 January 2021].

Smith, C. (2019) 20 Interesting flickr facts and stats 2019 by the Numbers. Available from: https://expandedramblings.com/index.php/flickr-stats/ [Accessed 22 January 2021].

Smith, C. (2017) Snapchat Statistics and Facts (June 2017). Available from: http://expandedramblings.com/index.php/snapchat-statistics/ [Accessed 22 January 2021].

Stahel, U. [no date] In Praise of Visual Literacy – A Plea. Available from: https://photography-in-switzerland.ch/essays/fotokompetenz-heute-ein-pladoyer?lang=en [Accessed 22 January 2021].

Stanford University School of Engineering’(2014) Stanford Engineering’s Fei-Fei Li explores visual intelligence in computers. Available from: https://www.youtube.com/watch?v=ylVsqXzlJqA [Accessed 22 January 2021].

Stansfield, K. (2009) *Practice scores, a toolkit for artistic research*. The University of Dundee.

Statisticbrain (2017) Instagram Company Statistics. Available from: http://www.statisticbrain.com/instagram-company-statistics/ [Accessed 22 January 2021].

Stengers, I. (2000) *The invention of modern science*. Minneapolis: University of Minnesota press.

Tagg, J (2009) *The Disciplinary Frame: Photographic Truths and the Capture of Meaning*. University of Minnesota Press.

Teachout, T. (2017) When It Comes to Art, How Seeing Less Is Seeing More. Available from: https://www.wsj.com/articles/when-it-comes-to-art-how-seeing-less-is-seeing-more-1491134402 [Accessed 22 January 2021].

Tensorflow [no date] How to Retrain an Image Classifier for New Categories. Available from: https://www.tensorflow.org/hub/tutorials/image_retraining [Accessed 22 January 2021].

The computer vision foundation (2020) Computer vision awards. Available from: https://www.thecvf.com/?page_id=413#LHP [Accessed 23 June 2021].

The Photographers' Gallery (2016) Decision Space. Available from: https://thephotographersgallery.org.uk/whats-on/digital-project/decision-space [Accessed 22 January 2021].

The Photographers' Gallery (2018) The Photographers’ Gallery Foundation and the Concerned Photographer. Available from: https://thephotographersgallery.org.uk/viewpoints/photographers-gallery-foundation-and-concerned-photographer [Accessed 22 January 2021].

The Photographers’ Gallery (2018a) All I Know Is What’s On The Internet. Available from: https://thephotographersgallery.org.uk/whats-on/exhibition/all-i-know-is-whats-on-the-internet [Accessed 22 June 2021].

The Photographers’ Gallery (2018b) Geekender: Towards a Feminist Internet. Available from: https://thephotographersgallery.org.uk/whats-on/event/geekender-feminist-internet [Accessed 23 June 2021].

The Photographers’ Gallery (2018c) Slow Art Day. Available from: https://thephotographersgallery.org.uk/whats-on/events/slow-art-day [Accessed 23 June 2021].

The Photographers' Gallery [no date] History, Mission & Vision. Available from: https://thephotographersgallery.org.uk/about-us/history-mission-and-vision [Accessed 22 January 2021].

The Photographers' Gallery [no date] Slow Looking: an introduction. Available from: https://thephotographersgallery.org.uk/whats-on/slow-looking-introduction [Accessed 22 January 2021].

Thread Magazine (2008) Go Against the Grain. Available from: https://www.threadsmagazine.com/2008/11/23/go-against-the-grain [Accessed 22 January 2021].

Toister, Y. (2019) PHOTOGRAPHY, Love’s labour’s lost, *Photographies*, 12 (1), pp. 117–133. DOI:10.1080/17540763.2018.1501726.

Tsuchiya, N. and Koch, C. (2005) Continuous flash suppression reduces negative afterimages, *Nature Neuroscience*, 8, pp. 1096. Available from: https://doi.org/10.1038/nn1500 [Accessed 22 January 2021]

Turner, S. and Coen, S. E. (2008) Member Checking in Human Geography: Interpreting Divergent Understandings of Performativity in a Student Space, *Area*, 40 (2), pp. 184–193. Available from: http://www.jstor.org/stable/40346113 [Accessed 22 January 2021]

Tversky, B. and Hemenway, K. (1983) Categories of environmental scenes, *Cognitive Psychology*, pp. 121–149.

Vapnik, V. (2000) *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.

Vee, A. (2017) *Coding Literacy: How Computer Programming is Changing Writing*. The MIT Press.

Venkatesan, M. (2018) Artificial Intelligence vs. Machine Learning vs. Deep Learning. Available from: https://www.datasciencecentral.com/profiles/blogs/artificial-intelligence-vs-machine-learning-vs-deep-learning [Accessed 22 January 2021].

Vijayanarasimhan, S. and Grauman, K. (2009) Multi-level active prediction of useful image annotations for recognition, in: *Advances in Neural Information Processing Systems*. pp. 1705–1712.

Walther, D., Rutishauser, U., Koch, C. and Perona, P. (2005) Selective visual attention enables learning and recognition of multiple objects in cluttered scenes, *Computer Vision and Image*

Understanding, 100 (1), pp. 41–63. DOI:https://doi.org/10.1016/j.cviu.2004.09.004.

Wikipedia [no date] Grain (textile). Available from: https://en.wikipedia.org/wiki/Grain_(textile) [Accessed 22 January 2021].

Wilke, A. and Mata, R. (2012) Cognitive Bias, in: Ramachandran, V. S. (ed.) *Encyclopedia of Human Behavior (Second Edition)*. Second Edition. San Diego: Academic Press, pp. 531–535.

Woolgar, S. (2014) It Could Be Otherwise: Provocation, Irony, and Limits. Available from: http://cstms.berkeley.edu/current-events/it-could-be-otherwise/ [Accessed 22 January 2021].

WordNet (2010a) WordNet entry: 'Parisian', *Princeton University*. Available from: http://wordnetweb.princeton.edu/perl/webwn?o2=&o0=1&o8=1&o1=1&o7=&o5=&o9=&o6=&o3=&o4=&s=Parisian&i=0&h=00#c [Accessed 22 January 2021].

WordNet (2010) WordNet entry 'ratatouille', *Princeton University*. Available from: http://wordnetweb.princeton.edu/perl/webwn?o2=&o0=1&o8=1&o1=1&o7=&o5=&o9=&o6=&o3=&o4=&s=ratatouille&i=0&h=0#c [Accessed 22 January 2021].

WordNet (2010b) WordNet entry: 'reformer', *Princeton University*. Available from: http://wordnetweb.princeton.edu/perl/webwn?o2=&o0=1&o8=1&o1=1&o7=&o5=&o9=&o6=&o3=&o4=&s=reformer&i=1&h=00#c [Accessed 22 January 2021].

Yang, K., Qinami, K., Fei-Fei, L., Deng, J. and Russakovsky, O. (2019) Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy.

Zylinska, J. (2017) *Nonhuman photography*. Cambridge, Massachusetts: The MIT Press.